CLASSIFYING COMPLEX TOPICS USING SPATIAL-SEMANTIC DOCUMENT VISUALIZATION: AN EVALUATION OF AN INTERACTION MODEL TO SUPPORT OPEN-ENDED SEARCH TASKS

A thesis submitted for the degree of Doctor of Philosophy

by

Timothy Cribbin, BSc.(Hons.), MSc.

School of Information Systems, Computing and Mathematics

Brunel University

September 2005

ABSTRACT

In this dissertation we propose, test and develop a novel search interaction model to address two key problems associated with conducting an open-ended search task within a classical information retrieval system: (i) the need to reformulate the query within the context of a shifting conception of the problem and (ii) the need to integrate relevant results across a number of separate results sets. In our model the user issues just one high-recall query and then performs a sequence of more focused, distinct aspect searches by browsing the static structured context of a spatial-semantic visualization of this retrieved document set. Our thesis is that unsupervised spatial-semantic visualization can automatically classify retrieved documents into a two-level hierarchy of relevance. In particular we hypothesise that the locality of any given aspect exemplar will tend to comprise a sufficient proportion of same-aspect documents to support a visually guided strategy for focused, same-aspect searching that we term the aspect cluster growing strategy. We examine spatial-semantic classification and potential aspect cluster growing performance across three scenarios derived from topics and relevance judgements from the TREC test collection. Our analyses show that the expected classification can be represented in spatial-semantic structures created from document similarities computed by a simple vector space text analysis procedure. We compare two diametrically opposed approaches to layout optimisation: a global approach that focuses on preserving the all similarities and a local approach that focuses only on the strongest similarities. We find that the local approach, based on a minimum spanning tree of similarities, produces a better classification and, as observed from strategy simulation, more efficient aspect cluster growing performance in most situations, compared to the global approach of multi-dimensional scaling. We show that a small but significant proportion of aspect clustering growing cases can be problematic, regardless of the layout algorithm used. We identify the characteristics of these cases and, on this basis, demonstrate a set of novel interactive tools that provide additional semantic cues to aid the user in locating same-aspect documents.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Professor Rob Macredie, for his generous and helpful advice and enduring patience over the course of this long journey. I would also like to thank my family and friends for supporting me over the years and for tolerating my obsession with this dissertation, which, I regret, has lead to my absence at many occasions that I would have liked to attend.

CHAPTER 1: INTRODUCTION

1.1. Introduction

This dissertation proposes, evaluates and develops a new approach to the general problem of answering an open-ended question using the results returned from an on-line information retrieval system. Such questions are traditionally difficult to answer because the problem scope may be broad and complex, consisting of multiple related ideas or aspects. Moreover, the searcher's mental model of this problem space will tend to evolve as the search progresses (Bates, 1989), due to encounters with unexpected information that generate new perspectives on the problem (O'Day and Jeffries, 1993).

The remainder of this chapter is organised as follows: We begin by describing how ill-defined information needs can be classified as either narrowing or expansive in nature. We then explain why these needs are difficult to satisfy using classic (query driven) information retrieval systems. We outline how query expansion tools can support the problem of specifying a query but explain why they are more useful for simple, narrowing needs than for more complex and expansive information needs. We then outline a solution path by introducing an interaction model originally proposed by Leuski (2001) that simplifies the process of isolating relevant documents within a retrieved document set by representing the inter-document similarity structure as a spatial-semantic visualization. As relevant documents tend to form a cluster within the visualization (Leuski, 2001; Allan et al., 2001), this allows a simple strategy whereby the cluster of relevant documents is 'grown', by following relative proximity cues from one or more known relevant exemplars. We argue that this interaction model might be extended to support an open-ended search, where the searcher must both discover distinct aspects of relevance and grow multiple distinct clusters of documents associated with each aspect, by allowing such a search to take place within a single, consistent visual representation. Potential issues are outlined and research questions presented. Finally, an outline of the methodological approach is presented.

1.2. Background and motivation

Despite a rapidly changing information landscape and user population, the classic information retrieval (IR) model is still a popular means of access, particularly for large or dynamic document collections. In this model documents are represented and accessed in concrete terms. There are many advantages to this approach, particularly if source documents are available in electronic format. The process of indexing and retrieval requires no human intervention and can therefore be automated, creating a fast and highly scaleable system. The method of document access, specification of a logical query statement, is also optimal for certain retrieval tasks (e.g., finding known or well-defined targets).

In the classic model, searches are conducted by specifying a logical statement of information need or query that the system then matches against an index of terms linked to documents. The system returns a list of documents that match or closely match the query. Normally this list is ranked in order of degree of match or relevance. In the standard interaction cycle the query is refined by changing, adding or removing terms, until the top results (i.e., the first page) contains the desired document or an acceptable number of relevant documents.

The classic model remains a highly effective method of satisfying well-defined needs. For example, the tasks of retrieving a known article or answering a closed question such as "How old was Benjamin Franklin when he died" are easily accomplished. This is because the key facets of a correct response are known and can be readily specified.

Yet, many information needs are initially only partially defined in the mind of the searcher (Belkin et al., 1982). Before the searcher can specify their need in logical terms, they must first refine their conception of the underlying problem (Taylor, 1968). Such needs are hard to satisfy in a classical system (Belkin et al., 1982) because a fundamental mismatch exists between the system requirements for a logical description of relevance criteria and the ability of the searcher to form such an expression; the searcher possesses gaps in their knowledge rather than a well-defined need and these gaps can only be bridged by exploring the contents of a suitable document collection.

Newby (1998) defines three types of information need. In addition to targeted or well-defined needs of the kind already described, a search can also be driven by both narrowing and expansive needs. In both cases the problem is ill-defined, but the two need types are qualitatively different. A single search episode, particularly with respect to an open-ended question, can involve the pursuit of all three of these types of need in varying combinations (Newby, 1998; Bates, 1989).

Narrowing needs occur when a searcher is looking for something in particular and will recognise it on sight (Toms, 1998), but cannot define this target (or set of related targets) precisely a priori. Although unable to express their need precisely, they are able to make incremental relevance judgements when presented with a series of imperfect but converging options (Newby, 1998). For this reason, hierarchical classifications or interactive menus are particularly useful for this kind of need as they allow the user to navigate through a sequence of ever more specific options until they arrive at a suitably focused and relevant document or sub-collection of documents.

Instantiations of narrowing needs would comprise closed or at least highly constrained questions. For instance, the searcher may want to find a simple explanation of how to sort a variable array. Using the Open Directory ™, a web directory, they would start by selecting the "Computers" node, then "Algorithms" and finally "Sorting and searching" to produce a short list which contains a high proportion of potentially relevant links.

Expansive needs are fundamentally opposed to narrowing needs in that the searcher is trying to broaden their knowledge within a topical domain (Newby, 1998), rather than refine their specification of a certain target. The motivating problem or question is open-ended and so the full range of relevant facets that define the problem may not be initially apparent to the searcher. Further, these facets may be quite diffuse as well as convergent (relevant documents will not necessarily be closely related in their content), whereby the solution to the problem can be divided into a range of distinct, yet topically-related aspects (Muresan, 2002; Muresan and Harper, 2004). O'Day and Jeffries (1993) use the term progressive searching to describe how searches can take in a broad range of aspects within the scope of the general motivating problem. Bates (1989) uses the notion of evolving/berrypicking search to describe this process where the query being pursued becomes a dynamic and shifting entity, rather than a static goal, with much relevant information being discovered accidentally or incidentally (Bates, 1989; Toms, 1998) rather than as a result of any systematic strategy. In a manual environment this kind of searching can often manifest as browsing the shelves or area scanning (Bates, 1989). This strategy is highly data-driven and thus requires a highly structured environment where rich patches of broadly relevant items can be easily identified and where it is possible to search in a non-linear fashion, making connections between similar yet non-proximal information items. The optimal electronic environment for such a search pattern is currently hypertext (Newby, 1998; Toms, 1998). Hypertext allows for the expression of a rich semantic topology; an embedded link can relate even a relatively minor concept within a document to another that describes that idea more completely.

An instantiation of an expansive need might be to form a general literature review on a topic or to answer an otherwise open-ended question. For instance, the searcher may want to learn about the full range of different garden plants that might suit different parts of his garden. In contrast to a narrowing need, the relevant facets that describe specific groups of plants may be diffuse; there may be several aspects to the problem. The searcher might seek both annuals and perennials, plants that like light and shade. Within a hypertext network, links from a single page or document can be made to a diffuse range of related concepts. The potential is there for the user to follow unorthodox paths that lead to accidental or chance discoveries (Toms, 1998). It is through the making and following up of chance discoveries that the user is able to define and resolve their ill-defined problem.

Newby (1998) stresses that these needs are not exclusive and can occur in various combinations; it is not uncommon for a search episode to consist of all three types. This view of information seeking is consistent with other well-known models of the search process (e.g., Bates, 1989; Xie, 2000, 2002) and is likely to be particularly true for cases where the question, or information problem is open-ended in nature.

As we have discussed, narrowing and expansive needs are ill-defined and thus favour browsing strategies whilst targeted needs are well-defined and most effectively satisfied through query specification. The interface requirements for browsing and searching are quite distinct, however. Browsing requires a rich semantic structure that can be efficiently scanned and navigated, whereas searching requires the facility to retrieve documents that are relevant to a specifiable need. We argue that answering an open-ended question requires an interface that can support both browsing and searching within a consistent context. Modern IR systems can support browsing as well as searching, albeit in a somewhat counter-intuitive way, providing appropriate strategies are applied. This may be necessary if suitable, structured interfaces are not available. We now discuss the benefits and limitations of strategies available to the user of a classic IR system involved in this kind of complex, evolving search.

1.3. Information retrieval strategies for complex needs

The classic IR interaction model was not designed to help users resolve ill-defined problems. In fact, standard interfaces to IR systems are based upon a model that is an extrapolation of database retrieval. In database retrieval, searches are typically fact-oriented, highly structured and performed by users who are expert in the database schema. In information retrieval, however, information needs tend to be less well defined (Belkin et al., 1993). Expert searchers (e.g., intermediaries) have adapted to the constraints of this model by developing a number of formal strategies that simplify the task of specifying a query for a non-trivial need. Many of these tend to be quite algorithmic in process, whereby the searcher breaks the problem down into distinct facets (concepts) and refines these sub-queries separately, by ORing synonymous or related terms, before systematically recombining them. For example, in the building blocks strategy, once all the facets are specified the intersection or conjunct of the sub-sets retrieved by each of these queries is found by ANDing them together to form a single query. The successive fractions strategy is slightly different in that the major facets are combined first then, if necessary, further, less specific facets are joined to the query until the desired recall and precision levels are achieved. A variety of other variations on the facet strategies exist (see Harter, 1986).

Such strategies are collectively referred to as analytical strategies as they are planned and algorithmic in execution. Facets are identified and defined before the interaction begins with the IR system. The retrieval sets that result from submitted queries tend to be evaluated more in terms of set size (e.g., a manageable or acceptable number of references) rather than content. These strategies are best suited to formal information environments (e.g., a library) where documents are catalogued and represented as semantically consistent bibliographic meta-data rather than their literal content, where the terms used to describe the same concepts can be highly variable (Furnas et al., 1987). The use of controlled vocabulary means that, with expert knowledge of subject descriptor schemes, precise and exhaustive facet expressions can be constructed relatively quickly and easily.

Hence, the classic IR model and the analytical strategies developed to exploit systems based on it are a legacy of the early days of online searching when searches typically took place in formal, structured environments and where expert intermediaries were on hand to help clients to define their problem and formulate suitable queries. Further, in the early days on-line IR access was relatively expensive so casual exploration of online resources was discouraged.

However, even in formal, mediated information environments, experts have since recognised there is a need for more informal, exploratory strategies that place a greater emphasis on interaction with, and learning from, document content (Marchionini, 1995). These strategies are particularly important if the information problem is poorly defined; perhaps just a general topic that must be explored, or idiosyncratic to the extent that the facet structure may not be easily defined, even by a search expert. There are two such strategies of particular interest here: pearl growing and interactive scanning.

The pearl growing strategy (Markey and Cochrane, 1981; Harter, 1986; Marchionini, 1995) involves examining the attributes of known relevant documents to build an exhaustive query. Relevant examples (the pearl) may be brought to the task or discovered through a brief search (Harter, 1986) where a simple query (e.g., 2 or 3 terms) can be used to identify one or two relevant 'example' documents. The query developed from the initial sample of exemplars is iteratively refined using new exemplars from each subsequent search. Hence, the layers of the pearl (known relevant documents) are slowly grown around the initial core by further specifying the facets of the query with key terms extracted from the exemplars. Given the bottom up approach, pearl growing has limited use for expansive searching, and is most suited to narrowing searches where the object of the search is quite specific (relevant documents are conceptually quite similar).

Interactive scanning (Harter, 1986; Marchionini, 1995) is another useful interactive strategy, particularly if there are no known relevant documents to use as a starting point. In contrast to pearl growing, this is a top down approach that begins with a tentative or high-recall query that is sufficiently broad to capture documents discussing most facets of the problem, albeit along with a large number of other, non-relevant items. The user examines the results of the initial query (or at least a top ranks sample of it) and notes key facets that appear relevant to the problem. A number of successive searches are then performed using these facets. In this sense pearl growing may be used in conjunction with this strategy to grow these facets from the contexts in which they are discovered. The top down nature of interactive scanning means that it is possible to identify quite diverse facets relating to the problem, hence this can also be a useful strategy for an expansive need.

Even these less formal analytical strategies can be too complex or involved for novice users to use in unsupervised settings (Harter, 1986; Marchionini, 1995). They involve continuous attention, and careful judgement of which terms are key. Further, many modern, particularly web-based, resources are indexed automatically from raw text and may have little or nothing in the way of consistent meta-data. The vocabulary mismatch between relevant document representations compounds the problem of facet specification (Furnas et al., 1987). In the case of a top-down exploratory approach like interactive scanning there is also the extra burden of building a model of facet structure (how terms logically relate to each other), a model that may well be complex and that will evolve and shift in structure as the search progresses (Bates, 1989; O'Day and Jeffries, 1993).

There have been efforts to support both of these kinds of interactive strategy. Such systems, however, have tended to focus on one strategy or the other and are thus quite distinct in nature. For instance, interactive scanning can be supported by document organisation techniques such as clustering (Hearst and Pederson, 1996) or spatial-semantic visualization (Lin, 1997; Wise et al., 1995; Skupin, 2002) that can provide a high-level overviews of themes within a given collection. These techniques, whilst superficially impressive have not been empirically evaluated as expansive search tools, due to the inherent methodological issues associated with measuring efficacy in unstructured search tasks (although see Chen et al., 1998). There is, however, a longer tradition in IR of supporting the pearl growing strategy (narrowing needs) with techniques that infer the user's intentions by means of document relevance feedback.

Automatic or semi-automatic query expansion (QE) has received a lot of attention within the IR field. Key terms are extracted automatically using statistical analyses that compare the content of known or assumed relevant documents to other documents. There are a number of different approaches. The conventional approach is for the user to provide explicit feedback on retrieved documents by marking relevant items. Relevance feedback approaches generally lead to good improvements in query precision (see for example, Salton and Buckley, 1990) when used correctly (i.e., the user indicates a sufficient number of relevant examples), although studies show that non-expert users are often reluctant to or fail to understand the importance of providing sufficient examples (Hearst, 1999).

Local feedback (Attar and Fraenkel, 1977), sometimes referred to as pseudo relevance feedback, avoids the requirement to evaluate and mark relevant documents by simply assuming that the top ranked documents are likely to be more relevant than lower ranked ones and examining the differences between these two sub-sets. Local feedback can be effective, although performance tends to be highly dependent on the precision of the initial query (see Xu and Croft, 1996; Xu and Croft, 2000). Incremental feedback (Aahlsberg, 1992) is an alternative approach that reduces the burden of identifying exemplars by presenting just one candidate, the next most relevant unseen document, at each QE iteration. This also alleviates another problem experienced by QE users – confusion and disorientation caused by constant re-ranking of both seen and unseen documents (Aahlsberg, 1992).

QE approaches also vary on the degree to which the user can control the reformulation process. Fully automatic QE approaches hide the query reformulation process completely from the searcher, whilst semi-automatic approaches let the searcher select candidate terms before they are added. Koenemann and Belkin (1996) have shown, for instance, that allowing users to 'filter' candidate terms results in fewer iterations being needed to achieve an optimal query.

Despite these advances, most QE techniques are of limited use for complex, evolving needs because the process is, by design a tool for narrowing a specific query – filtering relevant from non-relevant documents and exhaustively defining the shared and convergent facets that define relevance (Belkin et al., 2000). Even though the burden of explicit facet identification and specification is alleviated, there is the implicit assumption, as with many analytical strategies, that there is a single, optimal response, which is the intersection of the document sub-sets relevant to each specified facet.

If the information need is complex, however, facets may form multiple, diffuse intersections where each sub-set of relevant documents is semantically and thus lexically distinct from the other sub-sets; each sub-set therefore comprises documents that discuss or refer to a distinct aspect of relevance to the problem and thus have more in common with each than they do to other relevant documents (Muresan and Harper, 2004).

If the documents marked as relevant are diffuse in this way, the QE algorithm is likely to be unable to select a good set of expansion terms that are both common to most known documents yet discriminating enough to filter out non-relevant documents. Furthermore, if the aim is to search expansively, new aspects of the problem will emerge as the search proceeds (Bates, 1989; O'Day and Jeffries, 1993). Changing the focus of the query mid-way through a QE cycle would violate the core principle of query expansion (Bates, 1989), which is that there is a specific thing or set of closely related things the user wishes to isolate from the database – a one-time query (Bates, 1989) requiring a linear process of query refinement. If the user's query were to shift significantly at any point then multiple new examples would be required before the actual query statement converged on the new focus and, as a consequence of this shift, it is likely that items that were relevant to the earlier query would drop out of the top ranks of the retrieved set.

Searching for distinct aspects using relevance feedback is therefore likely to require multiple discrete episodes each focusing on a specific aspect. As the outcome of each aspect search will be a separate retrieval set and it is the responsibility of the user to understand the relationships (e.g., overlaps) between these sets. More importantly, this interaction model is counter-intuitive to the notion of an evolving query as semantically distinct new ideas (query shifts) triggered during a QE episode would need to be placed on hold and followed up during later QE cycles.

1.4. The cluster growing strategy

In light of these problems, a promising alternative to query expansion for complex/evolving search is inspired by the work of Leuski (2001), whose Lighthouse interface turns the task of locating similar documents into a simple visual search task (Leuski, 2001; Allan et al., 2001). Leuski's interaction model sees the searcher performing a brief search for their topic of interest. The system takes the top 50 documents from the retrieved documents and constructs a spatial-semantic model based on a model of inter-document similarities of these items. In a spatial-semantic model, documents are represented as nodes or points in 2D or 3D space and their relative semantic similarity is conveyed by their proximity in this space. Inter-document similarity is computed automatically by measuring the overlap between term frequency vectors of each pair of documents (see Salton and McGill, 1983).

The first strong assumption of spatial-semantic visualization is therefore that relevant documents will have more similar term usage profiles. The rationale for this is rooted in the cluster hypothesis of IR (van Rijsbergen, 1979), which states that similar documents tend to be relevant to the same requests. A second strong assumption of spatial-semantic visualization is that this inter-document similarity model, or at least the sub-set of this model that relates to relevant documents, can be effectively conveyed in visual space. The truth of the specific hypothesis for spatial-semantic modelling, at least for simple topics, has been supported by the work of Leuski (2001) and others (Rorvig and Fitzpatrick, 1998; Sullivan and Rorvig, 1998). Work with discrete clustering algorithms also supports the idea of organising relevant documents according to measures of their lexical similarity (Hearst and Pederson, 1996; Wu et al., 2001).

Leuski (2001) evaluated the performance of a strategy that begins with the searcher browsing the conventional ranked list from the top. When a relevant document is encountered the searcher selects this item in the list and switches their attention to the spatial-semantic visualization. The relevant node, which is highlighted, is used as the seed or anchor point from which to find further relevant items. The searcher examines nodes in proximity order. We call this the cluster growing strategy, because the user literally grows a cluster of relevant items guided by spatial-semantic cues within the visualization. Hence, cluster growing is a visual surrogate of QE and, more fundamentally, the pearl growing strategy. When further relevant documents are found, these are also marked. As more relevant documents are encountered, the distribution of the relevant cluster becomes more apparent, making it easier to find further items. Leuski (2001) found that the strategy was just as efficient as conventional QE, and sometimes more so, with additional benefit that there was no iterative cycle that restructured the view of potentially relevant documents. This is important because earlier studies of relevance feedback have found that searchers can become confused or disoriented by changes in document order that tend to happen at each successive iteration (Aalbersberg, 1992).

In this thesis, we view complex information needs as being composed of multiple aspects that are specific instances of the problem but relatively distinct from each other. According to the spatial-semantic principle, when a document containing multiple aspects of the same complex topic is visualized, documents discussing the same aspects should cluster more coherently than those that are relevant, but discuss other aspects. Hence, rather than building a single cluster within the visualization, the user is able to build a number of aspectually distinct clusters. A significant benefit of this approach is that the external model of documents, seen and unseen, that is presented to the user remains constant throughout the whole process.

This consistency of context is important for a complex, evolving search. Studies of hypertext browsing have advocated the use of context maps to support the formation of a mental model, and avoid disorientation between changes in focus (see Vicente and Willeges, 1988; Stanney and Salvendy, 1995; Hook et al., 1996). Studies also show that spatial-semantic models of large document collections appear to provide useful thematic overviews (Wise et al., 1995; Chen et al., 1998; Skupin, 2002) that can facilitate expansive browsing, for example finding an interesting item within a collection (Chen et al., 1998).

However, whilst evidence supports the use of spatial-semantic visualization to facilitate focused retrieval in small *ad hoc* document sets on the one hand, and general, expansive browsing in large collections on the other, it is unclear as to whether visualizations of more complex and somewhat larger *ad hoc* retrieval sets could support the kind of multi-aspect, evolving search required to resolve an open-ended question.

In the next section we propose that the cluster growing strategy can be adapted to support more complex, evolving information problems. We describe the essence of our interaction model and define the scope of the problems relating to this model that will be dealt with in the remainder of this dissertation.

1.5. Aspect cluster growing

Our general thesis is that Leuski's (2001) interaction model can be extended to support the resolution of an open-ended question by allowing multiple aspects to be searched, using the cluster growing strategy, within a consistent spatial-semantic context. An example of the kind of open-ended question considered might be:

"What are the most significant achievements made using the Hubble space telescope since its launch?"

This represents an information need that is certainly complex – relevant answers range from estimations of the age of the universe to the effects of gravitational lenses. If we assume the searcher is not particularly familiar with the topic then it is also an expansive, evolving query.

We propose a model where, following a high-recall (broad) query, the user explores the content of the retrieved set and identifies a range of distinct aspects of the relevant topic. At some point following the identification of an aspect, the user attempts to locate the sub-set of other available (retrieved), aspect relevant documents. These focused searches are achieved by applying the cluster growing strategy, beginning at the location of the document in which it was discovered. The relative proximities of nodes to the known exemplar (i.e., the pearl) are the cues that guide the order in which unseen documents are viewed. From hereon, we term this sub-task of locating same aspect documents within the context of the retrieved set visualization as *aspect cluster growing* and the strategy of following relative proximity cues as the *aspect cluster growing strategy*. Our revised interaction model is described in greater detail in section 2.2.

The focus of this work is on the ease with which, given a known exemplar document, the application of the aspect cluster growing strategy leads to successful and efficient retrieval of other aspect-relevant documents. We do not directly consider how these initial exemplars are discovered. These discoveries could equally result from browsing and marking sample of the top ranked documents, as in Leuski's (2001) original model, or by browsing the visualization directly (Chen et al., 1998).

Also outside of the scope of this work is formal consideration of the order in which the sub-tasks of aspect discovery and single aspect retrieval occur. Given the evolving nature of ill-defined and complex information problems (Bates, 1989; O'Day and Jeffries, 1993) this is likely to be a continuous cycle where new aspect instances are discovered at various stages of interaction with the retrieved set.

Hence, the primary focus of this work is on the process of aspect cluster growing and how to generate the structures required to support this strategy. In the next section we will outline the general problems that will be investigated in subsequent chapters.

1.6. Problem definition

Our interaction model makes two strong assumptions. The first assumption is that it is possible to automatically classify documents into a two-level relevance hierarchy based on the structure of the similarity matrix computed from text analysis of the retrieved document set. In other words we assume that document similarities effectively partition same aspect documents from those that are both non-relevant and, to a lesser degree, relevant but discuss different aspects of the topic.

The second assumption is that it is possible to map this modelled structure to a spatial-semantic layout using an unsupervised layout algorithm. This is not necessarily a given, as earlier work in document retrieval set clustering has demonstrated (Wu et al., 2001). Wu et al. (2001) studied the use of clustering to organise a complex topic within retrieved document sets. They found that although relevant documents tended to gather within one or two main clusters (within a six or seven cluster solution), this clustering did not effectively partition documents relevant to the same aspect(s) of relevance.

More pertinently, to date there has been no formal study of topical aspect clustering within retrieved document sets that are visualised using the spatial-semantic metaphor (although see Swan and Allen, 1998 for a related approach). Most work of this kind has focused on the extent to which these visualizations are able to simply partition relevant from non-relevant documents (e.g., Leuski, 2001, Rorvig and Fitzpatrick, 1998).

Whilst Leuski (Leuski, 2001; Allan et al., 2001) conducted a rigorous formal study of cluster growing efficiency for simple topic retrieval, even comparing the use of 2D versus 3D visualization, he only offers anecdotal evidence to demonstrate the potential with respect to the aspect level retrieval. In their final conclusions, Allen et al. (2001) comment on the difficulties associated with applying their strategy to complex topics and discussing the need for future research and development to adapt Lighthouse to support this kind of task.

Leuski (2001) does propose various interactive tools that might dynamically augment the static spatial structure in response to relevance feedback, to support aspect cluster growing. However, these are not formally evaluated either in the form of simulated or actual user studies. Furthermore, these tools are essentially simple adaptations of tools known to work for simple topic retrieval situations rather than being developed on the basis of a detailed understanding of aspect clustering behaviour within spatial-semantic visualizations.

In this thesis, we take the view that prior to developing interactive tools, it is important to attempt to optimise the spatial-semantic layout process so that, as far as possible, exploration and aspect cluster growing can occur as a seamless browsing process, much like browsing the shelves of a library, rather than being an activity that is heavily dependent upon secondary interaction tasks such as making document relevance judgements and identifying terms for query reformulation. The interactive tools that we eventually develop (Chapter 6) are based on our acquired knowledge of the factors or circumstances that tend to lead to failure of aspect cluster growing using spatial cues alone.

In the next section we translate the problems discussed into the three general research questions that form the main strands of this dissertation.

1.7. Research questions

The feasibility of our interaction model rests upon the assumption that it is possible to generate, using unsupervised procedures, spatial-semantic visualizations of retrieved document sets that effectively organise documents according to the structure of the intended topic. In particular, for the sake of the aspect cluster growing strategy, we require that documents relevant to the same aspects of the relevant topic form cohesive visual clusters. To this end Chapters 3, 4 and 5 provide a series of analyses that (i) show that inter-document similarity measures effectively classify the relevant topic, (ii) compare different approaches to spatial-semantic visualization for representing this classification and (iii) determine the conditions under which the aspect cluster growing strategy fails.

We examine two information problems or open-ended questions of the kind presented at the start of section 1.3. For each question we retrieve a set of documents from a larger collection using a simple query, based on the question or topic description, designed to capture the full breadth of the topic (from many aspects). As such the retrieved document sets are generally complex in their topical structure, containing many aspects of the relevant topic as well as many non-relevant topics. Amongst the relevant documents, there is likely to be considerable variation in, for instance, the extent to which specific aspects are discussed (i.e., document frequency), the consistency with which the same aspects are discussed within documents and the breadth of topical focus of specific documents (i.e., whether they focus on just one or several aspects of the topic).

Given the nature of these scenarios and the requirements of our interaction model, we ask three related questions:

1. To what extent can a standard text analysis procedure model the general semantic structure expected by our interaction model and particularly the low-level structure required by the aspect cluster growing strategy?

2. Given an adequate semantic model, which approach to spatial-semantic layout best preserves the general and, in particular, the low-level structure expected by our interaction model?

3. Under what conditions does the aspect cluster growing strategy tend to fail and how can we use this knowledge to guide development of interactive support tools?

The process of answering these questions is incremental in nature and the approach used to answer each question is, to a great extent, dependent upon the outcome of the previous stages of analysis. A detailed overview of the issues surrounding each question follows.

The **first question** relates to the validity of the cluster hypothesis (van Rijsbergen, 1979), which predicts that documents relevant to the same requests (queries) will tend to be more similar in their content than they are to other documents. For our interaction model to work it is necessary, if not sufficient, that the semantic models of complex *ad hoc* document sets should effectively classify documents to two levels of topical relevance. In other words documents that discuss the same general topic (are generally relevant to the question) must tend to be more similar to each other than they are to other, non-relevant documents. In turn, documents that discuss the same aspect of the relevant topic must tend to be more similar to each other than they are to other relevant documents.

We require, therefore, that aspect similar documents will be generally the most lexically similar documents within a retrieved set. If this is the case then the aspect cluster growing strategy will be efficient in that the user wastes a minimal amount of time examining non-relevant items and will not get overly distracted by relevant documents discussing different aspects. That said, given the evolving nature of the search a degree of accidental discovery of new distinct aspects is desirable, particularly as the user begins to exhaust their search as they approach the edge of the current aspect cluster. Hence, it is also desirable that the generally relevant items should be more similar to each other, even if they discuss different aspects, than they are to non-relevant items.

There is some evidence to suggest these requirements can be met (Muresan, 2002; Muresan and Harper, 2004). In their experiment, the authors examined a medium-sized collection (747 documents), mostly comprising documents known to be relevant to six defined topics, where aspects of each topic had also been defined and document relevance judged in each case. In addition to known relevant documents, the document set was 'polluted' by non-relevant items that were frequently retrieved by topical queries. The semantic model was created using a standard text analysis method where inter-document similarity is measured by calculating the correlation in term (words) usage. Three conditions were examined: all document similarities, same-topic document similarities and same-aspect document similarities. The authors found significant general trend in mean similarity in the expected direction (Muresan, 2002).

However, the semantic properties of the studied document set was highly contrived and somewhat different to the structure we would expect within an *ad hoc* retrieval set. Muresan's (2002) set comprised several well-defined and distinct topics. Whilst there would have been some semantic overlap between these different topic classes, in the majority cases one would expect inter-topic similarity to be relatively low. In a real high-recall, *ad hoc* retrieval set, however, the distinction between relevant and non-relevant may be less clear, given that all documents will be, to varying extents, relevant to the same request. Thus the set will comprise not only relevant documents and clearly non-relevant items but, between these sub-sets, there will be a third sub-set of documents that are somewhat related to relevant items but not relevant to the user's information need.

Therefore, to establish the feasibility of our interaction model, we first need to consider whether the high occurrence of similar yet not relevant items will distort the neat classification of documents by topic and aspect that was observed by Muresan (2002). To this end, in chapter 3 we develop a test bed of two complex topical scenarios, each comprising an *ad hoc* retrieval set retrieved using a broad, high-recall query. We then conduct a similar analysis to that of Muresan (2002) for both scenarios where we measure the degree of topical classification for the single relevant topic. Our method of analysis is adapted accordingly.

Assuming that the required classification of documents is present within such semantic models, the **second question** concerns layout or how best to represent this all-important structure as a visualization.

Given the fundamental principle of the spatial-semantic metaphor – that the proximity between a pair of nodes maps directly to the degree of similarity between the documents being represented - the most natural, and therefore common approach to spatial-semantic visualization is some form of multi-dimensional scaling (MDS). A form of MDS called force-directed placement is used by Leuski (2001) to create the Lighthouse visualization.

MDS algorithms work by accepting a matrix of inter-document proximities as input and using some iterative procedure that attempts to locate documents as points or nodes into a spatial configuration where the inter-node proximities match, as closely as possible the input proximities. Thus, these algorithms treat layout as a global optimisation problem, where all inter-document relationships are treated as equally important and preserved to the best extent.

A fundamental obstacle, however, for any approach to spatial-semantic layout is that a considerable amount of structural information represented in the semantic model will inevitably be lost due to the dimension reduction process. To understand the reason for this, let us first consider the problem of plotting a matrix of inter-city proximities to 2D visual space. This is a simple problem with a perfect solution because the proximities were calculated from an origin space of equal dimensionality. In contrast, a semantic model will normally have a high dimensionality equivalent to the number of unique terms (the vocabulary) used to represent the content of each document. After applying pruning heuristics to remove terms that are likely to be poor discriminators (see Salton and McGill, 1983), even quite small document collections are likely to be defined by a term-space of several thousand dimensions. Projecting such a space onto two- or three-dimensions results in considerable compromises in node location, because there are insufficient degrees of freedom to position every node at the appropriate relative distance to all other nodes. In mathematical terms, misplacements occur because dimension reduction gives rise to many situations where the triangle inequality is violated - where for three given nodes, the distance from A to C becomes greater than the sum of the distances between A to B and B to C. Such inequalities will often result in compromised location of nodes, or misplacements, where nodes that are closely associated in the semantic model may be located far apart in the output space or, in contrast, quite unrelated nodes may be located close together. The likelihood of misplacement is also intimately tied to the number of nodes that must be mapped, whereby the complexity of layout optimisation increases exponentially with set size.

Taking the view that information loss during dimension reduction is inevitable, it is important to select a layout algorithm that focuses on preserving and representing the elements of the underlying topology that are most relevant to the information need of the user (Skupin and Fabrikant, 2003). For our interaction model, we are most concerned with emphasising the relationships between same aspect documents. Given our hypothesis that these tend to be the stronger relations within the semantic model, we ask whether it is more appropriate to apply a layout algorithm that emphasises local (the strongest inter-document similarities), rather than global optimisation. In chapter 2, we discuss algorithms that follow the local approach along with previous work that has applied such techniques to document visualization. In chapters 4 and 5, using our topical scenarios (from chapter 3), we examine the relative efficacy of local and global optimisation techniques to create spatial-semantic models that retain the desired structure and therefore optimise the efficiency of the aspect cluster growing strategy.

Given that compromises in layout are inevitable, regardless of the algorithm applied, our **third question** asks why spatial-semantic cues fail to support aspect cluster growing in certain situations and what we can do to accommodate such failures. In this dissertation, a failure of spatial-semantic cues is broadly defined as occurring when few or none of the nearest neighbours of an exemplar discuss the aspect that captured the user's interest within that document. Such a failure would cause a user following the aspect cluster growing strategy to browse through an unacceptable number of non-relevant items before encountering a relevant item.

By building a model of the factors associated with failure situations we aim to develop appropriate interactive tools that will provide the kind of visual-semantic cues required to dynamically reintroduce lost structural information into the visualization. It is seen as important that the *ad hoc* cues can be elicited in a way that minimises the required input (particularly with respect to query specification) from the user.

There are two general reasons why spatial-semantic cues might be insufficient to afford efficient aspect cluster growing from a given exemplar document. First, this may occur simply because there is little or no correlation between the term vectors of the exemplar and aspect similar documents, in other words the aspect relationship is poorly encoded within the underlying semantic model. This situation would arise from limitations of automatic text analysis, e.g., vocabulary mismatch (Furnas et al., 1987), and is beyond the scope of this thesis.

Instead, we focus upon a second general situation where although the association between the exemplar and other aspect relevant documents may be quite strong, searching nodes in proximity order proves to be a relatively inefficient strategy. This situation may occur because the immediate locality around the exemplar is crowded with documents discussing various related concepts in addition to the aspect of interest. Hornbaek and Froekjaer

(1999) found this was a common problem experienced by users when searching and browsing within a spatial-semantic model of a heterogeneous document collection. Alternatively, compromises in the spatial-semantic layout process may cause the exemplar to become relatively isolated from the main aspect cluster. Such trade-offs might occur when the exemplar is relatively focused on the specific aspect of interest whilst other aspectually related documents tend to be more 'topical' discussing several aspects of the general topic (see Muresan, 2002). The converse situation, where the exemplar is highly topical and the aspect relations highly focused, would also present similar problems. Furthermore, we anticipate that such problems would be compounded if the other topical aspects discussed by the related documents happen to be represented relatively more prominently within the set (in more documents) than the user's current query.

We predict that whichever layout algorithm is used, however optimal it is, it is likely to be necessary at some stages of the search process to support the user in their aspect cluster growing by supplementing the static spatial-semantic cues with dynamic cues. The reasons for the failure of spatial-semantic cues are likely to be highly variable and difficult to model exhaustively. Hence, our more realistic aim is to model the distinctive features of problematic exemplars and to apply this model to the development of simple to use, but powerful interactive tools that provide additional cues that can help the user to resolve such ambiguity.

1.8. Summary of research goal

Having defined our three main problems, our general research goal can be summarised as follows:

To develop and evaluate the potential utility of a novel interaction model to support the answering of an open-ended question using documents retrieved by a high-recall query.

To recap our specific research questions are as follows:

1. To what extent can a standard text analysis procedure model the general semantic structure expected by our interaction model and particularly the low-level structure required by the aspect cluster growing strategy?

2. Given an adequate semantic model, which approach to spatial-semantic layout best preserves the general and, in particular, the low-level structure expected by our interaction model?

3. Under what conditions does the aspect cluster growing strategy tend to fail and how can we use this knowledge to guide development of interactive support tools?

1.9. Approach

Our approach is to follow the work of Leuski (2001) and others (e.g., Rorvig and Fitzpatrick, 1998; Sullivan and Rorvig, 1998; Swan and Allen, 1998) by examining our three questions within the framework of the Text REtrieval Conference (TREC) test collection (Voorhees and Harman, 1997; Voorhees and Harman, 1998). Specifically, we utilise topics and relevance data from the interactive track. The associated topics simulate an open-ended search task and provide benchmark relevance data that allows IR system evaluation to proceed effectively without the need for user studies. Each topic comprises an open-ended question (the topic description), definitions of distinct aspects of the topic known to exist within a specific source collection, and reasonably exhaustive relevance data linking one or more documents in the source collection to each defined aspect.

This resource allows us to create scenarios in which specific hypotheses, pertaining to spatial-semantic visualization structure and search strategy performance, can be tested using objective and reliable experiments. This is seen as preferable to user testing, where it is expensive and time consuming to test a sufficiently large sample and where measurements may be confounded by complex interactions between individual differences such as cognitive ability, reading speed, familiarity with visualization interaction, topic familiarity and general information retrieval experience. It also allows us to analyse clustering and the cluster growing strategy across a much larger, and broader range of situations and to reliably identify factors that may both facilitate and hurt the efficiency of our strategy. As we will show in chapter 6, we can use the knowledge gained from these factorial analyses to guide the development of appropriate interactive strategy support tools.

Using this approach, we examine topic-aspect classification, first within high-dimensional term space and simple discrete cluster solutions (Chapter 3), and then within spatial-semantic visualizations generated using different layout algorithms (Chapter 4). We will also analyse the efficiency of the cluster growing strategy (Chapters 3 and 5) for aspects using a simple adaptation of the strategy based evaluation (SBE) method developed by Leuski (2001) for his original investigations. This approach simulates user behaviour by calculating a search function, which predicts the order in which unseen documents are viewed (e.g., based on their similarity or proximity to a known, relevant document node). As indicated above, this method allows for objective comparison of cluster growing efficiency across different visualization schemes and different information problems without the need to consider the influence of user differences. It also allows us to identify factors that hurt aspect cluster growing efficiency for poorly performing exemplars (Chapter 5). This knowledge is then used to guide the design of interactive tools that are intended to help users to continue searching effectively when both inter-document similarity and spatial-semantic cues are sub-optimal (Chapter 6).

1.10. Structure of dissertation

In Chapter 2, we carry forward our three questions and consider the relevant literature in more detail. From this literature review we develop specific hypotheses for each question, which we carry forward to and test in our main analyses (Chapters 3 to 5).

In chapter 3, we develop our test bed of three topical scenarios based on two topics provided by the TREC-6 (Over, 1997) and TREC-7 (Over, 1998) interactive tracks. The selected topics differ significantly in the degree of document overlap between aspect sub-sets. For each topic a document set is retrieved using a high recall query and a text analysis is run to generate a semantic model. Relevant summary statistics for each of these elements of the test bed are provided. In particular, we determine whether the semantic models produced by text analysis classify documents into the structure required by the interaction model (question one). We also examine the effects of topical structure (overlapping vs. distinct aspects) and set size on classification efficacy. Discrete cluster solutions of our semantic models are presented and analysed to provide a preliminary insight into the feasibility of representing this structure effectively in low-dimensional space, to provide a comparison with earlier work that has examined aspect partitioning using this method (e.g., Wu et al., 2001) and to demonstrate the importance of interpreting the resulting structure of clustering/scaling algorithms within the context of the known structural properties of the underlying semantic model.

In Chapter 4, we move on to address question two. We begin the chapter by describing the creation of our spatial-semantic solutions using two distinct visualization schemes. We then present a visual analysis, using augmented versions of the solutions, that provides an illustrated overview of the comparative efficacy with which the schemes communicate key semantic features including the relevant topic, distinct aspects and also the discrete cluster structures produced by k-means in Chapter 3. Finally, we present a detailed quantitative analysis of the topic classification performance of both visualization schemes. Given our focus on the cluster growing strategy, we take a perspective where potential exemplar documents are treated as distinct cases or units of analysis (i.e., a relevant documents in a retrieval set). We compare the main and interactive effects of the visualization scheme and scenario specific attributes. The aim of this analysis is to compare, at a general level, the fidelity of relevant topic-aspect classification within the visualization schemes.

In Chapter 5, we draw conclusions with respect to question two and begin to answer question three. We simulate the aspect cluster growing strategy for both visualization schemes across all scenarios. This time, each exemplar case is considered once for each aspect it discusses, hence we measure strategy efficiency for all possible aspect cluster-growing 'micro-episodes' that might occur. Comparisons of the main and interactive effects of visualization scheme and topical scenario, allow us to determine the superior scheme for our interaction model. In the second part of the chapter, we begin to deal with question three. We then focus on the extent to which node-misplacements are responsible for poor aspect cluster growing performance, by comparing the relative performance of our strategy using pure similarity cues to the use of spatial-semantic cues. Finally, we perform an exploratory analysis to determine the correlates of overall poor exemplar cases i.e., aspect representatives that fail to make good exemplars, even when similarity cues are provided. This analysis helps us to form hypotheses with respect to the kinds of interactive tools that might best support aspect cluster growing.

In Chapter 6, we first reflect upon the findings of the earlier analysis chapters and discuss the implications for the design of interactive tools to support aspect-clustering growing. In the second part we complete our answer to question three by proposing a new conceptual approach called contexts in context that extends the principle of simple relevance feedback to address the observed limitations of the approach identified in Chapter 5.

In Chapter 7, we review the main outcomes of our analyses and summarise the contributions made. We then discuss the known limitations of this work and, on this basis, make suggestions for future work.

CHAPTER 2: FORMULATION OF APPROACH AND HYPOTHESES

2.1. Introduction

In chapter 1, we outlined the problems associated with answering an open-ended question using a classical information retrieval (IR) system and proposed an alternative interaction model. In this chapter we discuss the issues that need to be addressed in order to evaluate the general feasibility of the approach and to optimise its implementation. The outcome of this chapter is a set of formal hypotheses that will be used to answer our research questions. We propose two formal tests that will enable these analyses. We begin in this section, by restating the research problem, goal and questions of this dissertation before outlining the structure of the remainder of the chapter.

An open-ended question is characterised by an information need that is both complex, and so cannot be easily specified in a single, precise query and evolving, in that the relevant aspects of the problem are only partially known up front and tend to transpire during the course of the search process, as a result of accidental discoveries and inferences made as a result of interactions with retrieved document content and meta-data. We argued that this kind of searching is problematic in classic IR interfaces for several reasons. First, early queries are likely to be broad and ambiguous in nature. The lack of semantic structure afforded by the ranked list presentation format makes the process of identifying and defining key aspects of relevance within a long and diverse document set (i.e., interactive scanning) a tedious and cognitively demanding task. Second, as the search proceeds, the user must constantly reformulate their query as their perspective on the problem evolves and shifts. Third, as the query is shifting between conceptually diverse aspects of relevance, there is no single optimal set of results. Rather, relevant material is retrieved in bits and pieces across a number of query iterations. The onus is therefore on the user to collate this material, and to comprehend the significant relationships between information retrieved at different points during the process.

We proposed a potential solution to this problem in the form of an interaction model that uses spatial-semantic document visualization to organise documents, retrieved in response to a high-recall query, so that the proximities between document nodes are inversely proportional to their respective content similarity. Our model is based upon an existing model that was proposed and evaluated by previous work (Leuski, 2001; Allan et al., 2001). Central to this model is a strategy in which unseen, relevant documents are located by exploring unseen document nodes in order of relative proximity to the nodes of known relevant items. The notion that relevant documents are highly similar in their content similarity (compared to non-relevant documents) is based on a logical corollary of the cluster hypothesis of IR which states that similar documents tend to be relevant to the same requests (van Rijsbergen, 1979). Leuski's (2001) evaluation showed this to be the case for simple retrieval tasks; isolating relevant from non-relevant documents within the retrieved set. This 'cluster growing' strategy was equivalent to, or better than, following the initial ranked-list ordering and as good as automatically reformulating the query using relevance feedback.

Our thesis is that the cluster growing strategy will also prove effective for a complex evolving search, where multiple aspects of relevance must be searched. We see the user entering their initial, tentative query using just one or two 'topical' keywords from their ambiguous question statement and then growing multiple clusters within the context of a single, static visualization created from the retrieved documents. We refer to the process of growing a cluster of same-aspect documents, by following cues within the visualization, as the aspect cluster growing strategy.

The existing model (Leuski, 2001; Allan et al., 2001) has only been evaluated within the context of simple information retrieval problems (i.e., isolating a static set of closely related relevant documents), as opposed to a complex information need where relevant documents separate into many sub-clusters or 'pockets' of relevance (Leuski, 2001). We therefore need to evaluate the feasibility of our proposed model. Specifically, our proposed model makes two distinct assumptions that cannot be verified directly by previous experimental work. The first assumption is that it is possible to automatically compute a matrix of inter-document similarities from the texts of retrieved documents that classifies documents to two-levels of relevance (same-topic and same-aspect). The second assumption is that it is possible to visualise this represented structure as a spatial-semantic visualization.

Chapter 2: Formulation of methodology and hypotheses

The formal goal (presented in section 1.8) of this dissertation is *to develop and evaluate the potential utility of a novel interaction model to support the answering of an open-ended question using documents retrieved by a high-recall query.*

We will achieve this goal by answering the following research questions (introduced in section 1.7):

1. To what extent can a standard text analysis procedure model the general semantic structure expected by our interaction model and particularly the low-level structure required by the aspect cluster growing strategy?

2. Given an adequate semantic model, which approach to spatial-semantic layout best preserves the general and, in particular, the low-level structure expected by our interaction model?

3. Under what conditions does the aspect cluster growing strategy tend to fail and how can we use this knowledge to guide development of interactive support tools?

In this chapter we examine these questions within the context of the available literature. From this critical review, we develop a set of specific hypotheses in relation to each question, along with two formal tests that will enable us to test these hypotheses in subsequent chapters. The structure and specific aims of this chapter is as follows.

Section 2.2 outlines our interaction model in more detail, focusing specifically on the aspect cluster growing strategy. We define how this strategy fits into our interaction model and its relevance to the underlying search task. We then explain how our approach differs from that evaluated by Leuski (2001) and identify potential issues that need to be addressed.

Section 2.3 discusses the rationale and empirical evidence in support of the spatial-semantic metaphor, focusing particularly on its role in document presentation and information seeking. The section concludes by describing the process of spatial-semantic visualization, placing the three research questions within this context.

Section 2.4 discusses the how the semantic structure of a document collection might be modelled using an unsupervised approach based on the vector space model. The rationale for doing this is discussed within the context of the IR cluster hypothesis. We then discuss the empirical evidence that supports this general hypothesis before focusing on a recent, special case termed the aspectual cluster hypothesis (Muresan and Harper, 2004), which predicts the behaviour of inter-document similarities of relevant documents in the case of a complex information need.

Section 2.5 discusses existing methods for testing the cluster hypothesis within a given topology. We reflect upon these methods and propose two tests that enable us to test the feasibility of our interaction model from two perspectives. The aspect cluster separation test allows us to measure the extent to which average computed inter-document similarity increases as the semantic distance between documents decreases. The nearest aspect neighbours test simulates the aspect cluster growing strategy by measuring the rank-distance between any given aspect-relevant document and its nearest relevant neighbours. This provides us with an effective measure of maximum performance for the strategy in any given circumstance. We specify formal hypotheses that will allow us to answer research question one, using these two tests, within a range of quite distinct scenarios.

Section 2.6 focuses on the problem of representing the modelled structures using the spatial-semantic metaphor (research question two). We begin by providing an overview of the common approaches to spatial-semantic visualization. We then emphasise the fundamental problem of dimension reduction and the inevitable information loss that occurs during this process and that (global) approaches which attempt to preserve all inter-document relations are likely to be sub-optimal. We respond by suggesting that a spatial-semantic layout approach that emphasises local structure may prove to be the optimal technique for our interaction model. We posit formal hypotheses that will compare globally and locally optimised spatial-semantic visualizations in a range of distinct scenarios.

Section 2.7 deals with research question three. We explore the problem of how to support the aspect cluster growing strategy when spatial-semantic cues fail. We begin by discussing the potential of augmenting the visualization using relative similarity cues, specifying a hypothesis that most failures are due to node misplacement during the layout process. We then explore the problem situation of cases where an aspect exemplar is not sufficiently similar to its same-aspect relations to support the strategy, even using similarity cues. We discuss potential solutions to this problem, but caution that the final solution (presented in Chapter 6) is mainly dependent upon the outcome of our formal analyses.

2.2. Interaction model

In this section, we outline our interaction model. This model is derived from the model underlying the Lighthouse interface (Leuski, 2001). We begin by introducing Leuski's (2001) model before describing how we propose to extend the model for the kind of open-ended search task described at the beginning of this dissertation. Finally, we discuss key differences in the nature and underlying assumptions of our approaches.

2.2.1. Lighthouse

Lighthouse was developed by Leuski (2001) as a means of alleviating an inherent problem associated with locating relevant documents within a traditional ranked-list model of document organisation: that locating one or more relevant items provides the user with no direct cues as to the location of other relevant items. We now describe the rationale for Lighthouse and its use of clustering to support the process of locating relevant documents within a retrieved set.

In the ranked list, documents retrieved by the query (or more usually surrogates based on meta-data) are presented in order of their similarity to the query. The user begins their evaluation of the retrieved set at the top of the list (most relevant document) and searches for one or more relevant documents by browsing sequentially through items in the list, possibly interacting with document content if this facility is available (e.g., through hyperlinks). If the query is a precise description of the user's information need then this strategy is effective as most of the top ranking documents will be relevant to their need.

If the query is not particularly precise then the strategy becomes less efficient as relevant documents may be distributed irregularly across the ranked list, which may number tens to thousands of documents. As similarity to the query is the only organisational cue, the location of the first relevant document usually provides no clue as to the location of other relevant documents. If documents are scattered too thinly or irregularly within the top ranks of the list the user is likely to draw one of two conclusions: that the query needs to be reformulated or that the system cannot satisfy their query. Whichever conclusion is drawn, the user is unlikely to keep browsing through more than a 10-20 items (see Jansen et al., 2000), particularly if the potential rewards look slim. Non-expert searchers in particular are most likely to draw the latter conclusion as they may fail to appreciate why their query failed and/or understand the importance of query reformulation. More experienced searchers may reformulate their query (either manually or using relevance feedback tools) but this can be a time consuming process and also often, in one sense, an unnecessary process as many of the relevant documents may already have been retrieved, they are just difficult to locate in the ranked list format. In other words, the problem can be viewed as one of document organisation and presentation rather than a fault in the query or retrieval process.

Given this, Leuski (2001) considered the potential of using document clustering as an alternative means of representing the retrieval set. The idea of applying document clustering to support browsing of a retrieved set was not a new idea. What was new was the use of multi-dimensional scaling to convey clusters of similar documents, rather than discrete cluster allocation (e.g. Hearst and Pederson, 1996), and the combination of this visualization with the ranked list.

The rationale for clustering a retrieval set stems from the cluster hypothesis of information retrieval which states that documents that are similar tend to be relevant to the same queries (van Rijsbergen, 1979). This hypothesis emerged from studies of the vector space model of document representation, where documents are represented as vectors within a common high-dimensional term space, which showed that the similarity between documents relevant to a defined topic tends to be greater than those between the same documents and other documents that discuss different topics. By similar documents, we mean documents that exhibit a similar pattern of term usage. A logical corollary of the cluster hypothesis is that if the user has already found one or more relevant examples then clues that indicate which other documents are highly similar will guide the user more efficiently to further relevant documents (Leuski, 2001).

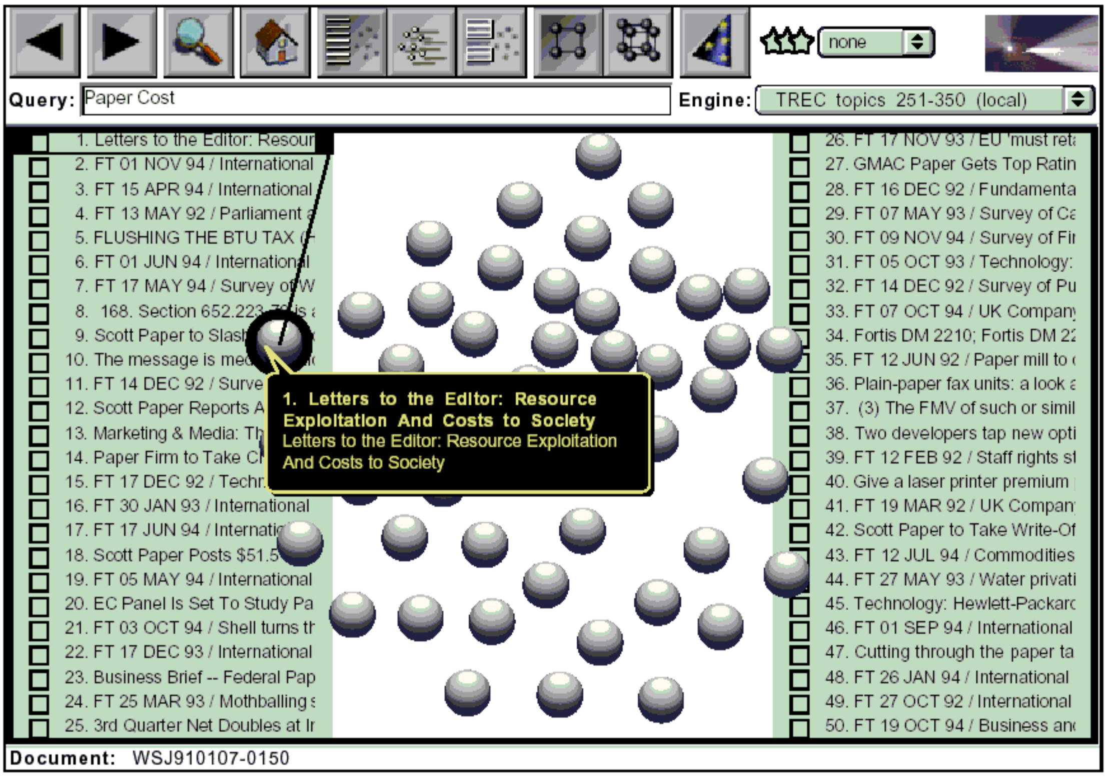
One way to provide such clues is to organise documents based on their similarity using a clustering algorithm. Clustering algorithms have two aims: to create sets of highly similar objects and to maximise the distance (or dissimilarity) between these sets. Previous studies of retrieval set clustering have shown positive results whereby relevant documents tend to converge on a small number of clusters within the solution (e.g., Hearst and Pederson, 1996; Wu et al., 2001).

Leuski (2001) also conducted an experiment that compared clustered representations of a retrieved document set with the traditional ranked list and also a ranked list enhanced by a relevance feedback tool (LCA: Xu and Croft, 1996) that re-ranked documents as relevant items were identified. These simulated user studies showed that applying a strategy whereby the user focused on clusters that were already known to contain relevant documents resulted in an increase in search efficiency (precision) of over 10% compared to browsing a traditional ranked list. Furthermore, performance was equivalent to that observed when relevance feedback tool was applied.

Whilst this experiment provided further evidence to support the use of post-retrieval document clustering, Leuski (2001) observed a key limitation of the model: Whilst documents can be assumed to be similar if they reside in the same cluster, a discrete cluster representation provides no clues as to the extent to which documents residing in different clusters are similar. This is critically important given the empirical evidence, which shows that although there is often a single cluster that contains a large proportion of relevant items, a significant number of the remaining relevant documents will be scattered across one or more other clusters (Hearst and Pederson, 1996). This fragmentation of relevant documents is likely to increase if the topic has more than one aspect (Muresan and Harper, 2004).

As a response to this Leuski (2001) proposed the use of multi-dimensional scaling (MDS) as a means of organising and representing documents by their similarity. In MDS representations, highly similar documents, represented as points or nodes, will tend to form coherent clusters. However, the aim of MDS is not to produce clusters *per se*; rather these features emerge from a process that simply seeks to find the best inverse match between input document similarities and output node proximities within a specified number of dimensions. A key benefit of this MDS is that whilst highly similar documents can form coherent visual features, as would be the case in a discrete cluster solution, documents that discuss multiple, key themes can be placed, for example, at the intersection between the respective clusters. Where two clusters that happen to contain several highly similar documents (e.g., the relevant documents), an emergent result might be the overlap of these otherwise distinct clusters within the visualization.

Based on this notion, Leuski (2001) proposed an interaction model where the system presents the retrieved set in both a ranked list and spatial-semantic format and the user exercises an effective browsing strategy that combines the cues provided by these two representations. Figure 2.1 shows the basic interface. The node of the currently selected document, the top ranked item, is highlighted with a black ring along with a context label showing summary details. The spatial-semantic visualization was created using a spring model variant of MDS originally proposed by Fruchterman and Reingold (1991). We discuss variants of MDS and other layout algorithms in more detail in section 2.6.



Query: Paper Cost

Engine: TREC topics 251-350 (local)

□ 1. Letters to the Editor: Resour
□ 2. FT 01 NOV 94 / International
□ 3. FT 15 APR 94 / International
□ 4. FT 13 MAY 92 / Parliament a
□ 5. FLUSHING THE BTU TAX (
□ 6. FT 01 JUN 94 / International
□ 7. FT 17 MAY 94 / Survey of W
□ 8. 168. Section 652.223
□ 9. Scott Paper to Slash
□ 10. The message is me
□ 11. FT 14 DEC 92 / Surve
□ 12. Scott Paper Reports A
□ 13. Marketing & Media: Th
□ 14. Paper Firm to Take C
□ 15. FT 17 DEC 92 / Techn
□ 16. FT 30 JAN 93 / International
□ 17. FT 17 JUN 94 / Internatio
□ 18. Scott Paper Posts $51.5
□ 19. FT 05 MAY 94 / International
□ 20. EC Panel Is Set To Study Pa
□ 21. FT 03 OCT 94 / Shell turns th
□ 22. FT 17 DEC 93 / International
□ 23. Business Brief -- Federal Pap
□ 24. FT 25 MAR 93 / Mothballing s
□ 25. 3rd Quarter Net Doubles at Ir

1. Letters to the Editor: Resource Exploitation And Costs to Society
Letters to the Editor: Resource Exploitation And Costs to Society

□ 26. FT 17 NOV 93 / EU 'must reta
□ 27. GMAC Paper Gets Top Ratin
□ 28. FT 16 DEC 92 / Fundamenta
□ 29. FT 07 MAY 93 / Survey of Ca
□ 30. FT 09 NOV 94 / Survey of Fi
□ 31. FT 05 OCT 93 / Technology:
□ 32. FT 14 DEC 92 / Survey of Pu
□ 33. FT 07 OCT 94 / UK Compan
□ 34. Fortis DM 2210; Fortis DM 2
□ 35. FT 12 JUN 92 / Paper mill to
□ 36. Plain-paper fax units: a look a
□ 37. (3) The FMV of such or simil
□ 38. Two developers tap new opti
□ 39. FT 12 FEB 92 / Staff rights s
□ 40. Give a laser printer premium
□ 41. FT 19 MAR 92 / UK Compan
□ 42. Scott Paper to Take Write-Of
□ 43. FT 12 JUL 94 / Commodities
□ 44. FT 27 MAY 93 / Water privati
□ 45. Technology: Hewlett-Packarc
□ 46. FT 01 SEP 94 / International
□ 47. Cutting through the paper ta
□ 48. FT 26 JAN 94 / International
□ 49. FT 27 OCT 92 / International
□ 50. FT 19 OCT 94 / Business an

Document: WSJ910107-0150

Figure 2.1: The Lighthouse interface (reproduced from Leuski, 2001, p. 47).

From his experience with retrieval engines and the test collection, Leuski (2001) had found that the good representatives of relevance tend to be quite highly ranked in the ranked list. Given this he proposed a strategy where the user begins by browsing from the top of the ranked list in the classical way. However, once a relevant exemplar is found, the user switches their attention to the visualization and continues browsing from there. Within the visualization the user employs a simple visual search strategy, which we will hereon refer to as the *cluster growing strategy*, to retrieve further relevant documents. The strategy proceeds as follows. The user marks the first relevant document before switching their attention to the visualization wherein the node describing the location of the relevant document is highlighted. Following the corollary of the cluster hypothesis (van Rijsbergen, 1979) that the other relevant documents will be more similar (and thus more proximal within the visualization) than non-relevant documents, the user proceeds to view documents in order of their node proximity to the exemplar. When another relevant document is found then the centre of the relevant cluster shifts to the spatial intersection and the unseen document that is most proximal to this point is viewed. This process is illustrated in Figure 2.2, which shows three sequential steps of the cluster growing process. In the figure, known relevant documents are shown in black, known non-relevant documents in white and unseen documents in grey. We can see how the centre point of the known relevant cluster (the black cross) shifts closer to the actual centre of the relevant subset (on the right) as more examples are identified. This process continues until the user decides, for whatever reason, to terminate the search.

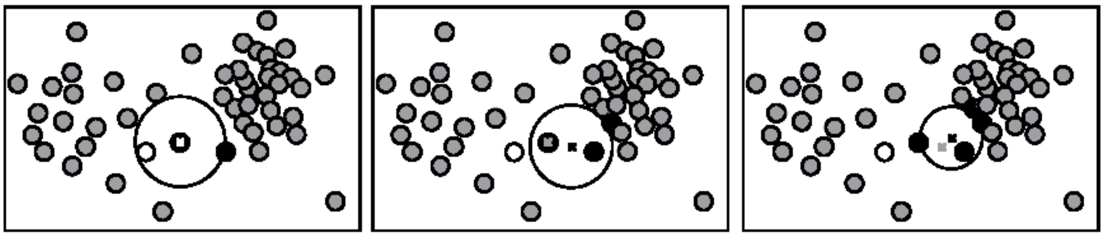


Figure 2.2: The 'cluster growing' search strategy (reproduced from Leuski, 2001, p. 34).

Leuski (2001) compared the precision of the cluster growing strategy to that of the ranked list strategy (documents browsed in their rank order) and a relevance feedback strategy where the query is iteratively refined when each new relevant document is found using local context analysis (LCA: Xu and Croft, 1996). The precision of the cluster growing strategy was also compared across different structures that used both the proximities of 2D and 3D spatial-semantic solutions and also the pure similarities computed between the documents in vector space representation. These comparisons were repeated across a 50 topics taken from the TREC-5 and TREC-6 conference test beds (Voorhees and Harmen, 1996,1997).

They found that the cluster growing strategy was, on average, around 20% more efficient at retrieving relevant documents than the ranked list strategy applied to a standard relevance ranking. The difference was significant regardless of the source of inter-document proximities. Furthermore, when using both the 3D (spatial-semantic) and D-space (pure similarity) structure, precision was significantly higher than applying LCA relevance feedback. The precision of the strategy using the 2D structure was not significantly different from the relevance feedback strategy. Furthermore, performance on 3D was significantly better then 2D.

Hence, it seems that the additional dimension provided by 3D over 2D allowed for a more faithful representation of inter-document similarities. However, subsequent user studies showed that the additional demands involved in comprehending a 3D structure outweighed the more accurate spatial-semantic cues. This effect was also found in another, unrelated study of spatial-semantic search (Westerman and Cribbin, 2000). The problems associated with the drastic dimension reduction required by spatial-semantic visualization and the trade-offs that must be made between using all spatial dimensions (3D) to map the similarity topology and the relative simplicity, for the user, of searching in 2D rather than 3D space are considered in more detail in section 2.6.

2.2.2. Aspect cluster growing

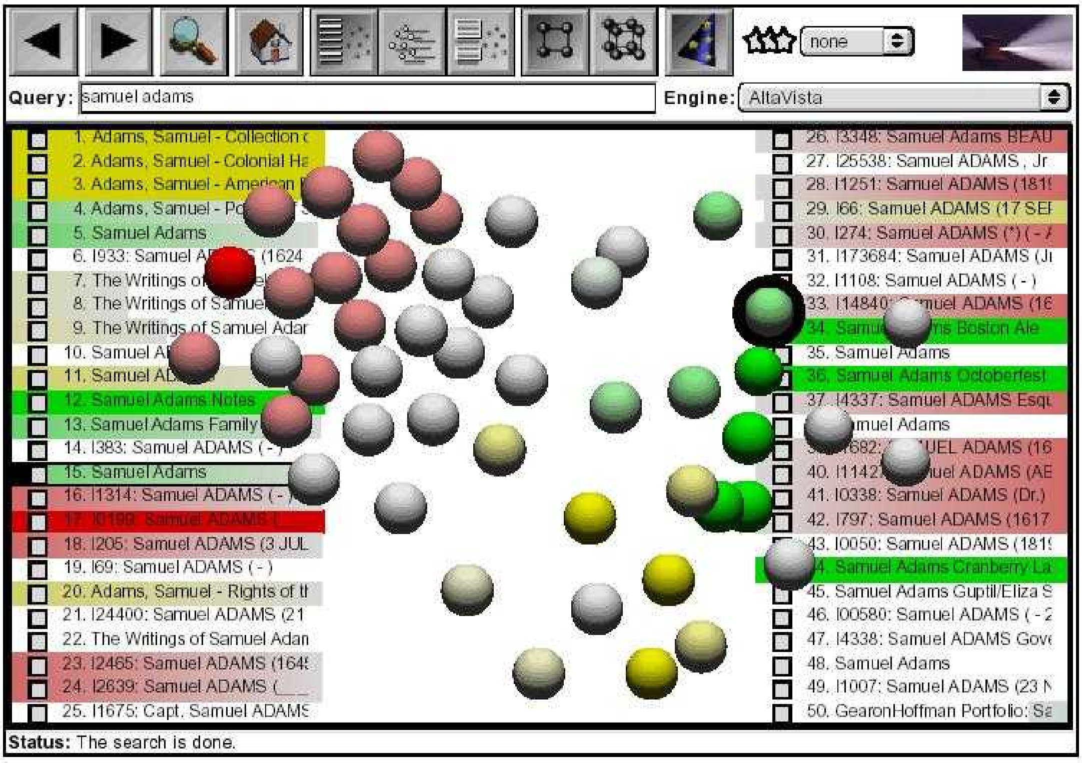
Towards the end of his thesis Leuski (2001) considers the potential for using the Lighthouse model and cluster growing to support retrieval of more complex, multi-aspect topics. He observes that certain queries that were clearly ambiguous in meaning were represented within the visualization by distinct 'pockets' or clusters of relevance. For instance the query "Samuel Adams" was used to describe documents about the legendry American beer maker and revolutionary. The fact that this man is famous for two distinct reasons was clearly represented in the visualization as can be seen in Figure 2.3, where documents known to be about Samuel Adams beer are shown in solid green whilst those about Samuel Adams the revolutionary are shown in yellow. Known non-relevant documents, for example about other people called Samuel Adams, are shown in solid red.

The lighter shades indicate documents that are estimated to be relevant to a known aspect, based on their similarity to the known relevant (or non-relevant) exemplars. The 'shade wizard' is based on an intelligent agent that uses relevance feedback to model the topic of interest and is one of several visual tools that Leuski (2001) implements to provide additional cues to support the basic spatial-semantic cue driven strategy. We return to discuss the need for and potential requirements of interactive tools later on in section 2.7.

We can see that the two aspects form quite distinct clusters within the visualization and both are well separated from non-relevant items. On this basis, Leuski's interaction model seems a promising means of supporting complex needs, for example an open ended question such as:

What are the most significant achievements of the Hubble space telescope since its launch?

We see the user beginning their search by formulating a simple free form query such as "hubble space telescope" or even just "hubble" and submitting this to an appropriate index for retrieval. Aspects exemplars could be identified either, as in Leuski's model, by using the top ranks of the ranked list, or alternatively the user could browse the visualized structure directly, perhaps aided by landmarks such as contextual key terms as demonstrated in larger, collection-wide thematic maps (see, for example Wise et al., 1995; Lin, 1997; Hornbaek and Frokjaer, 1999; Skupin, 2000).



Query: samuel adams

Engine: AltaVista

□ 1. Adams, Samuel - Collection c
□ 2. Adams, Samuel - Colonial Ha
□ 3. Adams, Samuel - American
□ 4. Adams, Samuel - Po
□ 5. Samuel Adams
□ 6. I933: Samuel A (1624
□ 7. The Writings of
□ 8. The Writings of Samue
□ 9. The Writings of Samuel Adar
□ 10. Samuel A
□ 11. Samuel AD
□ 12. Samuel Adams Notes
□ 13. Samuel Adams Family
□ 14. I383: Samuel ADAMS (-)
□ 15. Samuel Adams
□ 16. I1314: Samuel ADAMS (-)
□ 17. I0199: Samuel ADAMS (
□ 18. I205: Samuel ADAMS (3 JUL
□ 19. I69: Samuel ADAMS (-)
□ 20. Adams, Samuel - Rights of th
□ 21. I24400: Samuel ADAMS (21
□ 22. The Writings of Samuel Adan
□ 23. I2465: Samuel ADAMS (164
□ 24. I2639: Samuel ADAMS (
□ 25. I1675: Capt. Samuel ADAMS

□ 26. I3348: Samuel Adams BEAU
□ 27. I25538: Samuel ADAMS , Jr
□ 28. I1251: Samuel ADAMS (181
□ 29. I66: Samuel ADAMS (17 SEI
□ 30. I274: Samuel ADAMS (*) (- /
□ 31. I173684: Samuel ADAMS (Ju
□ 32. I1108: Samuel ADAMS (-)
□ 33. I14840: muel ADAMS (16
□ 34. Samu ms Boston Ale
□ 35. Samuel Adams
□ 36. Samuel Adams Octoberfest
□ 37. I4337: Samuel ADAMS Esqu
□ muel Adams
□ 682: UEL ADAMS (16
□ 40. I1142 uel ADAMS (AE
□ 41. I0338: Samuel ADAMS (Dr.)
□ 42. I797: Samuel ADAMS (1617
□ 43. I0050: Samuel ADAMS (181
□ 4. Samuel Adams Cranberry La
□ 45. Samuel Adams Guptil/Eliza S
□ 46. I00580: Samuel ADAMS (- 2
□ 47. I4338: Samuel ADAMS Gove
□ 48. Samuel Adams
□ 49. I1007: Samuel ADAMS (23 N
□ 50. GearonHoffman Portfolio: Sa

Status: The search is done.

Figure 2.3: The Lighthouse interface showing the different aspects of the "Samuel Adams" query (reproduced from Leuski, 2001, p. 68).

As a new aspect is discovered, the user would be able to employ the cluster growing strategy to find further, similarly relevant items. If other aspects are encountered during the process then the user can simply mark these in a different colour and return to them later once retrieval for the existing intention is seen as complete. They could equally temporarily abandon their current intention and return to it later as the distinct colour assigned to documents marked relevant to that aspect would allow the user to readily re-orientate to the earlier intention. In Chapter 1 we defined this application of the cluster growing strategy as *aspect cluster growing* to differentiate it from the simple task of general retrieval of a single homogeneous sub-set of relevant items.

The kind of search we are describing is very close to what Bates (1989) describes as berrypicking/evolving search and O'Day and Jeffries (1993) describe as progressive searching. This view of search as a non-linear, unpredictable and complex process is much closer to most real search episodes than the simple classical view that has dictated the design of the majority of information retrieval system designs and evaluations, including Leuski's (2001). The great potential strength of spatial-semantic visualization, and the aspect cluster growing strategy, is that together they allow both exploration of the topic and directed browsing of multiple, diverse aspects of the topic to take place within the same structural view. Users do not have to reformulate the query statement as their intentions change and they have a persistent history of their search progress and can easily distinguish between different intentions that they have followed. Adopting such non-linear behaviour imposes a huge cognitive demand on users of traditional interfaces such as hypertext or the ranked list because they lack this persistent overview context; users must construct their own, complex mental model, integrating between views as their intentions shift (Vicente and Willeges, 1988) and remembering how to command or navigate back to earlier intentions if they are left incomplete.

2.2.3. Proposed model

Hence, our proposed interaction model, of which aspect cluster growing strategy is an element and the focus of this thesis, can be summarised as follows:

1. The user specifies and issues a general query (e.g., one or two salient terms from their question) to the information retrieval system.

2. The system retrieves matching documents and downloads all or part of their full-text, computes an inter-document similarity matrix and uses a spatial-semantic layout algorithm to project the similarity structure on to a 2D or 3D space.

3. The system presents the retrieved documents as an interactive spatial-semantic visualization. Documents may also be represented as a ranked list, as in Lighthouse, although this is not an essential requirement of our model.

4. The user browses the visualization (or possibly the ranked list) with the intention of locating an unseen relevant document discussing a distinctly novel aspect of the topic.

5. If a new aspect is found then go to stage 6, else if the user decides that all relevant documents and aspects have been identified then go to stage 8

6. The user locates the relevant aspect exemplar in the visualization and browses unseen document nodes in proximity order, marking aspect-relevant documents as they are found until the decision is made to terminate the current aspect cluster growing intention.

7. If the user has terminated to pursue another aspect then go back to stage 6 else if the user considers that no further aspect-relevant documents will be found then go back to stage 4

8. End of search interaction

There are several key features/benefits of this approach. First, multiple aspectually distinct clusters of documents are grown over the course of the search episode. Second, as all clusters are grown within the same spatial-semantic structure of the retrieved set the user is able to become familiar with the thematic structure and use spatial cues to infer relationships between aspect clusters. Third, given this stable structure, the user is able to grow aspect clusters in parallel; for instance, if a new aspect is discovered, the current cluster can be temporarily abandoned whilst the user pursues this new query, yet easily relocated once the user decides to resume the old search. Fourth, despite a complex, evolving conception of information need, the user never has to explicitly reformulate their query; the search task is more akin to navigation rather than one of specification.

As already noted, this model is an adaptation of the Lighthouse model (Leuski, 2001). Leuski (2001) discussed the possibility of adapting Lighthouse to solve complex retrieval problems, but did not formally evaluate his system within this task context. He does provide an illustrative example, the Samuel Adams query, which demonstrates how spatial-semantic structure might support multi-aspect retrieval scenario. However, we intend our interaction model to be useful for far more complex search scenarios. We now outline the key differences in our approaches and discuss some of the additional challenges we will need to address.

2.2.4. Key differences in approach

The Samuel Adams example illustrates the basic essence of how our interaction model would work. However, this is a relatively simple example for a number of reasons. First, there are just two aspects. Second, these aspects are conceptually quite distinct. Third, each aspect is well represented in the retrieved set, forming a significant feature. Fourth, there is no overlap between aspects in terms of the documents that represent them; relevant documents are aspectually distinct. Fifth, the retrieved set is relatively small, just 50 documents.

Many open-ended search tasks involve topics that are significantly more complex, comprising both highly distinct and more closely related aspects. Some aspects may be discussed by a large sub-set of retrieved documents, whilst other aspects might be more esoteric or idiosyncratic in nature and therefore discussed by relatively few documents. In an ideal world, relevant documents would be highly focused on just one aspect, but in reality some documents might be more 'topical', discussing many relevant aspects. Furthermore, whilst one document might make only a brief, single sentence reference to a relevant aspect another might devote several paragraphs. Finally, in our interaction model we expect the query that retrieves the visualized document set to be quite ambiguous and broad in scope, retrieving document sets in the order of hundreds, or in the case of a web search, possibly thousands of documents. To ensure a representative sample of distinct aspects and associated documents, it would be desirable to visualize at least the top one or two hundred retrieved documents.

The potential utility of our approach increases inline with the complexity of the search problem. We therefore wanted to demonstrate the feasibility of our interaction model within the context of more demanding scenarios, rather than simple cases like the Samuel

Adams example. As such, both topics chosen for our test scenarios (see section 3.2) are highly complex, comprising at least 20 distinct aspects each. Further, all retrieved and visualized sets are in the order of hundreds of documents, rather than just a few dozen top ranking items. In Chapters 3, 4 and 5 our analyses specifically address the impact of aspect overlap (where documents discuss multiple aspects) and increasing document set size. The potential effects of both these factors are discussed later in sections 2.5 and 2.6 and specific hypotheses are presented.

From section 2.4 we begin formulate the hypotheses that will be directly addressed by our analyses in Chapters 3 to 5. Before we do this, we explore the general rationale for using the spatial-semantic metaphor to visualize document structure, and discuss the results from empirical studies that have evaluated the utility of the spatial-semantic approach as means of supporting a range of information seeking tasks.

2.3. Spatial-semantic metaphor

The purpose of this section is twofold. First, we introduce the spatial-semantic metaphor and discuss the rationale for its application to document organisation, reviewing empirical evidence that shows that people can understand the spatial-semantic metaphor and can utilise this understanding to support semantic browsing and searching tasks. We then conceptualise the process of implementing interactive document visualizations as a pipeline of inter-dependent stages of unsupervised modelling and user interaction, and explain how the three questions directly relate and justified by this process.

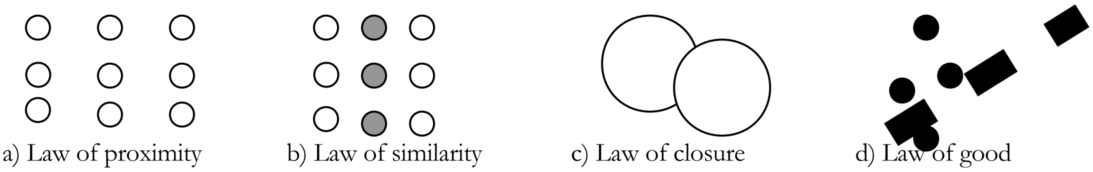
In this section, we explain why relative proximity is an effective cue to object similarity and present evidence that indicates users can readily equate object similarity with object proximity. Focusing on systems that use proximity to convey general similarity between documents we review the results of empirical studies that demonstrate how spatial-semantic cues can support information browsing and search.

2.3.1. Proximity as an organising cue

When searching a physical environment, the spatial organisation of items is a critical factor that governs task success. For example, we explore and retrieve known items by browsing the ordered shelves of a library (Bates, 1989) or supermarket. We organise our workspace by sorting and filtering incoming documents into arranged piles or trays on the desktop (Pirolli and Card, 1999). Often there is hierarchy to such organisation, whereby items are iteratively separated into more specific conceptual sub-sets of the overall collection. The use of hierarchical classification is intricately tied up with the notion of semantic distance between concepts within psychological space (see Brooks, 1998) and as such physical instantiations of a hierarchy tend to group objects progressively closer as the concepts that link them become more narrowly defined. Spatial-semantic visualization exploits our natural expectation that objects that are conceptually closer will be physically located more closely (Montello et al., 2003). In this thesis we refer to this principle of organising objects so that spatial proximity corresponds to semantic distance as the spatial-semantic metaphor. So why do we consider spatial-organisation to be such a powerful cue to conceptual similarity and what other cues do we use to structure our visual environment?

Whilst, we are able to consciously infer sense from complex or ambiguous visual images (as shown in studies of visual illusions), in order to make basic sense of the vast, changing flow of visual data that we receive moment by moment, much of our visual perception is achieved pre-attentively. The processes of pre-attentive perception are fast and effortless and their impact on our conscious experience of the world can be compelling. A red car in a car park full of blue cars will immediately attract our attention. We can differentiate immediately between a birds flying in formation and birds acting independently on the basis of their relative speed and direction of movement.

Early, seminal work by the Gestalt school proposed that there are number of fundamental laws or rules that govern the way we organise visual stimuli (Koffka, 1935; Eysenck and Keane, 1990). In addition to proximity, we also perceive structure based on the similarity, closure, good continuation and common fate of objects sensed in the visual field.



a) Law of proximity

b) Law of similarity

c) Law of closure

d) Law of good continuation

Figure 2.4: Gestalt laws of perceptual organisation

The law of proximity states that objects that are relatively near to one another tend to be seen as related. For instance, in figure 2.4a we see columns rather than rows of dots because the horizontal separation is greater than the vertical separation. The law of similarity describes how we tend to group together objects that are visually similar regardless of their spatial configuration. In figure 2.4b, for instance, even though nodes are equidistant we perceive three columns of dots, rather than a single grid. Likewise, in an array of mostly blue nodes a single red node will instantly 'pop out' as an anomaly. The law of closure states that if a pattern implies a coherent form, but is incomplete, the implied form will be perceived nevertheless. Figure 2.4c shows a good example where we perceive two overlapping circles even though only one circle is actually present. The law of good continuation states that visual elements that appear to follow the same path or pattern will tend to be associated together. In figure 2.4d, it is hard ignore a line running from bottom left to top right, despite the presence of a dense cluster overlapping the bottom end of this formation. Finally, the law of common fate says that elements that appear to be moving in the same direction will be grouped together. The bird formation example presented in the earlier paragraph illustrates the action of this law.

Fundamentally, spatial-semantic visualization is based upon the law of proximity. However, all of these laws can be incorporated into visualization design to emphasise conceptual groupings. Furthermore, these laws are by no means independent. As we shall see later, in section 2.3.2, research into the spatial-semantic metaphor shows how emergent features (e.g., clusters, lines) can interfere with the interpretation of relative proximity cues (Montello et al., 2003).

So far we have talked about how visual-spatial cues enable us to group objects together. Visual-spatial cues can also be used to encode ordinal and quantitative data. Related to this is the work of Bertin (1983) in the field of data graphics. Bertin (1983) explains how objects or marks within a graphic are organised according to their relative values with respect to one or more visual variables. Visual variables differ in terms of the level of data they can convey. For instance, whilst the length, area or location of a mark can communicate quantities, relative and absolute, the shape of a mark can generally only communicate nominal level attributes (e.g., discrete group membership).

Visual variables can be classified into two types: planar and retinal. Planar variables are those that utilise the spatial substrate, whereby distance along an axis might convey the absolute value of an object for specific variable, whilst relative proximity can convey associations between objects. Retinal variables are those that affect the appearance of marks, such as size or shape, and thus exploit similarity as an organising principle. Retinal and planer variables can be used in combination, but there is a limit to the number of variables that can be perceived pre-attentively and integrated into a single, coherent image or whole. According to Bertin (1983) this limit is two planar variables and one retinal variable. Interactive computer graphics may permit the use of a third spatial dimension for certain applications although, as we will discuss in section 2.3.2, the use of 3D can significantly increase the cognitive effort and ability required to interpret a visualization. If any more than one retinal variable is represented, then interpretation of the visualization also becomes more effortful and slow as the task of discriminating groups of related marks or perceiving correspondence between marks on basis of one retinal variable, is likely to be subject to interference from other, possibly more compelling organising cues. Bertin (1983) refers to this problem as a violation of the single image, which results from the fact there is no one simple, unambiguous form (Koffka, 1935).

Although, as already noted, proximity is by no means the only visual cue to similarity, it is possibly the most powerful and certainly the most flexible in terms of the type of information it can convey (Bertin, 1983; Card et al., 1999). The spatial substrate can be used to communicate all levels of correspondence between abstractly defined objects, from category membership, as discrete groups or clusters, to quantitative differences and ratios represented by relative distance (Bertin, 1983). One of the most common applications of spatial cues in this context is the scatter plot, or point display (Montello et al., 2003), where objects are projected as points onto a two- or three-dimensional plane. Each dimension of the plane represents a single common attribute of the set of objects and the location of an object along this dimension indicates its value. Given this scheme it is possible to not only interpret the absolute value of an object with respect to attributes but also to make relative judgements between objects along each attribute.

A particularly useful affordance of the scatter-plot, as an exploratory analysis tool, is the fact that straight-line distance between objects allows direct interpretation of their general similarity in terms of all spatially encoded attributes (i.e., dimensions). Objects that are similar in all respects will form coherent clusters, whilst objects that differ on one or more attributes will be more distal, with the magnitude of this distance increasing inline with magnitude of the difference. Objects that are particularly different in their attribute profiles (e.g., are distinct or erroneous cases) to most other objects become instantly detectable, appearing as isolated 'outliers' within the plot.

The application of scatter-plots to communicate general similarity can be taken a step further by using procedures that attempt to organise objects onto a visual plane based upon either their average similarity with respect to many different variables (e.g., vector similarity) or human judgements of their similarity. In other words the axes of a scatter plot do not necessarily relate to known or definable variables. This is often achieved using a class of techniques known as multi-dimensional scaling (MDS). MDS algorithms take a matrix of inter-object similarities (or dissimilarities) as input and output a low dimensional spatial configuration. The aim is to represent objects as a scatter plot of points in such a way that the relative distances between object points reflect empirical relationships in underlying data (Coxon, 1982). Although clustering is not a specific aim of the algorithm high-density regions of nodes will often emerge in the resulting configuration and will be perceived by the viewer as clusters. As we will see in the next sub-section, such features will tend to be perceived as groups of objects that are highly similar in some respect.

As indicated, the similarity data to be scaled may be acquired directly from subjective observations (e.g., asking people to rate the similarity of objects or concepts) or indirectly by measuring the correlation between objects with respect to a large number of defined attributes. For example, if the aim is to identify homogeneous groups of customers, then these attribute measures might relate to the frequency with which specific items or item types are purchased. More pertinent to this dissertation, if the aim is to model the structure of a large, heterogeneous collection of documents then objects (documents) might be defined in terms of their word frequencies. A measure of inter-document similarity can then be computed based on the assumption that documents with similar word usage patterns will tend to be similar in their content.

We apply a similar approach when creating our semantic models in Chapter 3. Before we do this, however, in section 2.4 we spend some time introducing the vector space model of document representation and discuss how measures of document similarity computed from such representations can create meaningful semantic models that can be used to automatically classify documents by topic.

Later in this section, we discuss how scatter plots of general similarity, for example those created using MDS, have been applied to the problem of browsing and searching document sets. First, we review the empirical evidence that supports the assumption that people can equate object similarity with node proximity within a scatter plot.

2.3.2. Comprehension of spatial-semantic structures

So far in this section, we have argued that spatial cues are key to the way in which we make sense of our environment. We explained how relative proximity is a powerful organisational cue that not only implies group membership, but also relative similarity between objects. We then described how scatter plots are traditionally used to convey abstract relationships between objects in terms of one, two or three distinct attributes. We then extended this traditional view of the scatter plot by explaining how techniques like MDS can allow this medium to be used in a way that can convey general similarity between objects in terms of a complex range of distinct attributes.

The idea of conveying a general similarity structures using relative, continuous proximity cues is a compelling one. This is the essence of the spatial-semantic metaphor. In this section, we review some key studies that elucidate the human response to spatial-semantic document visualizations.

Montello et al. (2003) describe the spatial-semantic metaphor (which they call the distance-similarity metaphor) as the most fundamental principle applied to any information visualization that exploits the spatial substrate. As such they embody this principle in what they call the *first law of cognitive geography*. The terminology reflects the author's geographic background and their aim to apply cartographic principles to spatial-semantic visualization. Their law derives from the first law of geography, which states that things that are relatively proximal within the environment tend to have similar properties (e.g., rainfall patterns, soil type etc.)

The first law of cognitive geography states that: "people believe closer things to be more similar than distant things." (Montello et al., 2003, p. 317). The authors were seeking to test the truth of this hypothesis by means of an experiment in which participants were presented with a series of scatter plots and told that points (nodes) represented documents (Montello et al., 2003). For each trial, a source node (A) and two target nodes (1 and 2) were highlighted in the scatter plot. Participants were asked to judge, along a continuous scale, which pair (A-1 or A-2) was more similar. They found general support for the hypothesis – participants did tend to rate the more proximal pair as more similar and rated equidistant pairs as equally similar.

However, they found that so-called feature emergent effects could override the effect of raw inter-node proximity cues, leading the participants to apply a feature rather a distance similarity metaphor. For instance, an emergent cluster effect occurred when A and one of the targets resided in the same high-density field of nodes (i.e., a visual cluster). The same cluster pair tended to be rated as more similar even if the second pair was more proximal. A similar effect resulted from linear features that emerged when a series of intervening nodes between one of the two pairs that was dense enough to create an effective pathway between the nodes.

This work has important implications for our interaction model. Firstly, it could partly explain why, in his user study, Leuski (2001) found that users often had trouble making accurate proximity judgements, a problem that lead him to implement the star wizard to elucidate the rank order of the three most proximal documents. Secondly, these results suggest that visualizations that create a structure that is rich in emergent features (i.e., with many coherent clusters and possibly pathways) may provide the strongest visual cues to guide cluster growing, providing these features indeed convey same-aspect relationships. Classical MDS algorithms attempt to convey the relationships between all node pairs. This can result in somewhat amorphous (feature poor) visualizations. Later, in section 2.6, we describe an approach that can create scaled solutions that emphasise only the most salient inter-node similarities. Given that same-aspect document similarities are likely to be the some of the strongest within a given collection (Muresan and Harper, 2004: see section 2.4.5), we propose in section 2.6.3 that the local optimisation approach can create a spatial-semantic visualization that is rich in task relevant emergent features.

Montello et al.'s (2003) study was carefully designed so as to purely test the effects of spatial variables on assumed similarity of given node pairs. They did not test the utility of spatial-semantic cues for supporting the location of an actual semantic target. Westerman and Cribbin (2000) conducted an experiment where participants searched a spatial-semantic scatter plot for target nodes representing concrete things. The main hypothesis was that participants would use the cues or 'scent' (Pirolli and Card, 1999) provided by the spatial-semantic structure (i.e., knowledge of the location of similar and dissimilar nodes) to incrementally direct their navigation towards the target.

The semantic objects used varied in their conceptual similarity, belonging to three distinct but potentially related classes: buildings (e.g., church, house), rooms (e.g., hall, lounge) and contents of buildings (e.g., chair, table). Trials were run under 2 and 3 dimensional scatter plots conditions. The spatial-semantic structure was derived from a consensus matrix of inter-object similarity judgements acquired from human judges. A second factor, variance, created a second set of conditions where the fidelity of the spatial-semantic structure of both 2D and 3D visualizations was manipulated by adding varying amounts of noise to the original scaled solutions. The specific hypothesis was that if participants were using proximity as a cue then performance would decrease as the match between semantic similarity and relative proximity decreased.

The results of this study showed a significant linear relationship, between spatial-semantic match and user performance, in the expected direction: performance decreased as the location of nodes became more random. A second important finding was that whilst the use of the third dimension allowed for a better spatial-semantic solution, this benefit was outweighed by the additional cognitive demands associated with navigating in 3D space.

Finally, and perhaps most pertinently, Leuski (2001) conducted a small user study to establish the viability of the visual cluster growing strategy. Similar to Montello et al. (2003), for each trial, participants were presented with a scatter plot and told the nodes represented documents, with one of the nodes already highlighted as relevant. They were told to locate all other relevant nodes in the space (these were actual spatial-semantic solutions of TREC topic retrieval sets). As they clicked on unseen nodes they would change colour to indicate whether they were relevant or non-relevant. Hence, whilst they could not read the content of the underlying documents, as more nodes were clicked, the distribution of relevant and non-relevant documents in the space became more apparent.

Leuski (2001) found that users understood the spatial-semantic visualization and that the visual cluster growing strategy enabled them to locate relevant documents more quickly than they would have done using the ranked list. However, similar to Westerman and Cribbin (2000) they found that the potential benefits of a 3D representation were outweighed by the additional cognitive demands of interpreting and navigating the extra dimension. Whilst Leuski's (2001) simulated user algorithm was able to exploit the better preservation of inter-document similarity information within the 3D solution, users could not; users actually performed better when only 2D cues were presented. The extra dimension seemed to make it harder for users to choose which unseen node was the next most proximal to the centre of the relevant cluster. We return to discuss the 2D versus 3D debate in more detail in section 2.6.

In summary, there is good evidence that people understand the spatial-semantic principle and are able to apply this principle to support simple information search tasks. We have also reviewed evidence that suggests that 2D visualizations are likely to be more comprehensible than 3D visualizations, despite the fact that 3D can produce a better spatial-semantic solution (a more faithful mapping of similarity to proximity). In the next section we consider empirical evaluations of interactive visualizations that have been applied to actual information retrieval tasks where users interact with the visualization in order to access and evaluate document content.

2.3.3. Information seeking and spatial-semantic visualization

Since the early 1990s there have been several notable attempts to apply and evaluate spatial-semantic techniques for document browsing and retrieval. A common application of such visualizations is to provide thematic overviews of document collections. A significant value of this approach is in the ability to represent large, complex topical structures within a compact space (Lin, 1997). For instance, there are examples of such overviews representing the topical structure, and the position of individual documents within this context, for collections of hundreds or thousands of items (Wise et al., 1995; Lin, 1997; Chen et al., 1998; Skupin, 2002).

These large-scale visualizations seem to be useful for providing users with an overview of a large collections, for instance to facilitate an understanding of the relationships between key terms and documents (Lin, 1997) and for providing users with clues to what topics are available and which terms might be used to begin more focused lines of enquiry (Chen et al., 1998).

Chen et al. (1998) conducted studies that examined browsing and retrieval behaviour within the Yahoo™ entertainment category and a self-organising map (SOM) visualization of the same documents. When participants were asked to find an 'interesting' page performance rates were comparable between the two browsing schemes, with 14 out 16 Yahoo hierarchy users found an interesting page within 10 minutes compared to 11 out of 15 SOM users. Subjectively, users liked the visualization because they were able to browse in a non-linear fashion, easily jumping from one part of the map to another, which is a difficult and time-consuming navigational task when using a menu-based system.

However, whilst participants responded well to the visualization for the exploration task, participants disliked the lack of explicit structure present within the SOM, particularly the lack of hierarchy. Participants also suggested that alphabetical listings of key terms be provided to support orientation. More recent work has gone someway to address these complaints. For instance, Skupin (2000) has applied hierarchical clustering to the document similarity data prior to layout, which results in a visualization that has a clear, three-level structure. This structure is made explicit through the use of legibility techniques borrowed from cartography such as proportional term label sizes and recursive bounding of zones to convey hierarchy and is similar in essence to the Treemap technique first proposed by Johnson and Shneiderman (1991). Also, Fabrikant (2000) demonstrated how term lists can be integrated with a spatial-semantic visualization.

Whether the negative feedback Chen et al. (1998) received on the SOM reflected a genuine lack of a logical structure or simply that users were more comfortable with familiar schemes of web directories (i.e., alphabetical listings and human generated categories) than the computer generated structure is unclear. However, it is interesting to note the result when participants were asked to relocate their interesting item in the other scheme. Participants switching from the hierarchy to the SOM were considerably less successful: only two out of 16 participants found the same page in the SOM compared to eight out of 15 participants who switched to the hierarchy. The authors concluded that in order to be a successful search tool, the system must be modified to integrate both querying and browsing.

A later study by Hornbaek and Frokjær (1999) evaluated a hybrid system that combined spatial-semantic visualization with querying. Like Chen et al. (1998), they found that participants valued the overview provided by a spatial-semantic visualization but found more directed browsing (e.g., finding similar documents) somewhat more problematic. The authors provided participants with a zoomable scatter-plot visualization of documents that was annotated with contextually placed key terms. They found that these terms were useful for inspiring more focused searches. They were also able to enter queries and see the results of highlighted in context. They particularly liked this feature as it enabled them to understand the distribution (and relations) of retrieved documents and the relationship of retrieved documents to other terms. As with the Montello et al. (2003) study, emergent features (e.g., dense patches or clusters) proved particularly attractive to participants when browsing, for instance a cluster of documents resulting from a query would immediately attract attention.

However, whilst they understood the idea of spatial-semantic cues, participants often had trouble understanding the relationships between adjacent documents. It is not clear how often this was due to misplacements and how often this was simply because items were not similar in the expected sense (i.e., current query). On the other hand, users were also prone to place too much faith in the spatial-semantic model, often assuming that a document adjacent to a relevant item must also be relevant when this was not the case. These issues suggest that users not only need to see which documents are similar but also need to understand why neighbouring documents are similar. This is likely to be a particular issue when the object of browsing is to locate documents that are similar for a specific reason, as would be the case for the aspect cluster growing strategy.

This seems to be somewhat contradictory to the outcome of Leuski's (2001) cluster growing experiments which found that directed browsing (retrieval) could be efficiently achieved using spatial-semantic cues. However, it is worth noting that the document sets visualized by Leuski were formed using reasonably precise queries, hence the ratio of relevant to non-relevant documents within the visualised set was high. Also, because the relevant topics would often have constituted major themes within their associated document set it is likely that the relevant sub-set would generally have formed a major cluster feature within the visualization. As no formal analysis of the cluster growing strategy for more specific sub-topics was conducted, it is possible that the problems experienced by Hornbaek and Frokjær's (1999) participants may have manifested amongst users of the Lighthouse interface (Leuski, 2001), had the focus been on growing more minor and distinct clusters of relevance.

In summary, there is good theoretical and empirical evidence that users can intuitively understand the principle of spatial-semantic mapping. There is also evidence to suggest that they can apply this understanding effectively to certain information seeking tasks, even retrieval tasks providing the relevant topic forms a major feature within the visualized collection. However, on balance of the evidence, spatial-semantic cues seem to be more useful and reliable for opportunistic searching (overview, exploration) rather than focused search tasks. The outstanding problem therefore seems to be that whilst spatial-semantic visualizations convey the general topical structure (major themes) of a document set or collection, more specific sub-topics are likely to be somewhat more obscurely represented within the structure.

The success of our interaction model depends first on whether it is possible to model the two-level relevance structure and then on whether a visualization layout algorithm can render this structure effectively to 2D space. A notable feature of all the studies we have reviewed is that they only evaluate the utility of the visualization itself. There is no data reported on the extent to which the underlying semantic model conveys the relevant structure or the extent to which node misplacement during the visualization process might have impacted negatively on user search performance. In this dissertation, we not only evaluate our interaction model and our core search strategy within the context of our visualizations but also within the context of the underlying semantic model. We now further explain our rationale for performing the latter analysis by explaining the process of spatial-semantic visualization.

2.3.4. The visualization pipeline

Card et al. (1999) present a reference model for visualization (see Figure 2.5), which defines a process or pipeline that begins with raw data and ends in a structured, interactive view of the data. The raw data is first transformed into structured data tables of cases represented by specified, common variables. The next step is to encode the values of specified data variables into appropriate visual variables in order to create a visual structure that will convey the desired information. Each data case is represented by a mark (visual object) within the visualization. The value of a case for a specific variable can be encoded into the mark either by varying its spatial location along a given dimension or by altering its appearance along some visual scale such as brightness or size (see section 2.3.2). Finally, a view transformation presents this visual structure on screen. This is not the end of the process, however. In an interactive system, the user can then modify the default view of the visual structure and even the structure itself. View transformations might include, for example, changing the point of view (e.g., zoom and pan), selecting cases (location probes) or filtering out unwanted cases. Changes to the visual structure itself by changing the visual mappings of existing variables or redefining which variables are to be mapped.

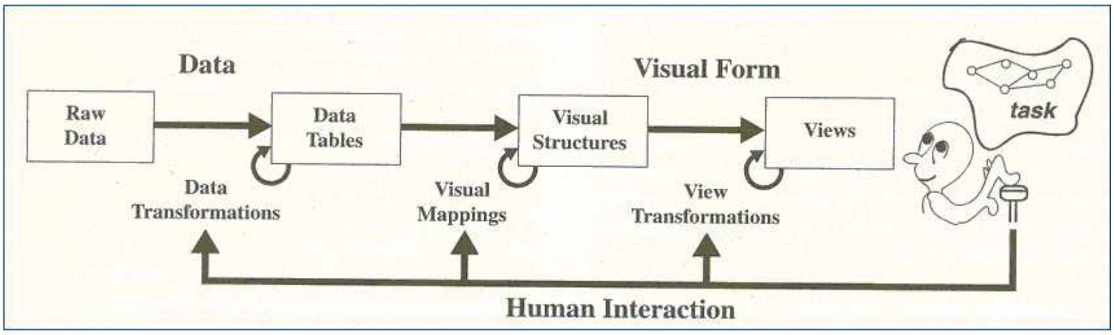


Figure 2.5: Reference model for visualization (reproduced from Card et al., 1999, p.17)

We can describe our approach to spatial-semantic visualization as a special case within the context of this model. Our raw data is the text of the documents retrieved as a result of the user's high-recall query. This is transformed into a semantic model (a set of data tables) by means of automatic text analysis. In this thesis we use a method based on the vector space model (Salton and McGill, 1983), which is described in greater detail in section 2.4.1. The important output of this process is an inter-document similarity matrix that contains values describing the lexical similarity (degree of term overlap) between all possible document pairs. The visual (spatial-semantic) structure is then created by inputting the similarity matrix into a layout algorithm, which represents each document as a node to be located in visual (in our case 2D) space. The algorithm attempts to place each node in a location such that its relative distance to other nodes is inversely proportional to their similarity. As already discussed, this approach is typically referred to as multi-dimensional scaling (MDS) and is discussed in more detail in section 2.6, where we consider potential layout algorithms for creating our spatial-semantic document visualizations. A default view transformation creates an initial view of the spatial-semantic structure. In our interaction model, we envisage that this should be an overview showing the entire structure. In an interactive system, however, view transformations will occur throughout the search process as a result of user selections and commands. It is important to note that, for our purposes, the spatial structure of nodes remains static in order to provide a consistent, learnable model of the retrieved document space. All changes to the visual structure are therefore augmentations of this persistent structure that make use of retinal variables (e.g., changing the colour or transparency of a node) or add contextual objects to the visualization (e.g., term labels).

In summary our visualization pipeline consists of the following stages:

1. Creation of an inter-document similarity matrix by means of automatic text analysis of retrieved documents.

2. Transformation of the similarity matrix into a 2D spatial-semantic visualization.

3. Augmentation of the spatial-semantic structure based on user interaction.

Our three research questions are both inter-dependent and intricately linked to the described pipeline of transformations. The feasibility of our interaction model and specifically the aspect cluster growing strategy is ultimately dependent upon the correspondence between inter-document similarity and the general and aspectual structure of the relevant topic as this serves as the sole input to the layout algorithm. It is then dependent upon the ability of the layout algorithm to preserve the key parts of this modelled structure. Finally, any deficits in spatial-semantic structure need to be resolved by means of interaction between user and system.

The underlying semantic model is critical to our approach. Relevant documents must be more similar to each other than they are to non-relevant documents and documents relevant to a specific aspect must be more similar to each other than they are to documents that discuss other aspects. The presence of this asymmetric, two-level hierarchical structure within the similarity matrix is key to our approach. Most importantly, for the purpose of the aspect cluster growing strategy, documents that discuss the same aspect of relevance must tend to be more similar to each other than they are to any other documents in the retrieved set.

Hence, question one asks: *To what extent can a standard text analysis procedure model the general semantic structure expected by our interaction model and particularly the low-level structure required by the aspect cluster growing strategy?*

We examine the literature relating to this question in sections 2.4 and 2.5. We consolidate what is known about the potential to model the structure of the relevant topic within a retrieved document set. We find that this question has not yet been addressed directly by previous work and therefore present a series of hypotheses that will be tested in Chapter 3 and propose two tests to enable this analysis.

Given a good similarity matrix, it is the responsibility of the layout algorithm to represent this structure faithfully as a 2D spatial-semantic structure. Information loss is inevitable during spatial-semantic visualization due to the dimension reduction involved. We need to ensure that we at least preserve the elements of this structure that describe the relationships between relevant documents.

Given this we proposed question two, which asks: *Given an adequate semantic model, which approach to spatial-semantic layout best preserves the general and, in particular, the low-level structure expected by our interaction model?*

In section 2.6 we explain the issues associated with spatial-semantic visualization and propose an approach that we anticipate will optimise the preservation of same-aspect document associations, whilst preserving the general two-level classification. A set of related hypotheses is presented that will be tested in Chapters 4 and 5.

Finally, user interaction with the spatial-semantic structure is likely to be highly important. As we discussed in section 2.2.4, modelling and conveying topical structure in the absence of user feedback is always going to be a challenging, if not impossible goal. The user must be able to indicate what is relevant and the system should respond to this feedback with cues that augment the spatial-semantic view in ways that support the search process.

As such, question three asks: *Under what conditions does the aspect cluster growing strategy tend to fail and how can we use this knowledge to guide development of interactive support tools?*

In section 2.7, we discuss how document search might be supported when spatial-semantic cues fail to adequately support the aspect cluster growing strategy. We suggest that many problems may result from compromises in the spatial-semantic layout process. We therefore discuss how the reinstatement of relative similarity cues might usefully support aspect cluster growing. We then reflect on our discussion in section 2.2.4 and suggest that in some cases the general similarity relationship between same-aspect documents may not be particularly strong due to conceptual diversity in the exemplar and/or its relations.

We propose the notion that exploring the influence of exemplar factors might lead to a better understanding of the conditions that result in low similarity between same-aspect documents. Exemplars are representatives of a specific aspect that may be used as the basis for aspect cluster growing. Exemplar factors describe the relationship between the exemplar and other aspect-relevant documents within the context of all documents within the retrieved set, for example the size of the aspect, and the complexity of the exemplar contents within the context of the topic. We suggest that by understanding the nature of documents that make poor exemplars we can develop effective interactive tools to support the aspect cluster growing strategy.

In this section, we first discussed the evidence that supports the use of spatial-semantic visualization as a means of supporting information seeking. We then outlined the visualization pipeline, which describes the process of spatial-semantic visualization and places our research questions within the context of this process. Hence, the remaining purpose of this chapter is to review the literature pertaining to the three main research questions. In the next section, we focus on question one and, through a review of the empirical evidence, demonstrate that it may be possible to model create the required two-level semantic model using a simple text analysis procedure.

2.4. Modelling topical structure

The kinds of spatial-semantic visualization described in the previous sections, and required by our interaction mode, are created using automatic (unsupervised) procedures. They therefore depend on the assumption that documents that discuss the same or closely related concepts tend to have quantitatively similar representations within the underlying data model. In most cases, inter-document similarity measurement is made possible by representing documents as high-dimensional term occurrence vectors. We now discuss the vector space model and examine the theoretical basis for its utility in modelling the topical structure of a document collection.

2.4.1. Vector space model

In the early days of online IR, documents were represented as bibliographic data, using a strictly controlled vocabulary for content descriptors. Modern systems now also index and match documents to queries according to their literal content, be it full text or abstract only. Retrieval is no longer dependent upon a perfect match; document relevance can be calculated on a continuous scale based on the similarity to the query (which might comprise dozens of terms). This allows two things: the use of more natural language or free form (as opposed to faceted Boolean) queries and more fine-grained relevance ranking of retrieved documents.

Similarity searching of primary content is made possible by the application of the vector space model (VSM) to document representation. The origins and development of the core principles and techniques of the vector space model approach can be traced to the SMART project, which began at Harvard University in 1961 (Salton, 1991; Salton and McGill, 1983). In this model, documents are represented as high-dimensional vectors where each dimension is associated with a unique term (e.g., word or phrase). A term vector for a given document therefore represents a profile, within a common space (term vocabulary), of term weights that can be directly correlated to either a query, represented in the same format, or other document vectors. This correlation provides a measure of general similarity between any two items. The most commonly used similarity metrics in IR are based normalised measures based on the dot product such as Cosine and Dice (see Korfhage, 1995). These produce continuous values in the range of 0 (no similarity) to 1 (identical).

The primary focus of this work is the use of vector space representations to build a *semantic model* of a given document collection or sub-set, rather than to perform retrieval *per se*. The term semantic model is used here to describe both the term-document vector space representation of documents (*term - document matrix*) and the matrix of inter-document similarity values (*similarity matrix*) derived from the comparison of document vectors. We will refer to the process of creating a vector space model and deriving a similarity matrix as *automatic text analysis*.

The creation of a semantic model and particularly the similarity matrix is a computationally expensive procedure. It normally begins by parsing the document texts to identify all unique terms that occur. The size of this 'vocabulary' can increase rapidly in as the number of documents considered increases, particularly if the nature of the content is quite diverse. For instance, in the test scenarios we build in Chapter 3, a set of 127 reasonably short newspaper articles contains over 5000 unique words.

Traditional indexing heuristics can be applied to reduce the size of the vocabulary (see Salton and McGill, 1983). For instance stop-words (e.g., and, their, also) can be removed as can low frequency words (e.g., those that occur in only one document) and stemming can be applied to merge morphological variants of words (see Porter, 1980). Even after such measures, however, the size of the common term space for a collection of even a few hundred documents is likely to remain in the thousands. Computing document similarities for all possible pairs from a vector space of this magnitude can be computationally expensive as *$(N^2 - N)/2$* comparisons, where N is the number of documents, must be made to create a complete similarity matrix. Hence, there have been many attempts to reduce the size of vocabularies even further, including the application of a statistical factor analysis on the term-document representation to represent terms as a smaller number of common factors (LSA: Deerwester et al., 1990) and the replacement of terms with concepts modelled using neural networks (Wise et al., 1995; Lin et al., 1991). In this dissertation, however, we do not explore the relative merits of these more advanced techniques, focusing instead on the potential of semantic models produced using a standard term vector space approach.

2.4.2. The cluster hypothesis

Having covered the application of the vector space model to semantic modelling, we now consider the theoretical rational for applying clustering and scaling techniques to these models in order to improve the representational structure of documents retrieved by a given query.

The cluster hypothesis of IR states that closely associated documents tend to be relevant to the same requests (Van Rijsgergen, 1979). In the classic information retrieval model, the goal of the query reformulation process is to move the query vector closer to the centre of the cluster of relevant document vectors. The corollary of the cluster hypothesis is that relevant documents, for any given query, should be more similar to each other than they are to other, non-relevant documents within a collection.

If this derived hypothesis is true, this leads to the possibility that applying cluster algorithms to either the document collection or a sub-set of it may be a valuable tool for IR system design (van Rijsbergen, 1979). Since the cluster hypothesis was proposed, there have been two main applications of clustering in experimental IR systems: improving recall and efficiency of retrieval by pre-clustering the collection and matching queries to the centroids (mean vectors) of document clusters rather than individual documents; and post-retrieval clustering as a means of improving the organisation of search results.

2.4.3. Cluster-based retrieval

The first attempts to exploit the clustering properties of document vectors applied a one time hierarchical clustering to an entire document collection (Willett, 1988). In cluster-based retrieval, when a query is issued, the system searches the cluster tree in either a bottom-up or top-down fashion retrieving all documents belonging to clusters that match the query above some threshold similarity score. There have been many variations on this strategy, however evaluations of experimental systems suggest that successful results only occur when the IR system comprises a relatively small document collection (Willett, 1988).

Voorhees (1985) compared traditional sequential searching with cluster-based searching across four different collections. The author also compared these strategies to a hybrid strategy where individual documents within matching clusters are matched to the query rather than just retrieved by default. The results showed that the cluster-based searching generally resulted in poorer performance than the other two strategies. Also, while all strategies were affected by the general extent to which relevant documents clustered within a collection, highly cohesive relevant sub-sets did not tend to favour the cluster-based retrieval strategy.

2.4.4. Retrieval set organisation

Hearst and Pederson (1996) hypothesised that the poor performance of cluster-based retrieval methods might be partly because inter-document similarity and therefore tendency to cluster was seen as a static property that could be computed once and independently of all possible query situations. They suggested that the relative similarity of a pair of documents would depend upon the context in which they were considered. The logic of this is sound: if document A is about cats and dogs and document B is about only dogs, then when considered within the context of a query focused on dogs they would potentially be quite similar, but dissimilar were the query focused on a cat related topic.

Based on their assumptions regarding the importance of context, Hearst and Pederson (1996) proposed the use of dynamic, post-retrieval clustering where only the frequency of terms that characterised the retrieval set were used to calculate document similarity. Evaluation of this interface, called Scatter/Gather, showed that clustering documents on the basis of similarity within the 'local context' of the query reliably produced solutions where the majority of relevant documents would tend to be assigned to the same 1 or 2 clusters within a 5-cluster solution.

The importance of regarding document similarity as a context-dependent property has been further emphasised by Tombros and van Rijsbergen (2001), who suggest that when clustering a retrieval set, the measurement of document similarity should be biased towards the co-occurrence of terms that appear within the user's query. Their approach yielded a more coherent clustering between relevant documents compared to traditional similarity measures that treat all terms appearing within the retrieval set as equal.

On a more pertinent note, Rorvig and Fitzpatrick (1998) have evaluated post-retrieval document organisation using MDS derived spatial-semantic solutions. They found that, using an appropriate scaling technique, most documents relevant to the query tended to converge, forming a single dense, 'bulls-eye' cluster within the centre of the visualization. The work of Leuski (2001), which we described in section 2.2, has also demonstrated the tendency of relevant documents to cluster within a scaled solution or retrieved documents, as evidenced by the promising results of his cluster growing strategy.

2.4.5. Aspectual cluster hypothesis

Wu et al. (2001) also conducted a study of post-retrieval clustering, similar to Hearst and Pederson (1996), but deliberately studied more complex topics. These topics, from the TREC interactive track (see Over, 1997), comprised relevance judgements that were sub-divided into distinct aspects of relevance to the topic, allowing them to test not only the extent to which relevant documents converged on the same cluster(s) but also the extent to which same aspect documents converged. They too found documents that were relevant to the topic as a whole tended to converge on one or two clusters (solutions generally comprised six or seven clusters). However, documents relevant to the same aspect did not generally tend to converge on same cluster, as one might expect.

The work of Muresan and Harper (2004) sheds some light over why clustering might have failed at the aspect level. Their studies showed that, for complex topics, there was a non-reciprocal relationship between relevance and similarity. This is summarised in their aspectual cluster hypothesis, which states that:

> *Similar documents tend to be relevant to the same requests, but documents relevant to the same requests are not necessarily similar. They tend to be dissimilar if they cover different aspects of the same complex topic*
>
> (Muresan and Harper, 2004, p.896)

Their experiments (which also used the TREC interactive test collection) showed that the distribution of computed similarities between relevant document pairs was positively skewed, with many values approaching zero. When they considered only document pairs that discussed the same aspect of the topic, they found that the skew, whilst still apparent was far less pronounced, and the mean similarity of same aspect documents was significantly greater than mean topic similarity and, in turn mean set similarity (all similarities).

They also studied a large range of clustering solutions. Like Wu et al. (2001) they found that most solutions comprised a small proportion of good clusters containing most of the relevant documents. However, they also noted that documents in the best clusters tended to be ones that were highly topical; they discussed more than one aspect of the relevant topic. The explanation for this is that the more aspects of the relevant topic that a document discusses, the more likely it is to be highly similar to another relevant document. A non-hierarchical (i.e., k-means) clustering algorithm will aim to find large groups (depending on the target number of clusters) of generally similar objects, hence highly topical (multi-aspect) documents stand the best chance of being allocated to the cluster that contains most of the relevant documents. The reverse consequence of this is that documents that are highly focused on only one aspect of a complex topic are likely to be, on average less similar to other relevant documents, and as such tend to be scattered over the cluster structure. This goes some way to explaining the poor outcome of Wu et al.'s (2001) clustering study.

Muresan's (2002) solution to the problem was to develop a system that assisted the user in generating multiple queries, each one being focused on a distinct aspect of relevance. Evaluation of this system produced positive results, however this mediated retrieval system, WebCluster, is dependent upon the availability of an appropriate, existing, structured resource that can be browsed in order to identify a set of aspect exemplars. These exemplars are then used to formulate a set of focused aspect-queries that are subsequently issued to a larger document index. In his study, this resource was manually constructed for the purpose of the experiments. In effect our approach is dealing with the problem of how to automatically generate such a useful structured resource from documents retrieved from an early, tentative query.

The implications of Muresan's (2002) findings, combined with those of Leuski (2001) and Wu et al. (2001) suggest that clustering may not be an appropriate document organisation technique for our needs. This is because the identification of a relevant aspect exemplar in a particular cluster would not necessarily help the user in locating further documents about that aspect, as they stand a good chance of occurring in other clusters. For example the user might be browsing a good representative cluster, the expected strategy when using a discrete clustering solution (Hearst and Pederson, 1996), and come across a document that discusses the novel aspect, A. It appears in the good cluster because it also discusses aspects B and C. Unfortunately, the two other documents that discuss aspect A discuss only that relevant aspect. According to the aspectual cluster hypothesis, the consequence is a high likelihood they will reside in another cluster: the cluster structure provides no clues as to where to find other documents that discuss aspect A. Furthermore, if they are highly distinct with respect to the relevant topic, and also discuss other non-relevant topics there is no guarantee that they will reside in the same cluster. In other words a sub-set of three documents discussing the same aspect could quite easily be scattered across three clusters.

A central hypothesis in this work is that spatial-semantic document organisation will be less affected by this problem because association between documents is represented along a two-dimensional, continuous scale rather than by discrete membership. In theory, a document that discusses more than one aspect can be placed at a point of inter-section between these aspect sub-sets. The anticipated consequence is that there will be a good chance that aspects comprising both highly topical and aspectually distinct documents will not be grossly separated within the organisational structure. There is currently no direct evidence to support this notion. Although previous studies have examined general topic clustering (i.e., Rorvig and Fitzpatrick, 1998; Allen et al., 2001), as far as we know there has been no work that has formally evaluated the tendency of distinct aspects to cluster within a spatial-semantic visualization of a retrieval set, although Swan and Allan (1998) have shown how spatial-semantic visualization can be used to determine which newly retrieved documents are most likely to be relevant, but aspectually-distinct from those already retrieved. Hence, such an evaluation is primary aim of this thesis (question two) and is discussed further, in this chapter, in section 2.6.

More pertinent at this stage is the question of whether the classification observed by Muresan and Harper (2004), where relevant documents that discuss the same aspect of the topic are more similar to each other than relevant documents that discuss different aspects, generalises to the context of our interaction model. Their test bed consisted of 175 documents known to be relevant to six different TREC Interactive topics, and 572 documents judged to be non-relevant to these six topics.

In our interaction model the set to be organised is retrieved with one topic in mind, but is likely to contain many documents discussing other topics. Whilst non-relevant documents might well form topical clusters, this possibility is not considered in our following analyses (Chapters 3 to 5). We are only interested in creating one main relevant cluster that is reasonably distinct from non-relevant items, and in organising the contents of this cluster according to aspects of the relevant topic. A major concern, that the trend observed by Muresan and Harper (2004) might not be observed in a query-retrieved set, stems from the fact that within such a set, many documents will be similar to relevant documents, despite being non-relevant to the intended topic. In Muresan's study (Muresan, 2002; Muresan and Harper, 2004) documents were manually selected on the basis of relevance and near relevance to several distinct topics. This would have almost certainly exaggerated the difference between the same-topic and all document distributions. Hence, there are no guarantees that the same hierarchical structure can be produced in this context.

Given this we now return to our first research question that we posed in Chapter 1: *To what extent can a standard text analysis procedure model the general semantic structure expected by our interaction model and particularly the low-level structure required by the aspect cluster growing strategy?*

By semantic structure we mean a two-level hierarchical classification. The first level of the hierarchy consists of relevant and non-relevant documents, and the second level is broken down into aspects of relevance. For our purposes this is a non-symmetric hierarchy, as we do not consider the topical or aspectual structure of non-relevant documents. Following the results of Muresan and Harper (2004), we would expect a general trend where documents discussing the same aspect will be most similar, documents discussing different aspects of the relevant topic to be significantly less similar, and documents discussing different topics to be least similar.

Muresan and Harper (2004) also found that the allocation of individual documents to the second level nodes is not exclusive: some documents will discuss more than one aspect of the topic. Whilst we would expect that documents allocated to the same node at the second level to generally be the most similar pairs within the collection, we would also expect the extent to which aspect sub-sets overlap to affect the degree of separation between the same-topic and same-aspect similarity distributions.

In the next section, we construct methodologies for testing both the general classification hypothesis and the potential success of the aspect cluster-growing hypothesis. We form general hypotheses relating to the expected success of these tests when applied to the semantic models associated with specific test scenarios. We also form a specific hypothesis relating to the effect of aspect overlap on relevant classification within the semantic model. Finally, we return to our earlier discussion of the importance of a topically-focused context by considering and formulating a hypothesis with respect to the effect of set size on relevant classification within a semantic model for a given test scenario.

2.5. Testing the cluster hypothesis

In evaluating the feasibility of our interaction model, we could just progress directly to a proof of concept by applying our semantic models to, and evaluating the results of, various spatial-semantic layout algorithms. Whilst this is an intuitive approach, this methodology alone is flawed because it ignores the variation in configurations that are possible from one clustering or scaling algorithm to another. For instance, Rorvig and Fitzpatrick (1998) only observed the characteristic bulls-eye effect for relevant documents when they applied a particular type of scaling that implemented a maximum-likelihood estimation procedure.

Following on from our discussion of the cluster and aspectual cluster hypotheses, in this section we argue the importance of testing the potential for relevant documents to be clustered, or their classifiability, by studying the clustering properties of documents within the vector space itself, before performing and evaluating any practical clustering experiments. This is important for two reasons. First, if relevant documents do not cluster in vector (similarity) space according to the expected topology, then it is unlikely that clustering will be successful and it may be beneficial to first look at alternative methods of modelling the semantic structure of the collection prior to attempting any kind of clustering. Second, if analysis shows evidence of the required classification structure in similarity space, then poor clustering performance of a particular algorithm should motivate attempts to first test alternative algorithms before considering an outright rejection of the cluster hypothesis.

There are two main, traditional approaches to testing the cluster hypothesis from this perspective. In this section, we begin by reviewing these methods. We then explain why these methods are inadequate, in their existing format, for testing our interaction model as they only consider document relevance as a simple binary property (relevant or non-relevant). We then introduce the test that Muresan (Muresan, 2002; Muresan and Harper, 2004) used to test the aspectual cluster hypothesis. We present a revised version of this test that reflects the goals of our hypotheses.

2.5.1. Cluster hypothesis tests

There are two well-known approaches that have been used to test the cluster hypothesis from this fundamental perspective. Although these apply a simple binary model of relevance, rather than the hierarchical model that we are interested in, it is worthwhile outlining these approaches first as they view the problem of testing the cluster hypothesis from quite different perspectives.

The original cluster hypothesis test, which we will refer to as the separation test, was proposed by Jardine and van Rijsbergen (1971) and is also discussed later by van Rijsbergen in his book (van Rijsbergen, 1979). Positive results from early applications of this test were used to demonstrate the potential of cluster-based searching within specific collections (Jardine and van Rijsbergen, 1971; van Rijsbergen and Sparck-Jones, 1973).

In this test, given a test collection and set of queries, two distributions of values are calculated. This first comprises all similarities between relevant document pairs (R-R). The second distribution comprises all similarities between pairs of relevant and non-relevant documents (R-NR). The operational hypothesis is that mean R-R will be significantly higher than mean R-NR, meaning that relevant documents tend to be more similar to each other than they are to non-relevant items.

Voorhees (1985) argued that the separation test was flawed because it concealed the effect of non-relevant documents that were also highly similar to relevant documents. For example, an R-NR sample may contain an equal number of highly similar document pairs to the R-R sample, but because the former sample is larger, these strong similarities contribute relatively little to the mean. Hence, it is possible for there to be a significant difference between the R-R and R-NR distributions even though there may be a significant number of non-relevant documents that are equally, if not more similar to relevant documents than other relevant documents. The extent to which relevant documents form exclusive clusters is clearly important when it comes to both cluster-based searching and dynamic clustering application. Given this, Voorhees proposed the nearest neighbour test.

The test involves taking all relevant documents, for all queries comprising more than one relevant document, and counting the number of relevant documents occurring in the top n (Voorhees used n=5) most similar documents for each case. Voorhees found that nearest neighbour values varied widely between collections, providing an explanation as to why cluster-based searching tended to be less successful in some test collections than others.

One can view these two tests as providing complementary data on the suitability of a document set for clustering. The separation test simply measures the extent to which relevant documents will tend to form a cluster in term space, whilst the nearest neighbours test provides a measure of the extent to which relevant documents tend to form exclusive clusters.

2.5.2. Testing the aspectual cluster hypothesis

Whilst these tests are useful for measuring relevant document clustering for simple topics, in this work we are interested in clustering complex topics. More specifically we seek an asymmetric hierarchical classification that distinguishes relevant from non-relevant documents at the top level and distinct aspects of the relevant topic at the second level.

Muresan (2002) proposed and applied a test for this kind of classification. Originally, Muresan intended to simply adapt the cluster separation test by including the similarity distributions for same aspect and different aspect pairs. However, the impact of aspect overlap, where documents discuss more than one aspect, created a problem. Aspect overlap would mean that many document pair similarity values would contribute to both aspect level distributions – a pair that were similar on aspect A could also be dissimilar with respect to aspect B.

Muresan therefore proposed a simplified version of the separation test. In this test three distributions are calculated. These are: *all similarities* between all document pairs within the set; *topic similarities* between all pairs of topically relevant documents within the set; and *aspect similarities* between all pairs of topically relevant documents that discuss the same aspect. The hypothesis was steady increase in mean from all, through topic to aspect similarity distributions.

Muresan's test treats each valid document pair as a single case. Inevitably, the distributions increase in size as the semantic focus becomes less specific, meaning that there are considerably more topic similarities than aspect similarities and considerably more set similarities than topic similarities, making analysis of variance comparisons problematic.

Given this and the nature of our interaction model, we modify Muresan's separation test slightly. The aspect cluster growing strategy demands that any relevant document should be a good aspect cluster growing exemplar, that is the document is more similar (therefore proximal in visual space) to documents that discuss the same aspect than to documents that discuss different aspects or different (non-relevant) topics. Hence, we treat each relevant document, rather than each similarity value *per se*, as a distinct case.

In our test, we generate three distributions of *k* cases, where *k* is the number of relevant documents for the given set. For each case we compute mean aspect similarity, mean topic similarity and mean set similarity. This test, therefore, measures the cluster separation of same-aspect, same-topic documents within the overall distribution of documents present in a given collection. If our cluster growing strategy is feasible then as the comparison set becomes semantically broader, we would expect mean similarity to drop. For continuity, we will refer to this test as the *aspect cluster separation (ACS) test*. Hence, our first hypothesis (H1), which considers all topical scenarios under study, is that:

H1: *The two level classification structure (topic and aspect cluster separation) will be evident for all scenarios whereby relevant documents will be, on average, more similar to the sub-set of documents that discuss the same aspect(s) than they are to the sub-set of generally relevant documents and, in turn, least similar to the retrieval set as a whole.*

This test has the same limitation of van Rijsbergen's (1979) test: it can prove that aspect similar documents tend to be more similar than documents that discuss different aspects or different topics, but it does not allow us to predict the potential precision of the cluster growing strategy, for instance how many relevant but aspect different documents, or non-relevant documents intermingle within the same aspect cluster. We therefore propose another, complementary test based on the nearest neighbours test (Voorhees, 1985). We call this the *nearest aspect neighbours (NAN) test*. In this test we measure, relative to each relevant document, the rank order position of the first and second same aspect documents. These raw measures can be analysed in pure form or we can calculate a variant on the R-precision measure used in the TREC evaluations (see Muresan and Harper, 2004). R-precision is the precision at rank R where R equals the number of relevant documents. However, as exemplars will vary widely in their aspect sub-set size, we apply a standardised precision measure, which we term R2-precision, to compare between cases. R2-precision is the precision of the explored sub-set at the point where the second relevant document is found.

It is difficult to set a concrete hypothesis for this test when applied to a given semantic model as we are considering a single distribution. However, Muresan and Harper (2004), when evaluating their mediated query techniques, report that nearest-neighbour R-precision values for single exemplar aspect queries in their test collection averaged around 0.18. Research also shows that searchers typically have little patience for browsing further than around 10 to 20 items in ranked list presentation format (Jansen et al, 2000). We therefore consider two positive finds within 10 documents (10-precision=0.2) to indicate a reasonable criterion for a successful search. As such for H2 we will be looking for an average R2-precision of at least 0.2 (the rank of second closest relevant document will tend to be equal to, or less than, 10):

H2: R2-precision for NAN in similarity space will be equal to or exceed 0.2 in most exemplar cases.

Naturally, if H1 and H2 are supported, we need to know how best to translate this classification faithfully to a spatial-semantic layout. We can apply these same tests to the inter-document proximity data associated with our spatial-semantic solutions. In particular, we find that our nearest neighbour test applied to spatial-semantic proximities provides a suitable means of simulating the basic aspect cluster growing strategy, where the user is expected to view documents in proximity order to the exemplar. We set hypotheses relevant to this question and discuss methodology later in section 2.5.

First, however, we consider two factors that are likely to influence the fidelity of the classification that we are seeking in our semantic models: these are aspect overlap and retrieved document set size.

2.5.3. Aspect overlap

The success of the aspect cluster growing strategy depends upon the extent to which documents relevant to each aspect form reasonably coherent and exclusive sub-sets within scaled space. In order for this to be possible, the necessary structure must at least be present within the semantic model. In the previous sub-section, we specified a general hypothesis that predicted a two level classification whereby relevant documents tend to form a general cluster within the general vector space of the retrieved set and, in turn, aspects form coherent and distinct sub-clusters within this general cluster. We outlined two tests that allow us to test this hypothesis with respect to vector space models.

The tendency of aspect sub-sets to form distinct clusters will depend upon the extent to which the document members are conceptually similar to each other and distinct from other documents in the retrieved set. Ideally, all relevant documents should be focused texts that discuss only one aspect of the topic. In reality documents may talk about more than one aspect of the relevant topic and many other concepts besides. Furthermore, topical structure is likely to vary within an aspect sub-set from one document to the next.

We need to study both the lower and upper bounds of conditions that might face our interaction model. As such, we will compare two topical scenarios - a *topical scenario* in this context comprises an open-ended question or topic, a set of known aspects of that topic, and a retrieved document set that contains one or more documents relevant to each of those aspects. We will choose one where the topical structure is conducive to aspect cluster growing and another where the use of the strategy is more challenging.

A conducive topical scenario is one where relevant documents (from our test collection) tend to focus mainly or only on one definable aspect of the relevant topic. In other words, aspects are relatively distinct within the context of the similarity matrix because these related documents will tend to be relatively similar to each other in comparison to other relevant and non-relevant documents. For the more challenging scenario, we will select a topic where many relevant documents tend to discuss several aspects of the topic. From the work of Muresan and Harper (2004) we know that relevant documents that are more topical in nature (discuss several aspects) tend to converge on to large thematic clusters and may thus become relatively segregated from other documents that discuss only one related aspect of the topic, particularly when dimension reduction algorithms (e.g., clustering) are applied.

This means that we would expect that in the more challenging scenario, same aspect documents will tend to be spread more broadly around relevant document nodes, particularly those specific documents that are known to discuss many aspect. In other words, we would expect the difference between mean aspect and mean topic similarities to be smaller.

Correspondingly we would also expect, for the challenging scenario, that the local neighbourhoods of relevant documents would be more likely to contain a mixture of different aspects, meaning that for a given document and specific aspect, the nearest relevant neighbours will be relatively less highly-ranked than would generally be the case in the more conducive scenario.

Hence, with respect to the first perspective, our third hypothesis (H3) is:

***H3**: In the overlapping aspect scenario, topic and aspect level cluster separation and mean R2-precision scores will be lower than in the distinct aspect scenario.*

2.5.4. Document set size

The main argument for dynamic rather than static document clustering is that document similarity is a dynamic quality that is dependent upon the context in which it is considered (Hearst and Pederson, 1996). Two documents that are highly similar within the context of a topically precise retrieval set may be relatively dissimilar when considered within the context of a large document collection.

In our interaction model, we assume our searcher is unable to specify a precise query but they are able to specify one or two key terms that broadly define their topic. Although a large number of non-relevant documents will remain, retrieving documents relevant to such a query will significantly increase the salience of the topic and its aspects within the set of documents to be browsed. Most importantly, this will be reflected within the vocabulary used to define document vectors, where terms that define the topical structure will form a much larger proportion of all terms and therefore play a larger role in defining inter-document similarity.

Potentially, such a broad query could still retrieve a very large number of documents. An important question is what proportion of the top ranking retrieved documents should be retained and visualised? First, there is a trade-off to be made between maximising recall of relevant documents and maximising precision, which is likely to fall as recall increases (Salton and McGill, 1983). Second, as precision drops so does the salience of the topic within the conceptual space. Furthermore, as set size increases the complexity of spatial-semantic layout increases exponentially (see section 2.6.2), resulting in more node misplacements and therefore potentially poorer aspect clustering. Furthermore, if documents are to be represented by distinct nodes (visual marks), then problem of displaying these nodes legibly also increases with set size.

We return to spatial-semantic issues relating to set size in section 2.6.5. With respect to classification fidelity within the semantic model we would expect this to decrease as set size increases, which leads to our fourth hypothesis:

H4: *In the smaller retrieval set scenario, topic and aspect level cluster separation and R2-precision scores will be greater.*

2.6. Optimising layout for aspect cluster growing

Having considered the importance of testing classification properties of the underlying semantic model and appropriate methods for doing so, in this section we take the next step forward to consider spatial-semantic visualization issues. We critically discuss different approaches to spatial-semantic visualization within the context of our interaction model. We suggest that an algorithm that focuses on optimising local structures may be more effective than more commonly used algorithms that attempt to create globally optimal solutions. We therefore propose a comparison between algorithms of each type.

The aim of spatial-semantic visualization is to represent the inter-document similarities described in high-dimensional vector space as accurately as possible as proximities in two or three dimensional visual space. More specifically, the resulting proximities must partition relevant documents from non-relevant ones and most pertinently of all, for the purpose of aspect cluster growing, partition the aspect sub-sets.

In this section, we consider the issues associated with translating the required structure from similarity to visual space. Hence, we assume that hypotheses 1 and 2 are supported and the problem is one of choosing the most appropriate layout algorithm. This relates directly to question 2, which asks: *Given an adequate semantic model, which approach to spatial-semantic layout best preserves the general and, in particular, the low-level structure expected by our interaction model?*

To this end, we review the common spatial-semantic visualization approaches, most notably multi-dimensional scaling (MDS) algorithms, hybrid approaches that combine discrete clustering with MDS and factor analysis. We then explain that the principle barrier to achieving a good translation is the fact that similarity information is lost by the dimension reduction process and that the main consequence of this with existing approaches is that only major features (clusters) are retained, at the expense of more minor conceptual relationships. We then present an alternative approach where we view layout as a graphing problem where, instead of trying to preserve all inter-document similarities, we focus only on presenting the minimum-spanning tree (MST) of the complete network implied by the similarity matrix. We hypothesise that MST will produce more appropriate visualizations for our interaction model. We then proceed to discuss the potential mediating factors that are expected to affect the success of the aspect cluster growing strategy in spatial-semantic visualizations.

2.6.1. Common approaches to spatial-semantic layout

The most common techniques used to create spatial-semantic visualizations belong to a class of algorithms that can be collectively referred to as multi-dimensional scaling (MDS). Although these approaches vary in the models that are applied, they are similar in that they all aim to optimise the mapping between input similarities and output proximities. This is identical to the goal of the spatial-semantic metaphor, making them an intuitive choice.

The traditional approach to MDS is to start with a random configuration and to make iterative adjustments to object locations in order to maximise the 'goodness of fit' between input similarities (or dissimilarities) and output proximities. There are several tests that can be used to measure this fit, the primary one being a stress function that measures the degree of disparity between input and output proximities. Additionally, fit can be measured in terms of the squared correlation coefficient (*r*-squared) between the input and output data that measures the percentage of total variance accounted for by the MDS configuration. Normally the algorithm continues until the observed improvement in the stress function for the last iteration drops below a certain threshold.

The development of MDS algorithms began in the 1950s (Torgerson, 1952), as computer technology made possible the complex calculations required to produce scaled solutions. Early metric approaches (e.g., Torgerson, 1952) were followed in the 1960s by non-metric MDS (e.g., Shepard, 1962; Kruskal, 1964), which relaxed the constraint on inter-object distances needing to be parametric in nature. This development was significant in that it allowed application of the technique to a much broader range of domains such as, for example, document similarity visualization where similarity data may not necessarily normally distributed. For a full discussion of the origins of MDS see Young and Hamer (1987). The most common algorithms in use today include Alternating Least Squares Scaling (ALSCAL: Takane, Young and De Leeuw, 1977) and PROXSCAL (Busing, Commandeur and Heiser, 1997). Both these algorithms support a full range of MDS models (including metric and non-metric scaling) and are available for use in statistical applications like SPSS and SAS. PROXSCAL, however, is a more recent evolution of MDS and is generally accepted as superior to ALSCAL, most notably because the criterion used for optimization is based on distances rather than squared distances.

The main use for MDS has been for exploratory analysis of either pure similarity data (e.g., human judgements of object or concept similarity) or similarity data derived from high-dimensional common attribute spaces (e.g., questionnaire responses). As discussed in section 2.3.1, in our application, inter-document similarities are generally measured by calculating the angle (e.g., Cosine, Dice) between high-dimensional document term vectors.

There are several examples of document visualizations created using this traditional approach to MDS (Wise et al., 1995; Hornbaek and Frokjaer, 1999; Westerman et al., 2005). Wise et al. (1995) used a metric MDS algorithm to generate the Galaxies visualisation, which represents the semantic space of medium sized document collection onto 3D space. The name was chosen because of the visual effect of a star-field that was produced, with thematically similar document nodes forming 'constellations' within the overall visual structure. Hornbaek and Frokjaer (1999) applied MDS to visualise a collection of 436 documents assembled from a bibliography of human-computer interaction resources. They further augmented the legibility of the visualization by selecting the top 20 most discriminating terms (terms that are common in a few documents) using a function provided by Salton and McGill (1983) and locating each term as a label at a location on the plane that represented the central point of its most common occurrence. As noted in section 2.3, users were attracted to the clusters that emerged from the spatial-semantic layout. Large, dense regions of nodes tended to attract the attention of users, particularly if there was an interesting term attached. More recently, Westerman et al. (2005) applied MDS (ALSCAL) to study the effects of dimensionality (2D vs. 3D) on topic retrieval in spatial-semantic visualisations.

In addition to traditional MDS techniques, there is an alternative sub-class of approaches we will call force-directed placement algorithms. These were created to solve the problem of producing aesthetically pleasing layouts of undirected graphs such as networks and trees but, as we will discuss, can also be applied to spatial-semantic visualization tasks by viewing the similarity matrix as a fully connected, or complete graph where similarities are equivalent to edge weights.

Essentially the graph, composed of nodes (often called vertices in the graphing literature) and links (often called edges), is treated as a physical system whereby nodes can be thought of as rings and those that are joined by edges are connected by springs. The springs exert an attractive force that pulls connected nodes together. These attractive forces are balanced out by repulsive forces that act between all pairs of nodes regardless of whether they are connected. When the algorithm is run, the sum effect of these forces is calculated at each iteration and vertices moved accordingly. These iterations continue until the system reaches a state of low energy (or stress).

The original approach to force-directed placement was called the spring-embedder model (Eades, 1984). This is a fairly simple system where the attractive force is equal for all edges, being calculated as a function of the log distance between edges multiplied by a constant. Repulsive forces are calculated as the inverse square of the distance between each pair of vertices. More recent refinements have improved upon Eades' (1984) algorithm. For example, the algorithm proposed by Kamada and Kawai (1989) incorporates Hooke's law into the force calculations, meaning that springs can have a natural length that the algorithm aspires to preserve. This is useful for graphs with weighted edges and thus particularly useful for representing similarity between document nodes.

By viewing the document similarity matrix as a complete, non-directed graph we can apply these algorithms to spatial-semantic visualization. A classic example of the application of force-directed placement for document layout is BEAD (Chalmers and Chitson, 1992; Chalmers, 1993). However, these algorithms are designed for partially rather than complete networks. Leuski (2001) warns that applying force-directed placement attempts to a complete, weighted graph can result in a very tight and somewhat amorphous formation that retains little spatial-semantic structure. His solution for Lighthouse, following Swan and Allan (1998), was to minimise the attractive forces between nodes where the similarity was below a certain threshold, by squaring these values. By effectively pruning the less important edges of the graph it was possible to produce visualisations where sub-sets of related documents formed distinctive clusters (see figure 2.3, for example). In section 2.6.3, we return the notion of edge pruning when we consider alternative approaches to optimising spatial-semantic structure for aspect cluster growing.

Another approach to spatial-semantic layout is factor or principal components analysis (PCA). In PCA the dimensionality of the vector space is reduced to a much lower number of factors. These factors are linear functions of the original dimensions (e.g., term weights or document similarities) that are independent from each other (account for separate portions of the overall variance). Documents are then plotted according to their weight along the top two or three dimensions. The end results of PCA can be very similar to the MDS techniques already described, although the approach to node placement is quite different. First, it is a definite, statistical procedure, unlike MDS where there may be multiple final solutions depending on the starting configuration of nodes and the number of iterations allowed before the configuration is accepted as a solution. Second, the aim of the procedure, as far as spatial-semantic visualization is concerned, is to map nodes to visual space according to the two independent factors that together explain the most variance, rather than to preserve the correspondence between document similarity and node proximity.

The benefit of PCA over MDS is that the dimensions (axes) of the visualisation tend to be more explicit and meaningful and can be labelled if required to support overview and comparison of features. The relative disadvantage is that PCA often does not scale well. As the complexity of the semantic model increases so the top two or three factors will account for less of the overall variance, which can result in a fan-like configuration where many documents reside at the origin because they have little or no relation to either of the principle factors (see Chen, 1999a; Cribbin and Chen, 2001).

A key problem with either traditional and force-directed MDS approaches is that the relationship between node set size and computational complexity associated with finding a globally optimal (low stress) solution is exponential in nature: each time the number of nodes doubles the number of calculations that must be performed at each iteration quadruples. This places significant limits on the feasibility of using MDS for dynamic or interactive applications. Furthermore, as set size increases so does dimensionality of the underlying vector space, which leads to greater information loss during dimension reduction and thus more node misplacements (see sub-section 2.6.2).

A solution to this is to divide the problem of layout into smaller chunks. There have been a number of recent examples of algorithms that combine clustering with MDS. The general aim is to place nodes in relation to thematic points of interest rather than trying to preserve all possible inter-relationships within similarity space. This is an evolution of the principle advocated by an earlier system known as VIBE (Olsen et al., 1993), where points of interest were specified query terms rather than derived concepts.

For instance the SPIRE project team found that the practical maximum for traditional MDS was around 1500 documents (Wise, 1999). In order to visualize larger sets (up to 6,000 documents) they applied the anchored least stress (ALS) algorithm to the Galaxies visualization (Wise, 1999). ALS applies clustering to the data first. The centroids (mid-points) of these clusters are then projected onto a visual plane (using PCA). Finally, documents are projected onto the same plane at location that best represents their relative similarity to each cluster, rather than each document. Wise (1999) notes that there are benefits of this technique, not only in terms of the reduction in computation time, but also because the algorithm places a greater emphasis on conveying the most important themes in the document node configuration, rather than focusing on small adjustments between document node pairs.

A similar approach was adopted by Andrews et al. (2002) with their InfoSky system. In this system documents are assigned to a hierarchical classification, which can be either pre-existing or dynamically computed. When the user selects a node in the hierarchy all documents sub-ordinate this node are visualised in the following steps. First, the centroids of all the sub-ordinate classes are mapped to visual space based on their similarity. A bias is introduced to ensure that sibling classes of the hierarchy tend to cluster. Second, documents belonging to these each sub-ordinate node are then organised by similarity within a bounded region surrounding their centroid.

Hence, the essence of both of these techniques is the same: to first organise themes or points of interest according to their similarity, then to locate individual documents according to their relationship to these points of interest. Both of these systems produce similar, galaxy type visualizations where documents that discuss a predominant theme form distinct 'constellations' within the configuration.

In summary, MDS approaches are generally most widely used due to their intuitive appeal and are most commonly applied for visualising moderately sized document sets. PCA can produce similar results to MDS, within more definable dimensions, but the value of this technique depends upon the size and topical complexity of the document set being visualised. Divide and conquer approaches that combine clustering with MDS can reduce computation time for large spaces and produce more distinctive, thematic structures.

2.6.2. Dimension reduction problem

A fundamental obstacle in spatial-semantic visualisation is that whilst we are limited to perceiving the correspondence between objects in at most three spatial dimensions, the dimensionality of the semantic models of interest can run into many thousands. Such drastic dimension reduction inevitably leads to compromises in node placement in the resulting spatial configuration whereby unrelated nodes may be located proximally whilst similar nodes are placed unexpectedly distally.

Following the example in chapter 1, mapping a matrix of inter-city proximities to 2D space is a trivial task as the dimensionality of input space is equal to the output space. All proximities are preserved perfectly. Now let us imagine the more problematic task of mapping the structure of an equilateral triangle, with vertices A, B and C to a single dimension (a line). By placing, A, B and C in sequence within equal distances between nodes, we can map the proximities AB and BC perfectly but the resulting distance AC is twice what it should be (AB + BC). If we attempt to resolve this by moving C to the same location as B this preserves AB and AC but the relationship BC is obscured. Whichever combination of node locations we try we always end up with a degree of disparity between the input and output proximities. We would get the same problem when we try to visualise the edges of a pyramid, a 3D structure, in 2 dimensions.

In order to create a perfect solution in 2D space, the rule of triangle inequality must be followed for all possible combinations of three nodes. This rule states that, for any three nodes, AC cannot be greater than the sum of AB and BC. In a similarity matrix of tens or hundreds of documents that is derived from a high-dimensional semantic model, there will be many instances where this rule is violated. The dimension reduction algorithm will attempt to find the best compromise, but inevitably disparities will occur regularly. Sometimes these will be quite minor and in other cases relatively drastic.

As previously noted, using 3D spatial representation can produce a solution of higher fidelity (Leuski, 2001; Westerman and Cribbin, 2000) but this comes at the expense of usability: nodes are occluded, absolute distances are difficult to judge in the z-plane, harder to build a cognitive map of the structure. Westerman and Cribbin (2000) found that the fidelity of spatial-semantic mapping of a 3D solution had to be at least 40% higher than a 2D solution, and for some measures twice as high before any net gains in user search performance were observed. Likewise, Leuski (2001) found that the benefits of 3D over 2D observed in the simulated user trials did not translate into superior performance amongst real users. In fact user performance was slightly but significantly poorer in the 3D condition.

In summary, whilst the additional dimension provided by 3D visualization can convey the semantic model more accurately, users cannot capitalise on this extra information. We need to seek an alternative strategy for creating more informative visualizations using only 2 dimensions. One approach is to utilise an algorithm that is selective in terms of the semantic features that are preserved during layout.

A major limitation with the MDS family of algorithms is that they seek a globally optimal solution. In seeking to reduce the stress in the solution, MDS places the same emphasis on all pairs of nodes. This not only makes the task computationally expensive, with $(n^2 - n)/2$ pairs to consider, but also means that the compromise in placement is spread equally across the whole of the structure.

In other words, MDS sees all document similarities as equally important. Given what our discussions of retrieval set clustering, it is apparent that this is not the case. According to the cluster hypothesis, relevant documents tend to be highly similar relative to the distribution of all document similarities (van Rijsbergen, 1979). According to the aspectual cluster hypothesis, documents relevant to the same aspect of the topic will be highly similar (Muresan and Harper, 2004). In our interaction model, we are most concerned with representing same aspect document similarities and, to a lesser extent, same topic similarities. If H1 and H2 are correct, then aspects represent distinct features or localities within similarity space. It may be more prudent, therefore, to select an algorithm that places an emphasis on retaining local features rather than global structure.

2.6.3. Local optimisation

In sub-section 2.5.1, we discussed the value of hybrid approaches that combine clustering with MDS (Wise, 1999; Andrews et al., 2002). The primary advantage of these techniques is that they scale well, reducing the complexity of the node layout problem by representing document nodes in terms of their similarity to thematic points of reference rather than each and every other node in the set. This makes them well suited to organising larger and topically diverse collections. However, we envisage our interaction model will generally be applied to just the top ranked portion of a retrieved set, a few hundred documents at most, rather than thousands of documents, so layout computation time is not a primary issue.

Although scalability is not a current issue in our case, the combination of clustering and MDS seems initially appealing as a means of biasing the spatial-semantic structure towards emphasising strong local features or emergent themes. Additionally the points of interest themselves, if labelled, would provide useful overview landmarks within the visualization.

However, the value of this approach would depend on the ability of the clustering algorithm to isolate the concepts of interest. We know from previous studies of document clustering that whilst major themes (e.g., the general relevant topic) are easily identified, more specific and minor themes are easily lost. As a reminder, both Wu et al. (2001) and Muresan and Harper (2004) found that documents relevant to specific aspects were often split across the cluster structure. The problem with clustering is that it is highly parametric in nature – the determination of values for factors such as the number of specified clusters (i.e., k-means) or the choice of partition level (i.e., hierarchical clustering) usually requires extensive trial and error. What we seek for our interaction model is a procedure that can run in an unsupervised fashion and still reliably preserve the most salient, intra-aspect document relationships.

A second, more promising path is to look again at the layout problem as a graph drawing problem. In our discussion of force-directed placement algorithms we noted how limiting the magnitude of certain attractive forces between document nodes produced more distinctive structures. Leuski (2001) adopted a threshold strategy where similarities below a specified value were squared to minimise their effect on the final configuration. By effectively pruning weaker links so that the configuration is based mainly on the effects of stronger similarities, Leuski (2001) was able to produce a structure where documents clearly divided into cluster sub-sets.

This approach is likely to require some trial and error to determine the optimal threshold value, with different optimal values likely for each topic and its associated retrieved set. So as with clustering and hybrid approaches, we are running into the problem of setting parameters that are likely to be moving targets, influenced by numerous variables such as topical complexity, document set size and so on.

An alternative, but related solution is to apply some absolute criterion when deciding which edges to retain. We have strong evidence to suggest that within the distribution of document similarities, same aspect document pairs will tend to be amongst the most strongly related document pairs (Muresan and Harper, 2004). We seek to confirm this characteristic within the context of our topical retrieved set scenarios by testing H1 and H2. This leads us to think of spatial-semantic layout as one of emphasising the shortest paths between documents in the set.

Minimum spanning trees (MST) are a class of algorithm that, given a connected, undirected weighted graph, seek to find sub-graph that is the spanning tree of minimum cost. A spanning tree is a sub-graph where all nodes are connected to at least one other node. There may be many spanning trees for a given graph, but the MST is the one where the summed total of retained edge weights is the lowest.

Prim's algorithm (Prim, 1957) computes the MST by growing a single tree until all nodes are connected. The algorithm begins by selecting the lowest cost edge, which forms the beginnings of our MST. For our purposes this would be the document pair with the highest similarity of all document pairs. The next iteration searches for the lowest cost edge that would connect an unconnected node to a node that is currently in the tree. This edge is selected and the edge selection iterations continue until all nodes are connected to the tree. An alternative to Prim's is Kruskal's (1956) original MST algorithm. The main difference is that Kruskal's (1956) algorithm proceeds to build a forest of trees that ultimately become connected into a single MST. Hence, on each iteration, the next lightest edge is selected regardless of whether either node is already part of the tree, providing it does not form a cycle (connect two nodes that are already indirectly connected).

An MST always has N-1 edges, just enough to connect each node to the resulting tree. There are no initial parameters to set, which means the algorithm can find an optimal (or near optimal) solution without any supervision. We add the near optimal clause because there may be branching points during the execution of the algorithm when it encounters ties, that is equally viable candidates, meaning that there may be more than one MST for a given graph.

Closely related are Pathfinder networks (PFNET: Schvaneveldt, 1989) which resemble MSTs when mapped to visual space. The main difference in appearance is that PFNETs tend to retain slightly more than N-1 edges of the original graph. This is because the algorithm allows cycles to occur in the structure (which is why they are networks and not trees) providing the triangle inequality condition is met. PFNET and MST are closely related, for example a minimum cost PFNET can be thought of as the set union of all possible MST solutions (see Chen, 1999b).

Spatial layout of an MST or PFNET can be easily accomplished using a force-directed placement algorithm. For example, later in this dissertation we introduce Neato, part of the GraphViz toolkit from the AT&T Laboratory (see North, 2002), which uses the algorithm developed by Kamada and Kawai (1989) to layout the undirected graph.

Both MST and PFNET have been applied to various document visualization tasks. For instance, Chen has applied PFNETs to author citation (Chen et al., 2002) and co-citation networks (Chen, 1999a; Chen and Paul, 2001) in order to visualise the structure and evolution of knowledge domains such as scientific fields. Chen has also compared MST and PFNET (Chen and Morris, 2003) for visualizing co-citation networks. Generally, Chen favours PFNET over MST (Chen and Morris, 2003; Chen, 1999a) for knowledge domain visualization because the cycles that emerge provide more complete communication of salient local features.

However, Cribbin and Chen (2001) compared MST, PFNET and PCA visualizations across a range of topics and associated information retrieval tasks. Participants browsed these spatial-semantic visualizations (200 newspaper articles) in search of documents relevant to range of increasingly specific queries (each subsequent query formed a sub-set of the previous query). Both MST and PFNET visualizations enabled better retrieval performance on the tasks than the PCA visualization and, correspondingly, participants felt these visualizations were easier to navigate and were less cluttered than PCA. However, there were some differences between MST and PFNET. Compared to the PCA condition, participants were much faster in locating the first relevant document when using PFNET but MST users were not significantly faster. In contrast, retrieval performance, as measured by the harmonic mean of retrieved document precision and recall, was significantly better compared to PCA user, for MST users but PFNET users (although there was a substantial mean difference between PFNET and PCA).

More interestingly, the results showed that for several measures, the differential between PCA and MST/PFNET performance was greatest for queries that required the location of just two closely related documents (both documents discussed the same event). Specifically, browsing was considerably more efficient in MST/PFNET compared to PCA. In terms of differences in actual retrieval success (relevant documents marked), MST users performed consistently better than PFNET and PCA.

This was only a small study (N=16) and it is difficult to draw any firm conclusions about the relative superiority of MST and PFNET. What we can conclude from Cribbin and Chen's (2001) study is that the spatial-semantic structure provided by MST is consistently more useful for a range of retrieval tasks than the PCA structure, and at least as useful as the structure provided by PFNET.

In the analysis that follows in Chapters 4 and 5, we will compare aspect clustering and simulated cluster growing performance in both locally and globally optimised visualization schemes. We will examine the utility of MST rather than PFNET. Only one is chosen, as any differences in structure are likely to be small between the two types of structure. MST is selected as it represents the most extreme level of edge pruning possible within a single graph. The globally optimised comparison scheme is PROXSCAL, a modern evolution of traditional MDS (Busing, Commandeur and Heiser, 1997). Force-directed placement was rejected due to reported problems associated with visualising complete (fully connected) graphs when similarities are derived from full-text vectors (see Leuski, 2001).

As a reminder, question 2 asked: *Given an adequate semantic model, which approach to spatial-semantic layout best preserves the general and, in particular, the low-level structure expected by our interaction model?*

Firstly, we will test the hypothesis that a faithful representation of the required two-level classification structure is present to a significant extent in at least one of the two visualization schemes. We will apply the ACS test applied to earlier hypotheses relating to structure of the underlying semantic models (H1, H3 and H4).

H5: *The two level classification will be effectively conveyed by spatial relations in (i) MDS and (ii) MST.*

Next, we predict that MST, due to its emphasis on preserving local structure, will locate relevant documents more closely to other same aspect documents than MDS:

H6: *Aspect level cluster separation will be greater for MST visualizations than for the MDS visualizations.*

Finally, we predict that the aspect cluster growing strategy will be more efficient in MST due to a higher chance of nearest neighbours of known relevant exemplars being also relevant. We will test this by simulating user performance of the aspect cluster growing strategy for a large range of exemplar/specific aspect cases. The simulated strategy function is a repeat of the NAN test applied to H2, H3 and H4, using scaled proximities rather than similarities.

H7: *Aspect cluster growing will be more efficient when using the MST visualizations compared to the MDS visualizations.*

2.6.4. Aspect overlap

In section 2.5.3 we discussed how the suitability of our interaction model, and particularly the aspect clustering growing strategy, might be affected by the extent to which the sub-sets of documents relevant to each aspect overlap. On this basis we proposed that we should evaluate the effect of two types of topical scenario that differ in terms of the average number of aspects discussed per relevant document.

Aspect overlap would not be a significant problem if the overlap between sub-sets was symmetrical, that is all documents that discuss a specific aspect discuss the same secondary aspects. The result would likely be a highly focused and topical cluster that would be easy to search for all of the discussed aspects.

A more realistic situation, however, is that members of the relevant sub-set for a given aspect will differ somewhat in their topic structures. Some may discuss only the one aspect whilst many may be more topical, discussing many different aspects that may or may not be shared within other aspect members. In the high-dimensional space of the semantic model, it is possible that such complexity can be accommodated and effectively represented.

In dimensionally reduced spatial-semantic representations, however, this kind of complexity is liable to cause certain compromises to be made in the layout of nodes. With respect to the use of clustering algorithms, Muresan (2002) concluded:

> *Clustering algorithms tend to group together documents that cover focused topics, or aspects of complex topic. Documents covering distinct aspects of complex topics tend to be spread over the cluster structure.*
>
> (Muresan, 2002, p.244).

In clustering solutions, at least, relevant documents behave differently depending on whether they are highly topical or relatively distinct in their content. Documents, which discuss several aspects tend to converge on the highly topical clusters because they tend to be relatively similar to a critical mass of relevant documents. Likewise, if a specific aspect is well represented within the set, so long as the relevant documents are highly focused on that aspect, they may converge on and dominate a particular cluster. However, if documents are distinct, discussing only a minor aspect of the topic, they could be assigned almost arbitrarily to a cluster.

It is known that highly topical documents tend to converge on a central dense cluster within retrieved set visualizations (Rorvig and Fitzpatrick, 1998; Leuski, 2001). It is not clear to what extent MDS or MST solutions can cope with situations where, for instance, the aspect sub-set is small and distinct, or where the sub-set is composed of both highly topical and highly focused documents.

Our next hypothesis predicts that the topical scenario where aspects overlap will be more challenging for both of our layout algorithms, resulting in poorer ACS (H8) and lower aspect clustering growing efficiency (H9).

H8: *Aspect level cluster separation will be lower in the overlapping aspect scenario than the distinct aspect scenario.*

H9: *Aspect cluster growing will be less efficient in the overlapping aspect scenario compared to the distinct aspect scenario.*

In terms of differences between the two layout schemes, we might conjecture that MST will cope well with distinct, focused aspect sub-sets due to its emphasis on preserving salient local relations whereas MDS, with its focus on global optimisation may do a relatively good job of organising sets comprising many highly topical documents perhaps finding a more balanced comprise in such situations. We expected MST to be generally better at aspect clustering than MDS (H6, H7), however we would expect that MST would have a greater advantage when mapping the scenario containing more distinct aspect sub-sets. Hence, we form the hypothesis:

H10: *The expected differences between MST and MDS will be greatest for the distinct aspect scenario.*

2.6.5. Document set size

As we discussed earlier in this section, the problem of spatial-semantic layout increases exponentially with document set size. We would therefore expect the fidelity of solutions to decrease in line with set size:

H11: *Aspect level cluster separation will be lower in visualizations of the larger retrieval set.*

H12: *Aspect cluster growing will be less efficient when using the larger retrieval set.*

Comparing the two schemes, we would expect MST to be more resistant to the complexity introduced by increasing set size, principally because the number of inter-document similarities that must be preserved increases only linearly, rather than exponentially as in the case of MDS. Furthermore, document pairs that were highly similar in the smaller set should also be relatively similar in the larger set so links that are present in the smaller set should, to a great extent be retained in the larger set. Hence, our hypothesis is:

H13: *The expected differences between MST and MDS will be greatest for the larger retrieval set.*

2.7. Refining local context cues

Spatial-semantic visualization is a process of dimension reduction. Whichever layout algorithm is used, disparities between the structure of the similarity matrix and that of the visualization are inevitable. In some cases, the user will find the spatial-semantic cues alone are sufficient to grow an aspect cluster. In other cases, relevant documents may be badly situated in relation to other relevant documents.

An obvious solution is to employ some sort of relevance feedback mechanism that can then be used to dynamically augment the visualization. There are two possible ways of providing this feedback to the system: 1) the user supplies some key terms based on key features of their aspect of interest or 2) the user simply asks the system which are the most similar documents to this exemplar.

In this section, we explore the possible solutions to the problem of what to do if spatial-semantic cues are inadequate to guide aspect cluster growing. We begin by explaining why it is inappropriate to expect the user to formulate a query indicating their intention, before discussing alternative, interactive strategies that might resolve the problem.

2.7.1. Query in context

Previous studies have demonstrated the principle of allowing users to see the results of a query within the context of a spatial-semantic visualization, for instance by highlighting document nodes that are relevant to the query. For instance, users in Hornbaek and Froekjaer's (1999) study were drawn towards particularly dense clusters of matching documents. However, in our interaction model, the emphasis is on a consistent mode of interaction, where users browse throughout the whole interaction episode. Studies including Hornbaek and Frokjaer (1999; Campagnoni and Ehrlich, 1989) have found that forcing users to switch between different interaction modes (i.e., referential to command line input) causes additional cognitive demands that break the flow of the primary information seeking task. Also, choosing good key words to query a full-text, uncontrolled index is not always an easy task. For instance, choosing terms that are too broad or polysemous could lead to the user being overwhelmed by highlighted, but non-relevant documents.

2.7.2. Resolving the effect of node misplacement

Instead, we opt for a strategy where the user simply needs to indicate to the system that the current document is relevant. A similarity search is then performed using the document vector as the query. This is familiar to users of web search engines where it is presented in the form of "show me more like this" and can be termed simple relevance feedback

(Hearst, 1999) in that it does not require the multiple document judgements required by conventional document relevance feedback.

The results of a document similarity search can be presented in the visualization in the same way as the manual query results were in Hornbaek and Froekjaer (1999), by highlighting the top ranked most similar nodes. The use of node highlighting (e.g., a colour change) has the advantage that the spatial-semantic cues remain stable. Furthermore, the user has the choice of following either cue individually or combining both together.

Leuski (2001) found that raw inter-document similarity cues as opposed to spatial-semantic cues lead to significant improvement in cluster growing performance, although the absolute differences in precision were quite small (a few percent). Of course, these results came from a study where the topics were quite homogeneous and formed relatively coherent clusters. In more complex topics, we would expect a greater amount of misplacement of nodes in relation to their aspect sub-sets. It will be interesting to see the extent to which augmenting the space with such similarity cue will increase aspect cluster growing performance. Our next hypothesis considers the utility of using similarity cues alone and is as follows:

H14: *The majority of problematic cluster growing cases are due to node misplacements and can thus be resolved by augmenting the visualization with relative similarity cues.*

2.7.3. When similarity cues fail

We anticipate the possibility that in some cases even pure inter-document similarity cues might be insufficient to guide the user in their search. Factors such as vocabulary mismatch (Furnas et al., 1987) and the conceptual diversity of aspect documents (including the exemplar) are likely to impact on the tendency for same-aspect documents to be identifiable by means of a measure of a simple measure of lexical similarity.

In this dissertation, faced with this problem we seek a solution that helps the user to specify their intention when nominating a document as an aspect relevant exemplar, as opposed to one that will increase the general similarity between same-aspect documents. This is not to ignore the possibility that intra-aspect similarity can be enhanced through more advanced methods of text analysis. However, we accept the reality that unsupervised text analysis will always produce cases where the topical relationship between documents is not appropriately reflected in their inter-document similarity score.

Our approach to this problem is to first perform an analysis that allows us to gain a clearer understanding of the conditions under which spatial-semantic and, more fundamentally, inter-document similarity cues fail. By understanding the nature of the exemplars or the specific aspect sub-sets that are associated with poor performance we aim to develop refinements to the strategy or tools that go beyond general similarity to provide more informative cues to the user.

For example, it may be the case that the best aspect exemplars are documents that are highly relevant to the original query. If this is the case, then the task of identifying distinct aspect exemplars would be best achieved by browsing the top ranks of the retrieved list (Leuski, 2001), rather than through browsing the visualization directly. Alternatively, we may find that poor exemplars are primarily those that discuss multiple aspects of the topic, and thus fail as exemplars for certain aspects because they tend to be more similar to other aspect sub-sets. If this is a typical case then we would need to develop tools that would simplify the task of specifying the salient conceptual facets of the exemplar that relate it to other documents within the retrieved set.

Given the latter observation, the problem would become one of query refinement. Given that we wish to avoid the need for the user to manually specify their query, our attention is turned to the field of automated and semi-automated query expansion. We have already introduced the nature and role of query expansion (QE) in Chapter 1. The classic approach works using document relevance feedback, whereby the user specifies a number of documents (i.e., from the retrieved list resulting from the current query) and the system extracts the discriminating terms from these documents and adds these to the query. This approach, however, relies on the user specifying multiple good examples of relevance and so may be problematic in situations where only one good exemplar is known or when the aspect is only represented by two or three documents within the retrieved set.

Promising alternative approaches are those do not require the user to provide any relevance feedback at all. Local feedback (Attar and Fraenkel, 1977) and local context analysis (Xu and Croft, 1996; Xu and Croft, 2000) work by assuming that the top ranking documents to the current query are mostly relevant to the intended query, thus saving the user from the responsibility of making document judgements. Local feedback works in a similar way to standard relevance feedback, expanding the query using terms that are relatively common within the local context of the query. Local context analysis (LCA) is a more sophisticated approach that selects new query terms based on the extent to which they co-occur with existing terms (Xu and Croft, 1996; Xu and Croft, 2000); the key assumption is that the best terms will be those that occur in the same contexts as all or most of the existing query terms. By paying attention to the context in which existing query terms occur within retrieved documents, a key benefit of LCA over local feedback is that it can select good expansion terms even when a large number of non-relevant documents appear within the top ranks of the retrieved set (Xu and Croft, 2000) and a recent, independent evaluation study concluded that it can perform comparably against traditional relevance feedback based query expansion, and that users preferred LCA because of the reduced effort involved (Belkin et al., 2000).

However, these approaches have only been proven in situations where the query expansion process begins with a manually defined query. In our case, the query is a document term frequency vector that may imply a broad range of concepts. Even a poorly defined user-defined query is likely to be more specific than an entire document vector. Furthermore, if problematic exemplars tend to be those documents that are most heterogeneous in content, we envisaged that this would present considerable problems for existing query expansion approaches. However, such an approach may work more effectively if the user is allowed to intervene in the query expansion process.

Term relevance feedback is a promising approach that might ameliorate the ambiguity associated with cases of where the only query is a single document exemplar. This approach can be based on existing query expansion approaches, such as those described above, as essentially it simply involves adding an extra step to the feedback process. Rather than automatically adding terms to the query, the user is allowed to choose, from the list of candidate terms identified by the system, those that are most relevant to their query and should therefore be added. Koenemann and Belkin (1996) compared term relevance feedback to standard 'opaque' document relevant feedback. They found that satisfactory queries were achieved in fewer feedback iterations if users were allowed to control which terms were added to the system. Search effectiveness when using the term relevance feedback system was significantly better than the control condition (manual query reformulation), and slightly better than the standard document relevance feedback system.

It is possible that some combination of local feedback or context analysis and term relevance feedback could provide a useful tool to support aspect cluster growing if problems do seem to arise from heterogeneity either in the exemplar or the local context of similar documents. Belkin et al. (2000) have already evaluated the combination of LCA with term relevance feedback and achieved promising results, however, to our knowledge, this combination has not yet been evaluated in situations where the query is a document. We envisage that the exact design of an effective term suggestion tool would depend upon the particular conditions associated with problematic exemplars. In chapter 5 (section 5.4) we model the conditions associated with problematic aspect cluster growing cases by exploring a number of variables relating to the structure of the exemplar itself and the retrieved documents that are semantically related to the exemplar. As this was an exploratory analysis, taking place within the context of the results of our previous analyses, it makes no sense to set *a priori* hypotheses at this stage of the dissertation. However the rationale for the variables explored during this analysis is outlined in detail in section 5.4.

In chapter 6, we use the findings from this exploratory analysis to develop a term suggestion algorithm, called local context distillation, which allows the user to pick the terms, from a suggestion list, that best specify the reason for their interest in the given exemplar. We also demonstrate two visual tools that provide two different applications of local context distillation terms.

2.8. Summary of questions and hypotheses

The purpose of this chapter was to define the conceptual framework that justifies and directs the programme of work reported in Chapters 3, 4, 5 and 6. We have defined hypotheses relating to our three main research questions and described the general methodological approach that will be used to test these hypotheses and explore related questions.

To summarise, we have formulated 14 hypotheses that will allow us to address the three research questions that were first put forward back in section 1.7. These hypotheses are as follows:

Research question one: To what extent can a standard text analysis procedure model the general semantic structure expected by our interaction model and particularly the low-level structure required by the aspect cluster growing strategy?

H1: *The two level classification structure (topic and aspect cluster separation) will be evident for all scenarios whereby relevant documents will be, on average, more similar to the sub-set of documents that* *discuss the same aspect(s) than they are to the sub-set of generally relevant documents and, in turn, least similar to the retrieval set as a whole.*

H2: *R2-precision for NAN in similarity space will be equal to or exceed 0.2 in most exemplar cases*

***H3**: In the overlapping aspect scenario, topic and aspect level cluster separation and mean R2-precision scores will be lower than in the distinct aspect scenario.*

H4: *In the smaller retrieval set scenario, topic and aspect level cluster separation and R2-precision scores will be greater.*

Research question two: Given an adequate semantic model, which approach to spatial-semantic layout best preserves the general and, in particular, the low-level structure expected by our interaction model?

H5: *The two level classification will be effectively conveyed by spatial relations in (i) MDS and (ii) MST.*

H6: *Aspect level cluster separation will be greater for MST visualizations than for the MDS visualizations.*

H7: *Aspect cluster growing will be more efficient when using the MST visualizations compared to the MDS visualizations.*

H8: *Aspect level cluster separation will be lower in the overlapping aspect scenario than the distinct aspect scenario.*

H9: *Aspect cluster growing will be less efficient in the overlapping aspect scenario compared to the distinct aspect scenario.*

H10: *The expected differences between MST and MDS will be greatest for the distinct aspect scenario.*

H11: *Aspect level cluster separation will be lower in visualizations of the larger retrieval set.*

H12: *Aspect cluster growing will be less efficient when using the larger retrieval set.*

H13: *The expected differences between MST and MDS will be greatest for the larger retrieval set.*

Research question three: Under what conditions does the aspect cluster growing strategy tend to fail and how can we use this knowledge to guide development of interactive support tools?

H14: *The majority of problematic cluster growing cases are due to node misplacements and can thus be resolved by augmenting the visualization with relative similarity cues.*

Hypotheses H1, H2, H3 and H4 (research question one) are dealt with in Chapter 3, where we begin by describing the development of our test scenarios and semantic models. Hypotheses H5, H6, H8, H10, H11 and H13 (research question two), which focus on the expected two-level relevance classification, are tested in Chapter 4, where we begin by describing the creation of our spatial-semantic visualizations. Chapter 5 tests the remaining hypotheses associated with question two, focusing on the potential performance of the aspect cluster growing strategy (H7, H9, H10, H12 and H13). Hypothesis H10 (research question three) is also tested in Chapter 5, where we conclude by performing an exploratory analysis that allows us to specify the requirements of the interactive solutions that are subsequently presented and demonstrated in Chapter 6.

We therefore begin our analyses in the next chapter by describing the construction of our topical scenarios and testing the hypotheses relating to research question one.

CHAPTER 3: MODELLING TOPICAL STRUCTURE

3.1. Introduction

In chapter 2, we discussed the challenges associated with the successful implementation of our interaction model, focusing particularly on the requirements for the aspect cluster growing strategy. The three key questions that drive this research are incremental in nature and are each related, in turn, to a successive stage of the spatial-semantic visualization pipeline (see section 2.3.4): modeling semantic structure using automatic text analysis, mapping the derived inter-document similarity structure to visual space and finally user interaction with and augmentation of the derived visualizations.

In this chapter we focus on the first stage of this pipeline, modeling semantic structure. Question one asked: *To what extent can a standard text analysis procedure model the general semantic structure expected by our interaction model and particularly the low-level structure required by the aspect cluster growing strategy?*

We apply text analysis to create a semantic model of a given set of documents. Our approach to text analysis works by converting document texts to a word term based vector space representation from which inter-document similarities are computed by measuring shared variance between document vectors. This is a long-standing approach (see Salton and McGill, 1983; van Rijsbergen, 1979) that has been consistently applied in several successful studies of general topic (e.g., Hearst and Pederson, 1996; Rorvig and Fitzpatrick, 1998) and aspect level (Muresan and Harper, 2004) document clustering.

This chapter does two things. Firstly, we describe the creation of the semantic models for the three topical scenarios that will form the context for our analyses. Secondly, we begin our analyses by applying cluster hypothesis tests to these semantic models to determine the potential success of trials to produce spatial-semantic visualizations that will support our interaction model and, in particular, the aspect cluster growing strategy.

This chapter is organized as follows: In section 3.2, we describe the creation of our topical scenarios. In section 3.3, we describe the text analysis procedure used to create our semantic models and summarise and compare the general distributions of inter-document similarity scores for each scenario. In section 3.4, we describe and justify the procedure used to collection our experimental data that we used to perform the two cluster hypothesis tests (ACS and NAN: see section 2.5) that form the core of our analyses. In section 3.5, we present the analysis from the ACS test, which provides us with an insight into the relevant classification properties of our semantic models. In section 3.6, we estimate the maximum performance of the aspect cluster growing strategy by performing the NAN test based on pure inter-document similarity data. Finally, in section 3.7, we present the solutions produced by a discrete clustering algorithm. A previous study of aspect level clustering, using a discrete clustering algorithm, showed disappointing results (Wu et al., 2001). The purpose of this evaluation is to demonstrate the importance of first verifying the cluster hypothesis in high-dimensional space; that layout algorithms can fail despite the relevant structure being present within the semantic model. These solutions also provide further benchmarks against which to compare the spatial-semantic visualizations that we create in Chapter 4.

In the remainder of this section, we outline the rationale for first verifying the cluster hypothesis for our semantic models, define the nature and origin of our topical scenarios, and finally present the formal hypotheses that we will test in our analyses.

3.1.1. Verifying the cluster hypothesis in similarity space

The main requirement for aspect cluster growing is that documents are organized as nodes in visual space such that those discussing the same aspect of the relevant topic form coherent clusters. As the only input to the layout algorithm, in the second stage of the pipeline, is the high-dimensional semantic model, it is critical that the desired classification structure is present within this structure.

This is because considerable information loss, in relation to the underlying, high-dimensional semantic model, is inevitable during clustering and spatial-semantic layout (see section 2.6). Much of the information within the semantic model is likely to be redundant or non-critical with respect to the intended purpose of the solution. Clustering and visualization algorithms deal with the dimension reduction problem by applying a wide range of optimization strategies and criteria. In many cases, key parameters (e.g., number of clusters, similarity thresholds) must be optimized by trial and error in order to achieve a satisfactory result (see Leuski, 2001; Rorvig and Fitzpatrick, 1998). For this reason, it is important that we experiment with and compare different algorithms and different optimization criteria in order to identify the layout algorithm that best preserves the required structure (in our case the two level relevant cluster separation). For this reason, in chapter 4 we compare two different approaches to spatial-semantic layout optimization.

However, in this chapter we argue that prior to comparatively evaluating layout algorithms, it is important to ensure whether (and the extent to which) the required structure is present within the high-dimensional semantic model. As the inter-document similarity matrix is the sole input to the layout algorithm, if the structure is not present for any given topic and document collection, then any layout approach is likely to fail and it will be fruitless to perform an extensive comparison. On the other hand, if the structure is present in the semantic model, but initial visualization approaches fail, then this would indicate that it is worth seeking and testing alternative approaches or refinements to the layout process. Muresan and Harper (2004) caution that studies in document clustering may sometimes fail not because relevant documents are not similar, but simply because the clustering algorithm or algorithms used were not able to organize documents in the required manner.

Given the complexity of the structure we wish to convey, we argue that it is particularly important that we first evaluate the properties of our semantic models. We achieve this using the cluster hypothesis tests (the ACS and NAN tests) that were developed at the beginning of section 2.5. These tests measure the extent to which the desired classification structure is present within the original high-dimensional vector space model. To reiterate, conducting such tests is important for two reasons: firstly, if the required topology is not present in the underlying model then it is unlikely that attempts to produce useful clustering or visualization models will be successful and would suggest the need to identify more appropriate semantic modeling techniques. Secondly, if the underlying topology is present but the visualization experiments are unsuccessful, then we know that the failure is due to the layout algorithm and can focus on identifying more effective methods in this respect. Our decision to perform this analysis is vindicated by the results of the analyses that follow. In section 3.7, we show that despite observing good clustering of semantically similar documents, both at the topic and aspect level of relevance, the solutions created by a discrete (*k*-means) clustering algorithm fail to aggregate many of the aspect sub-sets into the same clusters. Furthermore, in chapters 4 and 5 we show that the spatial-semantic visualization approach can produce much more coherent organization of same-aspect documents.

We now briefly describe the nature and origins of our test bed topical scenarios, before concluding this introduction with an outline of formal hypotheses to be tested.

3.1.2. Origins of the topical scenarios

Each of our semantic models and their respective spatial-semantic visualizations are created from a topical scenario. Our test bed consists of three topical scenarios. A scenario consists of a topic definition (an open-ended question), a set of topical aspect definitions (aspects of relevance), an ad hoc document set retrieved from a test collection using a simple, high-recall query, and a set of relevance judgments describing the relevance of each of the retrieved documents to the defined topic and aspect definitions.

This test bed will form the basis of the analysis we will use to seek answers the hypotheses set at the end of chapter 2. It consists of three scenarios each associated with a topic taken from the Text Retrieval Conference (TREC) interactive track (Voorhees & Harman, 1997, 1998).

TREC is an annual conference that provides a forum for testing and evaluation of experimental (and live) information retrieval systems. Each year the organising committee specifies a set of information seeking problems relating to a number of specified task types (e.g., question answering, ad hoc retrieval, cross-language retrieval). Participants compete, normally within the context of a specific task, to test their IR system or interface against other competing participants. Each participant applies their system to the same test collection of documents. For a given topic and task, the documents most commonly retrieved by participants' systems are pooled and evaluated for relevance manually by an independent judge or 'assessor'. The rich reference data that results from these activities provides a coherent set of benchmarks against which new systems can be evaluated and compared with earlier systems.

Whilst this reference data is derived from IR experiments, it can also be exploited for the purpose of evaluating document visualisation systems. Traditionally, evaluations of visualizations are often quite bespoke in design, where researchers test their own systems in isolation using bespoke tasks and document collections. This approach not only makes comparison of similar systems (across studies) difficult (see Chen and Yu, 2000) but is also requires considerable experimenter and user overhead making it expensive and time consuming. Following suggestions made in 1996 at the Second Annual Workshop on visual information retrieval interfaces (see Rorvig and Fitzpatrick, 1998), there have been several examples of where the TREC data collections have been used to evaluate document visualization systems and techniques. For example, published studies have emerged that exploit this comprehensive test bed to test either complete interface systems (e.g., Swan and Allen, 1998; Allen et al., 2001; Leuski, 2001; Wu et al., 2001) or specific layout algorithms (e.g., Rorvig and Fitzpatrick, 1998; Sullivan and Rorvig, 1998).

In this work, we use topics taken from the interactive track of TREC-6 (Voorhees and Harman, 1997) and TREC-7 (Voorhees and Harman, 1998). The interactive track is perfect for our purposes, because the associated topics come in the form of open-ended questions, where the task for the participants is to explore the test collection, using their IR system, to identify at least one instance of as many distinct aspects of the topic as possible. As such, the pooled relevance data not only specifies which documents are relevant, per se, but also defines distinct aspects of the topic and specifies which documents discuss each of the defined aspects.

Once we have constructed our scenarios (section 3.2), we perform a text analysis on each to generate a semantic model (section 3.3), comprising a term vector-space model of document representations and a derived matrix of inter-document similarities. It is the similarity matrices that we work with to evaluate the 'raw' potential for automatic document organization. We first perform two types of cluster hypothesis test (sections 3.5 and 3.6) before evaluating actual explicit document organization potential in the form of discrete (k-means) clustering solutions (section 3.7). As mentioned in the previous sub-sections, we include a discrete clustering solution to vindicate our decision to perform cluster hypothesis testing at the level of the semantic model, and to compare the properties of our scenarios to those of earlier studies that used discrete clustering to classify complex topics (e.g., Wu et al., 2001; Muresan and Harper, 2004). We identify some benefits of such a scheme, for our interaction model, but a number of limitations are also discussed, particularly in relation to the problem of performing focused aspect searches.

3.1.3. Experimental hypotheses

Question one asked: *To what extent can a standard text analysis procedure model the general semantic structure expected by our interaction model and particularly the low-level structure required by the aspect cluster growing strategy?*

The aim of our analysis in this Chapter is to test the following related hypotheses:

H1: The two level classification structure will be evident for all scenarios whereby relevant documents will be, on average, more similar to the sub-set of documents that discuss the same aspect(s) than they are to the sub-set of generally relevant documents and, in turn, least similar to the retrieval set as a whole.

H2: R2-precision for NAN in similarity space will be equal to or exceed 0.2 in most exemplar cases

H3: Topic and aspect level clusters will be less cohesive in the scenario where aspect sub-sets tend to overlap more and mean R_2-precision scores will be lower.

H4: Topic and aspect level clusters will be more cohesive for smaller retrieval sets of the same query and mean R2-precision scores will be higher.

3.2. Creation of topical scenarios

In this section, we describe how the topical scenarios were selected and created, along with a brief description of their characteristics. The test bed of scenarios we create is based upon topics and documents compiled for the purposes of the 6th and 7th TREC conferences (Voorhees & Harman, 1997, 1998). The test collection used for both years Interactive Tasks comprised 210158 documents sourced from the Financial Times (FT) Newspaper during the period 1991-1994.

3.2.1. Selection of topics

Two topics were selected, one from TREC-6 and one from TREC-7. The first general criterion for selection was that the topic should contain a reasonable number of distinct aspects (at least 10). This guarantees a challenging and realistic level of complexity (e.g., compared to the Samuel Adams example in section 2.2) given our task context of an open-ended question. It also provides us with a sufficient number of cases to conduct inferential statistics based on our cluster hypothesis tests. The second general criterion was that, for the sake of the intended cluster separation and aspect cluster growing experiments, a good proportion of these aspects must be discussed in two or more known (as judged by the

TREC assessors) documents. Finally, given the first step of our interaction model, the third general criterion was that it should be possible to retrieve a good proportion of relevant documents (high recall) using a short, simple query, whilst, at the same time, maintaining a reasonable level of precision (ratio of relevant to non-relevant documents). We decided that it should be possible to achieve such a set using just two super-ordinate key terms OR'd together as they might be in a typical tentative search by a naïve user (see Jansen et al., 2000). Finally, given our hypotheses, we required that the scenarios differed significantly in the degree of aspect overlap; the extent to which relevant documents refer to more than one aspect of the topic.

All interactive track topics from TREC-6 and 7 were evaluated by their relevance data and candidates that met the first two criteria were short-listed. Test retrievals were conducted for short-listed topics and recall levels evaluated. From the remaining candidates we chose topics 347i and 352i as they differed starkly in their degree of aspect overlap. Table 3.1 summarises these topics. A full list of aspect descriptions is included in appendix A1.

3.2.2. Retrieving the document sets

Simple software was developed to retrieve our test bed sub-collections. The aim was to retrieve sub-collections that contained most if not all relevant documents (as identified by TREC judges) for a given topic along with any other non-relevant documents that happened to satisfy the given query. As such, it was not seen as necessary to implement a full IR system, with term-document indexing. Instead, a simple sequential query-document matching procedure was used. To reduce search time, a working sub-collection of 26094 documents (approx 12.5% of the collection) was created to reduce the search time for each query trial. This sub-collection comprised all documents that were marked as likely relevant based on the pooled data submitted by participants of the interactive task tracks across the TREC-6, 7 and 8 conferences combined (TREC-8 relevant documents were included to allow this sub-collection to be reused for acquisition of further scenarios in future work). This created a sub-total of 2119 documents. The remaining 23975 documents comprised documents extracted randomly across the full temporal range of the FT archives, thus ensuring a good likelihood of matching, but topically non-relevant documents, would be retrieved by any given query.

These documents were decompressed and saved sequentially in plain text format. Topic querying was then performed using a simple term-matching algorithm where up to two terms (words or phrases) could be entered, separated by an OR operator to ensure maximum recall. The retrieved set of documents were ranked based on the sum of term frequencies: $tf_1 + tf_2$. To minimise a ranking preference biased towards longer documents, each *tf* value was multiplied by the inverse logarithm of the document length.

Topic	Properties	Topic Description
347i Wildlife Extinctio n	26 aspects identified within the test collection of which 12 are discussed in two or more documents	The spotted owl episode in America highlighted U.S. efforts to prevent the extinction of wildlife species. What is not well known is the effort of other countries to prevent the demise of species native to their countries. What other countries have begun efforts to prevent such declines? A relevant item will specify the country, the involved species, and steps taken to save the species.
352i British Chunnel Impacts	28 aspects identified within the test collection of which 21 are discussed in two or more documents	Impacts of the Chunnel - anticipated or actual - on the British economy and/or the life style of the British.

Table 3.1: Specifications of selected topics

3.2.3. Summary of the Extinction scenario

As already noted, the first scenario is based on Topic 347i of TREC-6 (Voorhees and Harman, 1997). The key specifications are detailed in table 3.1. The searcher is required to identify as many different countries as possible that have initiated active efforts to conserve an endangered native species.

In the original TREC-6 testing, an initial pool consisting of 86 of the most commonly retrieved documents were forwarded to the topic assessor for evaluation. Of these, half (43) were judged relevant to the topic leading to the identification of 26 distinct aspects.

Our key terms comprised “extinction”, as used repeatedly in the topic description, and “endangered species”. The latter alternative term was initially selected as a common expression used in relation to living entities that are at risk of extinction and, used in conjunction with term one, seemed to result in the highest recall. Of the 43 definite relevant documents within the queried set, 33 were retrieved giving an overall recall of

77%. Within these 33 documents references are made to 22 out of the possible 26 aspects. Another 94 'non-relevant' documents were also retrieved, giving a set size of 127 documents. The overall precision of the retrieved set was 26%.

Table 3.2 shows an even distribution of topically relevant documents across the rank distribution, with overall precision remaining at approximately 30% across most of the set. In fact 18% of relevant documents occur in the bottom 21% of the list. There is a slight peak in precision at the 20th rank point due to a concentration of six relevant documents within the 11-20 range.

	Rank 10	Rank 20	Rank 50	Rank 100
Precision	30%	45%	34%	27%
Recall	9%	27%	52%	82%
# of Aspects	3	11	16	21

Table 3.2: Relevance and Precision of Retrieved Set for Extinction

3.2.4. Summary of the Chunnel scenarios

As previously noted, the second scenario is based on Topic 352i of TREC-7 (Voorhees and Harman, 1998). The important specifications are detailed in table 3.1. The task is to explore the source collection for documents that discuss how the Channel Tunnel, opened on 6th May 1994, did or was anticipated to impact on the lifestyle and economy of British citizens. Hence, possible aspects could be both prospective and retrospective in nature.

131 documents were originally pooled for evaluation by the TREC-7 judges for this topic. Only 89 of these documents were actually confirmed relevant by the judges and associated with one or more of 28 distinct aspect definitions.

Selection of key terms was relatively simple in this case. The OR combination of "Channel Tunnel" and "Chunnel" is an obvious one and retrieved 87 out of the 89 known relevant documents within the collection. Another 131 'non-relevant' documents were also retrieved resulting in a total set size of 218 documents.

Recall for the topic was almost perfect (98%), whilst overall precision was 40% (87 / 218). As expected from the high recall, all 28 aspects are referred to within the retrieved collection.

Table 3.3 shows the mean precision and cumulative recall over different portions of the rank distribution. We can see, once again, that relevant documents span the full length of the list, with the last document ranked 213th out of 218. However, steadily-declining precision levels as the distribution increases suggest that the simple relevance ranking has been slightly more effective here than was the case for the Extinction topic query.

	Rank 10	Rank 20	Rank 50	Rank 100	Rank 200
Precision	90%	70%	76%	63%	43%
Recall	10%	16%	44%	72%	98%
# of Aspects	13	16	23	25	27

Table 3.3: Relevance and Precision of Retrieved Set for Chunnel Scenario

This scenario comprises a much larger document set than Extinction. In order to allow us to make a fair comparison of the effect of aspect overlap between scenarios without the potential confound caused by document set size, we created a third scenario, based on the existing Chunnel document set, comprising only the top 127 retrieved documents. This scenario also allowed us to isolate the effect of document set size as an independent variable. From hereon, we refer to the Chunnel based scenarios as Chunnel 127 and Chunnel 218. Chunnel 127 comprises 67 (87 for Chunnel 218) relevant document cases representing 25 (28 for Chunnel 218) distinct aspects.

The tendency for aspect overlap was significantly lower in the Chunnel scenarios, compared to the Extinction scenario. The analyses that examine aspect overlap will compare Extinction to Chunnel 127 (Sections 3.5.2 and 3.6.2). Based on the documents retrieved and retained, the mean number of aspects discussed per relevant document was 1.85 in Chunnel 127 and 1.18 in Extinction ($p < .001$).

3.3. Creation of the semantic models

In this section, we describe the method by which the semantic models were created for each document set. Section 3.3.1 outlines the general text analysis procedure. We then go on to describe the specific stages of the procedure and their outputs in sections 3.3.2 to 3.3.3. Section 3.3.4 describes our software implementation of the procedure and in section 3.5 we present a summary of the semantic models that were produced. The models will form the basis of all subsequent document layouts (clustering, visualization) and the evaluation conducted in sections 3.5 to 3.6.

3.3.1. Automatic text analysis procedure

The approach used here is based on vector representation schemes as utilised in experimental systems such as SMART (see Salton and McGill, 1983). The text analysis procedure is completely unsupervised; although certain parameters are set beforehand, the procedure itself runs without human intervention. First, a term vocabulary is derived by parsing the input text for unique word terms occurring within the document set. Documents are then represented as high-dimensional vectors where each dimension represents a vocabulary term. For a given document, the value along each dimension is calculated as a function of the importance of that term within a) the specific document and b) the document set as a whole. Once these term vector representations are formed, the similarities between these vectors are computed to determine the inter-document similarity between all document pairs.

The assumption of this approach is that documents that use the same terms to similar degrees are likely to be discussing the same concepts and topics. This "bag of words" approach assumes that it is not necessary to consider word order or grammar to determine useful measures of semantic similarity.

Hence, there are two key stages of our text analysis procedure producing two specific outputs: a term-document matrix describing the location of all documents within a common term space and an inter-document similarity matrix describing the general similarity between all pairs of documents. The following sections detail the particulars of each of these two stages.

3.3.2. Creating term vector space

The dimensionality of document vectors is determined by the number of unique and valid terms that occur within the whole collection. In this work, we use single words as terms, although many alternative terms schemes are possible, including n-grams (Dameshek, 1995) and higher order statistical 'concepts' (Deerwester et al., 1990; Karypis and Han, 2000). It is usual practice to automatically remove common 'stop' words from the vocabulary. Such terms are common words such as conjunctions and pronouns that tend to be of low information value (e.g., 'and' and 'before') and their inclusion simply adds noise to the resulting vector space model. The list of common terms removed from our semantic models contains 347 terms.

In a real-time application, it is desirable to keep vocabulary size to a minimum in order to both reduce storage and computational overhead and to remove terms that are likely to be poor 'discriminators'. In addition to stop word removal, we exclude terms that are shorter than four characters and occur in fewer than five documents. These constraints might seem a little strict, for instance many aspects are represented by fewer than five documents. However, we observed through our early trials that decreasing either of these parameter values resulted in an increase in vocabulary size that was disproportionate to any advantage gained in the structure of the resulting semantic model. Finally, we also remove all the terms that occur in all documents, as these will have no discrimination value.

Other methods of vocabulary reduction such as stemming (Porter, 1980), whereby grammatical variants of the same word (e.g., bank, banks, banking) are removed, and decomposition of raw terms into a lower number of statistical 'concepts' (e.g., LSI: Deerwester et al., 1990; Karypis and Han, 2000) are also possible but were not included in our algorithm. Fine-grained exploration and optimisation of the text analysis procedure was not a goal of this thesis.

A document vector is a T-dimensional array where T represents the number of unique terms in the common vocabulary. For each document the vector is populated with values representing the weight of term within that document. Term weights in this work are calculated by using the common TFIDF scheme (see Salton and McGill, 1983). This weighting reduces the impact of terms that occur more frequently across the collection, based on the assumption that common words have low discrimination value. The variant of this weighting formula used by our algorithm is shown in Figure 3.1, where TF is the term frequency, N represents document set size and n represents the document frequency of the term. Figure 3.2 provides an illustrative example of how document term vectors are represented as data table or matrix.

$$TFIDF = TF \bullet Log(\frac{N}{n})$$

Figure 3.1: TFIDF weighting scheme

	D1	D2	D3	D4
$Term_1$	1.23	3.76	0.00	0.00
$Term_2$	5.46	0.00	1.54	5.44
$Term_3$	0.00	0.00	2.33	2.66
:	:	:	:	:
$Term_k$	1.23	6.23	0.00	0.00

Figure 3.2: Example of a document-term vector matrix

3.3.3. Creating the similarity matrix

Document relations are represented mathematically as a matrix of inter-document similarities. The similarity matrix was computed by measuring the cosine between all pairs of document vectors. This was seen as preferable to measuring node proximity (e.g., Euclidean distance) *per se* which can be affected by variation in vector length, caused for example by variation in document length and key term weightings. There are several metrics that can be used for this purpose, all of which are based upon the simple dot product calculation (see Korfhage, 1995; van Rijsbergen, 1979). Here we choose the cosine metric, which is a commonly applied (e.g., SMART: Salton and McGill, 1983), normalised derivative of dot product that controls for differences in vector length. Cosine coefficients always fall within the range of 0 to 1, with 0 indicating no observable similarity and 1 indicating perfect similarity between data objects. The cosine measure is shown below in figure 3.3. The similarity matrix is represented formally as shown in figure 3.4.

$$\text{cosine}(doc_i, doc_j) \frac{\sum_{k=1}^{t} term_{ik} \bullet term_{jk}}{\sqrt{\sum_{k=1}^{t} (term_{ik})^2 \bullet \sum_{k=1}^{t} (term_{jk})^2}}$$

Figure 3.3: Cosine Similarity Metric (adapted from Salton and McGill, 1983)

	D_1	D_2	---	D_N
D_1	1.00	0.23	:	0.16
D_2	0.23	1.00	:	.45
:	:	:	:	:
D_N	0.16	0.45	:	1.00

Figure 3.4: Example of an inter-document similarity matrix

3.3.4. Implementation of the procedure

The process described above was implemented as a simple automatic text analysis program that transformed the documents for each scenario into a weighted term vector space representation and computed an inter-document similarity matrix. This implemented procedure can be broken down into four phases or sub-procedures: loading and cleaning documents texts, building the term list, computing the term-document matrix and computing the inter-document similarity matrix.

In the first phase, the program loads in the sequential, delimited file containing the retrieved document texts in their rank relevance order (see section 3.2.2). The document texts are then parsed to replace all punctuation with a blank space and to remove all common words (from a list of 347 words). The full text of each document is then parsed sequentially for all unique word terms (character strings delimited by spaces), creating an exhaustive term list or vocabulary for the set. During this phase document frequencies are also counted for each term. Once all text has been parsed, the vocabulary size is reduced further by removing all terms that appeared in all of the document texts (e.g., common SGML tags, the query terms), all those that appeared in four or fewer documents and all terms that had three or fewer characters.

The next phase creates the term-document vector space matrix, represented internally as a two dimensional (document by term) array. For each document, the frequency of each retained term (TF) is counted. Each TF value is then weighted by multiplying TF by the inverse of the document frequency (see figure 3.1 in section 3.3.2).

The final phase computes the similarity matrix for the document set, based on the term-document matrix. A similarity value (to three decimal places) is calculated using the cosine measure for all documents pairs. Similarity values are non-directional, hence the resulting matrix is symmetric and each pair is only calculated once (i.e., sim AB is the same as sim BA). This meant a total of *$(n^2$-$n)/2$* Cosine calculations for each run.

3.3.5. Summary of semantic models

In this final sub-section we provide an overview of the resulting semantic models created by this procedure for our topical scenarios. The semantic model for the Extinction scenario comprised 1648 unique terms and resulted in a similarity matrix with a mean inter-document similarity of 0.058. Chunnel 127 and 218 comprised 1289 and 2350 unique terms respectively with mean similarities of 0.074 and 0.061 respectively.

Further summary statistics are detailed in table 3.4 and the distributions are visualised as histograms in figure 3.5. All distributions were highly positively skewed, with the vast majority (75%) of similarity values falling within a few points of the mean (sim = 0.7 - 0.10). For all scenarios there is a long flat tail to the upper end of distribution (99th percentile) containing a minority of much stronger similarities (sim = 0.26 - 0.31). We anticipated that a significant proportion of these top percentile values would describe aspectual relationships within the document sets. Our analyses in sections 3.5 and 3.6 will confirm if this is the case.

	Extinction	Chunnel 127	Chunnel 218
N	8001	8001	23653
Mean	0.058	0.074	0.061
Median	0.043	0.061	0.048
Mode	0.030	0.070	0.030
SD	0.059	0.058	0.052
Skewness	4.308	2.266	3.120
75th percentile	0.071	0.096	0.076
99th percentile	0.307	0.286	0.262

Table 3.4: Summary Statistics for the Semantic Models

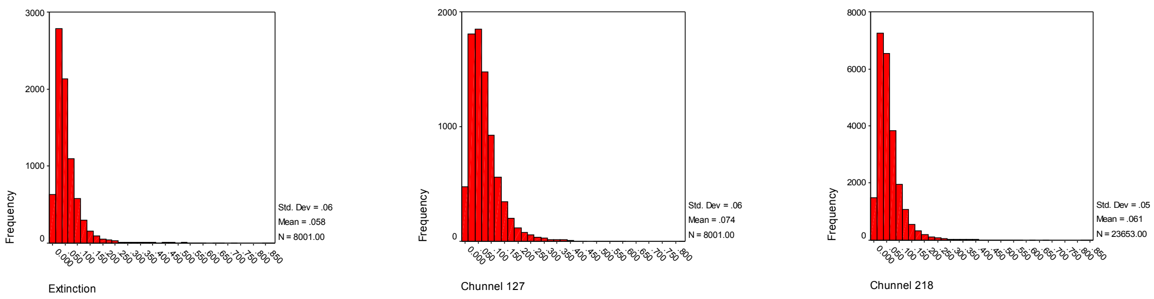


Figure 3.5: Distributions of all inter-document similarities for the three topical scenarios

3.4. Data collection methods

Data was collected from the scenarios and their semantic models, specifically the similarity matrices, for the purpose of the two experiments reported in sections 3.5 and 3.6. The first experiment, based on the ACS test, measures relative cluster separation between different semantic classes in our required two-level topical classification. The second experiment measures the upper bound potential performance of the cluster growing strategy using the nearest aspect neighbours (NAN) test. We now describe how these data were collected.

3.4.1. Aspect cluster separation test

Existing approaches to cluster separation tend to consider the similarities between all documents within a document class as single cases in each distribution (e.g., van Rijsbergen, 1979; Muresan and Harper, 2004). For instance, Muresan and Harper (2004) computed three distributions: all similarities; all similarities between relevant-topic documents; and all similarities between same-aspect documents.

Here, we adopt a somewhat different approach whereby each case is actually a mean similarity measure rather than a single inter-document similarity measure. The procedure for preparing the required data is as follows. For each relevant document we compute its mean similarity to same-aspect documents, same-topic (all relevant) documents and all documents. For brevity and consistency with continuity, these measures will be referred to respectively as R-AR, R-R, and R-ALL. We also use this approach in our analysis of spatial-semantic solutions (Chapter 4), the only difference being that the means are of proximities (distances) rather than similarities. If the desired two level hierarchical classification is present in the semantic model of a scenario, we should find a linear or quadratic trend occurring as we move from R-ALL through R-R to R-AR. In other words, the sub-set of same aspect documents around a relevant document will tend to be more similar to that exemplar than will the sub-set of same topic documents, which in turn will tend be more similar to the exemplar than the whole set containing documents that discuss other topics in addition to the relevant one.

The rationale for this approach is due to our focus being on the feasibility of cluster growing strategy which makes us more interested in the extent to each relevant document would make good aspect cluster growing exemplar, or pearl from which to grow an aspect cluster. Like Muresan and Harper's (2004) approach, our approach allows us to measure the tendency for increasingly similar classes of documents to form increasingly cohesive clusters. Additionally, it permits the conduct of between scenario analyses, without any extreme differences in sample sizes. It also allows for analyses that study the effects of independent variables within a single scenario (e.g., between aspect differences or between cases that make good and bad exemplars). We exploit the advantage of the former property in our analyses in sections 3.5 and 3.6.

In practice, our distributions are likely to be very similar in their statistical properties to the conventional all-pairs distributions used by, for instance, Muresan and Harper (2004). We are simply aggregating the data, perhaps losing some of the finer grained variance in the distribution in the process. In other words, for a given scenario, the difference in overall averages between our R-AR measure and, for instance, the all aspect similarities measure calculated by Muresan and Harper (2004) is likely to be quite small.

3.4.2. Nearest aspect neighbour test

The nearest aspect neighbour (NAN) provides us with a more direct measure of cluster growing strategy potential and is a variation on Voorhees (1985) cluster hypothesis test. It is also similar in nature and aim to the strategy functions employed by Leuski (2001) in his strategy based evaluation methodology. In Voorhees (1985) original test, the density of relevant documents appearing in the top k-most similar documents (the local neighbourhood) is computed for each relevant document. The end result is expressed as a mean percentage; the higher the percentage the better the support for the cluster hypothesis.

This original procedure would not be particularly informative for testing aspect-level clustering because each scenario has multiple distinct aspects of relevance and any given relevant document can discuss one or more of these aspects. As such, the pool of potential 'aspect relations' would vary in size for each relevant document and so averaging the sum of values would not provide a valid overall measure of aspect clustering.

We get around this problem with our NAN test, which only computes the rank positions of the two most similar same-aspect documents. By imposing this constant recall threshold, we standardise the measure for all cases, regardless of the number of aspect relations. Because documents can be associated with several aspects, we compute separate NAN scores for each relevant document and each distinct aspect discussed by each document. It could equally be used once for each document case, taking into account all related aspects, but in our case we wish the measure to reflect the potential for single-aspect cluster growing. This also allows us to compare the exemplar (starting point for cluster growing) value of given documents across different associated aspects. The value of this becomes more apparent in Chapter 5 (section 5.4) where we explore why documents cluster well to some related aspects but not to others.

Hence, a document that is associated with three distinct aspects would have three separate NAN scores. This means that a NAN dataset for a given scenario could potentially comprise many more cases than there are relevant documents. This is particularly common in our Chunnel scenarios, which contain many multiple aspect documents. However, to compute each case requires that there the respective aspect sub-set comprises three or more documents in addition to the exemplar case. Hence, the total number of cases may be suppressed where cases fail to meet this criterion. For this reason many cases were dropped from the Extinction scenario due to the large number of aspect sub-sets comprising just two documents.

3.5. Classification of topical structure

Having described how the ACS data was acquired, we now use this data to help us to understand the extent to which the expected document structure is present within our semantic models. In this section, we evaluate the extent to which the two-level topical classification structure that we seek is present within the semantic model of each of our scenarios and the variation in both topic and aspect clustering caused by aspect overlap and document set size. We directly test hypotheses H1, H3 and H4. We use the ACS test to measure general integrity of classification within each scenario. We then derive ratio measures from these mean document-class similarity scores and use these to compare the cluster cohesion of relevant documents across the scenarios.

3.5.1. General classification

Our first hypothesis (H1: see section 3.1.3) predicts that the two-level classification will be generally apparent in the structure of all scenario semantic models. Figure 3.6 clearly shows the expected trend across all three scenarios. We can see from table 3.5 that ANOVA and pair-wise contrast statistics confirm that both the general trends and the difference between adjacent class pairs are significant in every case. Figure 3.6 illustrates the general trends graphically. Most notably, for all scenarios, the slope of the curve increases significantly between R-R and R-AR, suggesting that same-aspect documents tend to be distinguished well by the inter-document similarity matrix.

Scenario	Overall	R-ALL v R-R	R-ALL v R-AR	R-R v R-AR
Exinction, 127 docs (n=24)	F(2,46)= 24.36***	***	***	***
Chunnel, 127 docs (n= 66)	F(2,130)= 115.59***	***	***	***
Chunnel, 218 docs (n=85)	F(2,168)= 235.91***	***	***	***
Overall (n=175)	F(2,348)= 244.78***	***	***	***

*** p<.001; ** p<.01; * p<.05

Table 3.5: ANOVA and pair-wise comparisons of mean similarity of relevant documents to all documents (R-ALL), topic (R-R) and same aspect (R-ALL).

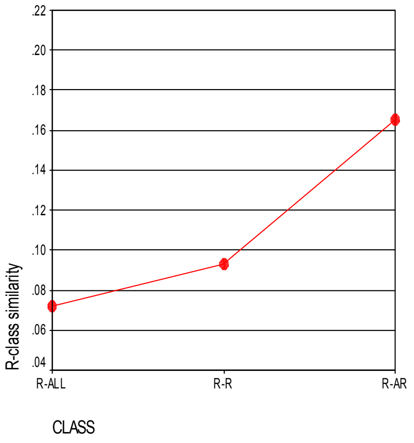


Figure 3.6a: Mean similarity of relevant documents to all documents (R-ALL), topic (R-R) and same aspect (R-AR)

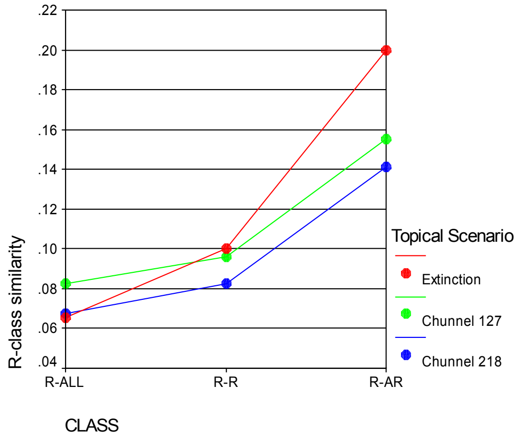


Figure 3.6b: Mean similarity, by topical scenario, of relevant documents to all documents (R-ALL), topic (R-R) and same aspect (R-AR)

3.5.2. Effect of aspect overlap

Extinction and Chunnel differ significantly in terms of aspect overlap, that is the extent to which relevant documents discuss multiple aspects of the topic. We wished to examine the effects of this factor on cluster cohesion and separation. H3 predicted that the two-level (topic-aspect) semantic classification would be weaker for the overlapping scenario.

Rather than simply measuring the general effects of aspect overlap on relevant document-class means per se, we felt it would be more meaningful to compare the relative cohesion of documents belonging to sub-ordinate classes (topic and aspects) to super-ordinate classes (whole set and topic). In terms of figure 3.6b, this means comparing the gradient of the edges. We are particularly interested in the relative cohesion of same aspect documents within the context of the topic and all documents in the retrieved set.

A two way mixed ANOVA (class by scenario), considering only Extinction and Chunnel 127, reveals a two-way interaction: F(2,176) = 9.065, p<.001. In figure 3.6b we can see a steeper incline for Extinction on both edges, indicating that the topic forms a more cohesive sub-set of the retrieved set and, in turn, aspects of the topic form more cohesive sub-sets of both the topic and retrieved set. Whilst both scenarios show a similar R-R mean, the R-AR mean is much higher for Extinction.

To analyse this difference in classification integrity in more detail, we introduce three new measures. These measure the relative mean similarity of between all pairs of class means: R-ALL:R measures the ratio of R-ALL to R-R, R-ALL:AR measures the ratio of R-ALL to R-AR and R-R:AR measures the ratio of R-R to R-AR.

Class comparison	Extinction	Chunnel 127	t-value
R-ALL:R	0.682 (n=33)	0.866 (n=67)	12.35***
R-ALL:AR	0.419 (n=24)	0.580 (n=66)	3.45***
R-R:AR	0.629 (n=24)	0.670 (n=66)	0.73ns

Table 3.6: t-test comparisons between Extinction and Chunnel 127 of sub-cluster cohesion for all class pairs in common vector space. Lower values indicate more coherent clustering of the latter document class within the context of the former class.

A low value would suggest strong clustering (in term space) of documents belonging to the sub-ordinate document class (e.g., same-aspect) in relation to the specified super-ordinate class (e.g., all documents). Calculating a ratio also provides a standardised measure, which allows us to make direct comparisons between scenarios even if the dispersion and range of their similarity distributions are quite different.

Table 3.6 shows a comparison of Extinction and Chunnel 127 whereby scenarios differ in the degree of aspect overlap but are equal in document set size (N=127). Highly significant differences (p< .001) for R-ALL:R and R-ALL-AR indicate that similarity values separate relevant documents from non-topical documents more completely in the non-overlapping scenario, both at the topic and aspect levels of relevance. This suggests that aspect cluster growing will be more impeded by instances of non-topical documents in the overlapping scenario. However, there is no significant difference between scenarios for R-R:AR, suggesting that there will be no difference in the extent to which topically relevant but aspectually non-relevant documents will impede the strategy.

The combined impact of these observations on the potential efficiency of the aspect cluster growing strategy is currently unclear and we hope to solve this conundrum in section 3.6.2. However, at this stage, H3 is supported.

3.5.3. Effect of document set size

H4 predicted that increasing set size would lead to poor separation of the aspect cluster within the topic and set clusters. To this end we compared the two versions of the Chunnel topic scenario: Chunnel 127 and Chunnel 218. We can see from figure 3.6b that although there are differences in the mean document-class similarities between the two Chunnel scenarios, the slope of the edges of the curve are relatively parallel. This suggests no general interaction and this is confirmed by ANOVA (F(2,298)=0.035, ns). However, to examine scenario differences more closely, we repeated the pair-wise comparisons of class ratios conducted using the same method applied in section 3.5.2.

Class comparison	Chunnel 127	Chunnel 218	t-value
R-ALL:R	0.866 (n=67)	0.828 (n=87)	3.366***
R-ALL:AR	0.580 (n=66)	0.515 (n=85)	2.182*
R-R:AR	0.670 (n=66)	0.623 (n=85)	1.453ns

Table 3.7: t-test comparisons between Chunnel 127 and Chunnel 218 of sub-cluster cohesion for all class pairs in common vector space. Lower values indicate more coherent clustering of the latter document class within the context of the former class.

Table 3.7 shows the results of this analysis. We can see that the effect of set size, whilst less significant, follows a similar trend to that seen for aspect overlap whereby the differences between the two scenarios are most significant in terms of the cohesion of topic and same aspect documents relative to the whole set (R:ALL:R and R-ALL:AR). Again, there is no significant difference in the ratio of mean topic similarity and mean aspect similarity (R-R:AR).

However, the differences are not in the expected direction, in that both the topic and same-aspect documents form more cohesive sub-clusters in the semantic model of the larger retrieved document set. Hence, H4 is not supported, but the reasons at this stage are unclear. It could be because decreasing the rank cut-off threshold adds a proportionally greater number of non-relevant documents (see table 3.3, section 3.2.4), thus making the relevant topic and associated aspects more distinct within the context of the common vector space. The observation that R-R:AR does not change significantly between scenarios certainly supports this idea.

This conclusion would suggest the interesting hypothesis that, within limits perhaps, increasing the recall-precision ratio may enhance the classification of topical structure. It is an interesting conjecture because it runs contrary to the views of Hearst and Pederson (1996) and Tombros and van Rijsbergen (2001) who suggest that document similarity measures are more meaningful when the context (common term space) in which they are computed is more focused on the user's query.

We must be cautious at this stage, however, because these observations are only true within the high-dimensional context of vector space. Whether this benefit translates to dimensionally reduced visual space remains to be seen in the following chapter (section 4.4.4). Prior to this, however, we will build upon these observations and those in previous sub-sections by applying our nearest neighbours analysis to determine the upper bound potential performance of the aspect cluster growing strategy.

3.6. Upper bounds of strategy performance

In this section, we use the NAN test to estimate the upper bounds of aspect cluster growing strategy performance across and within our scenarios. By using the NAN test, we are effectively simulating a user performing the strategy in high-dimensional space. As we did for the ACS tests, we look at general performance first, followed by the effects of aspect overlap and document set size. We directly test hypothesis H2 (section 3.6.1), and find further evidence to test hypotheses H3 (section 3.6.2) and H4 (section 3.6.3). We therefore begin by examining general performance of our strategy before examining the specific effects of aspect overlap and document set size.

3.6.1. General performance

H2 predicted that the aspect cluster growing strategy, guided by relative similarity cues, would result in an average precision of at least 0.2 at the point where the second relevant document is discovered. Table 3.8 shows the summary statistics for NAN analysis of our three scenarios. The most striking feature is the difference between the two topics. The local structures seem very similar for both of the Chunnel variants. In both Chunnel scenarios, just over 70% of all potential cluster growing exemplars have at least two same-aspect neighbours within the first ten nearest neighbours. Furthermore, at least 50% of exemplar cases have two same-aspect documents within the top five nearest neighbours.

In contrast, in the Extinction scenario only 17.6% of cases has two same-aspect documents within their ten nearest neighbours. In fact the for the worst 50% of cases, the rank position of the second relevant document is at least 22 and in the worst case of all the rank position is 70. The general likelihood across cases, however, of finding just one same aspect document in the top 10 nearest neighbours, however, was much better, with this criterion being met for 82.4% of exemplar cases.

Scenario	Average rank similarity of nearest aspect relevant neighbours		R2-Precision	% R2-P =< 0.2
	1st relevant	2nd relevant		
Exinction, 127 docs (n=17)	6.824 (6.000)	28.824 (22.000)	0.069 (0.091)	17.6%
Chunnel, 127 docs (n= 110)	4.255 (2.000)	10.364 (5.000)	0.193 (0.400)	72.7%
Chunnel, 218 docs (n=143)	5.007 (2.000)	10.322 (5.000)	0.194 (0.400)	71.3%
Overall (n=270)	4.815 (2.000)	11.504 (5.500)	0.174 (0.364)	68.5%

Table 3.8: Nearest aspect neighbours analysis for all three topical scenarios. For each cell means are shown first followed by median in brackets.

In summary, the potential for efficient aspect cluster growing seems very good for the Chunnel Scenarios but less so for the Extinction scenario. H2 is therefore only partially supported.

3.6.2. Effect of aspect overlap

According to H3, strategy performance should be better for the distinct aspect scenario (Extinction). The semantic model for the Extinction scenario is quite different to that of the Chunnel scenario (see table 3.8). Whilst the Chunnel scenario meets our expectations quite nicely, in the Extinction scenario aspect sub-sets seem to be somewhat fragmented. On the one hand, the first NAN tends to be relatively high ranking for most relevant document cases, yet on the other hand, for a similar majority of cases, the rank interval to the second NANs seems to be disproportionately large.

Non-parametric tests (Mann-Whitney) were used to compare the NAN scores for Extinction and Chunnel 127, due to the strong positive skew on the distributions and large differences in standard deviation on the 2nd NAN distribution. These confirm a significant difference between the scenarios for both the first NAN (U=490, p= .001) and the second NAN (U=363.5, p< .001), with performance being superior within the Chunnel scenario in both cases.

The direction of these differences is counter-intuitive. We would have expected that the potential for efficient aspect cluster growing would be poorer for Chunnel because of the greater tendency for documents to be relevant for multiple reasons. In trying to explain this result, the first question to we asked related to the fact that the sample of relevant document cases considered for Extinction is considerably smaller than for the previous analysis. This is due to the limited number of aspect sub-sets comprising three or more documents in this scenario. It seemed possible that the difference in aspect-set cluster separation (R-ALL:AR) that we observed in section 3.5.2 could be mostly accounted for by the high similarity coefficients between the smaller, two document aspects. To verify this we repeated the NAN comparison (for the first nearest aspect neighbour) with the data for two-document aspects (raising the number of Extinction case to 29). However, the addition of these cases has little effect on the observed difference between the two scenarios (U=1016, p=.002).

This is therefore an interesting problem that highlights the differences in the objectives of the two tests and, particularly, how topical structure may affect the sensitivity of the ACS method. Voorhees (1985) noted how differences in sample size between R-R and R-NR means could produce misleading results in the original cluster separation test. As the R-NR sample would generally be larger, the impact of a similar number of high similarities (relative to the R-R sample) will be lower. The same problem exists when we compare Chunnel to Extinction, where relevant documents tend to have a higher number of same-aspect relations (11.90 vs. 1.63). R-AR means for each case in the latter scenario are computed from a much larger sample and thus even though it seems that whilst there is a number of highly similar aspect relations in Chunnel, the mean is shifted further away from these high values by a relatively larger number of lower similarities.

It is possible that replacing the arithmetic mean with an alternative central tendency measure such as the median or mode might ameliorate the impact of differences in sample size. This would be an interesting question for future work. The implication for this dissertation, however, is that whilst the ACS test is a good preliminary check of the integrity of the general topical classification within a retrieved set, it is not necessarily a good predictor of between scenario differences in cluster growing performance when the respective relevant documents tend to differ grossly in terms of their topicality. As such, the observed differences need to be viewed with caution and interpreted within the context of NAN test results.

3.6.3. Effect of document set size

H4 predicted that increasing document set size would lead to less efficient strategy performance. In our ACS analysis, both topic and aspect clustering was stronger in the larger Chunnel scenario. We are interested to know what effect this has on structures local to relevant documents. From viewing table 3.8, it seems that document set size has little effect on NAN scores. Mann-Whitney U-tests confirm the reliability of this observation for both 1st NAN (U=7686.5, p=.74) and 2nd NAN (U=7722.5, p=.80).

Hence, even though both recall and precision vary between these scenarios, this has no effect on potential aspect cluster growing efficiency. Combined with our results from section 3.5.2, we must therefore reject H4. This is a promising result, which suggests that recall can be enhanced, by reducing the rank cut-off threshold, without incurring penalties on the precision of aspect clustering. However, we have yet to see the impact of the increased dimensionality associated with the larger document set on the fidelity of any resulting spatial-semantic solutions. We will examine this question in Chapters 4 and 5.

3.7. Discrete clustering

The general aim of this chapter is to run preliminary tests to check whether topical classification required to support our interaction model is present within semantic models created using a standard automatic text analysis method and to estimate the upper bounds of potential aspect cluster growing strategy performance. To this end, we have conducted low-level analyses of the similarity data, measuring and comparing cluster separation (ACS test) between the set and relevant sub-sets and also the relative similarity of relevant documents to other same-aspect documents (NAN test). Results were generally positive for the ACS tests and somewhat positive for the NAN tests, although in the latter instance performance was highly dependent on the topic under consideration.

In this section, we break briefly from our hypothesis testing to examine the extent to which our observed classification can be conveyed by a discrete clustering algorithm. The motivation for this is two-fold. First, previous work that has examined unsupervised organisation of retrieved documents relevant to a complex topic have focused on clustering (e.g., Wu et al., 2001; Muresan and Harper, 2004) as opposed to scaling (although see Swan and Allan, 1998 for a similar approach). The studies cited have found that clustering algorithms tend to produce poor results with respect to assigning same-aspect documents to the same clusters (Wu et al., 2001; Muresan and Harper, 2004). We wish to see whether the same problems occur when we attempt to produce a cluster solution for our semantic models. We also wish to extend these earlier findings by presenting a more detailed examination of the extent to which aspect clusters are accurately communicated. Our second motivation for this analysis is to provide an extra benchmark against which to evaluate the structure of our spatial-semantic visualizations in Chapter 4 and to vindicate the methodological decision to verify the truth of the cluster hypothesis in high-dimensional space prior to performing and evaluating document organisation algorithms.

Previous work has shown that relevant documents will tend to converge on a small number of clusters within a given solution (Hearst and Pederson, 1996; Wu et al., 2001), often with a single best cluster that contains most of the relevant documents (Hearst and Pederson, 1996; Muresan and Harper, 2004). This seems to be true for both simple and complex topics. However, the studies that have looked at more complex topics, have also found that aspect sub-sets often become fragmented across the cluster structure (Wu et al., 2001; Muresan and Harper, 2004). For instance Wu et al. (2001) found that although most relevant documents resided in one or two best clusters, documents relevant to the same aspect did not necessarily reside in the same cluster.

This seems initially like a counter-intuitive phenomenon: as same-aspect documents tend to be more similar than relevant documents discussing different aspects, we would expect them to be more likely to be clustered together. However, in reality, clustering algorithms, when creating a useably small set of clusters, necessarily focus on a high-level of organisation, seeking to maximise the thematic coherence of a significant number of documents (rather than pairs) within clusters and maximise the thematic distinction between clusters. Muresan and Harper (2004) demonstrated this effect on document organisation. They found that the topicality (i.e., number of same aspect relations) of documents has a major impact on clustering. Highly topical documents tend to be more similar to the relevant sub-set as a whole and are therefore more likely to be grouped together into a highly topical cluster. In contrast, documents that are distinct or highly focused in their perspective on the topic tend to be allocated to other clusters, sometimes apparently arbitrarily.

We begin in section 3.7.1, by presenting and evaluating the cluster solutions created for each scenario at the topic level of relevance, before examining, in section 3.7.2, the organisation of aspect sub-sets within the solutions.

3.7.1. 5-cluster solutions

Following Hearst and Pederson (1996), we created flat (non-hierarchical) 5-cluster solutions for each topical scenario. We used a standard k-means clustering algorithm as provided by SPSS v11.5. The input for each solution was a similarity matrix and all settings, apart from *k* were left on default. Tables 3.9 to 3.11 show the distribution of relevant and non-relevant documents in the solutions for each scenario. We will now briefly describe the key structural properties of these solutions as they relate to each topic.

We can see that relevant documents are scattered across at least four clusters in each of the solutions, although the extent to which relevant documents dominate each cluster varies considerably within each solution. The results of Cross-tabs (Chi-square) analyses for each solution confirm that clusters tend to vary in their relevance bias and this was a highly significant effect for all scenarios (p<.001). Hence, topical relevance is exerting a significant influence on the resulting cluster structures.

Extinction	Non-Relevant	Relevant	Pcl(Prl)
Cluster 1	41 (43.6%)	21 (63.6.%)	.34 (.29)
Cluster 2	4 (4.3%)	6 (18.2%)	.60 (.30)
Cluster 3	5 (5.3%)	5 (15.2%)	.50 (.30)
Cluster 4	8 (8.5%)	0 (0.0%)	.00 (.38)
Cluster 5	36 (38.3%)	1 (3.0%)	.03 (.41)
Chi-square = 24.257, df = 4, p< .001			

Table 3.9: Five-cluster solution for Extinction

Chunnel 127	Non-Relevant	Relevant	Pcl (Prl)
Cluster 1	12 (20.0%)	29 (43.3%)	.71 (.76)
Cluster 2	29 (48.3%)	8 (11.9%)	.22 (.76)
Cluster 3	9 (15.0%)	8 (11.9%)	.47 (.77)
Cluster 4	1 (1.7%)	11 (16.4%)	.92 (.92)
Cluster 5	9 (15.0%)	11 (16.4%)	.55 (.70)
Chi-square = 27.257 df = 4, p< .001			

Table 3.10: Five-cluster solution for Chunnel 127

Chunnel 218	Non-Relevant	Relevant	Pcl (Prl)
Cluster 1	15 (11.6%)	0 (0.0%)	.00 (.80)
Cluster 2	22 (17.1%)	5 (5.6%)	.19 (.78)
Cluster 3	8 (6.2%)	21 (23.6%)	.72 (.76)
Cluster 4	58 (44.3%)	23 (26.4%)	.28 (.60)
Cluster 5	28 (21.4%)	38 (43.7%)	.58 (.65)
Chi-square = 41.205, df = 4, p< .001			

Table 3.11: Five-cluster solution for Chunnel 218

Hearst and Pederson (1996) found that there was generally a best cluster containing a large proportion (in most cases >50%) of relevant documents. We can see this effect in our solutions, albeit to a weaker extent. In Extinction this is cluster 1 (63.6%), in Chunnel 127 this is cluster 1 (43.7%) and in Chunnel 218 this is cluster 5 (43.7%). However, cluster sizes vary within each solution and although these clusters contain the largest proportion of relevant documents within the cluster structure, being relatively large clusters they also comprise a significant number of non-relevant items.

In other words, although they are a rich source of relevant documents, they are by no means exclusively relevant clusters. To put this into perspective, the figures in the far right column describe two precision figures for each cluster: the first (Pcl) is proportion of relevant documents within the cluster and the second (Prl) is the proportion of relevant documents in the same number of top ranking documents. Hence, these two measures allow us to crudely compare a strategy of identifying, through whatever cues are provided by the interface (e.g., cluster size, key words), and looking at this 'best' cluster first, to a more conventional strategy of simply browsing systematically down the ranked list.

We see that whilst Pcl is marginally greater than Prl for Extinction (.34 vs. .29), the converse is true for both Chunnel solutions (Chunnel 127= .71 vs. .76; Chunnel 218= .72 vs. .76). Hence, in neither case does browsing the best cluster first provide the user with a significant advantage. In fact, if we look at the remaining clusters we can see that, in each solution, the most precise or 'topic rich' clusters tend to be the smaller clusters. Even so, browsing these more precise clusters first would only represent a more efficient strategy than browsing the ranked list in the case of the Extinction scenario.

Comparing precision on a 'by cluster' basis is probably a little unfair to the document clustering model as precision is always likely to be relatively high in the top rank intervals of the retrieved set. The real benefit of clustering is likely to be in locating relevant items further down the list, where they are more sparse. A fairer evaluation of is therefore to look at the broader, more realistic strategy where the user filters out the least relevant clusters, based on their meta-data, and devotes their attention to browsing the more promising ones. In other words, we will look at how cluster summary data might help the user to navigate more efficiently to the majority of relevant documents.

To demonstrate this strategy, we adapt our strategy function to simulate a user who decides to browse the top three most precise clusters together, rather than one cluster at a time. In doing so, we make the assumption that the most precise clusters will be summarised using terms that are clearly more relevant than those used to label other clusters (see for e.g. Hearst & Pederson, 1996). Hence, we are considering a much larger proportion of the retrieved set and compare the precision of this strategy to that of browsing the same number of top ranking documents. We find that benefits of document clustering are, again, most apparent for Extinction (Prl=.28; Pcl= .39, +39.3%). The advantage in the Chunnel scenarios is less pronounced, with moderate gains for Chunnel 127 (Prl=64; Pcl= .70, +9.4%) and relatively slim gains for the Chunnel 218 scenario (Prl= .45; Pcl= .47, +4.4%).

Hence, we have shown, for our complex topical scenarios, that clustering documents by their similarity can effectively separate relevant from non-relevant items although these solutions are by no means definitive; relevant documents do not tend to form large, exclusive clusters. Whilst there are some benefits to the user for general topic retrieval these seem to be significant only for the user who is prepared to browse through large numbers of documents in multiple clusters as opposed to the user who is only prepared to browse a few documents in the most promising cluster. It seems apparent that for our complex topics, Muresan and Harper's (2004) aspectual cluster hypothesis is correct in that relevant documents are not always highly similar. Moreover, most relevant documents are neither similar enough nor sufficiently distinct from non-relevant documents to form large, exclusive clusters.

Our analysis tells us that discrete clustering can, to a limited extent, effectively organise a document set retrieved for a complex information need in a manner that may, to some extent, facilitate the retrieval of relevant documents. However, the interaction model we proposed in section 2.2 assumes a task beginning with an open-ended question (an ill-defined information need) where search proceeds in a berrypicking/evolving style (Bates, 1989). In other words, the user is unable to define, up front, all aspects of relevance; indeed the relevance of some documents may not be apparent until the user has interacted with other documents.

We therefore expect that the user's query will evolve as they interact with documents (Bates, 1989; O'Day and Jeffries, 1993; Xie, 2000), meaning that their intentions will periodically shift from to new and different aspects of the topic based on accidental discoveries or insights (O'Day and Jeffries, 1993). Our interaction model assumes that in between these shifts the user will be temporarily focused on a specific aspect. This is the reason why the aspect cluster growing strategy is a central focus of our analyses throughout this dissertation. We are most interested in the extent to which document organisation algorithms can group not only generally relevant documents together, but also same-aspect documents. In other words, we ask to what extent can the user follow up their specific aspect intention without going beyond the cluster in which the first instance was discovered?

3.7.2. Aspect cohesion

At the beginning of section 3.7, we discussed the results of previous studies which suggest that discrete clustering seems to do a relatively poor job of assigning same-aspect documents to the same clusters (e.g., Wu et al., 2001; Muresan and Harper, 2004). This may be particularly true when the aspect sub-set comprises a mixture of both highly topical and highly focused or distinct documents (Muresan and Harper, 2004). We now look at the extent to which aspect sub-sets in our scenarios converge on the same clusters.

Aspect	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Cluster 1	1	2	1			1	2	1	1	4	1		1	
Cluster 2					1		1		1					
Cluster 3			1	1	1		1		1					
Cluster 4														
Cluster 5														
Aspect	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**		
Cluster 1	2	1	1	2		2	1							
Cluster 2									1	2		1		
Cluster 3	1				2									
Cluster 4														
Cluster 5					1									

Table 3.12: Extinction aspect distribution across 5-cluster solution

Aspect	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Cluster 1	1	2		4	2		2	1	1	12	6	3	2	
Cluster 2								1		1				
Cluster 3					2		3				2		1	
Cluster 4	2	1			3		1			1	6	3	6	1
Cluster 5	1	7									2	1		
Aspect	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**
Cluster 1	2	4	1	1		4	1	1	2	1	1		1	1
Cluster 2	1			2				1	2	4		2		
Cluster 3			1	1		1								
Cluster 4									1					
Cluster 5	6						1							

Table 3.13: Chunnel 127 aspect distribution across 5-cluster solution

Aspect	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Cluster 1														
Cluster 2			1	1				1						
Cluster 3	4	1			5	1	3			1	7	3	12	3
Cluster 4	3			3	1		3	2	1	1	2			
Cluster 5	4	10		1	2		3			13	9	6	2	
Aspect	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**
Cluster 1														
Cluster 2	1	1	1	1		2								
Cluster 3						1		1	2					
Cluster 4			1	4	1	1	1	1	2	4		2		
Cluster 5	9	4				2	1		2	1	1		1	1

Table 3.14: Chunnel 218 aspect distribution across 5-cluster solution

Tables 3.12 to 3.14 show, for each aspect, the distribution of associated documents across the cluster structure. Two main features are immediately apparent. The first feature, if we look within the rows of the tables, is the topical breadth of relevant clusters, particularly the smaller more, precise clusters. For instance in the Extinction solution, clusters two and three comprise only six and five relevant documents respectively, just under a third of all relevant documents, yet they account for six and seven distinct aspect instances, respectively and ten (approximately 45% of all represented aspects) collectively. Likewise, the larger relevant cluster comprises 64% of all relevant documents, yet accounts for 73% of all distinct aspects.

A similar pattern emerges from the Chunnel solutions. In Chunnel 127, the most precise cluster, cluster four, comprises 11 (16%) of relevant documents yet accounts for 10 (40%) of the 25 aspects represented by the modelled documents associated with this scenario. The largest proportion of relevant documents occurs in cluster one (43%), yet this cluster accounts for 92% of all represented aspects. Similarly, in Chunnel 218 the most precise cluster, cluster three, comprises 29 (24%) of relevant documents yet accounts for 13 (46%) of all represented aspects. The cluster with the largest number of relevant documents, cluster five, comprising 66 (44%) relevant documents, accounts for 18 (64%) distinct aspects.

Hence, in both scenarios relevant clusters account, proportionately, for more aspects of the relevant topic than their recall of documents relevant to the topic implies. In other words, although topically similar documents are being clustered, there is a lot of aspectual overlap between clusters. As expected, this overlap is greater in the Chunnel scenario where documents tend to span multiple aspects. The consequence of this can be seen as the second major feature of tables 3.12 to 3.14: the extent to which aspect sub-sets are fragmented across the cluster structure.

We can see that it is unlikely for aspect sub-sets of two or more documents to occur exclusively in the same cluster. Such fragmentation would be unhelpful for the searcher who, having discovered a new aspect, wished to locate all the other documents that discussed the same perspective on the topic. For Extinction, six out of the 11 aspects that are represented by two or more documents are spread over more than one cluster. For Chunnel, the proportion is higher with 19 out of the 21 multi-document aspects being spread over two or more clusters in Chunnel 218 and 16 out of 19 aspects in Chunnel 127.

Therefore, there is less aspect fragmentation in the Extinction scenario, which is consistent with the differences in aspect overlap between the scenarios and our hypotheses that predict better aspect separation and cluster growing performance within visual representations of the semantic models for the distinct aspect scenario (see section 2.6.4).

In summary, it seems that clustering is able to communicate high-level topical relations, but more specific aspect level relations are frequently lost in the document organisation process.

3.8. Conclusions

In this chapter we have described the process by which our test bed, comprising three topical scenarios and semantic models of these scenarios, was created. We then dealt with question one, which asked whether the required semantic structure would be present within our semantic models. To this end, sections 3.5 and 3.6 tested hypotheses relating to cluster separation and simulated aspect cluster growing performance using the inter-document similarities computed for our scenarios. In section 3.7, we explored the organisation of documents within a discrete cluster solution. Overall our analyses show positive support for the feasibility of our interaction model, although there were also some rather surprising differences observed between the different scenarios. The limited value of discrete clustering for organising aspects of a complex topic has been confirmed and demonstrates the importance of testing the cluster hypothesis within the semantic model prior to performing studies of clustering or visualization algorithms. We now summarise the key results and draw some interim conclusions.

3.8.1. Classification and potential strategy performance

Question one asked: *To what extent can a standard text analysis procedure model the general semantic structure expected by our interaction model and particularly the low-level structure required by the aspect cluster growing strategy?*

Our results show good support for the two-level cluster hypothesis both at topic and aspect levels. As predicted by H1, relevant documents becoming increasingly similar to other documents as the comparison sub-set becomes more closely related to its content. However, potential strategy performance seems variable, with good results (in line with H2) for the Chunnel scenario but not Extinction. This runs counter to our hypothesis (H3) and is even more surprising, given that the ACS tests indicated that the relative mean similarity of same-aspect documents was greater in the Extinction scenario. However, we conjecture that the superiority of Extinction in the ACS tests may have been an artefact caused by the gross differences in the average number of same aspect relations between scenarios, whereby mean intra-aspect similarities in the Chunnel scenario are more likely to be skewed by a larger proportion of relatively weak similarities, even when there are a similar number of strong similarities. This suggests that the use of median or modal similarity may be a more appropriate measure for the ACS test than the arithmetic mean.

Another surprising result is the effect of document set size, where relative mean similarity of topics and same aspect documents was actually greater for the larger set. This runs contrary to H4 where we predicted that the decreased focus on the relevant topic within the common term space would reduce impact negatively. Furthermore, our NAN comparisons did not show any differences in performance between the two Chunnel scenarios. This is interesting given that the overall precision value for the larger retrieved set was considerably lower and the size of the term space considerably higher. This suggests the interesting hypothesis that it is possible to automatically compute useful semantic models from relatively large (high-recall) and low-precision retrieval sets. The testing of this hypothesis is left for future work. Furthermore, we do not yet know the impact of increasing set size on the fidelity of spatial-semantic visualizations. It will be interesting to see the outcome when these comparisons are repeated, using our visualizations, in Chapters 4 and 5.

In summary, the structure expected by our interaction model does seem to be present within semantic models created using a simple text analysis procedure, although early signs indicate that aspect cluster growing performance may be problematic for the Extinction scenario.

3.8.2. Discrete clustering

We included an analysis of a discrete cluster solution in section 3.7, because we wished to confirm the same-aspect document fragmentation problems observed in earlier studies (Wu et al., 2001; Muresan and Harper, 2004) and demonstrate the importance of verifying the cluster hypothesis within the underlying, high-dimensional semantic model, prior to performing any clustering or visualization, so as to provide a ‘gold standard’ benchmark against which to judge the success or failure of a dimension reduction algorithm.

The general pattern in our analysis was that clustering solutions performed relatively well in terms of partitioning relevant from non-relevant documents, yet the more fine-grained aspect relationships were not well preserved and communicated in the cluster structure. This is consistent with the findings of previous clustering experiments (e.g Wu et al., 2001; Muresan and Harper, 2004). It seems likely from our results that discrete clustering has limited potential for aspect level clustering. This seems to be due to the tendency for clustering algorithms to communicate general themes rather than fine-grained inter-document relations. The observation that 84-90% of multi-document aspects were fragmented in the two Chunnel (overlapping aspect) scenarios compared to 55% in the Extinction scenario, where relevant documents tend to be more focused, supports this contention. The effect of topical diversity of relevant documents is likely to be further compounded by the extent to which other, non-relevant concepts are discussed. Unfortunately, such diversity is invisible in our test collections as, for obvious practical reasons, documents are only catalogued in terms of relevance to specified topics.

A key aim of this work is to demonstrate that spatial-semantic document organisation is better able to communicate complex topical structures than discrete clustering. The continuous nature of the organisational scheme along two dimensions should allow greater scope for representing complex, multi-faceted relationships between documents. In Chapters 4 and 5 we will examine the extent to which this is true.

CHAPTER 4: VISUALIZING TOPICAL STRUCTURE

4.1. Introduction

In the previous chapter we described the creation of our test scenarios and the evaluation of semantic models created from these scenarios. Cluster separation test results were positive for all scenarios, with relevant documents tending to be most similar to the same-aspect documents and least similar to non-relevant documents. However, despite the fact that the two-level topical classification was consistently detectable, when we applied the NAN test we found considerable variation in potential aspect cluster growing performance between topical scenarios. Consistent with a previous study by Wu et al. (2001), we found that whilst a discrete clustering algorithm can effectively partition relevant from non-relevant documents, same-aspect documents are frequently scattered across multiple clusters. In this chapter we begin to address our second research question (section 1.7) and its associated hypotheses (section 2.8) by evaluating the extent to which spatial-semantic visualization is able to convey the topical classification structure observed in Chapter 3. Our purpose is to begin to determine which layout optimisation approach, global (MDS) or local (MST), is likely to produce the best spatial classification our interaction model, and to gain an initial impression of the potential for each layout scheme to provide the cues to support the aspect cluster growing strategy.

This chapter is divided into three parts. We begin in section 4.2, by describing how our visualization solutions were created. In section 4.3 we conduct an initial, visual analysis of the solutions, examining the extent to which key semantic features and also discrete cluster membership are conveyed by the spatial-semantic structure. Finally, in section 4.4, we conduct a quantitative experiment where we apply the ACS test again to examine cluster separation of topic and aspect level sub-sets. Our methodological approach is almost identical to the previous chapter, except that proximity or distance measures (in visual space) are now the measure used for comparison rather than similarities.

4.1.1. Research question and hypotheses

Research question two asked: *Given an adequate semantic model, which approach to spatial-semantic layout best preserves the general and, in particular, the low-level structure expected by our interaction model?*

In Chapter 2 (see section 2.8), we defined nine hypotheses that we wished to test in relation to this question. In this chapter we will test the following six hypotheses:

H5: The two level classification will be effectively conveyed by spatial relations in (i) MDS and (ii) MST.

H6: Aspect level cluster separation will be greater for MST visualizations than for the MDS visualizations.

H8: Aspect level cluster separation will be lower in the overlapping aspect scenario than the distinct aspect scenario.

H10: The expected differences between MST and MDS will be greatest for the distinct aspect scenario.

H11: Aspect level cluster separation will be lower in visualizations of the larger retrieval set.

H13: The expected differences between MST and MDS will be greatest for the larger retrieval set.

Hypotheses H7, H9 and H12 relate to the evaluation of the aspect cluster growing strategy and will be addressed in Chapter 5.

4.2. Spatial-semantic visualization algorithms

Coherent clustering of relevant documents is critically important with respect to our interaction model. Most importantly, the cluster growing strategy relies on documents that discuss the same aspect of the topic occurring in close proximity to each other. We observed in the previous chapter (section 3.7) how discrete cluster structures, whilst able to communicate major themes, frequently fail to effectively convey more minor features such as aspect sub-sets. We hypothesise (H5) that the continuous, two-dimensional structure of a spatial-semantic visualization will afford better preservation of the topical classification, and particularly aspect clustering, that we observed in the high-dimensional semantic models. In section 2.6, we discussed the range of approaches available for creating these visualizations. We decided to compare two distinct approaches: one that aims to effectively map the relationships between all document pairs (global optimisation approach) and one that concentrates on preserving only the strongest relationships (local optimisation approach). We argued in section 2.6 that the latter approach would create better aspect-level cluster separation (H6), based on earlier findings by Muresan and Harper (2004), which show that mean same-aspect document similarity tends to be significantly higher than the mean of both all-topic and same-topic document similarity.

4.2.1. Global versus local optimization

Previous work has examined spatial-semantic visualizations created using multi-dimensional scaling (MDS) algorithms. These algorithms represent documents as a scatter plot of marks or visual points where the aim is to find the best inverse mapping between all input inter-document similarities and output inter-node proximities (e.g., see Wise et al., 1995; Hornbaek and Froekjaer, 1999). We compare MDS to a more restricted approach to node layout based on the minimum spanning tree (MST) of the similarity matrix, when it is considered as a fully connected graph. We make this comparison based on the hypothesis that a layout algorithm that focuses on preserving the strongest relationships between documents, or local optimisation, will result in more cohesive clustering of same-aspect documents than a global optimisation approach that attempts to produce proportionally accurate layout at all levels of similarity.

4.2.2. Algorithms

In this study our globally optimized solutions are created using the PROXSCAL algorithm (Busing et al., 1987) as implemented within SPSS v. 11.5. PROXSCAL represent several improvements over Alternating Least Squares Scaling (ALSCAL: Takane, Young and De Leeuw, 1977), which is also implemented within SPSS v.11.5 (base system), primarily because it aims to minimise Kruskal's stress, a measure that is based on distances, rather than squared distances. Given the nature of our strategy, an ordinal model was seen as sufficient and, in fact, seemed to produce more aesthetic and distinctive structures compared to interval or ratio models. Additionally, tied observations were left tied and all initial configurations were set to the simplex model. The inputs were the document similarity matrices described in the last Chapter, which were converted to proximities internally by the PROXSCAL algorithm prior to scaling.

Our local optimisation approach is to use force-directed placement to create a visualization of an MST of the similarity matrix. The MST is created by considering the similarity matrix as a fully connected graph, where all document nodes are connected by weighted edges (their similarity score). A MST is a weighted sub-graph that is created by pruning all but the most salient (lowest weight) edges, to leave a single tree (no cycles) of N nodes connected by N-1 edges. The set of document nodes is therefore connected as a tree structure where edges (document similarities) are of minimum weight (maximum similarity). A previous empirical study (Cribbin and Chen, 2001) showed that visual information retrieval performance is superior when using MST visualizations and that users find visualized MST structures more meaningful than the traditional MDS based scatter plot visualization.

We implemented a version of Prim's (1957) algorithm in Visual Basic 6. This program takes the similarity matrix as input and ranks all document pairs from lowest to highest weight (largest to smallest inter-document similarity). The algorithm begins by starting with the top ranking document pair and the rest of the tree is 'grown' by iteratively connecting the each remaining node. At each iteration, the highest ranked edge (inter-document similarity), that connects a node already within the tree with one outside of the tree, is added. This continues until all nodes are included within the tree.

The visualizations of our MST sub-graphs were created using the Neato program from the Graphviz suite (North, 2002). Neato uses the force-directed (spring model) placement algorithm by Kamada and Kawai (1989) to produce an aesthetic layout of nodes and edges with as few overlaps and edge crossings as possible. The input to Neato was a list of edges specified as node pairs along with edge weights (spring strength) specified as the square root of the similarity coefficient. The length of each edge within the visualized solution was set to constant. The list, as output by our MST program, was formatted in the "dot-language" required by Neato (see appendix B.1 for an example).

4.3. Visual analysis of solutions

In this section, we view the solutions created using the two layout approaches described above. We compare MST and MDS in terms of topic clustering (sub-section 4.3.1), aspect clustering (sub-section 4.3.3 and 4.3.4), and the cluster membership data (sub-section 4.3.2) produced by the k-means algorithm in the previous chapter (section 3.7). We show two samples of aspect clustering. In sub-section 4.3.3 we show the distribution of document nodes belonging to the aspect sub-sets that were poorly clustered by k-means (section 3.7.2) chapter. In sub-section 4.3.4, by way of contrast, we show the distribution of documents within the most cohesive aspect sub-sets (highest mean inter-document similarities).

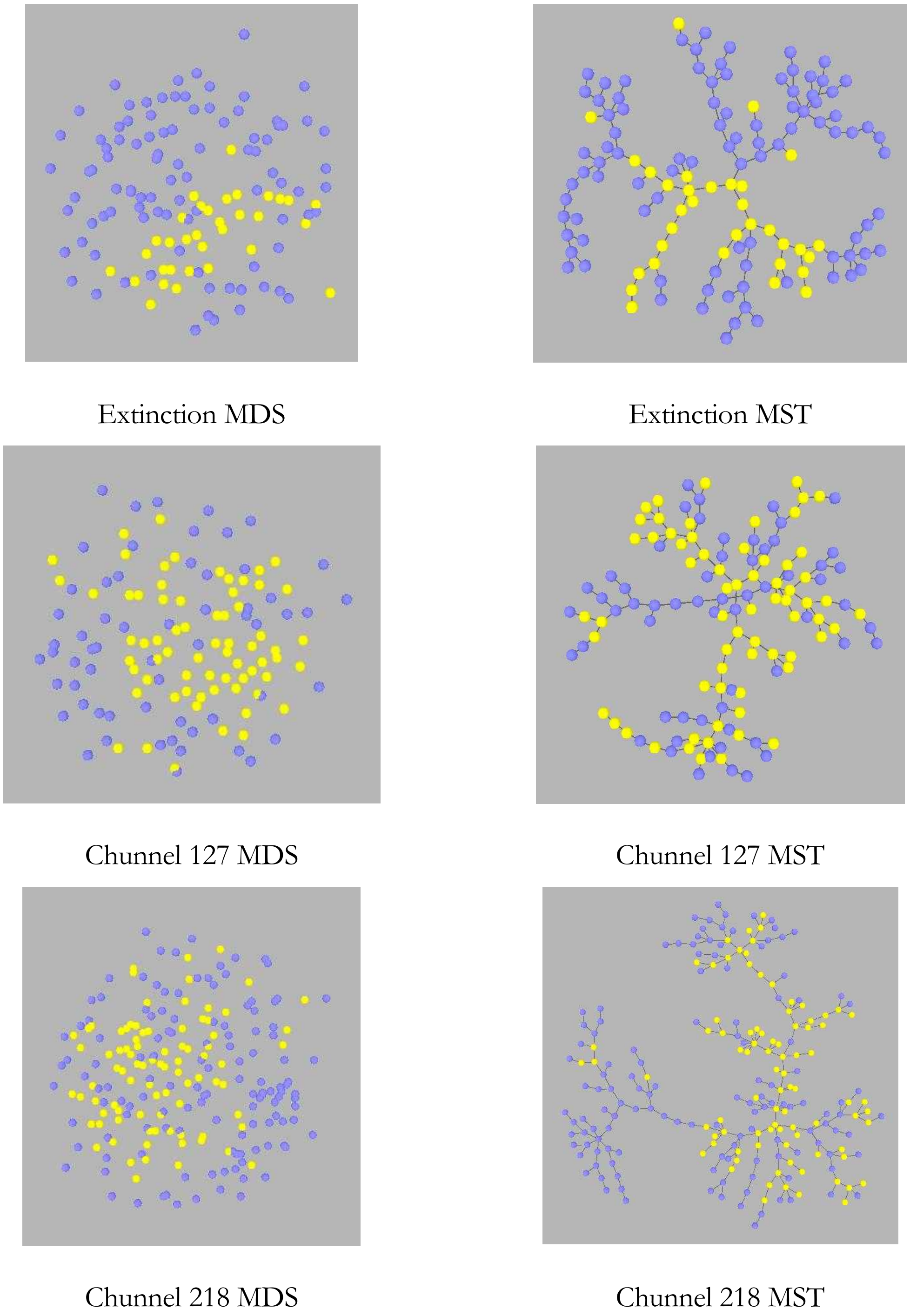
H5 predicts that both layout schemes will produce reasonable clustering of both the topic and specific aspects. However, H6 predicts that the integrity of the classification will be noticeably better for MST, particularly at the aspect level.

4.3.1. Topic clustering

Figure 4.1 shows the MDS and MST solutions for both document sets, with documents that are relevant to the topic (all aspects) marked up as yellow. Ignoring the topic augmentation briefly, we can see that the general structures created by MDS and MST are very different. The dendrite structure of MST creates visualizations with readily discernable features in the form of bunches and contiguous strings of document nodes. MDS on the other hand presents a more subtle structure, which on first inspection seems more uniform than the MST. Closer inspection, however, reveals considerable variation with a mixture of relatively dense and sparse regions of document nodes.

Returning to topic augmentation, in all of the MDS visualizations the topic forms a reasonably coherent cluster within the overall distribution of nodes. This clustering is notably more coherent in the smaller scenarios (Extinction and Chunnel 127). In Extinction, if the worst two outliers are ignored, the topic occupies a clear elliptical area towards the bottom of the visualization, within which relevant documents form several distinct clumps and only a very small proportion of non-relevant documents are located. In Chunnel 127 relevant documents occupy a distinct circular area just right of centre in the MDS solution. The lower half of this feature contains a dense concentration of relevant documents. The upper half forms a tail emanating from the lower half (like a comet) whereby the density of topical nodes decreases as the top of the main feature is approached. However, within this tail there are clear pockets of densely clustered relevant nodes.

In Chunnel 218, the relevant documents again mainly occupy a coherent circular area, this time offset slightly left of centre. If the worst outliers are ignored, this circular feature is fragmented into three main sub-clusters (one above and two below) separated by relatively fallow areas of space or non-relevant documents. Within these clusters, many small clumps of two or three documents are apparent, along with several larger clumps of five or more relevant documents.



Extinction MDS

Extinction MST

Chunnel 127 MDS

Chunnel 127 MST

Chunnel 218 MDS

Chunnel 218 MST

Figure 4.1: Topic clustering in the spatial-semantic visualizations (MDS and MST) of the three topical scenarios. Yellow nodes are relevant to at least one aspect of the general topic.

MST also clusters relevant documents coherently, but in quite a different way to MDS. In Extinction, for example, the topic appears to be distributed more broadly across the total node structure, however 29 out of the 33 relevant document nodes are connected as a continuous sub-tree of the main structure. There are at least six key branching points that seem to signify a clear change in topic from relevant to non-relevant. This suggests that labels that define the reason for definite branching points may be a useful aid to overview and navigation within this kind of visualization.

The Chunnel distributions are quite different from Extinction. In each case, unlike Extinction, there is no continuous sub-tree, but rather a number of dense patches of relevance, appearing as a combination of bunches and strings. The distribution of relevant nodes in the Chunnel 127 solution is quite broad covering most of the structure. The only distinctive, differentiating feature is a branch extending to the left of the structure which contains mainly non-relevant items. In Chunnel 218, there are also multiple dense patches of relevant nodes, which mainly occupy the right hand side of the structure. The left hand side of the tree contains 87 document nodes but only four of these are relevant.

In summary, for all scenarios and both visualization schemes, relevant document nodes clearly cluster together. MDS solutions seem to gather topical documents into a neater, more homogenous feature, whilst MST seems more prone to disperse relevant nodes into smaller pockets or dense sub-clusters of relevance. This is most true for the overlapping topic, particularly Chunnel 127, where patches of relevant documents are scattered quite broadly over the tree. This latter result is consistent with our analysis of similarities in the previous chapter (Tables 3.6 and 3.7) where we observed a relatively small separation of the topic cluster within the whole set cluster (R-ALL:R).

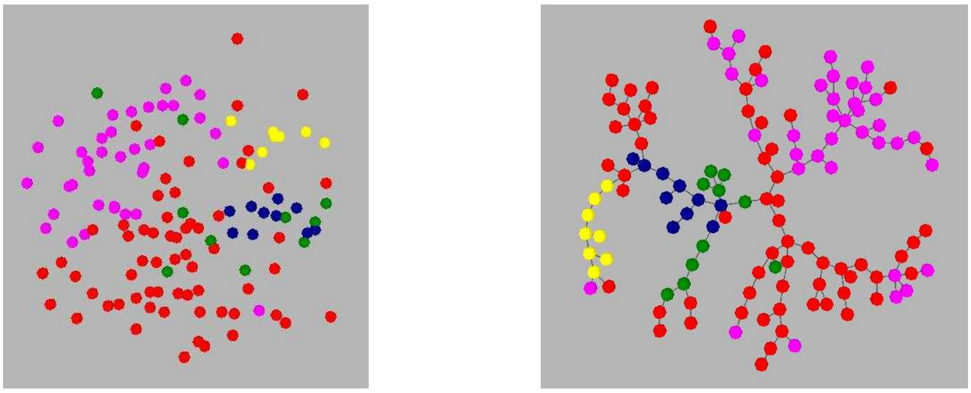
Whether the less cohesive clustering of the topic represents better separation of aspects within MST visualizations (H6) remains to be seen (in sub-sections 4.3.3 and 4.3.4) and later in our more comprehensive quantitative analysis of ACS scores (in section 4.4). Before we look at aspect level clustering, however, in the next sub-section we examine the similarity between spatial-semantic clustering and discrete clustering, by augmenting our visualizations with cluster membership information from the solutions reported in section 3.7.

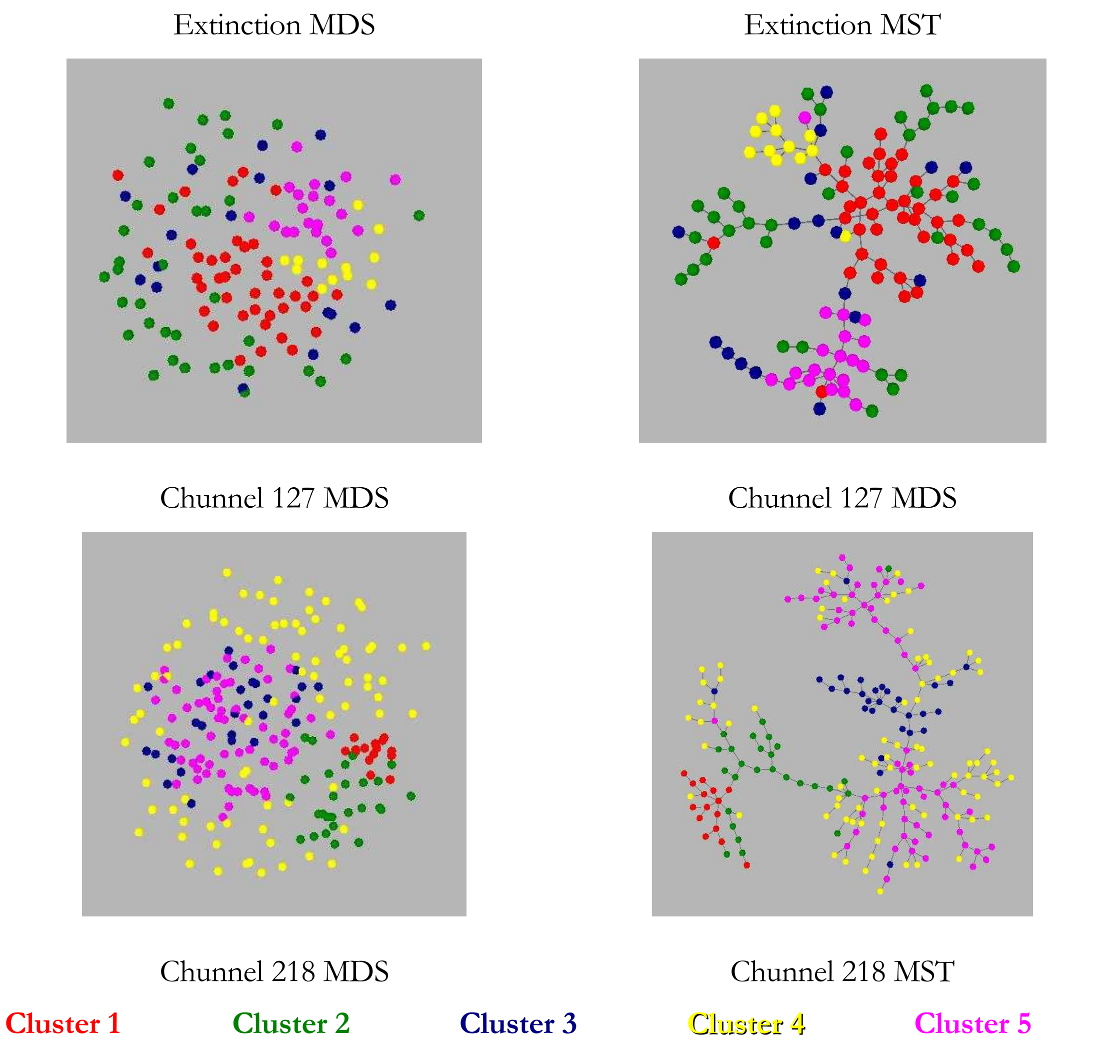
4.3.2. Compatibility with 5-cluster solutions

Despite the observed aspect fragmentation problems, the solutions discussed in Chapter 3 conveyed a useful structure, at least in terms of partitioning relevant from non-relevant documents. Hence, in addition to finding out whether aspect clustering tends to be better in spatial-semantic structures, we were also interested to see whether the discrete cluster structures would have anything in common with those of our visualizations. We examine the distribution of problematic aspects in section 4.3.3. First, in this sub-section we present and discuss versions our visualizations augmented with discrete cluster membership information.

Figure 4.2 shows our six solutions marked up to show document cluster membership. The correlations between the discrete and continuous solutions are immediately apparent. Many clusters, even relatively large ones, are represented as quite cohesive visual features. In MDS visualizations, some clusters, although coherent by themselves, tend to overlap significantly with other clusters. Clusters three and five for Chunnel 218 and, to a lesser extent, clusters one and two for Chunnel 127 are good examples of such clusters. In contrast to this, in the MSTs, no such merging occurs. Instead, the same clusters seem to be sliced-up and tessellated. The question of whether these overlaps and inter-sections represent same-aspect documents that were fragmented within the discrete cluster solutions is unclear at this level of analysis, but will be explored further in the next sub-section.

Some clusters are not so well represented in the spatial-semantic structures. It is not uncommon for a 'bin' cluster to emerge within a set of clusters where a residual sub-set of documents that do not fit the main theme of any of the other clusters tend to be consigned (Hearst and Pederson, 1996). MDS and MST handle these types of clusters quite differently. For instance, cluster two (green nodes) of the Chunnel 127 scenario can be thought of as such a cluster, MDS seems to scatter documents in a horseshoe shaped arc across the visualization. In contrast, MST seems to scatter these documents more purposefully into quite distinct sub-clusters, separated quite clearly by patches of nodes that belonging exclusively to other clusters. A similar effect can also be observed for cluster four (yellow nodes) of the Chunnel 218 scenario.





Extinction MDS

Extinction MST

Chunnel 127 MDS

Chunnel 127 MDS

Chunnel 218 MDS

Chunnel 218 MST

Figure 4.2: Discrete cluster augmentation of spatial-semantic solutions

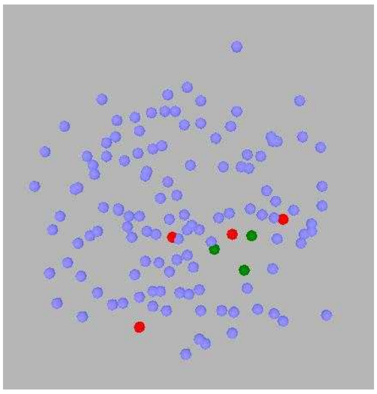
In summary, there seems to be a lot of correlation between discrete cluster membership and the grouping of documents within spatial-semantic solutions. However, some clusters seem to merge together or are spread over broader regions. Furthermore, MDS and MST seem to handle these clusters quite differently. First, whilst MDS might spread a cluster continuously across a large region of space, MST will segregate the same documents into coherent sub-clusters. Second, whilst two clusters might overlap within MDS, those same clusters will be partitioned and tessellated within the equivalent MST.

Following on from this last point, in the next sub-section we ask whether spatial-semantic visualization can resolve the observed organisational limitations inherent to discrete structures. We test investigate this by looking at how at the relative efficacy with which the aspects that were most badly fragmented across the cluster structures are organised by our layout algorithms.

4.3.3. Clustering of problematic aspects

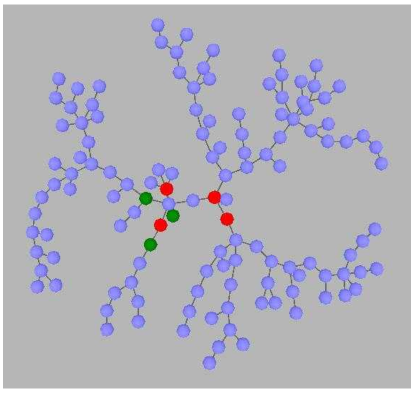
In Chapter 3, we observed how many aspects were grossly fragmented across the discrete cluster structures. In this sub-section we explore the relative utility of spatial-semantic visualization to resolve the fragmentation problems. To this end, we choose two of the mostly badly fragmented aspects from Extinction and three from each of the Chunnel scenarios and augment the visualizations with membership information for the aspects. Figure 4.3 shows the augmented visualizations. The numbers in round brackets after the aspect identifier show the clusters in which the aspect-relevant documents appear. The numbers within the square brackets show the number of relevant documents in each of these clusters.

Aspects 7 and 9 from Extinction were fragmented in the 5-cluster solution. Aspect 7 was distributed across clusters 1, 2 and 3. We can see that it is still badly fragmented in the MDS solution, with no document pairs occurring proximally to each other. We can also see that following the MDS cluster growing strategy from any exemplar would require the filtering of a considerable number of non-relevant documents before a same-aspect document would be encountered. In the MST, aspect clustering is somewhat better, with documents forming a reasonably coherent cluster and one pair of documents being directly linked and no document pairs separated by more than four links.



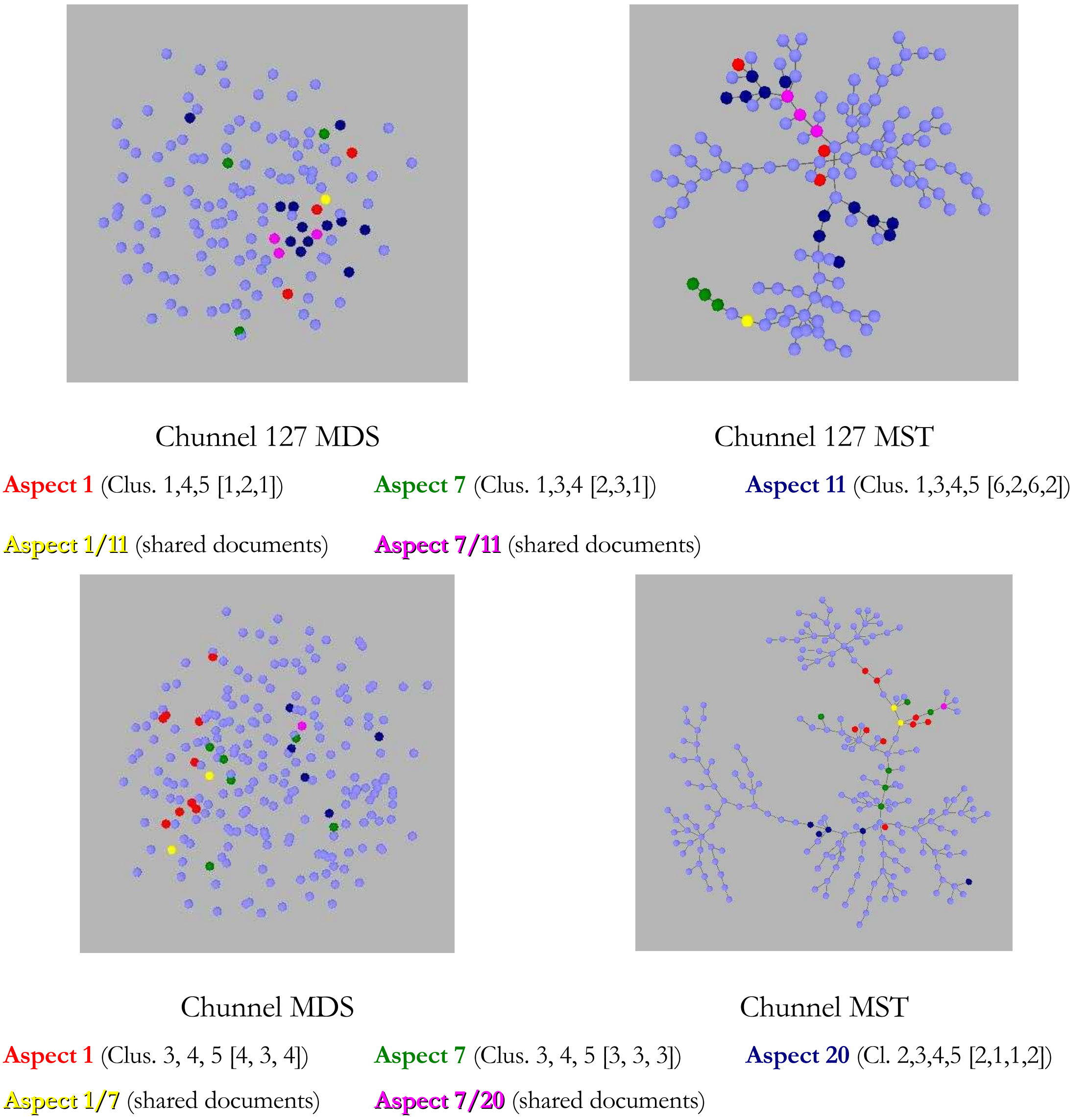
Extinction MDS

Aspect 7 (Clusters 1, 2, 3 [2, 1, 1])



Extinction MST

Aspect 9 (Clusters 1, 2, 3 [1, 1, 1])



Chunnel 127 MDS

Chunnel 127 MST

Aspect 1 (Clus. 1,4,5 [1,2,1]) Aspect 7 (Clus. 1,3,4 [2,3,1]) Aspect 11 (Clus. 1,3,4,5 [6,2,6,2])

Aspect 1/11 (shared documents) Aspect 7/11 (shared documents)

Chunnel MDS

Chunnel MST

Aspect 1 (Clus. 3, 4, 5 [4, 3, 4]) Aspect 7 (Clus. 3, 4, 5 [3, 3, 3]) Aspect 20 (Cl. 2,3,4,5 [2,1,1,2])

Aspect 1/7 (shared documents) Aspect 7/20 (shared documents)

Figure 4.3: Clustering of the most problematic aspects in spatial-semantic solutions

Aspect 9 is rendered better in MDS with the associated documents forming a neat equilateral triangular structure. However, this is not a particularly coherent cluster: for each aspect case there are still several nodes that separate them from their same-aspect relations. The MST rendering of aspect 9 is quite similar to MDS in terms of raw proximity, although there is less crowding in the proximity and none of the relevant documents are more than three links away from each other.

Three problematic aspects were selected from the Chunnel 127 cluster solution. Aspect 1, comprising four documents, was split over three of the five clusters with the best cluster containing two documents. We can see that the same problem remains in both MDS and MST with the three red nodes and single yellow node (shared with aspect 11) scattered across the structures. As in the clustering solution two of the documents are very close (one red, one yellow) in MDS but the others are isolated from all other representatives. The same pattern is evident in MST with two reasonably proximal nodes (although not quite as proximal as the cohesive pair in MDS) and two isolated nodes. Note that the yellow document that was part of the proximal pair in MDS is one of the isolated nodes in MDS. This highlights the grossly different approach to layout between the two algorithms.

Aspect 7 was also spread over three clusters in the k-means solution. The associated documents are shown as green and purple (shared with aspect 11) nodes in figure 4.3. This time the pattern is quite different between the two schemes. In MDS we can see that the three shared nodes are relatively cohesive but the other three green nodes are isolated and scattered across the visualization. The organisation is somewhat better in the MST solution. The three purple nodes are again proximal, but form a distinctive continuous string. Better still, the remaining three green nodes form an identical string, albeit in a separate region of the visualization.

Finally, aspect 11 is represented by a large sub-set of 16 documents. This aspect was particularly challenging for the k-means algorithm, which scattered relevant documents, over four of the five clusters. MDS performs a good job at organising this aspect, with a dense cluster of 13 out of 16 documents (purple, blue and yellow nodes) to the right of centre. This strong clustering may reflect a key strength of the global optimisation criterion for retaining major semantic features (i.e., themes), particularly when documents are highly topical, as evidenced by the fact that documents associated with this aspect frequently discuss other aspects, even within this restricted sample. In contrast, MST does a less impressive job by splitting the aspect into two main clusters and leaving the yellow node isolated.

Three aspects were selected from the Chunnel 218 cluster solution. Aspects 1 and 7 were distributed across three clusters, whilst aspect 20 was distributed across four clusters. Two documents are associated with both aspect 1 and aspect 7, and are differentiated through their yellow mark-up. One document is associated with aspects 7 and 20, and is differentiated by its purple mark-up.

We can see that aspect 1 is still badly fragmented in both MDS and MST solutions. In MDS, four documents form a relatively coherent cluster but the other seven documents are quite fragmented. MST renders the aspect slightly better overall with all but one of the relevant documents forming a relatively cohesive cluster, emphasised by a coherent chain of five relevant documents, which includes the two documents shared with aspect 7. This represented a considerable improvement over the discrete clustering solution (section 3.7.2) that was only able to partition four documents into a single cluster for this aspect.

Aspect 7 is also broadly distributed across both visualisations. In MDS, there is a reasonably coherent cluster of four document nodes (three green, one yellow) just left of centre. Whilst this is an improvement on the discrete cluster solution, all other documents are still quite distal. In MST, the relevant documents also form quite a broad distribution overall, however there is a reasonably coherent cluster of five documents (including both yellow and the magenta node) towards the top of the structure and a chain of three documents running vertically below the main clump, separated by three links (two non relevant nodes). The final node is completely isolated from the rest of the sub-set.

Finally, aspect 20 was the most fragmented of all aspects studied in Chapter 3 (section 3.7.2). The distribution of this aspect is quite broad in MDS, although three documents (purple node and two adjacent blue nodes) cluster reasonably coherently. The situation is similar in MST, where there is a coherent cluster of three documents (towards the bottom left) with a fourth reasonably proximal, whilst the remaining two documents are quite distal from the main clump of four and each other.

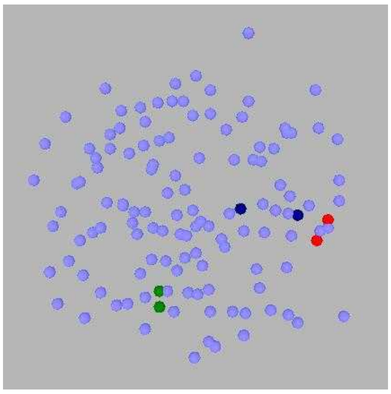
In summary, there is some evidence that spatial-semantic visualization can provide better aspect clustering than a discrete cluster solution of the same similarity data. Even where the fragment clusters are no larger than those found in discrete clusters, they do at least tend stand a good chance of being organised relatively cohesively, which may not be the case were they organised ‘within cluster’ according to query relevance or similarity to a conceptually higher-level cluster centroid (see Hearst and Pederson, 1996).

MDS and MST consistently produce quite different aspect sub-set configurations. MST has a tendency to split complex aspects into several tight clumps whereas MDS is more prone to producing a relatively tight main cluster with the remaining nodes being left isolated in apparently random parts of the visualization. In terms of H6, it is difficult from these first impressions to determine the general superiority of either algorithm for dealing with the more problematic aspects. On balance it seems that MST would be better for our interaction model because it appears that the discovery of a novel aspect exemplar could be readily followed, using our strategy, to locate at least one other same aspect document.

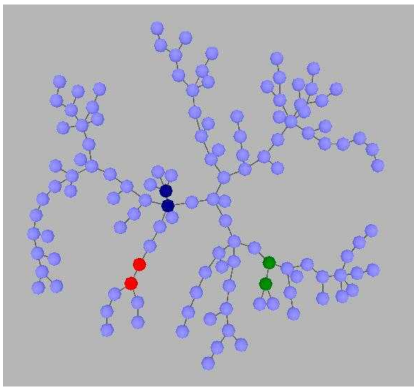
Given that these aspect examples were likely to be some of the most challenging aspects in the scenarios, we were not expecting perfect clustering in any of the cases. Many of these aspect sub-sets are likely to cluster poorly because inter-document similarities tend to be quite low. This could be due to vagueness of the aspect definition or vocabulary mismatch between same-aspect documents. We now look at more favourable aspect cases, where intra-aspect document similarity is known to be high and we would therefore expect better clustering.

4.3.4. Clustering of cohesive aspects

In this sub-section, we take three highly cohesive aspects from each scenario and evaluate the cohesion of their documents sub-sets within their respective spatial-semantic solutions. It is in this analysis that we expect the local approach, MST, to shine. Cohesive aspects were identified as those that appeared exclusively in a single cluster and had a relatively high mean intra-aspect similarity. We can see immediately, from figure 4.4, that for all scenarios, MST has the distinct advantage here. We would expect this given the bias of the MST approach to retaining the highest similarities. Whilst MDS renders the Extinction aspects reasonably coherently, those in the Chunnel set are considerably less so. For example aspects 4 and 26 are particularly poorly rendered in the MDS solution for Chunnel 218.

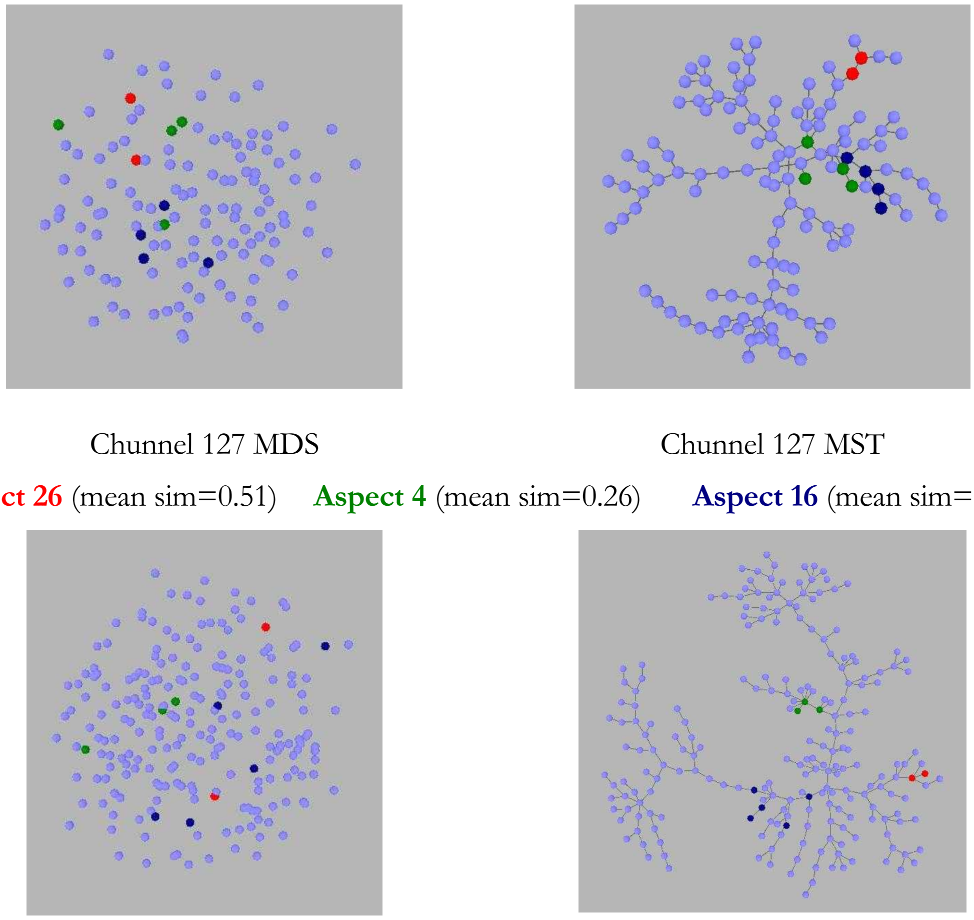


Extinction MDS



Extinction MST

Aspect 24 (mean sim=0.50) **Aspect 20** (mean sim=0.33) **Aspect 5** (mean sim=0.31)



Chunnel 127 MDS   Chunnel 127 MST

Aspect 26 (mean sim=0.51)   **Aspect 4** (mean sim=0.26)   **Aspect 16** (mean sim=0.24)

Chunnel 218 MDS   Chunnel 218 MST

Aspect 26 (mean sim=0.49)   **Aspect 14** (mean sim=0.46)   **Aspect 4** (mean sim=0.20)

Figure 4.4: Clustering of three cohesive aspects in spatial-semantic solutions

In contrast, MST handles all cases quite well or very well. The superiority of MST over MDS is most noticeable for Chunnel 218 where the three aspects form neat, well-separated clusters and Extinction where the two document aspects are always connected directly by a single link. Hence, as predicted by our H6 in section 2.6.3, the local bias afforded by MST has good potential to aggregate same-aspect documents, providing their lexical similarity is high. In contrast, MDS can isolate same-aspect documents despite their high similarity, which suggests that global optimisation can result in compromises that are counter-beneficial to our interaction model.

4.3.5. Summary

The aim of this section was to provide a preliminary overview of the organisational performance of our two visualization schemes. Although examination of only a small sample of cases is possible (without consuming vast amounts of space), key differences between the two schemes are already apparent.

The MDS solutions seem relatively amorphous compared with the distinctive tree structures of MST. However, both schemes seem to communicate a more useful classification than was possible using the k-means clustering algorithm (section 3.7). There seemed to be a high degree of correlation between spatial-semantic structure and the discrete cluster structures but in many of the sampled cases, the continuous structure allowed for far richer representation of the aspect and topic level relations.

MDS solutions tend to be very good at grouping the relevant topic but less capable of organising aspects into cohesive groups. It seems apparent that, like clustering algorithms, the global criterion of MDS favours the organisation of highly topical documents rather than more focused, aspectually distinct ones. In many cases, MDS is able to produce reasonable local configurations of same aspect documents, but often there is a tendency to isolate nodes within the same aspect sub-set, even when similarities are quite high. Topic clustering is noticeably superior to MST in the overlapping scenarios.

The local bias of MST means that aspect clustering tends to be superior, particularly for small highly focused document sub-sets. MST seems to fragment the larger, more overlapping aspects into two or more clumps of documents. This is not ideal, but for the purpose of aspect cluster growing is better than isolating many single nodes, as is often the case in the MDS solutions examined. In contrast to MDS, MST sacrifices topic clustering at the expense of preserving the strongest similarities. This is particularly noticeable in the overlapping scenarios where the topic fragments into a large number of distinct clumps. In this sense MST presents a more literal interpretation of the aspectual cluster hypothesis (Muresan and Harper, 2004), emphasising the differences within the relevant sub-set, whilst MDS focuses at the higher level and tries to find the common ground.

In the next section, we present a more comprehensive, quantitative analysis of topic classification structure using the ACS test that we presented in section 2.5.2 and have already implemented, on the inter-document similarity data, in section 3.5. Whilst this visual analysis has examined the clustering of a sample of whole aspect sub-sets, the analysis that follows considers clustering from the perspective of specific, relevant documents. Each relevant document is seen as a potential exemplar for aspect cluster growing. We are interested in examining the extent to which relevant documents tend to be located more proximally to same aspect and, to a lesser extent, same topic documents than they are to non-relevant documents. By considering this tendency within cases as a whole, we will be able to gain an impression of the extent to which the two-level classification, which is required by our interaction model, can be conveyed within a spatial-semantic structure.

4.4. Classification of topical structure

In this section, we will quantify the integrity of the topical classifications conveyed by our visualization solutions and to build a clearer picture of the relative cluster separation performance of our schemes within different scenario situations. To this end we repeat the ACS test, this time using inter-node proximities rather than inter-document similarities as our low level measures. These observations allow us to test directly hypotheses H5, H6, H8, H10, H11 and H13 (see section 4.1.1). We begin by examining general classification performance, before focusing on the effects of visualization scheme, aspect overlap and document set size.

4.4.1. General classification

The ACS test results (Tables 4.1 and 4.2) confirm that both same-topic and same-aspect documents tend to cluster more cohesively around relevant documents than non-relevant documents for all scenarios and both visualization schemes. Furthermore, relevant documents are consistently more proximal to same-aspect documents than they are to other topical documents. Figure 4.5 shows the differences between class means for all conditions.

Scenario	Overall	R-ALL v R-R	R-ALL v R-AR	R-R v R-AR
Exinction, 127 docs (n=24)	F(2,46)= 49.70***	***	***	**
Chunnel, 127 docs (n= 66)	F(2,130)= 138.93***	***	***	***
Chunnel, 218 docs (n=85)	F(2,168)= 178.92***	***	***	***
Overall (n=175)	F(2,348)= 328.64***	***	***	***

*** p<.001; ** p<.01; * p<.05 (2-tailed)

Table 4.1: ANOVA and pair-wise comparisons of mean relevance level similarity for MDS

Scenario	Overall	R-ALL v R-R	R-ALL v R-AR	R-R v R-AR

Exinction, 127 docs (n=24)	F(2,46)= 90.49***	***	***	***
Chunnel, 127 docs (n= 66)	F(2,130)= 55.43***	***	***	***
Chunnel, 218 docs (n=85)	F(2,168)= 72.80***	***	***	***
Overall (n=175)	F(2,348)= 161.16***	***	***	***

*** p<.001; ** p<.01; * p<.05 (2-tailed)

Table 4.2: ANOVA and pair-wise comparisons of mean relevance level similarity for MST

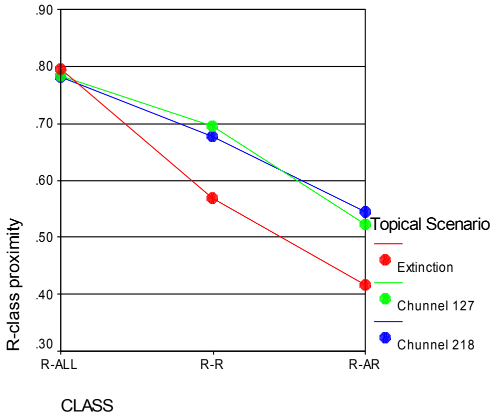


Figure 4.5a: Mean proximity, in MDS solutions, of relevant documents to all documents (R-ALL), topic (R-R) and same-aspect (R-AR)

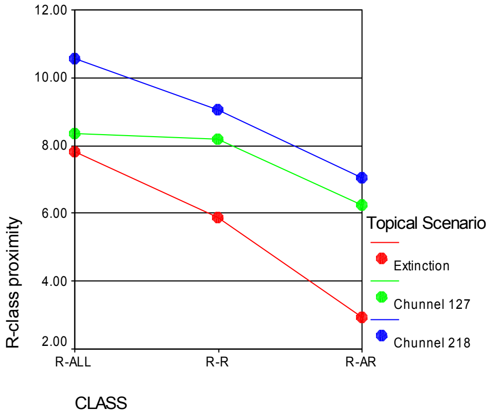


Figure 4.5b: Mean similarity, in MST solutions, of relevant documents to all documents (R-ALL), topic (R-R) and same-aspect (R-AR)

However, from the relative magnitude of the F-scores it seems that cluster separation is strongest in Extinction when MST is used but strongest in Chunnel, the overlapping scenario, when MDS is applied. Further, from figure 4.5 it also appears as if the variability in performance between scenarios is greater for MST than MDS.

However, until we standardise the scores it is difficult to gauge whether there are any key differences between visualization schemes and scenario type. We now perform ratio transformations of our observed means, as we did previously in sections 3.5.2 and 3.5.3, in order to make a fairer comparison of our layout schemes and also to examine the main and interactive effects of aspect overlap and document set size.

4.4.2. Effect of layout algorithm

The analysis in the remainder of section 4.4 uses the document-class cluster ratio measures, computed by the method first used in the previous chapter (see section 3.5.2). Each measure gives a value that describes the extent to which semantically more specific document clusters separate within a larger cluster describing a less specific relationship. Hence, R-ALL:AR measures, for a given relevant document, the ratio of the mean proximity to all documents to mean proximity to same aspect documents. If, as we would expect (H5), same-aspect documents tend to be closer to the exemplar document than other documents, then the ratio should be greater than 1; for instance a ratio of 2 would indicate than same-aspect documents are, on average, twice as close to the exemplar than all other documents.

Considering data from all scenarios together, table 4.3 shows the differences between the layout algorithms for all document class comparisons. We can see that aspect clustering is significantly superior within the MST visualizations. This is consistent with H6 (see section 4.1.1). However, this trend is reversed for topic-set cluster separation (R-ALL:R), which seems to be better within MDS solutions.

Layout algorithm	R-ALL:R	R-ALL:AR	R-R:AR
MDS	1.21	1.76	1.45
MST	1.15	1.99	1.69
t-test (df=174)	5.39***	2.84**	3.79***

Table 4.3: Comparison of MDS and MST visualizations using cluster ratio measures. Higher values indicate better separation of the semantically more specific document sub-set.

It is possible that differences exist between scenarios. In particular, there is the danger that this overall difference is skewed by the larger samples associated with the Chunnel scenarios. We will therefore now examine the effect of the different scenarios on general and specific visualization performance.

4.4.3. Effect of aspect overlap

In Chapter 3 (section 3.5.2) we found, as predicted by H3, that both topic and same aspect level clustering was more cohesive in relation to the overall set distribution for the Extinction semantic model. H8 predicted that aspect level cluster separation would be poorer in the visualizations of the overlapping scenario. We will therefore now compare Extinction (distinct aspects) with Chunnel 127 (overlapping aspects). Both scenarios have equal document set sizes (N=127). In all analyses, the standard deviations differed significantly between the scenario samples (higher in Extinction) and this difference was taken into account when interpreting the results of the independent t-tests.

Class comparison	Extinction MDS MST	Chunnel 127 MDS MST	t-value MDS MST
R-ALL:R	1.42 (n=33) 1.30	1.14 (n=67) 1.03	9.46*** 10.21***
R-ALL:AR	2.79 (n=24) 3.97	1.62 (n=66) 1.52	2.59* 5.44***
R-R:AR	1.98 (n=24) 2.91	1.41 (n=66) 1.48	1.72ns 4.42***

Table 4.4: Comparisons of Extinction and Chunnel 127 for cluster ratio measures across the two layout schemes. Higher values indicate better separation of the more semantically specific document sub-set. All significance values are based on an assumption of unequal group variances.

We can see that cluster separation performance is consistently poorer for the overlapping scenario, regardless of visualization scheme. All computed differences for MST are highly significant (p<.001). The differences for MDS are slightly less conclusive, particularly in terms of aspect cluster separation. Although same-aspect and same-topic documents tend separate well within the overall set, separation of the same-aspect documents within the topic cluster was not significant. Hence, we find general support for H8 although the significance of the predicted difference depends on the visualization scheme.

The lower significance of the observed differences for MDS suggests an interaction between visualization and scenario. H10 predicted that the aspect clustering superiority of MST (predicted by H6) would be greatest for the Extinction scenario. The rationale for this was that the local optimisation approach would work best when documents are more focused on specific aspects of the topic (i.e., aspects are more distinct).

If we look at table 4.4, we see for Extinction that the cluster separation of same-aspect documents tends to be greater for MST both in relation to the topic (R-R:AR) and all document (R-ALL:AR) distributions. Both these differences were significant (R-ALL:AR: t(23)=3.11, p=.005; R-R:AR: t(23)=3.34, p=.003). However, when we compare the visualizations for the Chunnel 127 scenario we find no significant differences between MDS and MST (R:ALL:AR: t(65)=1.13; R-R:AR: t(65)=.862). Furthermore, the observed difference for R-ALL:AR is in the opposite direction to that predicted. Hence, H10 is supported in that it appears that MST is superior, in terms of clustering same aspect documents, for the distinct aspect scenario but no better than MDS for the Chunnel scenario.

On an extra note, in the previous sub-section we observed that MDS held the advantage for general topic clustering. We can observe that this advantage is consistent across both scenarios studied in this section (Extinction: t(23)=4.59, p<.001; Chunnel 127: t(65)=9.38, p<.001).

4.4.4. Effect of document set size

In this sub-section, we compare the cluster ratio measures for the two Chunnel scenarios. H11 predicted that aspect separation would be poorer in the larger document set, due to the greater complexity of the layout problem. However, we observed in section 3.5.3 that H4 was not supported because both topic and aspect separation (relative to the set) was greater in the case of the larger version of the scenario.

Surprisingly, despite the dimension reduction process, we see the same unexpected trend here (Table 4.5). However, document set size has a significant effect only on the structure of MST solutions. Topic separation is better in the larger scenario (p<.001). This effect was clearly illustrated in the two MST Chunnel solutions shown in figure 4.1. Aspect cluster separation from the whole set is also better (p<.05). However, aspect separation within the topic cluster remains more or less the same.

Class comparison	Chunnel 127 MDS MST	Chunnel 218 MDS MST	t-value MDS MST
R-ALL:R	1.14 (n=67) 1.03	1.17 (n=87) 1.19	1.71ns 12.18***
R-ALL:AR	1.62 (n=66) 1.52	1.58 (n=85) 1.79	.60ns 2.40*
R-R:AR	1.41 (n=66) 1.48	1.33 (n=85) 1.52	1.55ns .33ns

Table 4.5: t-test comparisons between Chunnel 127 and Chunnel 218 of cluster ratio measures across both visualization schemes. Higher values indicate better separation of the more semantically specific document sub-set.

H13 predicted that MST solutions would provide the greatest classification benefits over MDS for the larger document set because the complexity of the layout problem only increases linearly rather than exponentially, as is the case for MDS. This is confirmed by our t-test analyses. Whilst there are no significant differences between MST and MDS within the Chunnel 127 scenario with respect to aspect clustering, the superiority of MST is significant in Chunnel 218, both with respect to the set (t(84)=2.39, p=.02) and the relevant topic (t(84)=2.41, p=.02). Furthermore, whilst topic clustering (R:ALL:R) is better in MDS for Chunnel 127 (t(65)=9.38, p<.001) this difference disappeared for the larger set (t(86)=.82, ns).

4.5. Discussion and conclusions

It appears that the strength of MDS, given its global optimisation criterion, is in preserving the high-level structure (major themes) of the semantic model. As with k-means clustering, documents that are broadly similar (e.g., topically relevant but about different aspects) tend to get grouped or rather thrown together, on this high-level basis. In other words, the finer aspect relations are sacrificed. MST, on the other hand, is working from the opposite perspective and sacrifices the higher-level relations somewhat in favour of preserving more distinct, albeit minor semantic features within the semantic model (see figure 4.4).

The observed differences in relative performance of the layout schemes between the distinct and overlapping scenarios seem to be consistent with their differing approaches to optimisation. When aspects are distinct and represented by focused, single-aspect documents, the local criterion of MST tends to be superior. In the overlapping scenario, MST has more problems. This is because whilst documents may cluster cohesively with some same-aspect documents, it is possible that the aspect sub-set will have been fragmented across the visualization space into several smaller clusters (see figure 4.3). If relevant documents tend to discuss many such aspects then the problem is likely to be compounded.

In MDS, on the other hand, although specific same-aspect sub-sets tend to be less cohesive (see figures 4.3 and 4.4), the grouping of the topic sub-set as a whole is more cohesive in relation to the whole visualization space. Therefore, the potential maximum distance between same-aspect (as well as different aspect) documents will be, on average, lower than in an MST solution. In other words, MDS appears to resolve the local contexts of relevant documents equally well in the overlapping scenario not because specific aspect clusters are more cohesive, but because the topical cluster is more cohesive.

The comparison of the small and large versions of the Chunnel scenario produced unexpected results, with cluster separation increasing in the larger scenario for MST, particularly for topic-set separation, yet remaining unchanged for MDS. The former result is particularly interesting. If we look at figure 4.1 again, we can see that there is a large branch (>80 documents) extending out to the left-hand side, which contains mostly non-relevant documents. This could be an artefact that skewed the aspect and topic clustering means towards higher values. However, the improvement in performance is consistent with the cluster separation differences found between the Chunnel semantic models in Chapter 3 (section 3.5.3). Either way, whilst there were no differences in aspect cluster separation between schemes in Chunnel 127, the advantage of MST seen in the Extinction scenario returns for Chunnel 218. Furthermore, whilst MDS consistently produced more cohesive topic clustering for the two smaller scenarios, this advantage is eroded in visualizations of the larger scenario. Together, these results provide strong support for H13 and the general hypothesis that MST is a more scaleable visualization algorithm than MDS. The question is why does document clustering improve for MST when the set size increases?

A possible explanation is that even though precision drops as recall increases the strength of intra-aspect similarities remains. We observed from our visual analysis of cohesive aspects (Figure 4.4) that mean aspect similarities for aspects (specifically aspects 4 and 26) tend to remain equally high in the larger set. We also know from section 3.3.5 that, if anything, average inter-document similarity decreases as set size increases. Given this, the proportional increase in the number of retained similarities results in more of the strong aspect similarities being retained. An interesting focus for future work would be to determine how precision-recall ratios interact with the strength of intra-aspect similarity values. Another hypothesis is that pathfinder networks (PFNET: Schvaneveldt et al., 1989), which are similar to MSTs but retain more than N-1 edges so long as the triangle inequality is upheld, may convey aspect relations more completely than MST as more key similarities would be taken into account during layout. A strong advocate of the merits of PFNET over MST, albeit in a distinctly different domain, is Chen (Chen and Morris, 2003). In his study of knowledge domain visualization he found that key relationships, depicting higher-order shortest paths between documents were often dropped by MST.

Overall our results show that MST is better than or equal to MDS in its ability to separate same-aspect documents from other, less strongly related or unrelated documents within the visualized set. When the algorithm does perform poorly, at least it is relatively (compared to MDS) rare for an exemplar to be completely isolated from all same-aspect relations. Hence, we anticipate that aspect cluster growing in the overlapping scenario, at least for the first one or two nearest neighbours, will be superior in MST even though the farthest neighbours may be more distal than in MDS.

Our interaction model aims to support the user in what is ostensibly a preliminary and highly exploratory information-seeking task. The user has only an open question and lacks the knowledge to formulate useful, focused queries. We anticipate a berrypicking (Bates, 1989) pattern of search, in which the focus of the query is highly dynamic. The user is most interested in expanding their knowledge (finding new aspects of the topic). Whilst the user will want to follow up new aspects as they are discovered, these focused searches are likely to be opportunistic and transitory intentions (Bates, 1989; Xie, 2000) rather than systematic and exhaustive searches. Having identified one or two good aspect examples, the user will be keen to find new instances of the topic or will be more likely to be unintentionally distracted by new instances and so the query will shift again. As we discussed in Chapter 1, our interaction model is supporting an information-seeking goal that would traditionally be accomplished using the interactive scanning strategy (Harter, 1986; Marchionini, 1995). Only once the exploration of the high-recall retrieval set is complete would the user then proceed, armed their new knowledge of the topical structure and a few good examples of key aspect (pearls), to perform a series of more focused, exhaustive searches within the IR system.

Returning to our visualization problem, it can therefore be seen as most important that each potential aspect exemplar is reliably located proximally to at least one or two similarly relevant documents. We have presented evidence here that spatial-semantic visualization holds the potential to remove the burden of query reformulation and cognitive integration of changing document views. Currently, MST seems like the algorithm that would most reliably and effectively satisfy this relaxed criterion. However, we need to test the potential utility of the aspect cluster growing strategy in a more explicit way.

Whilst the ACS test provides a broad impression of classification integrity, as we discussed in Chapter 3, it is a crude test that may be a misleading predictor of cluster growing performance particularly for scenarios where the aspect sub-sets are large or where unusually strong inter-documents similarities skew the mean. In the next chapter we see whether our original conjecture (H7), that MST will enable more efficient aspect cluster growing performance, is correct by repeating our NAN test on the inter-node proximity data. This instantiation of the test is equivalent to a simulation of the user performing the aspect cluster growing strategy and is similar to the strategy based evaluation method used by Leuski (2001). Combined with our observations in this chapter, the results from the analyses of the NAN test data will allow us to definitively choose the optimal layout scheme.

CHAPTER 5: ASPECT CLUSTER GROWING STRATEGY

5.1. Introduction

So far we have examined the extent to which both high-dimensional semantic models and spatial-semantic visualizations of these models are able to classify the structure of the relevant topic within the context of a retrieved set containing both relevant and non-relevant items. Results show that the structure of both high and low (visual) dimensional models can effectively separate relevant documents from non-relevant items at two-levels of relevance to the topic: general and aspect. In chapter 3, we evaluated the potential upper-bound performance of a key strategy afforded by this kind of spatial-semantic visualization: the aspect cluster growing strategy. This was achieved by performing the NAN test which simulated a user performing a focused aspect search, starting from a single known, relevant exemplar and examining unseen documents in relative similarity order. We found that this strategy, on average, enabled the user to identify two relevant documents in less than 10 viewing steps in nearly 70% of potential exemplar cases.

In this chapter, we repeat the strategy simulation, using the NAN test, only this time we assume a user who is searching within a visualised representation of our semantic models. As such, aspect cluster growing is guided by spatial-semantic cues (relative proximity to the exemplar) rather than pure similarity cues. The first aim of this chapter is to conclude our analysis in relation to question two by determining whether the aspect cluster growing strategy can be performed efficiently using spatial-semantic cues present within visualizations of the semantic models and to determine which layout approach, MDS or MST, is optimal for this purpose.

Our second aim is to begin to address question three, where we seek to characterise the conditions associated with cases where the aspect cluster growing strategy fails. We will use the outcome of this analysis later, in Chapter 6, to guide the design of interactive tools that provide extra support to the user engaged in the strategy. Our analysis approach has two stages. First, in section 5.3, we investigate the extent to which cases fail due to compromises in the layout process (node misplacements) by comparing NAN scores in similarity space (from Chapter 3) with those achieved in the spatial-semantic structures. We find that a proportion of poor cases are due to node misplacement and can thus be resolved simply by dynamically encoding relative similarity information into the visualization when an exemplar is selected. However, we find a significant proportion of cases where neither similarity nor spatial-semantic cues are sufficient to allow efficient performance of the strategy. In section 5.4, we take these residual problem cases and attempt to characterise the nature of these exemplar documents, their relationship to the topic and the retrieved set in general. From our analysis we identify key differences between good and poor cases that provide us with a basis from which to develop interactive tools that will provide more complete support for the aspect cluster growing strategy.

We begin by restating the questions that are dealt with in this chapter along with the specific hypotheses.

5.1.1. Research questions and hypotheses

Question 2 asked: *Given an adequate semantic model, which approach to spatial-semantic layout best preserves the general and, in particular, the low-level structure expected by our interaction model?*

In chapter 4 we performed the first stage of analysis in relation to this question by testing hypotheses relating to general classification of the topic within the spatial-semantic visualizations. In this chapter, we proceed to the second stage of analysis of spatial-semantic structures where we evaluate the potential retrieval precision of the aspect cluster growing strategy, comparing performance between our two layout approaches and between scenarios. Our specific hypotheses for the following analyses are as follows:

H7: Aspect cluster growing will be more efficient when using the MST visualizations compared to the MDS visualizations

We expect MST to be generally superior due to its emphasis on preserving the strongest relations within the spatial-semantic structure. We have already observed that, as predicted by H6, aspect sub-set separation tends to be greatest with the MST visualizations.

H9: Aspect cluster growing will be less efficient in the overlapping aspect scenario compared to the distinct aspect scenario.

We will compare the distinct aspect scenario (Extinction) with the equal-sized overlapping scenario (Chunnel 127). We expect the overlapping scenarios to be more challenging to both layout schemes due to the multi-lateral nature of its topical document relations. We have already observed that, as predicted by H8, aspect cluster separation tends to be greater within the distinct aspect scenario.

H10: The expected differences between MST and MDS will be greatest for the distinct aspect scenario.

We have already found support for this hypothesis in relation to ACS. As MST focuses on the strongest relations, we would expect MST to perform better in the distinct aspect scenario, as the differential between same aspect and same topic similarities seems to be greatest.

H12: Aspect cluster growing will be less efficient when using the larger retrieval set.

As the complexity of node layout increases with set size, we would expect cluster growing efficiency to drop as set size increases. However, the related hypothesis, H11, was rejected in the previous chapter because aspect cluster separation was unaffected by set size for the MDS approach, and separation was actually better for the larger set when using the MST scheme.

H13: The expected differences between MST and MDS will be greatest for the larger retrieval set.

We expect that MST will handle the larger set better because the rate at which the complexity of the layout problem increases is considerably lower than for the global, MDS scheme (linear as opposed to exponential). We have already found partial support for this hypothesis whereby significant difference in aspect separation only occurred for the larger scenario.

Once we have concluded our analyses for question 2 we begin to answer our final question. Question 3 asked: *Under what conditions does the aspect cluster growing strategy tend to fail and how can we use this knowledge to guide development of interactive support tools?*

Our related hypothesis (H14) is as follows:

H14: The majority of problematic cluster growing cases are due to node misplacements and can thus be resolved by augmenting the visualization with relative similarity cues

If node misplacement is responsible then dynamic augmentation of the visualization with relative similarity cues, on selection by the user of an aspect exemplar should easily resolve a problematic aspect cluster growing case. We test this hypothesis by comparing the relative utility of following similarity cues as opposed to spatial-semantic (proximity) cues for all cases.

We anticipated that in some extreme cases the strategy fails because the exemplar is simply not similar enough to the other aspect documents. In our final analysis section, we examine cases where the exemplar is both distal and relatively dissimilar to other aspect documents. We need to understand the nature of these cases, so that we can develop appropriate interactive tools to help the user help the user orientate to more profitable region of the visualization. We compare good and bad cases across a number of 'exemplar factors' that describe, from a number of perspectives, the relative importance of the target documents with respect to the exemplar and the retrieved set as a whole and also the conceptual ambiguity of the exemplar. This analysis was exploratory so there are no formal hypotheses. The exemplar factors we consider are: aspect size (number of documents relevant to the current aspect query), aspect relations (number of documents relevant to at least one aspect associated with the exemplar), aspectual diversity (number of distinct aspects discussed by the exemplar), aspect salience (ratio of aspect size to aspect relations), relevance ranking (of exemplar to the original topic query) and aspect similarity (mean similarity of exemplar to other documents relevant to the current aspect query).

Hence, in the next section we present stage two of the analysis for question two, before proceeding to address question three in the remaining sections.

5.2. Strategy performance

As in the NAN analysis in Chapter 3, each data case constitutes a simulation of the user performing the aspect cluster growing strategy, using a particular relevant document as the reference point for locating two further relevant documents. For each case, NAN scores are calculated by sorting all documents according to their relative proximity to the exemplar document within the respective visualization and observing the rank position of the 1st and the 2nd relevant documents for the aspect under consideration.

Hence, if a document discusses three relevant aspects it will constitute three distinct cases with the sample. Naturally cases were only calculated for aspects of three documents or more. This means that the sample size for Extinction (n=17) is somewhat smaller than those for Chunnel 127 (n=110) and 218 (n=143) where aspects tend to be much larger and documents are more likely to discuss multiple aspects.

Tables 5.1 and 5.2 summarise NAN score averages, R2-precision averages and the percentage representing the proportion of cases where R2-precision is less than or equal to 0.2 for MDS and MST respectively. R2-precision is a simple conversion of the 2nd NAN score to a format more familiar to the IR community and is calculated by dividing 2 (the number of retrieved documents) by the 2nd NAN score. The percentage in the far right column therefore represents the proportion of cases where R2-precision exceeds our threshold criterion of 0.2 (see section 2.5.2).

Scenario	NAN score (Average rank similarity of nearest aspect relevant neighbours)		R2-Precision	% R2-P =< 0.2
	1st relevant	2nd relevant		
Exinction, 127 docs (n=17)	14.35 (10.00)	28.94 (26.00)	0.069 (0.077)	5.9%
Chunnel, 127 docs (n= 110)	10.15 (4.50)	18.08 (11.50)	0.111 (0.174)	45.5%
Chunnel, 218 docs (n=143)	14.42 (7.00)	31.02 (17.00)	0.064 (0.118)	34.3%
Overall (n=270)	12.68 (7.00)	25.62 (15.50)	0.078 (0.129)	37.0%

Table 5.1: MDS nearest neighbour analysis for all three topical scenarios. For each cell means are shown first followed by median in brackets.

Scenario	Average rank similarity of nearest aspect relevant neighbours		R2-Precision	% R2-P =< 0.2
	1st relevant	2nd relevant		
Exinction, 127 docs (n=17)	11.88 (7.00)	30.06 (13.00)	0.067 (0.154)	41.2%
Chunnel, 127 docs (n= 110)	10.55 (3.00)	18.34 (6.50)	0.109 (0.308)	65.5%
Chunnel, 218 docs (n=143)	14.15 (3.00)	28.90 (7.00)	0.069 (0.286)	63.6%
Overall (n=270)	12.54 (3.00)	24.67 (7.00)	0.081 (0.286)	63.0%

Table 5.2: MST nearest neighbour analysis for all three topical scenarios. For each cell means are shown first followed by median in brackets.

The key trend in both tables is one where median NAN scores are considerably lower than the means. This positive skew is caused by a small number of particularly poor NAN scores at the upper end of each distribution. Hence, the median average provides a better indication of typical performance than the mean average.

5.2.1. Effect of layout algorithm

The first notable feature observed in tables 5.1 and 5.2 is that whilst performance means are quite similar from one visualization scheme to the other, median performance scores are considerably lower in the MST scheme. Given these skewed distributions and the difference in skewness between the distributions a non-parametric difference test was chosen. The overall performance (considering all 270 cases) of the two schemes was therefore compared using the Wilcoxon signed ranks test.

Assuming an aspect cluster growing strategy guided by relative node proximity cues alone, the 1st relevant document was found sooner or equally soon within the MST for 63.7% of cases. This difference was significant (z=2.83, p=0.005). Likewise, for the 2nd relevant document MST was also superior, with the user locating the document sooner or equally soon in 62.6% of cases (z=2.71, p=0.007). Furthermore, in 63% of cases within the MST distribution two relevant documents are found in 10 viewings or less, compared with just 37% of cases within MDS. Hence, we can conclude that H7 is supported.

However, although MST is generally equal to or better than MDS, it also provides the worst aspect cluster growing exemplars. If we take the worst 10% of cases for each scheme distribution, we find that MST is the poorer performer. The range of second NAN scores for MDS is 40-157 compared to 78-199 for MST. In other words, whilst MST generally offers superior cues for the aspect cluster growing strategy, it also provides the worst cases.

5.2.2. Effect of aspect overlap

So far we have found that MST generally provides the best support for our strategy when all cases for all scenarios are considered. We now consider the differences at the scenario level, comparing the extent to which each scheme supports our strategy when the aspects are either distinct or overlapping.

In the previous chapter we found in our comparison of Extinction and Chunnel 127 that classification performance of both schemes was negatively affected by aspect overlap. We also found that whilst MST provided superior aspect clustering in Extinction, this advantage disappeared in the overlapping scenario.

For MDS we find that cluster growing performance is superior in the overlapping scenario. This is true for both the first (z=2.73, p=.006) and the second relevant document (z=2.88, p=0.004). Further, the proportion of exemplar cases where two documents are located within 10 viewings is much higher in the Chunnel 127 scenario (45.5%) than in the Extinction scenario (5.9%). We find the same trend for MST, again both for the first (z=2.53, p=0.01) and the second relevant document (2.43, p=0.02). The proportion of good cases is also higher in the Chunnel scenario (65.5%) than the Extinction scenario (41.2%), although we can see that the difference is less extreme. Hence, H9 is rejected because the observed effect, for both schemes, is in the opposite direction to that predicted.

Given that MST provided better aspect cluster separation than MDS for Extinction (see section 4.4) and that the proportion of good cases is higher (41.2% vs. 5.9%) we expected that MST would be the better visualization scheme for this scenario. However, Wilcoxon signed ranks test shows no general difference in the performance of MST and MDS (z=0.24, ns), with MST being better or equal in just 53% of cases. Again we suspected this would be due to a small number of extremely poorly performing exemplar cases in MST. This is confirmed if we look at the 90th percentile of NAN (2nd retrieval) scores where MDS (57.6) is considerably lower than MST (80.8).

Likewise, for Chunnel 127, we find no difference between the two visualization schemes (z=0.94, ns) with MST being superior to MDS only 50% of the time and equal in 6.4% of cases. Also, although the proportion of good cases was higher in MST (65.5% vs. 45.5%) the most poorly performing exemplars were worse in the MST distribution (90th percentile = 61.9) than MDS (90th percentile = 44.7).

Hence, contrary to our observations in the previous chapter, H10 is not supported in relation to aspect cluster growing performance because no significant differences occur between the two layout schemes in either of the two scenarios.

5.2.3. Effect of document set size

We found in the last chapter that increasing the set size had either a non-significant effect (MDS) or a beneficial effect (MST) on aspect cluster separation. This went contrary to

H11, which predicted that cluster separation would be greater for the smaller set, but consistent with the observations in chapter 3 where we also rejected H4 because topic and aspect cluster separation within the set cluster was significantly higher for the larger set.

Here, we find that aspect cluster growing is less efficient in the larger set when using the MDS scheme both for the 1st (z=2.28, p=0.023) and 2nd retrieval (z=3.32, p=0.001). Further, the proportion of good cases (R2-precision <=0.2) is also lower in Chunnel 218 (34.3%) in comparison to Chunnel 127 (45.5%). In contrast for MST we find that cluster growing performance does not change for either the 1st (z=0.143, ns) or 2nd (z=9.55, ns) retrieval as set size increases. Likewise, the proportion of good cases is almost equal (65.5% vs. 63.6%). As mentioned earlier in this sub-section, this is consistent with the results of our analysis of clustering growing in similarity space (H4). Hence, H12 is only partially supported, in that it is true for MDS but not for MST.

We expected that MST would cope better with the increased layout demands of the larger set. We have already observed that there is no difference between the two visualization schemes for Chunnel 127 (z= 0.94, ns). However, the difference between MDS and MST is highly significant for Chunnel 218 (z=2.69, p=0.007). For both scenarios the proportion of good cases is higher (Chunnel 127: 65.5% vs. 45.5%; Chunnel 218: 63.6% vs. 34.3%). Hence, H13 is supported.

5.2.4. Summary

Overall we find that that aspect cluster growing is, on average, at least if not more efficient when using a MST visualization compared to using a MDS visualization. Furthermore, we find that a much larger proportion of MST cases meet our R2-precision criterion. However, the differences between the two visualization schemes are attenuated by the tendency for MST solutions to comprise a small number of very bad aspect exemplars; whilst MST seems to provide the better visualization scheme for our interaction model, there are a significant number of cases where the aspect cluster growing strategy cannot be effectively guided by spatial-semantic information.

Overall there are 37% of cases where the criterion for locating two documents in less than 10 viewings is not met. In the Extinction scenario this proportion increases to 68.8%. Part of this shortfall is likely to be due to information loss during the dimension process. In the next section we evaluate the extent to which cluster growing performance for the worst

MST cases is due to node misplacements that occur during the layout process by comparing aspect cluster growing performance in MST with that in high-dimensional similarity space.

5.3. Effect of node misplacement

In this section, we determine the utility of replacing spatial-semantic cues with the original relative similarity information. We take the data used for the NAN testing in Chapter 3 and compare this to the MST data. By comparing these two distributions we can measure the extent to which compromises during layout limit the success of our strategy. We can also gain an impression of the relative benefits of augmenting the visualization with relative similarity information when an exemplar is identified to the system. Such a tool would be akin to the 'show me more like this' relevance feedback tool that is available in some web search engines. This is a simple approach to relevance feedback that requires on a single example of relevance and is therefore less demanding on the user than full document relevance feedback. The user simply indicates that a document is relevant and the system uses the document vector as a query to retrieve similar documents and present these to the user in rank similarity order (see, for example, Jansen et al, 2000; Hearst, 1999).

We will now compare the relative efficacy of similarity and spatial-semantic cues to establish the extent to which such a dynamic augmentation of the visualization would resolve the sub-sets of problematic cases we have observed.

5.3.1. Comparison of similarity and spatial-semantic cues

H14 predicted that the majority of problematic aspect cluster growing cases would be attributable to node misplacement. In other words, the replacing spatial-semantic cues with similarity cues will resolve most, if not all, problematic cases (i.e., increase R2-precision to 0.2 or higher).

Table 5.3 repeats the results of the NAN test reported in Chapter 3 (section 3.6). If we compare these data with those of MST reported in section 5.2 (table 5.2) we see that overall following similarity order from a given aspect exemplar is generally more effective than following proximity order within an MST solution, both for the 1st (z=9.045, p<.001) and for the 2nd (7.362, p<.001). R2-precision median is 27.2% higher overall and the mean is over twice as high (+214%). Furthermore, a higher proportion of cases meet our precision criterion for the second nearest neighbour, however this benefit is only 9% (68.5 vs. 63%) when considering all scenarios together.

Comparing the data within each scenario we see that whilst performance is generally much better for the Chunnel scenarios, although following MST proximity cues actually appears to be more efficient than following similarity order in the Extinction scenario, at least for the 2nd nearest neighbour. Whilst this difference is not significant (z=0.426, ns) we can see that considerably more cases met the precision criterion for the second neighbour when following spatial-semantic cues in MST (41.2% vs. 17.6%).

Scenario	Average rank similarity of nearest aspect relevant neighbours		R2-Precision	% R2-P =< 0.2
	1st relevant	2nd relevant		
Exinction, 127 docs (n=17)	6.824 (6.000)	28.824 (22.000)	0.069 (0.091)	17.6%
Chunnel, 127 docs (n= 110)	4.255 (2.000)	10.364 (5.000)	0.193 (0.400)	72.7%
Chunnel, 218 docs (n=143)	5.007 (2.000)	10.322 (5.000)	0.194 (0.400)	71.3%
Overall (n=270)	4.815 (2.000)	11.504 (5.500)	0.174 (0.364)	68.5%

Table 5.3: Inter-document similarity nearest aspect neighbours analysis for all three topical scenarios. For each cell means are shown first followed by median in brackets

Hence, in many cases, particularly those of the Chunnel scenarios, following similarity order seems likely to prove a useful alternative strategy to following proximity cues. We could envisage, for example, an interactive tool where the user selects the exemplar document and the system highlights and possibly labels with rank position, the top 10 most similar documents. We now determine the extent to which this simple relevance feedback approach would resolve problematic aspect cluster growing cases in MST.

Problematic cases are defined as those where the strategy of following proximity cues falls below a precision of 0.2. We find that although this strategy is generally more efficient than following cues provided by the MST structure, providing relative similarity cues can does not resolve all of the cases that are problematic when using spatial-semantic cues. If we consider all aspect exemplar cases (n=270) and select only those cases where MST failed to meet the 20% precision criterion we can identify 89 out of the original 270 cases (33%)

where the first relevant document cannot be located in five or fewer viewings. Measuring the proportion of these 89 cases that satisfy the criterion when similarity, rather than MST proximity cues are used we find that 29.2% of the problematic MST cases can be resolved by substituting relative proximity cues with relative similarity cues. Repeating the same analysis for the second relevant document, there are 100 out of the original 270 cases (37%) where MST fails the precision criterion. We find that 30% of these problematic cases can be resolved by substituting proximity cues with relative similarity cues.

Hence, the strategy of substituting spatial-semantic cues with similarity cues only resolves around 30% of problematic cases in the MST solutions to our topical scenarios. This still leaves a large proportion of all cases where aspect cluster growing cannot be supported by either proximity or relative similarity cues. Specifically this is 63 cases (23%) for the 1st retrieval and 70 cases (26%) for the 2nd retrieval.

In these cases same aspect documents are not sufficiently similar (in terms of general similarity) to cluster cohesively around the exemplar in either visual or high-dimensional term space. Clearly, a more appropriate and powerful alternative to simple 'more like this' relevance feedback is required for such cases. In the next chapter we propose an approach for enhancing the simple relevance feedback strategy. This approach is inspired by the analysis that follows in the next section, where we model the specific correlates and potential causes of poor exemplar performance.

5.4. Correlates of combined strategy performance

Given that 23-26% of document exemplar cases in our three scenarios fail to meet our 20% precision criterion when either proximity or similarity cues are used to guide aspect cluster growing, we need to understand why general similarity values are an insufficient cue to guide the user's search. In this section, we compare the properties of problematic exemplar cases to those that are able to meet the precision criterion in either spatial-semantic or similarity space. In doing so we gain a clearer understanding of why some documents make poor exemplars, this enables us to hypothesise (later in section 5.5) how simple document relevance feedback approach can be enhanced to provide a more informative cues without incurring excessive, additional demands on the user.

The fundamental cause of our problematic cases is that they are not similar enough to the other documents discussing the aspect of interest to constitute a good exemplar for cluster growing. An obvious recourse here is look for ways of improving intra-aspect document similarity at the text analysis stage of the process. Potentially useful avenues include analysing text at the sub-document level, perhaps by dividing documents into topically coherent passages (e.g., see Hearst, 1997; Ostler, 1999; Larocca Neto et al., 2000; Kleinberg, 2002) and term-vector dimension reduction approaches such as LSA (Deerwester et al., 1990; Karypis and Han, 2000). However, in this work we are most interested in developing techniques and strategies that make the most of the information already available in a given semantic model rather than looking to optimise semantic modelling per se. Hence, in this section, although we include relative similarity within the analysis for completeness, we do not view it as an explanatory variable *per se*.

We compare the properties of cases where aspect cluster growing precision drops below 0.2 for both SIM and MST with all other cases. We examine a number of 'exemplar factors' that were introduced in section 5.1.1. These variables describe, from various perspectives, the relationships between the exemplar, the aspect subset and other documents in the set. If good and bad cases can be distinguished with respect to one or more of these variables, then this will provide us with clues to the kind of interactive support (additional cues) that might enhance the aspect cluster growing strategy. In the next sub-section we briefly justify our rationale.

5.4.1. Outline of exemplar factors

As a reminder, the exemplar variables to be studied are aspect size, aspect relations (of the exemplar), aspectual diversity (of the exemplar), aspect salience, rank relevance (of the exemplar to the original query) and aspect similarity (relative to the exemplar). Aspect similarity is the mean inter-document similarity between the exemplar and all aspect relevant documents and is included simply for reference and comparison. A significant difference between good and poor exemplars on this variable almost goes without saying as, by definition, poor cases are those where relative inter-document similarity is not high enough to guide the location of relevant documents. As a reminder, our aim is to characterise the nature of poor cluster growing situations so that we can go beyond simple similarity cues to provide additional, more specific cues (e.g., key terms) that can orientate the user more efficiently towards unseen relevant documents. Hence, it is the observed differences on the remaining exemplar factors that will be of primary interest. We now define each factor and explain how significant differences between good and poor cases might inform the design of alternative interactive strategies and tools.

Aspect size is simply the number of documents relevant to the topical aspect under consideration. If poor cases tend to be associated with a smaller aspect size, then our aim would be to develop a means of supporting the identification of distinct but minor features within a generously sized local context surrounding the exemplar.

Aspect relations is the number of same aspect relations that the exemplar has within the document set. A significant difference between good and poor cases on this variable would be ambiguous by itself, because a high number of aspect relations might be due to high aspectual diversity within the document or a large aspect sub-set size, or both. Hence, the implications would depend on the co-occurrence of differences on one or more other variables.

Aspectual diversity is the number of defined aspects associated with an exemplar case. Rationally, the likelihood of an exemplar becoming isolated from the main cluster of the current aspect will increase if it discusses several distinct aspects of the topic, particularly if the document tends to be more similar to documents about another aspect. Augmentation of members of a largish local context within the spatial-semantic structure might, for instance, separate the local context into distinct clusters. What would be needed is a means of differentiating these emergent features using discriminating labels.

Aspect salience describes the salience of the aspect of interest within the local context of all documents that discuss the same aspect or aspects as the exemplar. It is calculated as the ratio of aspect size (minus one to allow for the exemplar) to all aspect relations. Hence, if a document focuses on only one aspect then salience equals one. If the exemplar discusses several aspects but most of the aspect relations are about the current aspect of interest then aspect salience would be greater than one-half. Clearly this measure is similar to aspect diversity as salience is likely to drop as exemplars become more diverse. However, it is slightly more sensitive in that it accounts for the relative size of the current aspect in multi-aspect exemplar cases. If poor performance can be associated with lower salience, then it would mean that, if forced to make compromises the layout algorithm tends to locate a document within or near to the larger cluster of highly similar documents. Again, a solution that describes and distinguishes between emergent local context features might be an appropriate solution here. In particular a successful approach would be particularly sensitive to minor features or clusters within the local context.

Finally, rank relevance is the rank position of the exemplar in the original retrieved document list. More highly ranked documents are likely to be better representatives of the general topic and so may make better exemplars. If poor performance is associated with very low rank relevance, this would support Leuski's (2001) approach of combining the ranked list information with the spatial-semantic visualization. The interface might therefore encourage the user to identify distinct aspect instances from the top ranks of the list before switching to the visualization, rather than exploring the visualization directly.

5.4.2. Factors that discriminate good and poor exemplar cases

Table 5.4 shows the means and results of Mann-Whitney U-tests computed for each of the outlined variables between poor cases (where both MST proximity and similarity fail the 0.2 precision criterion) and all other cases. Non-parametric tests were chosen due to a non-normal distribution for the majority of the examined factors.

Aspect similarity aside, we see that the most significant and consistent differences between the poor and the good aspect exemplar groups occur as a result of aspect size and aspect salience. The poorest exemplars are characterised by a smaller relevant aspect sub-set size an exemplar that is related to a relatively high proportion of documents discussing other aspects of the topic.

Exemplar factor	1st NAN			2nd NAN		
	Poor (N=63)	Good (N=207)	Sig.	Poor (N=70)	Good (N=207)	Sig.
Aspect size	8.02	9.93	**	7.66	10.12	***
Aspect relations	15.25	16.93	Ns	16.23	16.65	Ns
Aspectual diversity	2.38	2.31	Ns	2.50	2.27	*
Aspect salience	0.558	0.632	+	0.523	0.647	**
Rank relevance	53.16	62.15	Ns	48.51	64.09	*
Aspect similarity	0.114	0.174	***	0.115	0.176	***

+ $p<.10$, * $p<.05$, ** $p<.01$, *** $p<.001$ (2-tailed)

Table 5.4: Differences between good and poor exemplar cases with respect to our specified exemplar factors. Poor cases are exemplars that fail the 0.2 precision criterion for both relative similarity and proximity (MST) cues. Good cases are all non-poor cases.

There is also a weak effect of rank relevance and aspect diversity, although differences are only significant for the 2nd NAN. Poor exemplars tend to be ranked higher in the retrieved set and tend to be slightly more topical than good exemplars although the effect size for the latter variable is very small. The former result is interesting in that it suggests that the strategy of identifying relevant exemplars by browsing the top ranks of the retrieved list (Leuski, 2001) may not be the optimal approach for the user. At least this appears to be the case when the object is to find aspect level, as opposed to topic level, exemplars.

5.5. Discussion and conclusion

In this chapter we applied the NAN-test to the spatial-semantic structures provided by our visualization solutions, effectively simulating the use of the aspect cluster growing strategy for a large proportion (270 cases) of potential document exemplars across all three topical scenarios. We began in section 5.2, by comparing the relative performance of the aspect cluster growing strategy, using only spatial-semantic cues, within our two visualization schemes. These results combined with those obtained in Chapter 4 allow us to draw a conclusion on question two. We conclude that overall, MST produces a better topical classification and facilitates more efficient cluster growing performance. In the second part of this chapter (sections 5.3 and 5.4) we addressed question three. In section 5.3, we found that whilst following similarity cues generally results in more efficient aspect retrieval than following spatial-semantic cues, many problematic cases are not due to node misplacements, but rather more fundamental failures in aspect clustering within similarity space. In section 5.4, we asked whether these most severe problem cases have common characteristics that would help us to develop appropriate interactive support. We identified two variables, aspect size and aspect salience that effectively distinguish good from bad cases. We will now discuss the implications for these results.

Research question two asked which visualization scheme provides the optimal layout for our interaction model. We have compared two distinct approaches, one where the layout algorithm seeks a globally optimal representation of underlying similarities, and another where a priority is placed on preserving local features (strong inter-document similarities). We can conclude that the local optimisation approach, based on an MST representation of the similarity space, provides solutions that generally enable more efficient cluster growing in the majority of exemplar cases than the global approach, MDS.

Even though MST is superior to MDS in most cases, there are still around 37% of exemplar cases where MST proximity cues are insufficient to support efficient aspect retrieval. The dispersion of NAN scores in these problematic cases is also very broad with the worst cases requiring a considerable proportion of the visualization to be searched. This phenomenon is illustrated clearly, for instance in figure 4.3 (section 4.3.3). The yellow node in the Chunnel 127 MST screenshot is grossly isolated from the main clusters of its two related aspects (blue and red nodes).

We found that a higher proportion of problematic cases occurred in the non-overlapping scenario, Extinction. This was a curious result given that the results of our ACS tests in Chapter 4 indicated that same aspect documents formed generally more cohesive clusters in MST space. However, it is worth bearing in mind that many of the relevant aspects were not included the NAN analysis as they were represented by two or fewer instances within our retrieved set. This meant only 17 cases, covering five (out of 22) distinct aspects represented within the set were considered in the analysis. It is possible that the strong overall aspect clustering observed in the ACS test was largely due to the strong cohesion of the many small, two document aspects (see figure 4.4, section 4.3.4).

In Chapter 4 we made the interesting observation that topical classification performance of both MDS and MST was not negatively affected by increasing document set size (increasing the rank cut-off point in the retrieved list). In fact topic and aspect separation relative the whole set was actually greater within the MST visualization of the larger Chunnel scenario. In this chapter, we found that aspect cluster growing performance within MDS was negatively affected by increasing, but MST performance was unaffected. Furthermore, whilst there was no significant difference cluster growing precision between the layout schemes for the smaller Chunnel scenario, as predicted MST had the advantage for the larger set. This is an encouraging result, which supports our initial expectation that MST would be more scaleable than MDS, given that the complexity of layout increases linearly rather than exponentially with increasing document set size.

Research question three was concerned with identifying the conditions under which the spatial-semantic cluster growing strategy fails and identifying ways of providing additional support to the user performing aspect cluster growing in these situations. Given that loss of structural information was inevitable due to the dimension reduction involved in spatial-semantic visualization, our first recourse was to determine the extent to which poor exemplar cases are due to node misplacements within the visualization and can therefore be supported by substituting spatial cues with similarity information. Even with the aid of similarity cues, it seems that the 23-26% of cases observed across our three scenarios remain problematic, with aspect precision falling below 0.2.

Finally, we identified two major correlates of poor performance: aspect size and aspect salience. We found that these problematic cases can be differentiated from non-problematic cases in that they tend to occur when the size of the target aspect sub-set is smaller and the salience of this sub-set with the local semantic context (all topically related documents) is relatively small. This suggests that it may be possible to support the user by providing a means of identifying more minor, yet distinct concepts that relate the exemplar to documents that, whilst not highly similar, are at least reasonably similar in their general content.

We also found a weak effect of rank relevance of the exemplar. Counter to our informal expectations, documents that are ranked more highly in the initial retrieval set do not appear to make the best exemplars. Leuski (2001) proposed the strategy of identifying the first relevant example from the top ranks of retrieved documents before visually growing the relevant cluster from the location of this document within the visualization. Whilst this was shown to be effective for a simple topic retrieval task (Leuski, 2001) our observation suggests that it may not be the optimal strategy for identifying distinct aspect instances. It may be that direct browsing of the visualization, supported by useful overview cues to orientate the user towards dense topic-relevant patches, would be a more effective exploration strategy. This question must remain, however, for future work.

In the next chapter we apply what we have learnt in this chapter to the development of a prototype interface and interactive tools to support aspect cluster growing in problematic exemplar situations. We propose a means of identifying and describing the range of exemplar relevant concepts represented within the local context of a given document. We call this approach to term suggestion: Local Context Distillation. We propose two ways in which terms suggested by this method can be implemented as visual cues to support aspect cluster growing in problematic situations.

CHAPTER 6: SUPPORTING ASPECT CLUSTER GROWING USING LOCAL CONTEXT DISTILLATION

6.1. Introduction

In the previous chapter, we demonstrated how the aspect cluster growing strategy can produce acceptable performance in a large proportion of cases across our three topical scenarios. However, in a significant proportion of cases the precision criterion (P=0.2) was not met. In some of these cases, this was due to misplacements and compromises in layout caused by the extent of dimension reduction. However, in a vast majority of cases, the problem could not be resolved by simply replacing spatial-semantic cues with pure relative similarity cues. For some reason, in these cases, computed similarity to documents related by the aspect of interest is too low (relative to other related documents) for these documents to appear within the local neighbourhood of the exemplar. This could be because the documents share no common key terms. However, we also reasoned that this could because other relating key terms were more salient either within the exemplar or across documents within the collection. To this end we looked for correlates of problematic cases, variables that tend to differ significant between the two groups (good and poor exemplars). We identified, two salient variables in particular: aspect size and aspect salience. In this Chapter we propose and demonstrate a solution that is based on our findings; an approach for supporting aspect cluster growing in the most problematic cases.

We begin by discussing the problem within the context of existing approaches to query formulation and refinement (section 6.2). We then introduce our approach, which has two parts: an algorithm, called Local Context Distillation, that applies a weighting function to terms to select potential query terms based on the common content of the exemplar and its near neighbours (section 6.3) and an interface integrating two visual tools, Concept Signposts and Concept Pulses, that use these terms to augment the visualization with additional navigational cues (section 6.4). Finally, in section 6.5 we demonstrate the potential of these visual tools using a series of problem case walkthroughs. Section 6.6 discusses the implication of this approach and avenues for future research and development.

6.2. Related work

In Chapter 5, we found that spatial-semantic cues are often sufficient to support aspect cluster growing, meeting or exceeding our precision criterion of 0.2 in 41% to 65% of potential exemplar cases, depending on the scenario. Across all scenarios, of the cases that fail, around 30% are due to misplacements in the layout (dimension reduction) process and can be resolved by adopting a simple relevance feedback strategy where the user is provided with cues describing relative similarity of documents to the exemplar.

However, that still leaves a total of around a quarter of all observed cases where same aspect documents are not similar enough to the exemplar for the simple relevance feedback strategy to be successful. We examined these problematic cases in fine detail to determine which variables were able to differentiate good from problematic cases. From this analysis we found a number of variables that characterised poor exemplars.

Predictably, exemplars that were less similar to their aspect relations did not enable efficient aspect cluster growing. This might be due to a number of possible factors such as vocabulary mismatch (the same aspect is discussed using different terms) or low salience of the aspect within either the exemplar or related documents. These problems could be directly addressed by looking at alternative semantic analysis techniques. Whilst this avenue of research is beyond the scope of the thesis, some relevant ideas are discussed for future work in Chapter 7.

More interestingly, we found that problematic cases tend to occur when the aspect of interest is relatively small (it is represented by relatively few documents) and particularly when this the aspect is 'competing' with one or more other aspects discussed in the exemplar and the documents sought form only a small proportion of all topically related documents (low aspect salience). The most extreme cases are likely to occur when the other exemplar aspects are represented by larger and typically more coherent and/or proximal clusters within the semantic model. We therefore perceive a significant part of the problem as one where the local neighbourhood of the exemplar is polluted with a large number of documents that, whilst similar, and possibly more similar to the exemplar, are related for reasons other than the current aspect of interest (i.e., they discuss other aspects). What is needed is a means to allow the user to specify their reason for selecting the document as an exemplar more precisely.

A simple solution, from the implementation perspective, would be to include a within-set query box feature so that the user could specify the nature of the aspect of interest by specifying one or more key words. Previous work in this area has combined manual within set query functionality as a complement to spatial-semantic cues (e.g., Chalmers and Chitson, 1992; Hornbaek and Frokjaer, 1999), whereby matching documents are emphasised within the context of the visualization, for example by changing colour or brightness. However, whilst users find it useful to see the results of their queries 'in context', particularly when the highlighted documents form distinct clusters (Hornbaek and Froekjaer, 1999), the requirement to shift mode between referential (browse) and command line styles of interaction creates an additional cognitive demand that can cause users to lose focus on their primary, information-seeking task (Campagnoni and Ehrlich, 1989; Hornbaek and Froekjaer, 1999). We wanted to protect the user from this additional demand.

Further, having just identified an aspect of interest, selecting even just one or two good discriminating terms may not always be a trivial task, especially when the aspects of the document are quite close or overlapping in their semantics and thus terminology. As such, we seek a solution where the system attempts to guide the user towards their aspect of interest by making evidence-based guesses about why the selected exemplar might be relevant and allowing the user to select the closest match.

But how can the system infer the user's information need simply from their indication that a document is a relevant exemplar? This is an impossible expectation, particularly if the document discusses many different concepts and possibly topics. What it can do, however, is to speculate on why the document might be relevant based on the overlap between its semantic features and those of its nearest neighbours. Our solution attempts to explain the topical diversity of the exemplar by performing an analysis of the overlap of term usage within the exemplar and its near neighbours. The most highly weighted terms are returned to the user who can then select the most discriminating terms that best explain their reason for selecting the document as an exemplar. We call this approach Local Context

Distillation (LCD) as we are trying to separate the potential user queries that the exemplar might represent, given the context of the retrieved set.

Given the nature of problematic exemplar cases, we propose that the system sometimes needs to consider a broad sample of near neighbours to ensure that at least one representative of the aspect of interest is captured. In the implementation, therefore, the user is able to dynamically adjust this threshold until the optimal keyword list is found. The essential aim of our solution is for the system to generate a set of key words that adequately describes and discriminates the full thematic spectrum of the exemplar; its minor as well as major concepts.

In section 6.3 we describe our algorithm in more detail. Later in the chapter (sections 6.4 and 6.5), we describe how LCD terms might be applied, interactively to the spatial-semantic visualization in order to support aspect cluster growing. Before we introduce our approach, we will discuss previous work, mainly from the field of interactive information retrieval, which was considered during the development of LCD.

6.2.1. Term suggestion and relevance feedback

In this section, we propose a novel extension to the simple relevance feedback strategy that enables the user to recognise rather than specify the reason for their interest in a given relevant document. We begin by providing an overview to existing techniques that can be used to elicit key terms that can be used to refine an initial query. We break these down into two types: query expansion and document cluster labelling.

Our first avenue of enquiry was to look at term relevance feedback as a technique used for query expansion. Term relevance feedback is a refinement to the classical document relevance feedback approach. In document relevance feedback, the user browses the retrieved documents and identifies several relevant examples. The system analyses this sample and weights occurring terms according their importance (e.g., the extent to which they discriminate known relevant documents from other retrieved documents). Highly weighted terms are then added to the query or, in some systems, the existing query terms are re-weighted based on their computed salience.

In early systems this was an automatic or opaque process as far as the user was concerned. All they did was indicate a sample of relevant examples and the new query was formed and sent, leading to a revised retrieval list. However, a study by Koenemann and Belkin (1996) found that relevance feedback was more effective when the users were able to moderate the terms selected by the system for query expansion. Not only were queries slightly more precise than when the process was automated, but also good queries were achieved in significantly fewer relevance feedback iterations.

We will refer to the strategy of suggesting terms, and allowing the user to have the final say on which ones are appropriate to add as *term relevance feedback* (Roussinov and Chen, 2001), to distinguish it from classical document relevance feedback where the query expansion process is hidden from the user. A key value in term relevance feedback is that the user is given control over the process of refining the query, specifying only those features that actually define their need, without having to think of terms themselves (Koenemann and Belkin, 1996). This makes it a promising field to learn from in the development of our approach.

One key drawback of the term suggestion approach examined by Koenemann and Belkin (1996) is that, like standard document relevance feedback, it requires multiple document relevance judgements. This technique only produces useful terms if supplied with a sufficient sample of relevant examples (Hearst, 1999; Hancock-Beaulieu and Walker, 1995). However, a requirement of LCD is that it must be able to suggest key terms immediately from only one relevant example document.

A useful alternative approach to traditional document relevance feedback, that avoids this requirement, is local or pseudo relevance feedback (Attar and Fraenkel, 1977; Xu and Croft, 1996). In a local feedback system the top k (e.g., 10 or 20) retrieved documents are assumed to be relevant. Using the same methods of document relevance feedback, the most salient terms can be identified through an analysis of the discriminating properties of this sub-set. This idea neatly extends our simple ‘more like this’ relevance feedback approach, as the top *k* most similar documents to the exemplar can be assumed relevant and accordingly mined for good query terms. However, whilst local feedback minimises the demand on the user to evaluate document relevance and indicate good examples, it is also dependent upon the precision of the initial query (Hearst, 1999, p.308). Although good results are possible if many of the top-ranked documents are relevant, if this is not the case then local feedback can produce erratic and unexpected results (see Xu and Croft, 2000).

For this reason a standard local feedback approach does not seem suitable for our needs. Our problematic cases are problematic precisely because the aspect of interest is represented by a relatively small number of relevant documents and that these tend to be ranked relatively low in terms of similarity to the query (the exemplar document). Local feedback would likely suggest helpful terms in less problematic aspect cluster growing cases, where the aspect of interest already forms a salient feature within the top ranking documents, but this approach does not address our problematic cases. What we need is a means of identifying distinct, yet relatively minor semantic features shared between the exemplar and the local context documents.

A possible solution is to perform document clustering on the sampled local context and to describe the topical structure by selecting the key terms associated with each computed cluster. Clustering interfaces usually select terms for a given cluster based on the centroid or average term vector of all cluster members. A simple approach is to select the most frequent or highly weighted terms from the centroid of each cluster (see Carey et al., 2000; Skupin, 2002), although a more effective approach for our needs is likely to be one where the best terms are those that are not only highly weighted within a cluster but also relatively rare outside of the cluster (Lundquist et al., 1997). This latter approach would ensure that a bias is placed on more distinctive rather than broadly topical terms.

However, discrete clustering is, by nature a trial and error process and identifying optimal parameters (e.g., number of clusters, similarity threshold) can require significant and knowledgeable human intervention (see Xu and Croft, 2000). We also know from our own analyses in Chapter 3 and those of others (see Wu et al., 2001; Muresan and Harper, 2004) that when cluster solutions tend to focus on the major themes to the detriment of more minor themes, documents discussing minor themes, such as our problematic aspects, can easily be separated, particularly if the documents vary in the breadth and nature of their semantic content. In other words, given that problematic cases tend to occur when the aspect of interest is small or has low salience in the local context, clustering may conceal the very features we are seeking to extract.

An alternative to approach to local context clustering is local context analysis (LCA: Xu and Croft, 2000). This approach is more discriminating than local feedback but does not rely on clustering. Like local feedback, the top ranks of the retrieved set are assumed to be generally relevant, but the algorithm judges terms based on the extent to which they co-occur, that is occur within the same contexts as the existing query terms. Hence, there is the potential to select a range of terms that are quite diffuse with respect to the query, rather than just terms that focus on the main feature of the local context (see Belkin et al., 2000).

However, in its original format as a query refinement device, term co-occurrences that are more common in the local context (compared to the global context) are also seen as more important. To avoid problems of multi-topicality in long documents, the procedure begins by separating the local context into passages. The algorithm then assigns weights to terms based on the extent to which they tend to co-occur in the same passages as each query term. Terms that co-occur with all query terms are seen as most important and terms that co-occur with only one query term are seen as least important. Additionally, terms that are relatively infrequent in the global (whole collection) context are also weighted higher. A full description of the procedure with metrics and formulae can be found in Xu and Croft (2000).

Recently, Belkin et al. (2000) compared LCA with the document relevance feedback method of suggesting query terms for the purpose of an interactive search task. Users were performing the classic TREC interactive task of seeking an instance of as many different aspects of the topic as possible. Hence, in line with our requirements, suggested terms needed to be diffuse in nature, rather than focused on the main theme of the topic.

In line with this expectation, they found that LCA suggested more unique terms and, in turn, LCA users selected more of these suggested terms for query expansion. However, a complaint from users was that many of the suggested terms were quite ambiguous, for example unusual proper nouns or numbers. In this thesis we believe this ambiguity is a necessary consequence of the goal of explaining topical diversity in the local context of a query but that such ambiguity can be resolved to a great extent by presenting terms in some sort of context. In section 6.4.2, we demonstrate how representing terms within the spatial-semantic context of the document visualization can, to a degree, alleviate single term ambiguity.

Despite these observations, overall Belkin et al. (2000) found that instance retrieval performance was roughly equal between groups using the two approaches. However, they concluded that LCA was better on balance because less cognitive effort was required from users (i.e., evaluating documents for relevance). The important result was that it showed that multiple, human document relevance judgements are not necessary to form a rich context of potentially useful query terms.

The general LCA approach is therefore a promising one. Unfortunately it fails as a solution for our problem, as does local feedback, in one key respect: it is a query expansion algorithm. In our problem, the user has not defined a short query to express their need; they have simply nominated a document text as an exemplar because somewhere, and to some extent, it describes an aspect of their information need. In other words, whilst the document is the query, more precisely it can be seen as a collection of terms that contains the intended query.

Hence, our goal is somewhat opposed to that of traditional term suggestion approaches in that we wish to narrow the query, rather than expand it. In other words, our aim is to identify possible queries based on the extant relationships between the exemplar and documents in the local context. It seemed that whilst the LCA approach could be used to provide a diffuse set of terms using a document as the query, the number of unique terms in the query would make it too computationally expensive for real-time interaction.

We have considered both QE techniques and document clustering. Neither adequately fulfils our needs. QE techniques fail because they require a query that is already quite specific and local context co-occurrence analysis of all document terms would be too computationally expensive. Local clustering is not a good option either, due to its parametric nature and bias towards preserving major features. In response to this we propose Local Context Distillation (LCD), a novel approach to term suggestion. LCD is similar in some ways to LCA in that it selects good terms by analysing the local document context of the query. However, our approach is far simpler and more efficient, but is still capable of producing a conceptually diverse set of terms. Further, the weighting function is biased, but not exclusively so, towards minor features that relate the exemplar to the local context. This satisfies the main requirement identified in our analysis in Chapter 5. We now describe the development and implementation of our term suggestion algorithm.

6.3. Local context distillation

To recap, our problematic exemplar cases tend to have the following characteristics:

1. The nearest aspect relevant neighbours are quite distal in term of rank similarity to the exemplar (many other documents are more similar to the exemplar) and possibly scattered quite widely within this distribution.

2. The aspect is represented by a relatively small sub-set of documents.

3. The exemplar is related to other aspects of the relevant topic and the proportion of documents discussing the aspect of interest within the global (retrieved set) context, compared to those discussing other aspects is relatively low.

The consequence of the first characteristic is that the optimal size of the local context is a moving target – in some instances, a context of the top 20 documents might capture several relevant documents; in other cases a context of 50 or more documents might be required to capture a good sample of same aspect documents.

In our solution, therefore, the user is able to dynamically adjust the size of the local context if none of the suggested terms adequately describe their intention. On selection of an exemplar, the local context size is set relatively low (top 10 documents). If the initial term suggestions are unhelpful this may be because the context is too small to capture a sample of relevant documents. To accommodate this possibility, the user is able to incrementally increase the context size and view the resulting impact of these increments on the term selections.

The combined consequence of all three characteristics is that, even assuming the local context to be analysed for terms is large enough to capture all of the relevant documents, the discriminating features that relate these documents to the exemplar may represent only a minor feature of the local context. Hence, relevant terms are likely to be suppressed by a large number of more salient, non-relevant relating terms.

To support problematic aspect cluster growing cases, we therefore need a function that allows for the selection of terms specific to minor relating concepts with the local context sample. The ability to also identify major relating concepts is also potentially useful, but less critical because such information is likely to be readily represented by spatial proximity and/or general similarity cues.

6.3.1. Term weighting function

We view this problem as one of distilling potential queries (distinct concepts) from the exemplar based on the characteristics of the local context (top k most similar documents to the exemplar). Our approach is therefore similar to those of local feedback and LCA where the top ranking documents are assumed to be a rich source of potentially useful new query terms. It is different in that we are not helping the user to expand their query, but to specify which aspect of the document exemplar best corresponds to their intentions.

We achieve this by selecting distinct features of the local context, placing an extra weighting on features that directly relate the exemplar to these documents. Hence, the most important terms are those that are both present in the exemplar and distinctive to the local context.

In developing our weighting function, we followed a standard premise of IR which is that for any given query there will be a set of optimal terms that effectively discriminate relevant from non-relevant documents within the collection (Salton and McGill, 1983). This leads us to form the strong assumption that:

> *For any given aspect of the relevant topic, in the retrieved set there will be at least one term that occurs in all the relevant documents and only in the relevant documents.*

Hence, we are looking to place a high weight on terms that are exclusive to the exemplar and closely related documents. Given the characteristics of our problematic exemplars, it is important that terms that are rare in the local context stand an equal if not better chance of being selected than those that dominate it. Given this we reasoned that an effective function might be one that simply measures how completely a term has been captured within the local context. Lundquist et al. (1997) found, in their experiments with local feedback, that the best weighting metric for selecting query expansion terms was one that considered the ratio of term frequency within the local context to its frequency in the global context. Specifically, they found the best terms were selected from a function that divided local document frequency by the log function of global document frequency.

As we wanted to emphasise the effect of terms that are globally rare (i.e., exclusive terms for small aspect sub-sets) we removed the log transformation on global document frequency to create the following simple function, F, which describes how the weighting for a given term is computed:

$$F(term) = \frac{df_{local}}{df_{global}}$$

Hence, df_{global} is the frequency of the term within the retrieved set, whilst df_{local} is the frequency of the term within the local context sample, including the exemplar. Our informal experiments found that removal of the log function did seem to produce intuitively better term suggestions, particularly in problematic exemplar cases. Remember that the user is able to adjust the size of the local context until they are satisfied with at least some of the suggested terms. As soon as the local context is large enough to comprise most if not all relevant documents then any term, following our strong assumption, which is exclusive to the all and only relevant documents, would be assigned a maximum weighting of 1.

Hence, even if only two other documents discuss the same aspect and they are both scattered and relatively distal to the exemplar, as soon as the context completely encapsulates them, any aspect exclusive terms will be assigned the maximum weight. Furthermore, smaller aspects are somewhat favoured because the fewer the number of aspect documents, the greater the impact each encapsulated document has on the importance of exclusive terms.

In an ideal situation, this assumption would hold for all aspects of the topic. However, this perfect situation is unlikely to be the case. Vocabulary mismatch is common between documents that discuss the same topics (Furnas et al., 1987). For operational purposes we therefore make the more relaxed assumption that:

> *For any given aspect there will be at least one key term that occurs in most relevant documents and only occurs in a small number of non-relevant documents.*

Even so, our rationale remains sound. The best terms will be those that tend to mostly occur within the local context of the exemplar, even if this is a relatively large sub-set of the whole context (e.g., in cases where nearest aspect neighbour are fairly distal). Terms that are least suited to describing specific relationships to the exemplar will be those that are scattered across the entire global context; those that do not discriminate the relevant aspect in any way. Our observations in Chapter 4, and those of others (Muresan and Harper, 2004; Wu et al., 2001), show that even complex relevant topics generally form a distinct sub-set of all documents, so even in problematic cases caused by an isolated exemplar or poorly-clustered aspect, if relating key words exist they should be identifiable from a local context that is significantly smaller than the entire context.

Note that F also does not include the weight of the term (TFIDF) within the exemplar or the local context. This is deliberate because we wanted to avoid inflating the weight of terms that have already contributed strongly to inter-document similarity measures. However, we found through informal experimentation that better results are obtained when F is moderated according to the presence or absence of the terms in the exemplar. In order to suppress the weight of terms that are absent from the exemplar, we divided F for these terms by a constant of 2, as this seemed to result in intuitively better term selections.

This approach is better than simply excluding terms that do not occur in the exemplar. It means that whilst terms that occur in the exemplar are more likely to be selected, terms that are exclusive to the local context by higher order association also stand a good chance of being selected. Such terms may represent useful substitutes for exemplar terms when the exemplar shares few terms with other relevant items, for example when the exemplar or reference to the aspect within the exemplar is very brief.

Hence, all terms are assigned a quantitative weight. For the demonstrations that follow we set an arbitrary threshold of the weight equal to the 15th mostly highly weighted term. This means that sometimes more than 15 terms are selected if there are a number of tied weights at the threshold. Having settled on our distillation term weighting function, we now describe its application within the visual context.

6.4. Applying local context cues to the interface

In this section, we present two novel tools that utilise LCD derived terms to support the aspect cluster growing strategy. We then present a series of walkthroughs that demonstrate how these tools might aid the user in locating same aspect documents in problematic exemplar situations.

The first tool, Concept Signposts, assigns each term to its best representative within the local context. The aim is to lead the user to the centre of the aspect cluster, assuming that it is well captured by the local context. The second, Concept Pulses, is an interactive tool that allows the user to gain an overview of the distribution of interesting terms, not only within the local context but also the whole context. To place these two tools in context, we first describe the prototype design of an interface that might accommodate them.

6.4.1. Implementation of prototype interface

The aim of this prototype was to implement a working interface that can integrate all of the main concepts that we have discussed over the course of this dissertation. These are the spatial-semantic visualization of the retrieved document set, simple 'more like this' relevance feedback, local context distillation, dynamic adjustment of local context size, and two visual tools that exploit LCD terms: Concept Signposts and Concept Pulses. In this section, we give a brief overview of how these concepts fit together within the interface.

As this is an early prototype, we have used the MS Visual Basic 6 programming environment for development. The use of a visual, high-level language has allowed ideas to be implemented, tested and refined quickly and simply. Whilst VB6 does not provide optimal performance for computationally demanding tasks (e.g., 3D rendering), the dynamic features of the interface are quite usable on our modestly specified development PC (Athlon XP1800+, 512MB RAM). The visualization was implemented as a virtual environment object, using the freely available WildTangent 3D API (http://www.wildtangent.com). Representing the visualization as a model within a virtual environment, particularly using the relatively high-level API provided by WildTangent, greatly simplified the management of visual elements, allowing simple control over a range of visual (translucency, animation) and interactive effects (e.g., zoom and pan) using the built in objects and methods. Figure 6.1 shows a paper landscape (Brath, 2003) of the interface at its current stage of development. There are four main elements: the visualization view; the document view; the local context view; and the aspect view.

In the visualization view, each document node is initially represented as a blue, translucent sphere. Nodes are mapped to the X and Y coordinates computed for the earlier analysis and node object size scaled and camera distance set accordingly so that that node overlap is minimised and the entire visualization is visible. Camera angle is orthogonal to the XY plane, looking along the Z-plane. For MST visualizations, it is possible to also show the retained links between nodes.

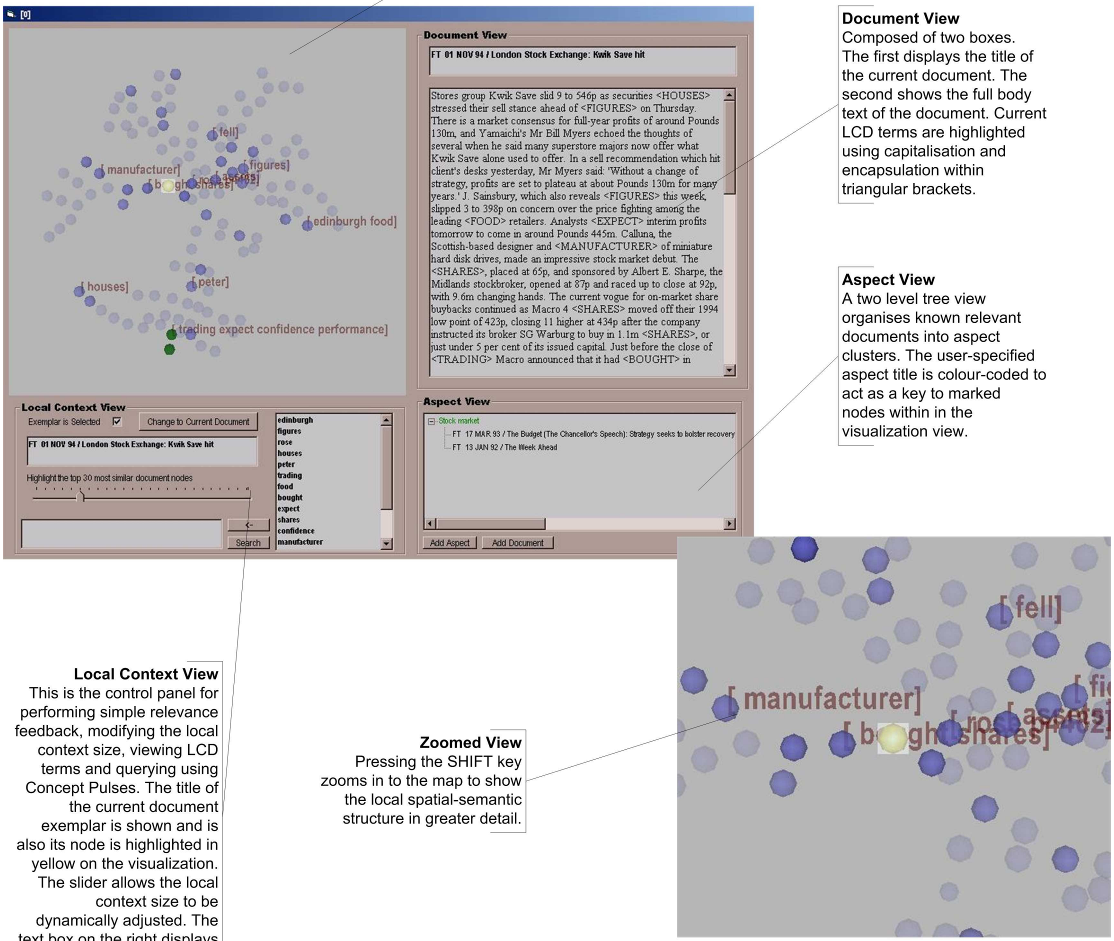
Prototype of a document visualization interface incorporating Local Context Distillation terms

Supporting multi-aspect discovery and retrieval within a persistent visual context

Presents an interactive spatial-semantic visualization of documents retrieved from a high recall query to guide topic orientation, topic exploration and aspect retrieval within a persistent visual context. The spatial-semantic structure is static, to aid formation of a mental model of the topic domain. Aspect retrieval is achieved by cluster growing from a known relevant exemplar. Spatial-semantic cues, where necessary, can be supplemented by transiently highlighting the most similar nodes to the current exemplar document and linking these nodes to key terms derived from Local Context Distillation (LCD).

Spatial-Semantic Visualization View
Documents are represented as translucent blue spheres (nodes). Concept Signpost label cues are shown in translucent red text.

Brushing a node causes the title to be displayed (as a tool tip and in the Form title). The node changes to dark blue to show selection.
Clicking a node causes the document text to be show in the Document View. The visual node becomes contained by a translucent box to show it as the current document in view.
Marking a node changes it to a different colour (e.g., green in this example), which can be varied to allow fast discrimination between topical aspects. The mid-blue nodes are those belonging to the local context, composed of the most similar documents to the current exemplar (30 in this example).



FT 01 NOV 94 / London Stock Exchange: Kwik Save hit

Stores group Kwik Save slid 9 to 546p as securities <HOUSES> stressed their sell stance ahead of <FIGURES> on Thursday. There is a market consensus for full-year profits of around Pounds 130m, and Yamaichi's Mr Bill Myers echoed the thoughts of several when he said many superstore majors now offer what Kwik Save alone used to offer. In a sell recommendation which hit client's desks yesterday, Mr Myers said: 'Without a change of strategy, profits are set to plateau at about Pounds 130m for many years.' J. Sainsbury, which also reveals <FIGURES> this week, slipped 3 to 398p on concern over the price fighting among the leading <FOOD> retailers. Analysts <EXPECT> interim profits tomorrow to come in around Pounds 445m. Calluna, the Scottish-based designer and <MANUFACTURER> of miniature hard disk drives, made an impressive stock market debut. The <SHARES>, placed at 65p, and sponsored by Albert E. Sharpe, the Midlands stockbroker, opened at 87p and raced up to close at 92p, with 9.6m changing hands. The current vogue for on-market share buybacks continued as Macro 4 <SHARES> moved off their 1994 low point of 423p, closing 11 higher at 434p after the company instructed its broker SG Warburg to buy in 1.1m <SHARES>, or just under 5 per cent of its issued capital. Just before the close of <TRADING> Macro announced that it had <BOUGHT> in

Stock market
- FT 17 MAR 93 / The Budget (The Chancellor's Speech): Strategy seeks to bolster recovery
- FT 13 JAN 92 / The Week Ahead

FT 01 NOV 94 / London Stock Exchange: Kwik Save hit

edinburgh
figures
rose
houses
peter
trading
food
bought
expect
shares
confidence
manufacturer

Document View
Composed of two boxes. The first displays the title of the current document. The second shows the full body text of the document. Current LCD terms are highlighted using capitalisation and encapsulation within triangular brackets.

Aspect View
A two level tree view organises known relevant documents into aspect clusters. The user-specified aspect title is colour-coded to act as a key to marked nodes within in the visualization view.

Local Context View
This is the control panel for performing simple relevance feedback, modifying the local context size, viewing LCD terms and querying using Concept Pulses. The title of the current document exemplar is shown and is also its node is highlighted in yellow on the visualization. The slider allows the local context size to be dynamically adjusted. The text box on the right displays the top 15 LCD terms

Zoomed View
Pressing the SHIFT key zooms in to the map to show the local spatial-semantic structure in greater detail.

Figure 6.1: Paper Landscape (Brath, 2003) of the prototype interface integrating spatial-semantic visualization view, document view, local context view and aspect view.

Although the default node colour is blue, this can be varied to show different states and properties in response to user interaction. Each node can be clicked on to view the document contents in the document view frame. The currently viewed document node is then encapsulated within a translucent white cube. If a node is selected as an exemplar, it turns opaque and yellow. In turn, all local context nodes become more opaque (appearing darker) and nodes that fall outside of the current local context become suppressed by turning more transparent (appearing lighter). The opacity of a local context node varies as a logarithmic function of the rank similarity. This is a subtle effect that only becomes noticeable as the local context size becomes quite large, and is simply intended to help the user differentiate between strongly- and weakly-similar documents. As we shall discuss later in this sub-section, the colour of nodes can also be changed to represent aspect membership of known relevant items. The text labels in the visualization are Concept Signposts, which we introduce in the next sub-section. The user can also zoom and pan within the visualization, which can be useful when exploring dense regions of nodes and Signposts.

The document view simply shows the title and text of the currently selected document (bounded by the white cube). Within the text, all occurrences of LCD terms are capitalised and bounded by triangular brackets in order to facilitate scan-browsing within longer documents.

The aspect view is similar to the aspect windows system presented by Swan and Allan (1998). Its purpose is to help the user keep track of their search progress by showing documents marked as relevant and to discriminate between these documents by their aspect. This view links to the visualization using colours. Each aspect is headed by a distinct colour, which is used to colour the nodes marked as relevant for that aspect within the visualization. In figure 6.1, one aspect is recorded and corresponding marked nodes are shown in green within the visualization. Currently, although documents can be assigned to multiple aspects in the aspect view, each document node can only be assigned one colour in the visualization. There are many possible solutions to this problem. For instance, Allan et al., (2001) use a pie-slice metaphor to visually encode multiple aspect membership into document nodes. Another possibility is to cycle node colours between aspects or even to only show node-aspect membership on demand. However, this problem has not been the focus of our evaluation and must be left for future work.

Finally, the local context view frame handles user-system interaction relating to simple relevance feedback and LCD. The user nominates the current document as an exemplar by clicking on the button at the top of the frame. On doing so, the visualization immediately updates by emphasising the *k* most similar document nodes in the visualization and suppressing all other nodes. The size of the local context, *k*, can be dynamically adjusted by the user using the slider bar. In figure 6.1, *k* is set to 30 documents. The list box to the right of the frame is where the current local context terms are presented. Double clicking on individual terms initiates a Concept Pulse whereby nodes of documents that contain the term are animated within the visualization. We discuss the details of this tool in sub-section 6.4.3. Multiple terms can also be selected and pasted into the query string box at the bottom of the frame. The matches of the query string are also represented within the visualization as Concept Pulses.

Having provided an overview of the main interface elements, we now introduce the two novel visual tools that integrate LCD terms into the visualization in order to support the user during non-trivial aspect cluster growing episodes.

6.4.2. Concept Signposts

The aim of Concept Signposts is to express to the user how the current LCD terms relate to spatial-semantic features associated with the local context documents (i.e., highlighted features within the visualization). More specifically, in situations where the exemplar is isolated from a more coherent cluster of relevant documents, Signposts provide cues to guide the user towards this feature. Even if all relevant documents are fragmented, a discriminating aspect term stands a good chance of being attached to another relevant example.

This is achieved by showing the LCD terms 'in context' by assigning each term to its best representative within the local context. Currently, the best representative is simply the document that has the highest weight (TFIDF) for a given term. In most cases, terms tend to be distributed relatively evenly across the distinct local context features and so serve to emphasise different reasons for exemplar similarity.

Figure 6.2 shows an example from the Extinction scenario where the local context is set to 40 documents. Local context nodes are highlighted in dark blue. We can see that key terms are spread quite broadly across the spatial-semantic structure of the local context, indicating that LCD has identified a reasonably diverse range of concepts. Most of the sub-clusters of dark blue nodes are reasonably proximal to a signpost string although the nature of some of the more fragmented peripheral nodes is somewhat ambiguous. The exemplar document is entitled:

How we saved the rhino with rifle and chainsaw: Elizabeth Robinson watches a desperate attempt to beat the poachers

The article is about a project to save Rhinos from poachers by safely removing the prize that they seek – the animal's horn. The project is based in Zimbabwe but was inspired by a similar project in Namibia. We can see that LCD has identified major exemplar key terms (e.g., Rhino, horn, Zimbabwe, poachers) that relate it to the local context documents. It has also identified relatively minor concept terms such as "Safari" and "Mozambique" that are both only mentioned once in what is a relatively long document (1307 words).

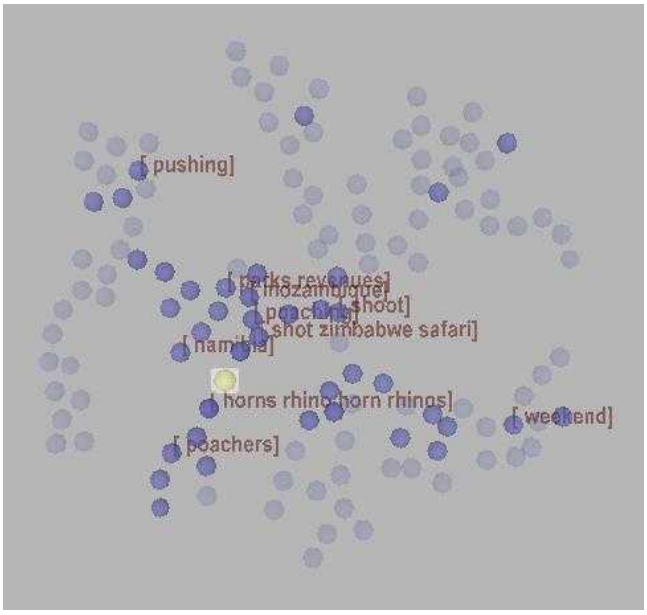


Figure 6.2: Concept Signposts example within the MST visualization of the Extinction scenario.

This strategy has a useful corollary whereby terms that are closely associated tend to be assigned to the same, or highly proximal documents within the visualization. Hence, terms that may, by themselves, be ambiguous tend to disambiguate each other through their relative proximity. This 'magical' term clustering emerges from the inherent semantic structure of the visualization and the fact that documents that focus on the same concepts tend to cluster coherently. In figure 6.2, for instance, the terms "rhino" and "horn" are assigned to the same node whilst the terms "poaching" and "shoot" occur proximally.

The Concept Signposts strategy differs from typical approaches to local context labelling in spatial-semantic visualizations in that most, if not all, terms will be either directly or indirectly related to a specific document, the selected exemplar, rather than simply context-free representatives of the major features within the current structural view (e.g., see Horbaek and Froekjaer, 1999).

One current limitation with Concept Signposts is that each term can only appear once in the visualization. This may be a problem if it is a highly key term within two distinct clusters of the local context. Clearly, it is not feasible to attach each term to all the documents in which it occurs due to the clutter and overlap this would create. Possible strategies might include assigning a term twice if the 1st and 2nd best representatives were relatively distal in the visualization, particularly if the 2nd representative did not yet possess any Concept Signpost string.

On a related note, even with the 'one term one assignation' strategy, overlap between signpost strings can cause legibility problems, particularly when vertically adjacent nodes both have Concept Signpost strings. Strategies we have tried to combat this have included reducing the size of the font (although this causes readability problems at the overview level) and rotating the text object on its x-axis, which gives an effect akin to writing the text on a roller and spinning it along its long axis. The text object is one-sided, so it becomes invisible for half of the rotation phase. If all Signpost strings rotate at the same speed, but at different phases, then they no longer obscure each other.

6.4.3. Concept Pulses

Concept Pulses are intended to complement Concept Signposts, although user interaction and system response are quite different. Concept Signposts appear in direct response to simple relevance feedback and provide the user with a general overview of emergent feature characteristics. Concept Pulses, on the other hand, allow the user to engage in a form of *ad hoc* dynamic querying (Williamson and Shneiderman, 1992). By selecting the most indicative terms, the user is provided with immediate visual feedback that shows the distribution of those terms within the spatial-semantic visualization and their relative salience within matching documents.

The rationale for Concept Pulses stems from an inherent limitation of the Concept Signposts method: that each term can only appear in one location. This limitation is inherent because multiple presentation of terms leads to excessively long label strings for specific nodes that tend to overlap and thus obscure each other. Also as labels become very long (more than four average length terms) it becomes difficult to associate the contextual origin of the tail end terms.

Given this inherent limitation of Concept Signposts, it is likely that in many cases a useful term will not be applied to a document that is relevant to the current aspect of interest, even though the term might be clearly discriminating and present within such documents. Concept Pulses directly address the limitation by allowing the user to see immediately which documents contain the selected term.

In their most simple usage, the user can select (by double clicking) any distilled term in the LCD term list. The system response to a pulse request is to rapidly inflate each document node to a size proportional to the weight of the selected term in the document. Hence, documents that discuss the term most frequently will become the largest nodes in the visualization (see figure 6.3). This is a dynamic animation within the visualization, where nodes rapidly inflate to a size proportional to their query match and slowly deflate again to their normal size.

Concept Pulses provide three types of information about the term's usage: how often it occurs across the set (its document frequency); where it occurs and particularly where it seems to be a relating feature of a document cluster; and in which document(s) it is most salient (heavily used). Nodes deflate at a constant rate, hence those that were inflated the most will remain over-sized for a relatively longer period than other nodes. This further aids the user in identifying the most salient documents and clusters, as these will be the last nodes to return to standard size.

Figure 6.3 shows an example within the Extinction visualization where the term 'forest' has been pulsed. The local context is the top 20 most similar documents. We can see in this case that most of the over-sized nodes are close neighbours (dark blue), however Concept Pulses also affect non-local nodes as can be see by the pair of oversized light blue nodes in the bottom left of the visualization. The exemplar document is in the middle of a cluster situated in the top left of the visualization. We can immediately see the importance of using translucency with nodes. The fact that inflated nodes encroach on each other's space does not affect the user's ability to discriminate between nodes, even when one completely occludes another.

Looking first at the local context nodes, we can see three documents within this cluster that are good representatives of this term, one of which is clearly the strongest representative in the set. However, we can also see two other distinct clusters of documents that clearly talk about forests, one of which is almost central to the visualization (four documents) and the other which is situated just a little further below by the 'logging' signpost (two documents). Closer inspection reveals that these three forest features are quite distinct in nature. The top left (most local) cluster discusses the argument for the preservation of temperate forests (as well as the traditionally popular tropical forests) due to environmental concerns, for example a common theme in this cluster is the reduction in numbers of the spotted owl in northwest American forests. The central cluster mostly discusses the forests of Africa, in particular their regeneration. The bottom cluster discusses the arguments of environmental groups (e.g., Greenpeace and WWF) for preserving forests, particularly the tropical forests.

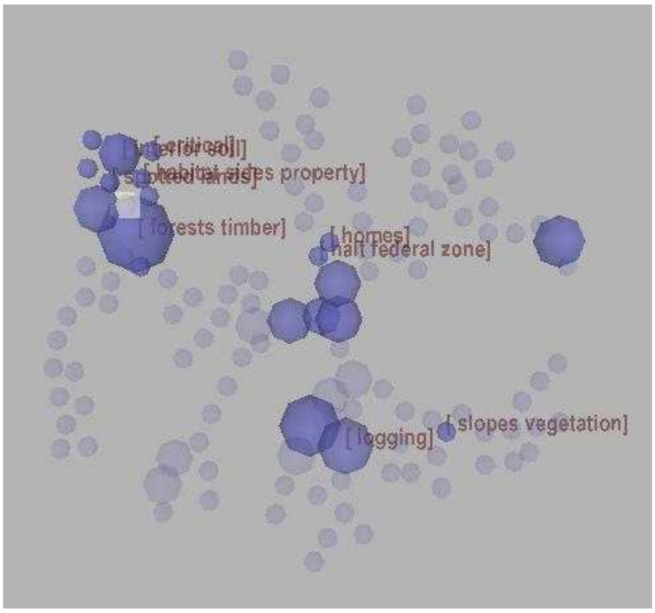


Figure 6.3: Concept Pulse for the term 'forest' within the MST visualization of the Extinction scenario.

Also notable are two clustered light blue (non-local) nodes in the bottom left of the visualization. These are about illegal trade in specific forest dwelling animals such as tigers in India and parrots in Paraguay. This shows another benefit of Concept Pulses, which is to highlight the occurrence of key concepts occurring in documents that are currently outside the specified local context.

Concept Pulses also allow the user to rapidly build a query, by selecting a number of distilled terms from the Listbox and then clicking a search button. The search routine calculates a match function whereby the weight of each term that occurs within a document has a cumulative effect on its pulsed size. Hence, the largest nodes will tend to be those that discuss more terms, but may also be nodes that discuss a small proportion to a great extent.

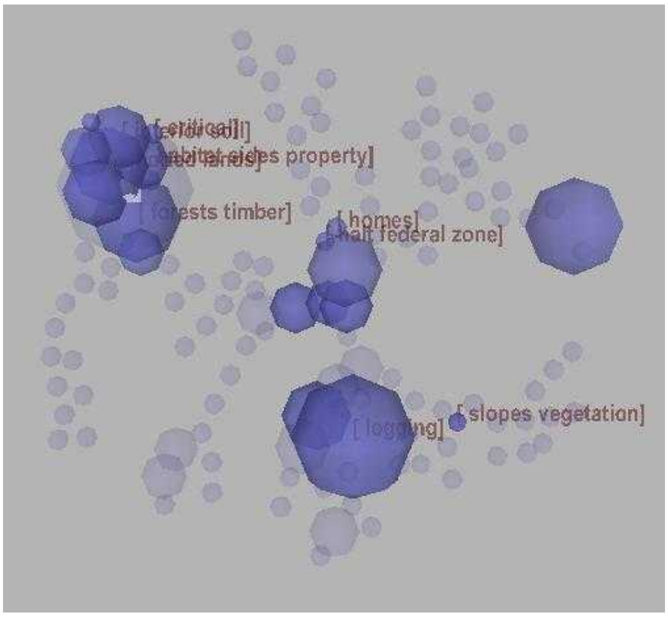


Figure 6.4: Concept Pulse for the terms "forest" and "logging" within the MST visualization of the Extinction scenario.

Figure 6.4 shows the effect of adding the term 'logging' to the simple 'forest' query shown above. This produces some interesting effects. The importance of the top left cluster, which surrounds the exemplar, is emphasised further. This region mainly discusses the problems caused by the timber trade in temperate regions, particularly the United States. The bottom cluster of two documents has also become more salient, as predicted by the 'logging' signpost. A particularly interesting effect is considerable increase in the salience of the isolated node on the far right. Closer examination reveals this node discusses a distinct aspect of the query: the tensions between the importance of timber trade in the greater Soviet economy and the impact this has on local communities and wildlife habitats.

6.5. Using LCD terms to support aspect cluster growing

In this section, we select four of the problematic aspect cluster growing cases from the sample that came to our attention in Chapter 5. We begin with an example of how Signposts alone can quickly support orientation in an aspect cluster growing situation. We follow this with another, more complex example, where both signposts and single keyword pulses are used together to solve the problem. Our third example demonstrates the value of combining LCD terms to support aspect cluster growing. Finally, we present an example where LCD terms fail, both in their signpost and pulse application. However, we resolve this by showing how the pulse principle can be extended to allow query by phrase or passage.

In each case, the exemplar was problematic for both first and second NANs; in other words these are extreme cases where the exemplar is likely to be isolated from any main aspect cluster or sub-clusters.

6.5.1. Discriminating two distinct exemplar themes

Our first example case is one where the exemplar document is clearly split between two distinct themes. It is taken from our Extinction scenario, document #3, and is shown in full in appendix C.1. The first half of the document discusses the impact of a ruling by the convention on international trade in endangered species (CITES) to protect the elephant through controls on the ivory trade. This is a clearly relevant aspect of the Extinction topic (aspect 19) and the document is associated with two other documents based on this aspect, documents #59 and #14. The second half of the document, however, is a large table detailing international balance of payment figures, and is clearly non-relevant. The general economic theme seems to have diluted the importance of the environmental theme of the document and its relationship to two other documents. In similarity space, the nearest aspect relevant document is ranked 9th in the order of most similar documents and the second nearest is ranked 34th.

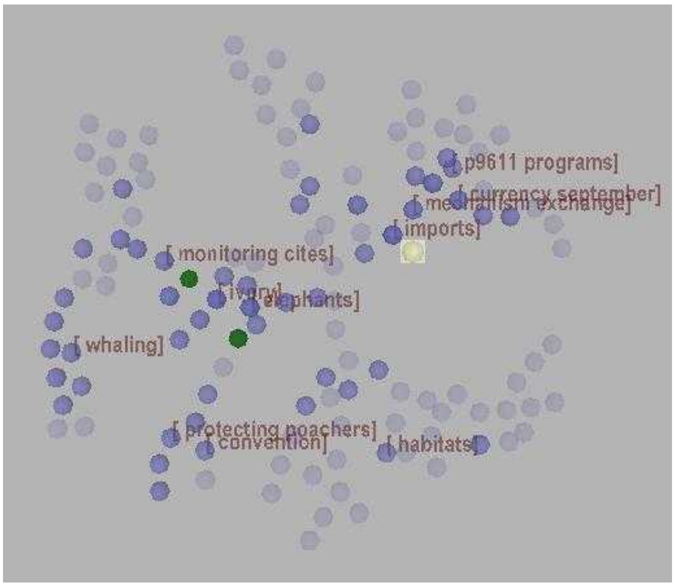


Figure 6.5: Concept Signposts for document #3 of the Extinction scenario. The user has selected this document (marked in yellow and surround by a translucent white box) as relevant because it discusses the efforts of CITEs in the protection of Elephants (aspect 19). Relevant nodes are marked in green. The exemplar document is split between discussions of endangered species and general economics news (see Appendix C.1). Signposts clearly show that document #3 has been located near to documents about the latter topic and that the main cluster for CITEs and elephants resides over to the left-side (aspect relevant documents highlighted in Green).

The effect is even more apparent when we view the distribution of these three documents in MST space. The two other relevant documents are situated some way to the left of our exemplar (see figure 6.5) and have been highlighted in green.

Selecting a local context size of 50 documents reveals a number of distinct clusters within the visualization. The application of signposts immediately explains the two themes of the exemplar and provides clear cues to the user as to where to focus their attentions (see Figure 6.5). The proximal cluster of the local context (dark blue nodes) is clearly about economic matters, whilst terms about environmental concepts dominate the local context clusters to the left, which comprise the two relevant documents. One of the related documents (document #59) is immediately proximal to the node that has the 'CITES' signpost label and the other document is within two nodes distance of the 'Elephants' term.

6.5.2. Using Concept Signposts and Concept Pulses in combination

In this example we demonstrate the complementary value of Concept Signposts and Pulses. We have selected an exemplar of aspect 9, document #31 (see appendix C.2), from the Extinction scenario, which again proved to be problematic for the aspect cluster growing strategy, even when similarity cues were followed. The local context size is 50 documents. Following spatial-semantic cues, the first relevant document is ranked 9th in proximity to the exemplar and the second is ranked 13th. Following similarity cues, the first and second relevant documents are ranked 9th and 34th respectively. In other words, strategy performance is poor using spatial-semantic cues but actually better than using similarity cues.

Aspect 9 has the definition "Zimbabwe, Rhino, Elephants". Remember, the Extinction question is to identify the efforts made by as many different countries as possible to protect endangered species. Hence, the user will be primarily searching for documents that discuss Zimbabwe. There are two other relevant documents for this aspect, documents #116 and #96. Both the exemplar and #96 discuss the country's efforts to preserve elephants, whilst #116 focuses on the Rhinoceros.

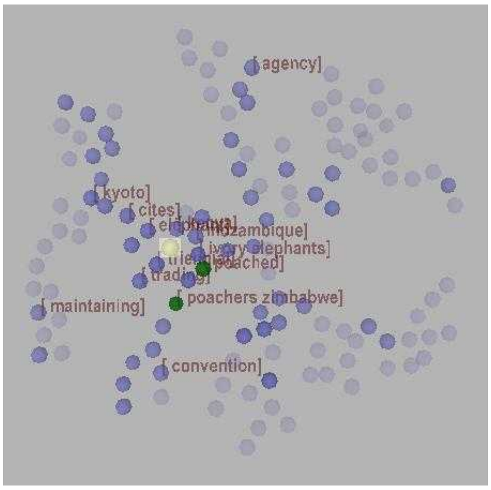


Figure 6.6a: Concept Signposts for Extinction document #31, which is an exemplar for aspect 9 (Zimbabwe, Rhino, Elephant). Local context size is 50 documents. Relevant nodes are marked in green

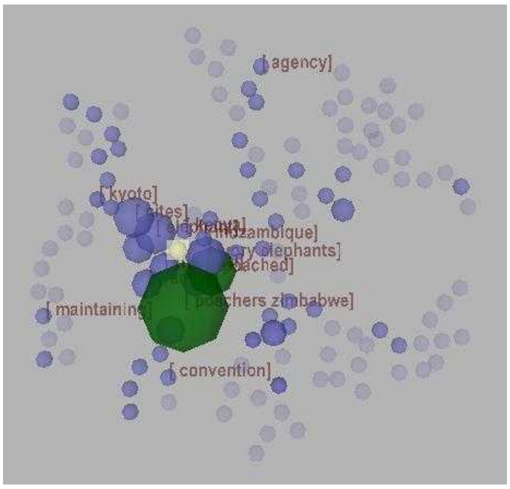


Figure 6.6b: Concept Pulse for Extinction scenario document #31 using the LCD selected term "Zimbabwe".

Looking at the left-hand screenshot in figure 6.6a we can see that LCD has selected Zimbabwe as a key term, along with "elephant(s)". Other related terms include "ivory", "poached", "poachers", "CITES", which is a major international conference that discusses policy on animal trade, and "trading". Note, however, that "rhino" is not present, most probably because it has been suppressed by the weighting procedure, as it is not an exemplar term.

We can see that the term "Zimbabwe" is associated with one of the relevant documents (#116). As a primary key term, this should allow the user to identify this document immediately. However, document #96 is less easy to find using Signposts alone. It is the representative of "poached" and is proximal to the representative of "ivory" and "elephants", but this area of the visualization has a dense concentration of documents discussing elephants and the ivory trade in a number of African countries and there is no clue here that document #96 may discuss Zimbabwe's role in Elephant preservation.

By pulsing "Zimbabwe" the visualization reveals that #96 is the second best (second largest node) representative of this term (Figure 6.6b). Concept Pulsing, in this case, therefore provides a strong cue that potentially allows the user to locate both relevant documents within two viewings, a maximum precision of 1.

6.5.3. Pulsing multiple terms

So far, we have demonstrated the successful usage of both Concept Signposts and Concept Pulses for facilitating aspect cluster growing in problematic cases. Both these examples, however, have focused on the Extinction scenario where aspect definitions are quite distinct in nature. In most cases aspects are distinguished clearly by the nation or organisation of interest and in many cases the species of interest. In the Chunnel scenario, however, aspect definitions are less distinct and often somewhat broad in definition (see appendix A.2), which accounts for both the larger size and overlapping nature of the aspect sub-sets. For instance, aspect 11 is somewhat diverse in definition as relevant documents can talk about both improvements and harm to local economies caused by the new rail link. Furthermore, there are a number of closely related sub-topics, such as aspect 13 ("Changes in Kent economy/employment") and aspect 7 ("Changes in real estate market"), that one would expect to be, and indeed are discussed regularly discussed in the same documents.

Hence, we were expecting the Chunnel scenario to be a more challenging test of our solution. The aforementioned aspect 11 is a good example, even though it is a relatively well-represented aspect (18 documents). Document #197 proved to be a poor cluster growing exemplar for this aspect, with the 1st and 2nd nearest neighbours within the visualization ranked, respectively, 9th and 18th most proximal. A key problem is that the exemplar is relevant to several aspects of the topic, including the closely related aspect 7 ("Changes in real estate market"). It focuses primarily, however, on aspect 1 which is about environmental impacts of the Chunnel. The text of the document #197 can be found in appendix C.3.

The relevance of this document to aspect 11 is principally due to a brief reference to the local regeneration and a new shopping centre development in Stratford, East London, which has grown up around the new rail line. However, this is simply a lead-in mechanism to the primary topic – the negative impact on the Kent countryside. As such most of the relevant nearest neighbours are about this aspect.

Figure 6.7 shows the location of the exemplar (node marked in yellow surrounded by a translucent white box) and the distribution of the other relevant documents (marked in green) within the MST visualization. The local context size is 50 documents. LCD has identified "Stratford" as a key term, but there are no terms that clearly relate to regeneration or commercial developments. Signposts has attached the term "Stratford" to a document just above the exemplar (see figure 6.7). Whilst this is not relevant to either that aspect or the topic generally, we can see that if the user continued in this direction to the next node, they would find another aspect relevant document. Unfortunately, no further relevant documents are located in this region. Seven of the remaining 16 relevant documents are located in a dog-leg shaped formation that begins just below the exemplar and stretches downwards and out towards the left side of the visualization. Another dense cluster of five documents occurs further down around the "mile" Signpost (see figure 6.7).

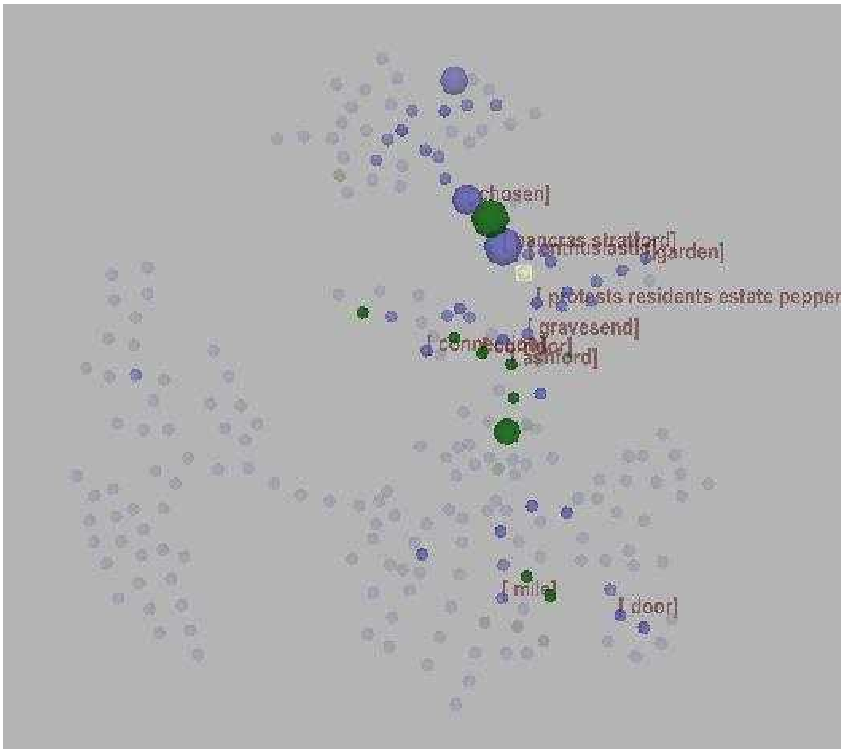


Figure 6.7a: Concept Pulse using the term "stratford", for a user interested in Chunnel aspect 11. This term has been selected by LCD based on the exemplar document #75 and a local context size of 50 documents. Relevant nodes are marked in green.

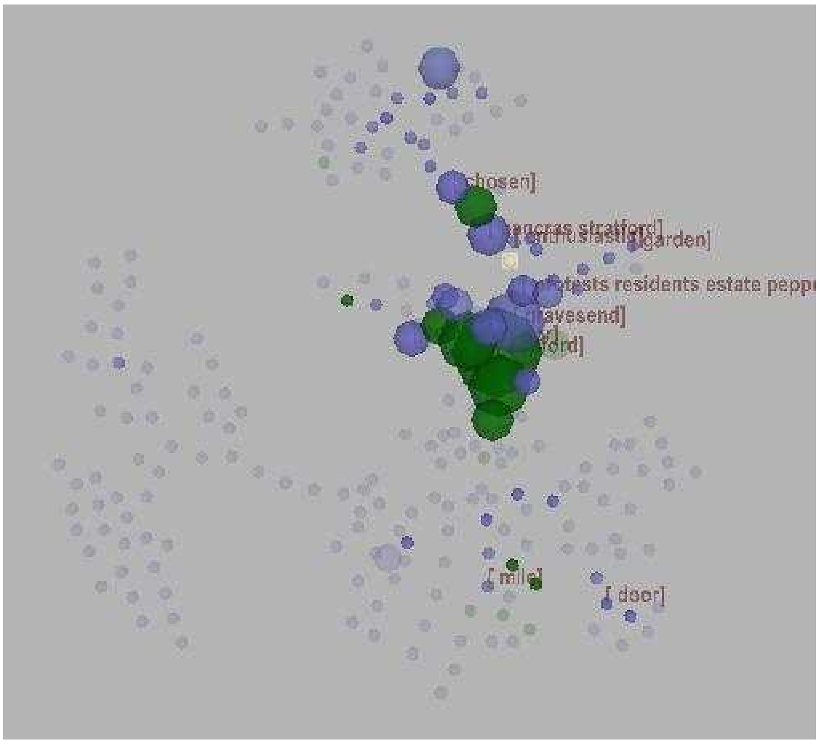


Figure 6.7b: Multi-term Concept Pulse, for a user interested in Chunnel aspect 11, applying the terms "ashford", "stratford" and "gravesend" in combination.

Let us assume that the user decides to pulse “Stratford” to see if any more information is available on the commercial redevelopment of Stratford or neighbouring areas. Figure 6.7a shows that the most salient nodes are situated around the Concept Signpost representative for this term. With the exception of the one relevant document already noted, documents in this area are focused more on aspects relating to the construction of the rail-link and its environmental impact. We can see one further relevant document, however, situated at the bottom end of the main dogleg feature. Examination of this document reveals that economic growth is expected, not just in London but also all along the proposed route, which will run through the county of Kent. In particular, Ashford is mentioned as an area of expected high growth. “Ashford” is already in the LCD term list. Additionally, the user might now note that “Gravesend”, another Kent town, is also mentioned in the LCD term list. The user might therefore consider it worth expanding the query to include the names of towns situated on or around the rail-link.

Concept Pulses allow the user to select multiple LCD terms and assign a cumulative visual weighting to each node. Figure 6.7b shows the result when the query “ashford stratford gravesend” is pulsed. We can see that a dense cluster of nodes, that encapsulates the dogleg formation of relevant nodes, becomes the most salient region in visualization. The two largest, unseen nodes in this region are both relevant to aspect 11.

6.5.4. Pulsing a selected passage

Our experimentation with the system revealed that many of the problematic cases in our three scenarios could be adequately resolved using single, and particularly multi-term Concept Pulses that were formulated using LCD selected terms. However, there are several cases where LCD failed to identify sufficiently discriminating terms.

One such example is the use of document #75 as an exemplar of aspect 2 of the Chunnel 218 scenario. This aspect focuses on how the high-speed rail line, from London to the Chunnel, was financed (see appendix A.2), with most relevant documents focusing on the relative contributions of public and private finance.

Document #75 is a possible exemplar of this aspect, but is located somewhat distally from the main cluster near the top of the visualization (see figure 6.8a). Whilst it makes several brief references to the rail link plans, it differs from the majority of documents judged relevant in that it focuses mainly on the proposed construction of a station complex near Dartford by a private investor called Blue Circle, that will support the line.

Taking a local context size of 50 documents, as with previous examples, is sufficient to capture eight out of the 10 other documents relevant to aspect 2. However, whilst LCD selects distinctive terms such as “Blue”, “Circle” and “Dartford”, the only identified key term that is relevant to the aspect as defined by the stimulus extract shown above is “Financed”. Figure 6.8a illustrates the problems faced by the user trying to locate the main cluster from document #75 using spatial-semantic or Concept Signpost cues. Nine out of a total of 11 (including the exemplar) relevant documents (green nodes) are organised into a dense, roughly T-shaped bunch of nodes at the top of the MST visualization. However, the exemplar (yellow node) is completely isolated and distal from this cluster and the user would need to view 88 non-relevant documents before finding the nearest relevant neighbour if spatial-semantic cues alone were followed. It is encouraging that the Concept Signpost for the LCD term “Financed” is located near to the main cluster, but unfortunately the Signposted document itself is not relevant.

Pulsing using the term “Financed” produces a more positive result (see figure 6.8b). Although the top representative of this term is not relevant, the next largest node is adjacent and represents a relevant document. There are two other slightly smaller nodes in the vicinity, one of which is also relevant. However, there are no clues to alert the user to the rich patch of relevant documents situated to the right of these nodes.

We asked why LCD might fail to select appropriate key terms, even in a situation like this where most of the aspect sub-set has been captured by the local context. We conjectured that this problem might stem, in some cases at least, from the fact that LCD focuses on individual word terms and takes no account of the contextual co-occurrence of terms within documents. In cases like this one, it is combinations of terms, rather than individual words that seem to best describe the key concepts. For instance, where this exemplar document makes relevant references to the aspect, phrases like “Union Railways”, “rail link”, “high-speed” and “Pounds 2.5bn” occur that are common within and reasonably exclusive, to the other relevant documents in the main aspect cluster. However, by themselves, the component words of these phrases are likely to occur broadly across the global context of the set, so LCD does not consider them important.

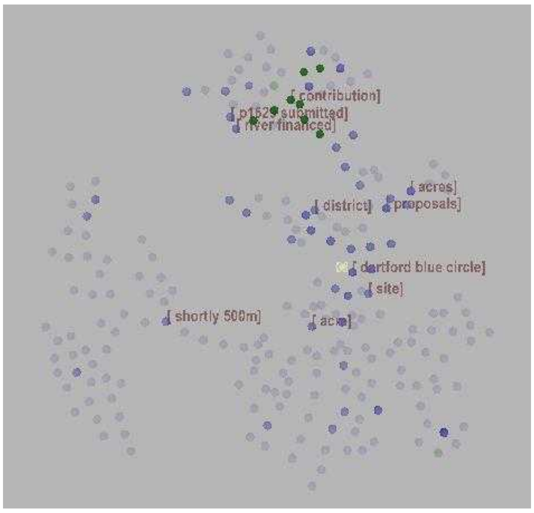


Figure 6.8a: Local context of document #75 from the Chunnel Scenario, which is an exemplar of aspect 2 (Financing of high-speed rail line). Relevant nodes are marked in green.

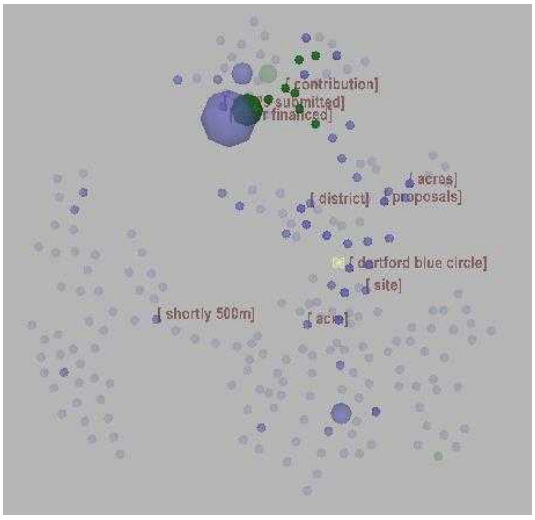


Figure 6.8b: Concept Pulse from the term "Financed"

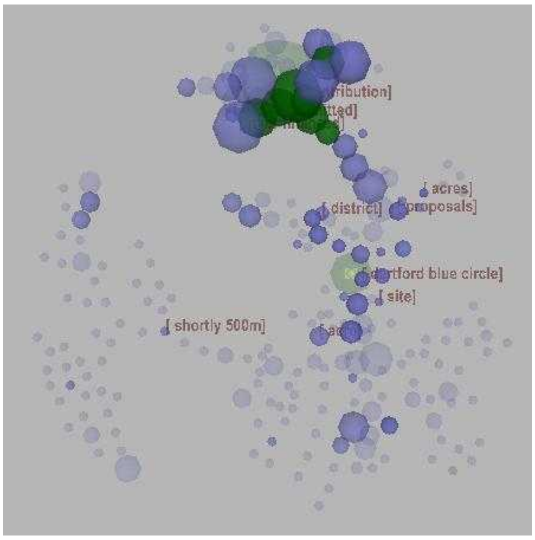


Figure 6.8c: Concept Pulse using stimulus passage selection

We reasoned that it is therefore necessary to consider the sum of several terms that tend to occur together, either as phrases, or nearby (e.g., within the same sentence) and that define a good aspect query, even within the relatively constrained context of a retrieved set. Redesigning LCD to identify phrases creates a non-trivial set of problems such as whether to build phrases dynamically, so that they are specific to the local context, or to modify the global text analysis procedure to include phrase terms. The former solution would incur considerable computational overhead, which is likely to reduce the responsiveness of the LCD procedure significantly. Likewise, phrase identification would also slow down the initial semantic modelling process and the increased vocabulary size would cause proportional increases in computation time for both inter-document similarity analysis and LCD term weighting.

Given this, we decided to trial a simpler solution to the problem: query by passage selection. This simple extension to the Concept Pulse tool provides a neat solution to the problem of a poor LCD response, by allowing the user to directly indicate the stimulus for their interest from within the document text itself. This strategy is similar to that supported by the TELLTALE (Pearce and Miller, 1997) and VOIR (Golovchinsky, 1997) dynamic hypertext systems. In our system, the user is able to select the relevant passage and submit it as a query. The system parses the string and extracts all terms that occur in the vocabulary of the semantic model (the common term space). This is then passed to the Concept Pulse routine, which provides visual feedback in the regular way.

For instance, the first and most notable reference to the rail link in document #75 is the following passage:

> *"The land is on the route of the Pounds 2.5bn rail link, to be financed jointly by the private and public sectors, which was announced by the government earlier this week."*

This passage is highlighted in bold in appendix C.4. If we pass this string to Concept Pulses, the following terms are extracted:

land route pounds rail financed jointly private public announced government earlier

Figure 6.8c shows the visual array resulting from the Concept Pulse. The relevant T-feature is clearly exaggerated and encapsulated within a dense region of significantly inflated nodes. Close inspection reveals that the largest node, which is relevant, actually falls outside of the local context, as indicated by its high transparency (light shading). The next largest green node is equally proportioned to the largest blue, non-relevant node. In total, eight out of a possible 10 relevant nodes are significantly inflated. Hence, viewing the cumulative effects of several marginally relevant terms within a clearly relevant passage can produce a useful query.

6.6. Discussion and conclusion

In this Chapter, we reported the development of an approach to supporting aspect cluster growing in the kind of problematic situations identified in Chapter 4. Problematic situations are defined as those where the exemplar is isolated from same aspect documents in both spatial-semantic and high-dimensional vector space. Core to our solution is the concept of Local Context Distillation. The LCD algorithm aims to identify potential query terms in response to the nomination of an aspect exemplar document and to suggest these to the user. Whilst previous work in the area of term suggestion (e.g., Attar and Fraenkel, 1977; Koenemann and Belkin , 1996; Xu and Croft, 2000) has focused on expanding an existing user-specified query, our problem is different as the aim is to narrow the query, to distil from the intended query from a single nominated document exemplar. Our solution is an algorithm that looks for terms that are exclusive to documents occurring in the local context document sample, placing a higher weighting on terms that occur in the exemplar itself. The user can manipulate the size of this sample until the optimal set of key terms is presented.

We then introduced a prototype interface that integrates the LCD with a spatial-semantic view of the retrieved set. This interface also incorporates two novel visual tools that apply the terms suggested by LCD to the visualization context, providing additional cues to support the process of aspect cluster growing. Concept Signposts use LCD terms to augment the spatial-semantic visualization. Each term is applied as a label to the best document representative within the local context. A useful consequence of applying terms to the spatial-semantic context is that related terms tend to congregate, providing further disambiguation of their meaning and reinforcing the description of salient document clusters within the local context.

Concept Pulses provide the user with an alternative strategy, which supports search when the key LCD terms have multiple senses within the local context or when there is no coherent main cluster of aspect relevant documents. On selection of one or more LCD terms, the system responds by rapidly inflating each document node within the visualization to a size proportional to the importance of the selected key terms within that document. Nodes gradually return to normal size over a period of a few seconds. The visual array and flow effects created by this animation support search by allowing the user to quickly identify the best matches and dense regions of good matches within the stable and familiar context of the spatial-semantic overview.

In the final part of this Chapter, we presented four examples of how these tools are able to facilitate aspect cluster growing in problematic cases. Our demonstrations show that LCD works best with the Extinction scenario. This is likely to be due to the more distinct and concrete nature of the aspects in this topic. Aspects are clearly distinguishable by the particular country or organisation discussed by relevant documents and, in many cases, the species of interest. LCD was less successful in the Chunnel scenario, where aspects are more broadly defined and closely related.

From our informal trials it was evident that, for many aspects in this scenario, useful terms were more likely to be phrases or other non-contiguous combinations of terms that by themselves are ambiguous words like adjectives that tend to co-occur within relevant passages as opposed to single unambiguous keywords. Such words are not likely to be selected by LCD in its current implementation because, by themselves, they are not good discriminators as they are commonly used in a range of different contexts. Only when considered together do they become important query terms.

We discussed the potential benefits of adapting LCD to identify key phrases in addition to single word terms and concluded that whilst identifying useful LCD phrases on an ad hoc basis is likely to prove a difficult problem to solve in an efficient manner, it is still an interesting avenue for future research. Existing approaches to phrase identification, such as lexical (term) co-occurrence (e.g., Xu and Croft, 2000; Lund and Burgess, 1996) or identification of noun-compounds (Anick and Vaithyanathan, 1997) are likely to be too computationally expensive for use in a real-time, interactive system. LCA, discussed earlier in section 6.2, has shown that term co-occurrence analysis can be computationally feasible when terms are restricted to a local vocabulary and comparisons only need to be made between a small number of query terms and the local context vocabulary. However, in our problem the query is very long and mostly redundant (see section 6.2). The computation time would therefore be significantly increased for long exemplar cases or when a large context size is required to capture the key phrases.

Suffix tree clustering (STC) is an interesting approach that is worthy of further investigation (Zamir and Etzioni, 1998). In this approach, documents are grouped into 'base' clusters based on a shared contiguous sequence of terms. Base (single phrase) clusters are then combined to produce larger clusters. This has been show to be a fast procedure for dynamic document clustering. Potentially this procedure could be applied to identify key phrase strings within an exemplar, based on their co-occurrence within the local context. Another benefit is that it a suffix tree can be built incrementally, which means that if necessary it can be stopped mid-way once a sufficient set of good phrases have appeared, or simply extended if the size of the context is increased.

Whilst adopting phrases as the term unit could potentially bring benefits to the LCD approach, we have already shown that this limitation could be alleviated to an extent by a much simpler solution: query by passage selection. This approach allows the user to override the constraint of suggested terms by allowing them to highlight, directly from the exemplar, the phrase or passage that stimulated their current query. The system extracts, from this more specific relevance exemplar, all the terms that occur within the vocabulary of the semantic model and executes a Concept Pulse from this query string. A similar approach to querying was used in the TELLTALE (Pearce and Miller, 1997) and VOIR (Golovchinsky, 1997) dynamic hypertext systems. However, this strategy currently requires an extra interaction step and analysis from the user and in many cases there may not be a single coherent passage that provides a definitive exemplar. One way to alleviate the analysis required by the user might be to visually organise the exemplar document into homogeneous or distinct passages. Existing work in the area of document summarisation (e.g., Hearst, 1997; Ostler, 1999; Larocca Neto et al., 2000; Kleinberg, 2002) could usefully inform the development of such a feature.

To conclude, we have presented a solution approach to deal with the problematic exemplar cases identified in the previous chapter. We have been able to demonstrate a number of cases where the application of LCD terms to the visual context clearly facilitated the aspect cluster growing process. This is an open problem, however, and we have also proposed a number of avenues for further work, particularly with respect to LCD, that could enhance this solution approach still further.

CHAPTER 7: CONCLUSIONS

7.1. Introduction

The goal of this dissertation was *to develop and evaluate the potential utility of a novel interaction model to support the answering of an open-ended question using documents retrieved by a high-recall query.* In this chapter we discuss the achievement of our goal, drawing conclusions based on the analyses we have presented in this dissertation. We begin with a brief review of the presented dissertation (section 7.2), followed by a summary of research outcomes (section 7.3) where we discuss, within the framework of the three main research questions and their associated hypotheses, the extent to which the aims of this dissertation have been met. We then outline the general and specific contributions of this work (section 7.4). This is followed by a discussion of the limitations of the reported work (section 7.5). Finally, recommendations for future work are presented (section 7.6).

7.2. Review of dissertation

In Chapter 1, we introduced our thesis, by proposing a novel interaction model to support the problem of answering an open-ended question using an indexed, full-text document collection. In this interaction model, the user performs a high-recall query, which retrieves a broad cross-section of documents relevant to the intended topic, discussing many distinct aspects, along with many non-relevant documents. Spatial-semantic visualization is applied to provide a structured, interactive representation of retrieved documents, which allows the user to browse documents in an associative fashion, much like within the ordered shelves of a library. The utility of spatial-semantic visualization to support expansive searching (exploration) and narrowing (query refinement) search on a well-represented topic is well supported by the results of previous work (Chen et al., 1998; Allan et al., 2001; Cribbin and Chen, 2001). We focus specifically on a key strategy associated with our interaction model: the aspect cluster growing strategy. On discovery of a novel aspect of the relevant topic the user applies this strategy to find other similarly relevant documents. This strategy simply involves searching unknown documents in proximity order from the known relevant document node. This strategy of cluster growing has been shown to be effective for retrieving further documents relevant to a topic that is well represented within a visualization of a retrieved document set (Allan et al., 2001).

What was not clear was whether this success will transfer to situations where relevant documents form relatively minor features within the spatial-semantic model. We hypothesised that the structure of a spatial-semantic visualization can adequately support this strategy. We formulated three specific questions that relate to the general problem. We needed to know how to create an interactive spatial-semantic context that will support the aspect cluster growing strategy whilst maintaining a stable global context that allows the user to monitor the progress of their search and build a mental model of the relationships between different aspects of relevance. To create a useful spatial-semantic visualization we needed to be able to automatically generate an underlying semantic model that organises retrieved documents in a way that corresponds to the aspectual structure of the relevant topic without any prior knowledge of document relevance.

Our first question asked whether a standard approach to text analysis can create such a semantic model. Our second question asked which layout algorithm best conveys the required structure. Finally, anticipating that spatial-semantic structure might not always provide good cues to support the aspect cluster growing strategy, question three asked what the conditions would be under which spatial-semantic cues tend to fail and how can we apply this knowledge to develop appropriate interactive tools to support the strategy. Our approach has been to perform this investigation by measuring, in objective terms, the extent to which relevant topical structure can be communicated by spatial-semantic visualization and by simulating user performance of the aspect cluster growing strategy. This objective approach allows us to measure the upper bounds of potential performance within different visualization schemes across a range of search scenarios and without the potentially confounding effects of individual differences. This approach is feasible because of the algorithmic nature of the aspect cluster growing strategy, which is dependent on well-defined and objectively measurable properties of spatial-semantic visualizations and the availability of appropriate topics and relevance data made available from past Text Retrieval Conference (TREC) experiments.

Chapter 2 reviewed the literature relevant to our three questions and formulated hypotheses that were to be tested in Chapters 3 to 5. We developed two different tests that allowed us to measure semantic document clustering from two perspectives. The aspect cluster separation (ACS) test measures the degree of classification conveyed by document clustering at two levels of relevance – the general topic and specific topical aspect. The procedure involves computing three distributions for each scenario that describe, for each topically relevant document, the mean similarity or proximity between that document and same-aspect, same-topic and all documents within the document set. This allowed us to quantitatively test the hypothesis that the tendency for documents to cluster, in similarity (Chapter 3) and spatial-semantic space (Chapter 4), will increase as the semantic distance between them decreases. The second test is the nearest aspect neighbours (NAN) test. This is adapted from Voorhees' (1985) nearest neighbours test, to provide a fair test of theoretical or potential aspect cluster growing performance. Given that aspect sub-sets can vary widely in size, this test measures the rank distance between any given relevant exemplar and the first and second nearest aspect neighbours only. Additionally, from our discussions of spatial-semantic visualisation issues and specifically those associated with information loss (document node misplacement) due to dimension reduction, we identified two diametrically opposed approaches to document node layout that we subsequently compared in order to select the optimal visualization scheme for our interaction model (research question two). The first is a classical approach to spatial-semantic layout, called multi-dimensional scaling (MDS) whereby the algorithm seeks to find the best correspondence between all document similarities and node proximities. We contrast this global approach to optimisation with a local approach, whereby only the most salient inter-document similarities are considered during layout. This is achieved by considering the similarity matrix as a complete network. The minimum spanning tree (MST) of this network is computed prior to document node layout. We hypothesised that this will produce more cohesive clustering of aspectually-related documents, as evidence suggests their similarities will be relatively high within the distribution of all document similarities (Muresan and Harper, 2004).

The aim of Chapter 3 was to answer research question one, where we sought to determine the extent to which the structure of a relevant, but complex topic within a retrieved document set can be modelled using a standard text analysis algorithm. We began by describing the creation of our test bed, which comprises three distinct scenarios derived from two topic descriptions (open-ended questions). Scenarios were derived from topics and data made available from past TREC Interactive tracks (Over, 1997; Over, 1998). Each scenario comprised a single topic description, topic-aspect definitions and associated document relevance data, and a set of documents retrieved from the source collection.

Topics were both open-ended questions but were selected to be quite different in their answer structure, with 'Extinction' being composed of relatively distinct aspects and 'Chunnel' of relatively overlapping aspects. The source collection comprised articles from the Financial Times newspaper for the years 1991-94. Bespoke document sets for each scenario were retrieved from the source collection using a simple high-recall queries derived from the topic descriptions. Two scenarios were created based on the Chunnel topic, comprising the top 127 and top 218 documents from the same query and one scenario based on the Extinction topic, which also comprised 127 documents. The retrieved document set for each scenario was subject to an unsupervised text analysis using a typical approach based on the vector space model of document representation (Salton and McGill, 1983). Each of the resulting semantic models comprised a term-document matrix and an inter-document similarity matrix. The similarity matrices formed the basis for all ensuing analyses. The remainder of this chapter was devoted to answering question one where we applied the ACS and NAN tests to the similarity data for each scenario. Finally, we demonstrated the problems associated with attempting to convey the observed topical structure in the semantic models using a discrete clustering algorithm.

The aim of Chapter 4 was to begin the resolution of research question two, which sought to determine which layout algorithm produced the optimal spatial-semantic structures for our interaction model. We focused on two distinct approaches: Multi-dimensional scaling (MDS), where the algorithm seeks to find a globally optimal fit between true inter-document similarities and inter-node proximities in visual space; and minimum spanning tree (MST), a local optimisation approach where only the most salient inter-document similarities are intentionally preserved. We began by describing how the spatial-semantic visualizations were created, followed by a comparative visual analysis of the semantic structure conveyed by these visualizations at various levels including topic, aspect and discrete cluster membership. We then performed a comparative quantitative analysis of the topical structure conveyed by the respective visualizations, using the ACS test as developed and described in Chapter 2.

The aim of Chapter 5 was to resolve research question two and to answer the first part of question three, which sought to determine the conditions under which the aspect cluster growing strategy fails. We reported the results of simulated user trials for the aspect cluster growing strategy, conducted using the NAN test developed and described in Chapter 2 and previously applied to document similarities in the semantic model in Chapter 3. Potential performance of the strategy, using proximity cues, within both MDS and MST visualizations of the same semantic models of all scenarios, was measured for all aspects represented by two or more relevant documents and all possible cluster growing exemplars for those aspects. The effects of visualization scheme, aspect overlap and document set size were analysed. We then determined the extent to which ordinal level node misplacement, caused by compromises associated with dimension reduction, impacts on the efficiency of the cluster growing strategy. In particular, we sought to determine the extent to which problematic cluster growing cases could be resolved by substituting spatial-semantic cues with true document similarity cues. Finally, we identified two key factors that distinguish poor aspect exemplars, where neither spatial-semantic nor true similarity cues are sufficient to allow acceptable cluster growing performance, from those that provide good or acceptable support for the strategy. We discussed the implications of these identified factors for the design of interactive strategy support tools.

The aim of Chapter 6 was to resolve the second part of research question three, which asked how can we use knowledge of problematic cases to develop useful interactive tools to support aspect cluster growing. We introduced a term suggestion approach called Local Context Distillation (LCD) which, based on relevance feedback of just one known relevant exemplar, aims to identify key terms that describe potential reasons for the user's interest in that document. This is achieved by identifying terms that both occur in the exemplar and are highly exclusive to the local context (nearest neighbours) of this document. This produces a set of keywords that can be used either as contextual cues or query terms. We presented two tools that demonstrate each of these potential methods of application. Concept Signposts augment the existing spatial-semantic visualization by attaching each term as a label to the nodes whose associated document forms the best representative of that term within the local context. Concept Pulses, on the other hand, provide a form of dynamic querying whereby the user can select any combination of one or more of the suggested terms of interest and instantly gain an overview of their usage within the context of the visualization. Our demonstrations showed how these tools can support aspect cluster growing when the exemplar is dislocated from the remaining relevant documents. Limitations of the current implementation of LCD were also identified and possible avenues of improvement discussed.

Having reviewed the structure and content of this dissertation, we now review the research outcomes for each of the three questions and their associated hypotheses.

7.3. Research outcomes

This section summarises our conclusions for each research question. Tables 7.1 to 7.3 are provided for reference and summarise the specific results, by research question, for the specific hypotheses that were tested.

7.3.1. Question one

Question one asked: *To what extent can a standard text analysis procedure model the general semantic structure expected by our interaction model and particularly the low-level structure required by the aspect cluster growing strategy?*

Question and hypotheses	Outcome
Question 1: To what extent can a standard text analysis procedure model the general semantic structure expected by our interaction model and particularly the low-level structure required by the aspect cluster growing strategy?	**Cluster separation was significant for all scenarios. Acceptable precision was observed for nearly 70% of cases. Aspect overlap resulted in poorer overall cluster separation but closer nearest aspect neighbours. Larger set size resulted in better cluster separation but had no effect on the similarity of nearest aspect neighbours.**
H1: *The two level classification structure (topic and aspect cluster separation) will be evident for all scenarios whereby relevant documents will be, on average, more similar to the sub-set of documents that discuss the same aspect(s) than they are to the sub-set of generally relevant documents and, in turn, least similar to the retrieval set as a whole.*	**Supported** for all scenarios both for main effects (p<.001) and pair-wise comparisons between parent-child clusters (p<.001).
H2: *R2-precision for NAN in similarity space will be equal to or exceed 0.2 in most exemplar cases*	**Supported.** 20% precision at the point of locating the 2nd nearest aspect neighbour satisfied in over 68.5% of exemplar cases. Median rank of 1st and 2nd nearest aspect neighbour is 2 and 5.5 respectively.
H3*: In the overlapping aspect scenario, topic and aspect level cluster separation and mean R2-precision scores will be lower than in the distinct aspect scenario.*	**Partially supported.** Cluster separation is better for distinct topic but cluster growing is more efficient in overlapping topic. Aspect cluster separation within the set cluster is lower within the overlapping scenario. Topic cluster separation within the set cluster is lower for the overlapping scenario. Proportion of cases achieving 20% precision at the point of locating the 2nd nearest aspect neighbour is greater for the overlapping scenario (72.7% vs. 17.6%). Additionally, the cluster growing is generally more efficient in the overlapping scenario (p<.001) both for the 1st nearest (2 vs. 6) and 2nd nearest (5 vs. 22) relevant documents.
H4: *In the smaller retrieval set scenario, topic and aspect level cluster separation and R2-precision scores will be greater.*	**Rejected.** Aspect cluster separation tended to be better for the larger set. Topic cluster separation was significantly better for the larger set. No difference in the proportion of cases achieving 20% precision at the point of locating the second nearest aspect neighbour (72.7% vs. 71.3%). No general difference between scenarios in strategy efficiency.

Table 7.1: Summary of results relating to research question one

By general semantic structure we mean a two-level hierarchical classification whereby documents relevant to the general topic tend to be more similar to other relevant documents than to non-relevant documents and that in turn tend to be most similar to documents that are relevant to the same aspect or aspects of the topic. By low-level structure we mean the extent to which the nearest neighbours of a document tend to discuss the same aspects of the topic. We answered this question using two different tests. The ACS test considered the extent to which the hierarchical structure was apparent for across the sample of known relevant documents. The NAN test effectively simulated the user performing the aspect cluster growing strategy in high-dimensional vector similarity space. Hence, it provides a measure of maximum performance for the strategy. This is a theoretical maximum, however, as it is generally unlikely that any layout algorithm would be able to perfectly preserve the ordinal relationships between all relevant documents and their nearest neighbours.

The results of the ACS test showed that the expected hierarchical classification was evident for all three scenarios, with a highly significant linear effect on mean inter-document similarity as the comparison sub-set became more specifically related to a given relevant document. The NAN tests revealed that aspect cluster growing showed the potential to be an effective strategy in the majority of potential cases, with two same aspect documents being retrieved by the 10th nearest node in 68.5% of cases (n=270).

Looking more closely at the differences between scenarios, as expected, both aspect and topic cluster separation (within the set as a whole) was stronger for the more distinct topic. Unexpectedly, our comparison of the smaller and larger retrieval sets showed that both aspect and topic cluster separation was stronger within the semantic model for the larger document set. Our comparison of scenarios in terms of potential cluster growing performance produced results that were somewhat inconsistent with those of the ACS test. Cluster growing was more efficient in the overlapping scenario. Only a small minority of exemplar cases meeting the 20% precision criterion within the distinct aspect scenario and the differences in the rank positions of both the 1st and 2nd nearest neighbours differed significantly between the two topics. The effect of set size was unexpected, but less controversial, with no significant difference in potential cluster growing performance between the larger and smaller versions of the same topic. Combined with the results of the ACS test comparison these results are very encouraging and suggest that there is potential for our interaction model to work for even larger retrieval sets.

The conflicting effects of aspect overlap emphasise the key differences in the methods and objectives of the two tests. The ACS test is a high-level test that aims to provide a high-level measure of structural fidelity, whilst the NAN test focuses more on local structure. Also, the former test considers the model at the quantitative level, whilst the latter test considers the ordinal relationships between documents within the structure. The observed differences can be partly explained by the fact the aspects in the overlapping scenario tend to be much broader in scope and thus typically have a much higher number of same aspect relations. It seems that although, generally, same-aspect documents cluster less cohesively in the semantic model of the overlapping scenario, the most similar relatives seem to be relatively more similar than those of the distinct aspect scenario. It seems possible that the impact of these stronger similarities is outweighed by a relatively larger proportion of weaker similarities. In chapter 3, we suggested that if the ACS test is to be used to compare scenarios that differ grossly in this way that the median average may provide a fairer assessment of general document clustering than the arithmetic mean.

Finally, we examined the fidelity of aspect clustering in discrete clustering solutions. This analysis was included for completeness, to both verify whether the aspect fragmentation problems observed in previous efforts to cluster complex topics (e.g., Wu et al., 2001; Muresan and Harper, 2004) were also a feature of our semantic models and to provide a benchmark that more clearly demonstrates the superiority of spatial-semantic visualization for the purpose of our interaction model. In line with previous work (e.g., Wu et al., 2001), we found that whilst *k*-means clustering was reasonably successful in partitioning relevant from non-relevant documents within the set, despite being relatively more similar, specific aspect sub-sets tended to fragment across multiple clusters.

7.3.2. Question two

Question two asked: *Given an adequate semantic model, which approach to spatial-semantic layout best preserves the general and, in particular, the low-level structure expected by our interaction model?*

We began Chapter 4 with a visual analysis of some of the more interesting features of the spatial-semantic visualizations that were created for our analysis. Initially encouraging was the coherence of topic clustering, particularly within the MDS visualizations. MST tended to fragment the main topic cluster into multiple sub-clusters, particularly for the overlapping scenarios. Also encouraging was that both layout schemes were able to duplicate and build upon the cluster structure produced by the k-means algorithm, whereby relevance rich clusters tended to gather and overlap.

Question and hypotheses	Outcome
Question 2: Given an adequate semantic model, which approach to spatial-semantic layout best preserves the general and, in particular, the low-level structure expected by our interaction model?	**On balance, MST provided superior cluster separation for aspects, and equal or superior support for the aspect cluster growing strategy. However, MST created a small proportion of extremely poor aspect cluster growing cases. Also of note, MDS provided superior separation of the topic within the visualization.**
H5: *The two level classification will be effectively conveyed by spatial relations in (i) MDS and (ii) MST*	**Supported** for all scenarios and both visualization schemes.
H6: *Aspect level cluster separation will be greater for MST visualizations than for the MDS visualizations*	**Supported** for aspect separation both within set and topic clusters. However, MDS tended to organise the general topic more cohesively within the visualization.
H7: *Aspect cluster growing will be more efficient when using the MST visualizations compared to the MDS visualizations*	**Supported.** Of all cases studied the simulated user found the first two relevant documents faster in over 60% of cases when using the MST visualization. 20% precision criterion achieved in almost twice the number of cases compared to MDS.
H8: *Aspect level cluster separation will be lower in the overlapping aspect scenario than the distinct aspect scenario.*	**Supported.** Both topic and aspect cluster separation within the set cluster was greater within the distinct scenario for both layout schemes. Aspect separation within the topic cluster was greater for MST but not for MDS.
H9: *Aspect cluster growing will be less efficient in the overlapping aspect scenario compared to the distinct aspect scenario.*	**Rejected.** Significantly better performance within the overlapping scenario for both MST and MDS.
H10: *The expected differences between MST and MDS will be greatest for the distinct aspect scenario.*	**Partially supported.** Aspect cluster separation was better in MST for the distinct aspect scenario but not for the overlapping scenario. No significant general difference between schemes in rank analysis of nearest aspect neighbours for either scenario. MST was better or equal for 53% of Extinction cases and 56.4% of Chunnel cases. This was despite the observed ratio (MST to MDS) of good cases being higher for the distinct scenario (8.17) compared to the overlapping scenario (1.44). Seems that MST produced extremes, with a larger proportion of good cases than MDS, but a small proportion of very bad cases.
H11: *Aspect level cluster separation will be lower in visualizations of the larger retrieval set.*	**Rejected.** No effect of set size for MDS. Topic and aspect cluster separation within the set was significantly better in the larger set for MST.
H12: *Aspect cluster growing will be less efficient when using the larger retrieval set.*	**Partially supported.** Supported for MDS but not for MST where there was no difference.
H13: *The expected differences between MST and MDS will be greatest for the larger retrieval set.*	**Supported.** Significant general differences in aspect separation and cluster growing support between layout schemes for the larger set, but not the smaller set.

Table 7.2: Summary of results relating to research question two

We also found that MST, with its local bias, produced better clustering of the most cohesive aspects in all scenarios. In contrast, of the aspects that fragmented badly in the discrete cluster solution, sometimes MST did a better job, but in other cases MDS was superior. A notable tendency of MST was to organise problematic aspects into two or more tight clumps, whereas MDS would either produce a single cluster or simply scatter the individual nodes.

Our quantitative analysis of the spatial-semantic solutions began with a repeat of the ACS tests. The procedure was almost identical to before (for question one), except that the measure used was inter-document spatial proximity rather than similarity. Overall, we saw that cluster separation was consistently complete to a significant level for both layout schemes in all scenarios. Comparison between the two layout schemes revealed, as predicted, that MST was superior at clustering same-aspect documents both within the set cluster and the topic cluster. However, MDS was more effective at clustering the general topic within the set cluster, which is likely to be the effect of the global bias, which causes more major themes to be conveyed most effectively.

The NAN tests were also repeated in a similar fashion and revealed that upper bound aspect cluster growing efficiency was significantly better when using the MST visualization, with equal or better performance in over 60% of all cases considered. Furthermore, the 20% precision criterion for the second nearest relevant neighbour was met nearly twice as frequently for MST (37% vs. 63%).

However, the more detailed analysis that compared visualization performance between scenarios revealed a more complicated picture. As predicted, aspect cluster separation was better for the distinct aspect scenario. However, against our predictions but consistent with the results of our analysis with the underlying semantic model, aspect cluster growing was more efficient for the overlapping scenario. As expected, the local optimisation afforded by MST meant that the biggest differences between the two schemes, in terms of cluster separation, occurred for the distinct aspect scenario. However, there was no significant general difference between the two schemes for either the distinct aspect scenario or the equivalent sized overlapping scenario, although MST did tend to be equal or better in slightly more cases than MDS (53% and 56%). The observation that MST was not significantly better than MDS, at least for the distinct aspect scenario, was a curious one, especially given that the proportion of cases meeting the 20% criterion for aspect cluster growing was over eight times higher for the MST visualization. The reason for the lack of an observed general difference was attributed to the fact that although MST provided the best cluster growing situations it also provided the worst cases where the exemplar was particularly isolated from the main aspect cluster.

The effect of document set size was also interesting. We expected that increasing the size of the document set would produce generally poorer visualizations, given the corresponding increase in the dimensionality of the semantic model. Contrary to our expectations, aspect cluster separation was unaffected in MDS and actually improved for MST. Increasing the set size resulted in poorer aspect cluster growing performance in MDS, as predicted, but not for MST. As predicted, the local optimisation bias meant that MST was more resilient to the increasing complexity of larger semantic models. Whilst there was no general difference in upper bound cluster growing performance for the smaller scenario, there was a highly significant difference between the two schemes for the larger scenario. To reinforce our conclusions made with respect to question one, the implications of this are that our interaction model might be feasible, for much larger document sets, if MST is used as the layout scheme.

To conclude, MST provides better or equal aspect separation and cluster growing support in all scenarios. Hence, we can provide an answer to question two with reasonable confidence. The benefits for this scheme are particularly notable for the larger retrieval set. However, the use of MST seems to come with drawbacks. Although the algorithm ensures that highly similar documents cluster well, it can make some gross compromises when aspectual relations are less strongly encoded within the semantic model, causing extreme outliers that are likely to cause problems for cluster growing, particularly when such an outlier is the first discovered instance of an aspect. This leads us neatly on to our answer to question three.

7.3.3. Question three

Question three asked: *Under what conditions does the aspect cluster growing strategy tend to fail and how can we use this knowledge to guide development of interactive support tools?*

We began to answer this question in section 5.3 and concluded at the end of chapter 6. In answering this question, we focused on the MST visualization on the basis of our earlier findings, which suggested that it provided the optimal spatial-semantic cues for cluster growing in most situations.

We first compared the efficiency of a simulated user following either MST proximity cues or pure similarity cues. We found that similarity cues generally lead to better performance, although not in the Extinction scenario. In fact mean and median performance scores for the 2nd nearest neighbour were better, but not significantly so, when following proximity cues. In the Chunnel scenarios similarity was generally a more reliable cue. However, despite these general differences, along with modest increases (9%) in the proportion of all cases meeting the 20% criterion, we found that many of the worst exemplar cases remained problematic even when similarity cues were applied.

Question and hypotheses	Outcome
Question 3: Under what conditions does the aspect cluster growing strategy tend to fail and how can we use this knowledge to guide development of interactive support tools?	**Found that a significant proportion of cases were due to fundamental limitations of the document similarity matrix within the semantic model, rather than node misplacements alone. Problematic cases were associated with smaller aspect sub-set size and lower aspect salience. Developed the LCD approach to term suggestion, which attempts to elucidate minor related themes. Demonstrated the utility of LCD terms to support problematic cases by means of two visual tools: Concept Signposts and Concept Pulses.**
H14: *The majority of problematic cluster growing cases are due to node misplacements and can thus be resolved by augmenting the visualization with relative similarity cues*	**Partially supported.** Whilst following similarity cues was generally more efficient than using MST proximity cues, the majority of problematic cases were not just due to node misplacement but due to fundamental failure of the semantic model. 33% and 37% of all cases failed the 20% precision criterion in MST for 1st and 2nd NAN respectively. For each NAN, only 30% of these cases failed the criterion due to misplacement alone. Remainder of cases still failed the criterion even when proximity was substituted for similarity. This left 23% of cases still failing on 1st NAN and 26 failing on the 2nd NAN. Explored potential correlates of poor exemplar performance, by comparing universally problematic cases (p<0.2 for MST and SIM) with remaining cases:
Aspect size	Significant difference, where problematic cases tended to occur when the aspect of interest was smaller
Aspect salience	Significant difference, whereby problematic cases tended to occur when the exemplar discussed more than one aspect and the sub-set of the aspect of interest was in the minority to all aspectually related documents.

Table 7.3: Summary of results relating to research question three

The next stage of our analysis split our data into two independent groups: problematic cases that failed the 20% precision criterion when spatial-semantic cues were followed; and good cases that met or passed this criterion. We found that only a small proportion of problematic cases were due to misplacements alone. In fact 70% of all exemplar cases that failed the precision criterion when using MST proximity also failed when pure similarity was substituted. Hence, many aspect cluster growing problems seemed to occur due to a fundamental failure of similarities computed from the semantic model, rather than or in addition to compromises in the node layout process.

To explore this further, we created two new groups from the data: cases that failed the 20% criterion for both cues and those that met or passed on at least one of the cues. We examined a number of variables that describe a potential exemplar's topical content and its relationship to the aspect of interest. Predictably, mean similarity to aspect relations differed significantly between the good and the bad cases. However, we also found that the variables of aspect size and aspect salience also reliably discriminated good from bad aspect cluster growing exemplars. We found that problem cases occurred when the aspect of interest was relatively small. We also found a difference in aspect salience. This measured the proportion of current aspect relations to all aspect relatives of the exemplar. In other words, aspect salience is a measure of the relative importance of the aspect of interest within the document space in comparison to other closely related, but non-relevant documents. We found that aspect salience was significantly lower for problematic cases. Together, these results indicated that aspect cluster growing is problematic when the aspect of interest is relatively small and competing with many other, probably more closely related documents, for proximity to the exemplar.

On this basis, we suggested a new strategy whereby the user nominates a single known relevant document as an exemplar and in return the system suggests a range of possible reasons for relevance that link that document to its nearest neighbours within the document space. Given the known correlates of poor performance, greater emphasis is placed on terms that describe relatively minor relating themes. Three distinct tools enable this strategy: local context distillation, concept signposts and concept pulses. Local context distillation (LCD) is a term suggestion tool that is loosely based on the pseudo relevance feedback approach. LCD examines the exemplar and a user-specified local context sample (top *k* most similar documents). Terms that are exclusive to the exemplar and its local context are most likely to be suggested. The rationale is that potential query terms are those that exclusively occur in documents discussing the aspect of interest. Given that relevant items may be quite distal from the exemplar, the user can increase the size of the local context until good query terms are selected.

Suggested terms can be used in one of two ways. Concept Signposts take the selected terms and attaches each one to the node of the best document representative, within the local context. Related terms tend to be attached to the same or proximal document nodes thus forming clearer conceptual definitions. The user's attention is drawn towards the region of the visualization containing the most promising terms. Concepts Pulses is a dynamic querying tool that allows users to rapidly test out different queries using both single and multiple LCD terms. The query matches are shown using animation, whereby nodes expand to a size proportional to their match before slowly deflating. This creates a compelling visual array that clearly indicates documents and clusters that are most relevant to the query.

We provided examples of how each of the two visual tools can support various problematic cluster growing situations. However, it is noted that LCD is not always able to produce good discriminating terms, particularly when aspect definitions are either conceptually broad or closely related to other aspects. For these situations, it is suggested that short (e.g., two term) phrases or passages would be more appropriate term units. We demonstrated how allowing the user to jump the rails of the LCD algorithm, by selecting passages or phrases directly from the exemplar for Concept Pulsing, could partially resolve the problem but suggested that development of LCD to support phrase suggestion is a logical next step in its development.

In conclusion to question three, our analysis has identified the characteristics of problematic aspect cluster growing exemplars. This knowledge has been applied to inform the development of interactive tools that used in combination can demonstrably resolve problematic cluster growing cases.

7.4. Contributions

Based on our thesis and the research outcomes summarised in the preceding discussions, the general and specific contributions of this dissertation can be summarised as follows:

1. **A novel interaction model to support open-ended search tasks:** We have proposed a novel interaction model that aims to simplify the process of exploring a

complex and unfamiliar topic (e.g., answering an open-ended question). This is achieved by organising documents retrieved using a tentative (high-recall) query using a technique known as spatial-semantic visualization. We view the search process as one where the user begins with only a vague conception of their information need and so their query evolves as novel and interesting information is discovered, with the focus shifting between multiple and sometimes diverse, yet related aspects of the problem (Bates, 1989; O'Day and Jeffries, 1993). This requires an interface that simultaneously supports both expansive and narrowing search needs (Newby, 1998), emphasises browsing strategies rather than query specification, and allows the user to maintain an overview of their search progress. In our model the user begins by retrieving a topically broad base of documents, using a simple, high-recall query (e.g., one or two key words or phrases). The system then presents these items to the user as an interactive spatial-semantic visualization. The associative structure of the spatial-semantic model, where documents (represented as nodes) are organised spatially according to their relative similarity, allows the user to browse the retrieved documents in a non-linear order and immediately follow-up an interesting discovery (discover similar documents) simply by examining neighbouring nodes in the visualization (the cluster growing strategy); no query reformulation is required and the global structural view of the retrieved document set and the user's search progress is persistent and stable. Our model is similar to that of Leuski (2001), but has been significantly adapted to support complex, evolving queries (Bates, 1989; O'Day and Jeffries, 1993) as opposed to conceptually simple, well defined and static queries.

2. **Empirical data that supports the feasibility of our interaction model:** We have demonstrated that an inter-document similarity matrix (of retrieved documents) can classify a complex topic at both the general-topic and aspect levels of relevance and that this structure can be preserved and usefully conveyed within a spatial-semantic visualization. Previous work (Rorvig and Fitzpatrick, 1998; Leuski, 2001; Allan et al., 2001) had demonstrated that relevant documents tend to form a coherent cluster within a spatial-semantic visualization of an ad hoc document set. Given this tendency, Leuski (2001) was able to demonstrate the utility of the cluster growing strategy for isolating relevant documents. However, most of these topics were simple in structure; there was only one aspect of relevance and so all relevant documents were highly similar. Until now, no study

has formally evaluated the potential of applying this strategy to a more complex, relevant topic represented within much larger, more diffuse (i.e., high-recall) retrieval set. Our interaction model fundamentally requires that inter-document similarity increase as the semantic distance between documents decreases and that this structure can be reliably conveyed by the structure of the spatial-semantic visualization. Muresan and Harper (2004) provided evidence that documents relevant to a complex topic also tend to be relatively dissimilar to non-relevant documents and that relevant documents that discuss different aspects of a topic tend to be less similar than those that discuss the same aspect. We have extended their results to provide evidence that this trend also occurs within the context of a high-recall retrieval set where there is only one topic of interest. We have demonstrated that even simple measures of inter-document lexical similarity can be used to classify such a document set into this two-level hierarchy of relevance. Moreover, the feasibility of preserving this two-level structure within a spatial-semantic visualization had not been studied before this dissertation, highlighting a further contribution of our work. We then demonstrated that this classification remains when the high-dimensional model is projected on to two-dimensional space as a spatial-semantic visualization. Finally, we have also shown that, using an appropriate layout scheme, spatial-semantic cues are sufficient to support efficient aspect cluster growing from a large proportion of all possible starting points (relevant exemplar cases).

3. **Formal evaluation of spatial-semantic document visualizations without the need for feedback from human subjects:** We have demonstrated an objective evaluation approach that measures the presence of a complex relevance classification within a semantic or spatial-semantic model and the efficiency of the aspect cluster growing search strategy using an existing benchmark test collection. We have performed our analyses by means of pre-defined topics and relevance judgements, rather than direct feedback from human subjects. This both reduced the time and cost of testing and allowed better control over random error. This objective approach was made possible by the availability of TREC interactive test-collections, which provide realistic search scenarios and two level (topic and aspect) document relevance data that provide a benchmarks for performance evaluation. The use of TREC data is not a new approach to evaluating either visual or non-visual IR interfaces; For example, Leuski (2001) evaluated his form of the cluster

growing strategy using data from the "ad hoc" task. However, the contribution of this dissertation is the first reported example of TREC interactive data, being applied in this way, to (i) objectively evaluate spatial-semantic visualizations of complex topics represented within query-retrieved document sets, and to (ii) concurrently evaluate both general structural fidelity of a visualization and potential search strategy performance.

4. **The development and evaluation of two new tests of the IR cluster hypothesis:** Our evaluation of classification and strategy performance has been achieved using two new tests, which we believe will be of future value to the research community. These tests were bespoke adaptations of existing tests of the IR cluster hypothesis (van Rijsbergen, 1979; Voorhees, 1985; Muresan and Harper, 2004) that provide methodologies for evaluating both the presence of a complex, hierarchical relevance classification structure and the efficiency of a cluster-dependent search strategy. The aspect cluster separation (ACS) test measures the relative mean size of clusters surrounding known relevant documents at three levels of semantic distance: same aspect, same topic and all retrieved documents. This approach provides a simple, statistically testable measure of the integrity of the expected two-level classification of the relevant documents. Treating each relevant document as a single data case allows straightforward statistical comparison of hierarchical classification between scenarios (e.g., using class ratio transformations). The nearest aspect neighbours (NAN) test effectively simulates a user growing an aspect cluster from a known relevant exemplar using a simple cue driven strategy. This is based on Voorhees (1985) test and is similar to Leuski's (2001) strategy based evaluation method. However, this test is specifically designed to accommodate complex topics, where documents may discuss more than one aspect of the topic, where aspect sub-sets are likely to vary considerably in set size and each relevant document is seen as a potential cluster growing exemplar for all the aspects that it discusses. A further contribution of this dissertation was our demonstration of how the results of the NAN test can allow for the identification of factors that discriminate between good and bad cluster growing exemplars and how their observed effects can be used to inform the development of the tools and strategies (see contribution five) that can improve cluster growing performance.

5. **The proposal and demonstration of novel interactive tools to support *ad hoc*, focused searches within spatial-semantic document visualizations:** This dissertation has reported the development of novel interactive tools to support focused aspect searches within spatial-semantic document visualizations. These tools provide useful and transferable alternatives to designers of both graphical and non-graphical IR interfaces. The design concepts and specifications of these tools were motivated by known limitations of spatial-semantic cues identified during the NAN test analysis. Local context distillation (LCD), which is based on the principle of local feedback (Attar and Fraenkal, 1977), analyses the relationship between a single known relevant document and its local and global contexts to suggest terms that allow the user to recognise (rather than think up) terms that specify their current query. Concept Signposts augment the local context of the exemplar, as represented within the visualization, using contextually located LCD term labels. These help the user to understand how different topics relevant to a selected exemplar are organised within the visualization. Concept Pulses provide a form of dynamic querying that combines user selectable LCD terms with animation within the visualization to help the user to rapidly experiment with different queries and understand how matches for these queries are distributed across the visualization. We have demonstrated the utility of these tools to support aspect cluster growing strategy episodes that proved problematic using either similarity or spatial-semantic cues. We also believe that the application of the LCD term suggestion concept is not limited to our interaction model but can also be easily and usefully transferred to classical, non-graphical IR interface contexts (see section 7.6)

6. **An objective comparison of two diametrically opposed approaches to spatial-semantic layout optimisation within the context of a specific search task:** We have demonstrated that both global and local optimisation approaches can effectively preserve the modelled two-level relevance hierarchy that is required by our interaction model for an open-ended search task. However, we have also found that there are key differences in their emphasis. Global optimisation (classic MDS), where the layout algorithm attempts to preserve all inter-document similarities, seems to better preserve the top-level relevance structure, suggesting it may be a better scheme to use for simple topic-cluster growing tasks. The local approach, on the other hand, where only the MST graph of the similarity matrix is

used as input to the layout algorithm, seems more appropriate for supporting an open-ended search characterised by a complex, evolving query. Our data show that MST separates distinct aspects and clusters same-aspect documents more coherently than MDS. Furthermore, our evidence indicates that MST is considerably more scaleable than MDS, allowing larger, more diverse retrieval sets to be visualized and searched within a single interaction episode.

7. **A demonstration of methods and the importance of cross-verification of the cluster hypothesis testing at both high-dimensional and visual levels of representation:** We have argued that researchers engaged in document visualization experiments should understand the underlying structure of the semantic model before interpreting the results of clustering or scaling procedures. This argument has been vindicated by our observations, which show the extent to which key structures are preserved, particularly same-aspect (low-level) relations, varies considerably from one technique to another. ACS and NAN tests are directly transferable and comparable between high-dimensional similarity space and low-dimensional spatial-semantic space. Results from analyses at the level of similarity space provide a benchmark that can avoid false rejection of the cluster hypothesis for a given scenario should initial visualization/clustering trials fail. This approach also provides a means of estimating key structural information loss during dimension reduction, as opposed to general information loss as would be measured by traditional correlation or stress measures of match between input and output proximities.

7.5. Limitations

This work was deliberately limited to a single test collection, containing only one type of document, newspaper articles, from a single source publication. These documents are therefore relatively homogeneous in writing style and quality. Using such a collection was desirable as it minimised potentially confounding influences on the semantic modelling process such as vocabulary mismatch, misspelling or variation in spelling and so forth. To accommodate such variation properly would have required, amongst other measures, consideration of different approaches to text analysis, which as stated in Chapter 1, was not an objective of this thesis. However, it is advisable to consider this limitation before generalising the reported findings to other document types and collections.

The lack of a formal user study to confirm the success of our objective analyses could also be perceived as a limitation. However, brief user studies, particularly when novel metaphors or interaction styles are involved, are known to be confounded by the effects of individual differences such as cognitive ability, experience with interactive graphics, that can mask out the effects of independent variables under study (e.g., see Swan and Allan, 1998). We argue that the analysis reported here allowed us to verify fundamental and objectively testable assumptions and inform key design decisions prior to the introduction of any complex and also costly and time consuming user studies. The results of our analysis provide a benchmark against which to interpret user success or failure when using the prototype interface for real. For instance, given a result that showed equally poor user performance of the cluster growing strategy when searching from both known good and problematic exemplars, the experimenter could immediately rule out spatial-semantic structural failure as a possible causal explanation.

7.6. Future work

The results of our analyses have provided encouraging support for some key assumptions of our interaction model. We have demonstrated that it is possible, even with a relatively simple text analysis procedure, to model the required semantic structure and that this structure can be adequately preserved despite dimension reduction and presented in spatial-semantic form. We have also shown that the simple strategy of aspect cluster growing is also feasible in a large proportion of potential cases. However, despite these achievements, we have only begun to evaluate and develop this interaction model. Many questions remain, several of which have emerged as a direct result of the analysis conducted in this work.

Further research should focus on the problem of optimising the semantic modelling method. Whilst a simple word term vector comparison approach has been shown to produce somewhat acceptable inter-document similarity matrices, there are clear limitations. First, a document can discuss several aspects of the relevant topic along with other non-relevant topics and the concept of interest may represent only a small part of the whole document. For this reason, the document as the unit of similarity analysis is probably too coarse for the purpose of modelling complex, topical structure. Leuski (2001) has also stressed this issue when considering the potential problems associated with 'multi-topic' documents. Breaking documents into passage units, however, has big implications for the size of similarity matrix and, in turn, introduces the difficult question of whether the spatial-semantic visualization should represent each passage as a distinct node, which would clearly present significant problems with respect to visualization legibility. Several potential solution paths can be envisaged. For instance rather than splitting documents into passages based on rigid criteria (e.g., paragraphs or every 100 words), a pattern analysis technique, for instance, along the lines of burst detection (Kleinburg, 2002) or TextTiling (Hearst, 1997) where sudden changes in feature occurrence can be used to detect the start of new topics, might provide more effective, and potentially economical, criteria for document partitioning. With respect to the presentation problem, the inter-document similarity matrix could remain as the input to visualization, with each pair-wise similarity being represented as, for instance, the closest matching passages occurring between the two documents. The implications of this approach for general high-dimensional and spatial-semantic classification and aspect level clustering would need to be carefully evaluated using the same, or a similar approach to that used in this dissertation.

A second problem that likely affected our results, particularly within the Chunnel scenario where aspect definitions were relatively broad and overlapping, is vocabulary mismatch. Vocabulary mismatch refers to the tendency for different people to describe the same concepts using different terms, and is a well-recognised problem with the field of IR (see Furnas et al., 1987). Extant approaches to dealing with mismatch include concept decomposition, where terms are replaced by higher-order derived concepts (e.g., Latent Semantic Indexing: Deerwester et al., 1990; Concept Indexing: Karypis and Han, 2000).

Vocabulary mismatch is a problem for LCD, our term suggestion tool. The facility for the user to identify aspect matches by referring explicitly to terms that occur in the exemplar, but not within other relevant documents could be highly effective i.e., whereby the suggested term is substituted with the relevant underlying concept feature when, for instance, the concept pulse routine is executed. Also, with respect to LCD, we discussed the utility of suggesting phrases instead of, or in addition to single word terms. Particularly when the aspect definition is broad or abstract in its subject, short phrases would be more meaningful and potentially easier to identify as exclusive to the local context. We gave one example, for instance, in the Chunnel scenario where the phrase "rail link" was considerably more meaningful and salient than the single words considered independently. We suggest that one interesting approach to phrase identification might be suffix tree clustering (Zamir and Etzioni, 1998) which has been proven to be an efficient means of identifying phrases of varying length that are common to two or more documents.

Future work should extend our work by examining further, alternative approaches to spatial-semantic visualization. In this dissertation, we have considered two diametrically opposed approaches to spatial-semantic layout. As predicted, treating the similarity matrix as a graph and radically pruning less salient inter-document similarities by computing the minimum spanning tree (MST) prior to spatial node placement lead to a layout that better conveyed aspect level relations in a good majority of cases, compared to a standard global optimisation approach (MDS). MST is only one method of graph edge pruning, however, and other techniques like Pathfinder network scaling (Schvaneveldt et al., 1989) are worthy of investigation. Chen and Morris (2003), for instance, give an interesting comparison of MST and Pathfinder networks for different spatial-semantic visualization application - co-citation analysis. Also, Leuski (2001) found that identifying the optimal inter-document similarity threshold, to ensure optimal topic clustering, was an important consideration in the development of Lighthouse (see Leuski, 2001). However, Leuski does not give details of the effects of varying thresholds either within or between topical scenarios, and of course the focus was on the clustering of well-represented topics, rather than complex aspect level structure. A study that examined the effect of manipulating this threshold, across multiple topical scenarios, would be worthwhile and interesting, particularly if this lead to heuristic functions that could be used to optimise topical clustering based on statistical properties of the semantic model that are observable prior to any significant level of relevance feedback.

We have already noted that there are alternative, potentially useful applications of the LCD approach. It would be worthwhile, for example to test the implementation of this tool within more traditional (i.e., non-graphical) search interfaces. One interesting avenue is the use of this term suggestion approach to support the 'more like this' function already available in many Web search systems.

Finally, whilst many of the outstanding issues can be dealt with, at least initially, using an objective experimental approach like the one followed in this work, user studies are ultimately required to fully confirm the validity of the interaction model as a means of support for open-ended question answering. In particular, the utility of the proposed, and future interactive support tools can only be fully evaluated and developed through analysis of user's subjective responses within controlled and realistic search task situations.

REFERENCES

Aalbersberg, I. (1992, June 21-24). *Incremental relevance feedback*. Paper presented at the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark, 11-22.

Allan, J., Leuski, A., Swan, R., & Byrd, D. (2001). Evaluating combinations of ranked lists and visualizations of inter-document similarity. *Information Processing & Management, 37*(3), 435-458.

Andrews, K., Kienreich, W., Sabol, V., Becker, J., Droschl, G., Kappe, F., et al. (2002). The InfoSky visual explorer: Exploiting hierarchical structure and document similarities. *Information Visualization, 1*(3-4), 166-181.

Attar, R., & Fraenkel, S. (1977). Local feedback in full-text retrieval systems. *Journal of the ACM, 24*(3), 397-417.

Bates, M. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review, 13*(5), 407-431.

Belkin, N. J., Oddy, R., & Brooks, H. (1982). ASK for information retrieval: Part 1, background and theory. *Journal of Documentation, 38*(2), 61-71.

Belkin, N. J. (1993). *Interaction with texts: Information retrieval as information-seeking behavior.* Paper presented at the German Computer Society, Information Retrieval Special Interest Group's 1993 Conference, Universitaetsverlag Konstanz, 55-66.

Belkin, N. J., Head, J., Jeng, J., Kelly, D., Lin, S., Park, S. Y., et al. (2000). *Relevance feedback versus local context analysis as term suggestion devices: Rutgers' TREC-8 interactive track experience.* Paper presented at the Eighth Text REtrieval Conference (TREC 8), 565-574.

Bertin, J. (1983). *The semiology of graphics*. Madison, Wisconsin: University of Wisconsin Press.

Brooks, T. (1998, October 25-28). *The semantic distance model of relevance assessment.* Paper presented at the 61st Annual Meeting of ASIS, Pittsburgh, PA, 33-34.

Busing, F., Commandeur, J., & Heiser, W. (1997). PROXSCAL: A multidimensional scaling program for individual differences scaling with constraints. In W. Bandilla & F. Faulbaum (Eds.), *Advances in Statistical Software* (Vol. 6, pp. 67-73). Stuttgart: Lucius & Lucius.

Brath, R. (2003, May). *Paper landscapes: a visualization design methodology.* Paper presented at Electronic Imaging 2003: Visualization and Data Analysis, Santa Clara, CA, 125-132.

Campagnoni, F. R., & Ehrlich, K. (1989). Information retrieval using a hypertext-based help system. *ACM Transactions on Information Systems, 7*(3), 271-291.

Card, S., Mackinlay, J., & Shneiderman, B. (1999). *Readings in information visualization: Using vision to think*: San Francisco: Morgan Kaufmann.

Carey, M., Kriwaczek, F., & Rüger, S. M. (2000, Nov 10-11). *A visualization interface for document searching and browsing.* Paper presented at the CIKM 2000 Workshop on New Paradigms in Information Visualization and Manipulation, Washington, DC.

Chalmers, M., & Chitson, P. (1992, June 21 - 24). *Bead: Explorations in Information Visualization.* Paper presented at the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen Denmark, 330-337.

Chalmers, M. (1993). Using a landscape metaphor to represent a corpus of documents. *Lecture Notes in Computer Science, 716*, 377-390.

Chen, H., Houston, A. L., Sewell, R. R., & Schatz, B. R. (1998). Internet browsing and searching: User evaluations of category map and concept space techniques. *Journal of the American Society for Information Science, 49*(7), 582-608.

Chen, C. (1999a). Visualising semantic spaces and author co-citation networks in digital libraries. *Information Processing & Management, 35*(3), 401-420.

Chen, C. (1999b). *Information visualisation and virtual environments*. London: Springer.

Chen, C., & Yu, Y. (2000). Empirical studies of information visualization: a meta-analysis. *International Journal of Human-Computer Studies*, 53(5), 851-866.

Chen, C., & Paul, R. (2001). Visualizing a knowledge domain's intellectual structure. *IEEE Computer, 34*(3), 65-71.

Chen, C., Cribbin, T., Morar, S. S., & Macredie, R. (2002). Visualizing and tracking the growth of competing paradigms: Two case studies. *Journal for the American Society for Information Science and Technology, 53*(8), 678-689.

Chen, C., & Morris, S. (2003, October 19-21). *Visualizing evolving networks: Minimum spanning trees versus pathfinder networks.* Paper presented at the IEEE Symposium on Information Visualization 2003, Seattle, Washington, 67-74.

Coxon, A. (1982). *The user's guide to multi-dimensional scaling*. London: Heinemann.

Cribbin, T., & Chen, C. (2001, January 21-26). *Visual-spatial exploration of thematic spaces: A comparative study of three visualisation models.* Paper presented at Electronic Imaging 2001: Visual Data Exploration and Analysis VIII, San Jose, CA, 199-209.

Damashek, M. (1995). Gauging similarity with n-grams: Language-independent categorization of text. *Science, 267*, 843-848.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391-407.

Eades, P. (1984). A heuristic for graph drawing. *Cogressus Numerantium, 42*, 149-160.

Eysenck, M., & Keane, M. (1990). *Cognitive psychology: A student's handbook*. Hove and London: Lawrence Erlbaum Associates.

Fabrikant, S. (2000). Spatialized browsing in large data archives. *Transactions in GIS, 4*(1), 65-78.

Fruchterman, T., & Reingold, E. (1991). Graph drawing by force directed placement. *Software Practice and Experience, 21*, 1129-1164.

Furnas, G., Landauer, T., Gomez, L., & Dumais, S. (1987). The vocabulary problem in human-system communication. *Communications of the ACM, 30*(11), 964-971.

Golovchinsky, G. (1997, April 6-11). *What the query told the link: The integration of hypertext and information retrieval.* Paper presented at the Eighth ACM conference on Hypertext, Southampton, UK, 67-74.

Hancock-Beaulieu, M., & Walker, S. (1995). An evaluation of interactive query expansion in an online library catalogue with a graphical user interface. *Journal of Documentation, 51*(3), 225-243.

Harter. (1986). *Online information retrieval: concepts, principles and techniques*. London: Academic Press.

Hearst, M., & Pedersen, J. (1996, August 18-22). *Reexamining the cluster hypothesis: Scatter/gather on retrieval results.* Paper presented at the 19th Annual International ACM SIGIR Conference, Zurich, Switzerland, 76-84.

Hearst, M. (1997). TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Lingustics, 23*(1), 33-64.

Hearst, M. (1999). User interfaces and visualization. In R. Baeza-Yates & B. Ribeiro-Neto (Eds.), *Modern information retrieval* (pp. 464). New York: Addison Wesley Longman.

Hook, K., Sjölinder, M., & Dahlback, N. (1996). *Individual differences and navigation in hypermedia.* Paper presented at the Eighth European Conference on Cognitive Ergonomics (ECCE-8), Spain.

Hornbæk, K., & Frokjær, E. (1999, August 30 - September 3). *Do thematic maps improve information retrieval?* Paper presented at the IFIP TC.13 International Conference on Human-Computer Interaction (INTERACT '99), 18-25.

Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing & Management, 36*(2), 207-227.

Jardine, N., & van Rijsbergen, C. (1971). The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval, 7*, 217-240.

Johnson, B., & Shneiderman, B. (1991). *Treemaps: a space-filling approach to the visualization of hierarchical information structures.* Paper presented at the 2nd International IEEE Visualization Conference, San Diego, 284-291.

Kamada, T., & Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters, 31*(1), 7-15.

Karypis, G., & Han, E-H. (2000, November 6-11). *Fast supervised dimensionality reduction algorithm with applications to document categorization & retrieval.* Paper presented at the Ninth International Conference on Information and Knowledge Management (CIKM 2000), 12-19.

Kleinberg, J. (2002, July 23-26). *Bursty and Hierarchical Structure in Streams.* Paper presented at the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 91-101.

Koenemann, J., & Belkin, N. J. (1996, April 13-18). *A case for interaction: A study of interactive information retrieval behavior and effectiveness.* Paper presented at the ACM Conference on Human Factors in Computing Systems, CHI '96, Zurich, Switzerland, 205-212.

Koffka, K. (1935). *Principles of gestalt psychology.* New York: Harcourt-Brace.

Korfhage, R. (1995). Some thoughts on similarity measures. *SIGIR Forum, 29*(1), 8.

Kruskal, J. B. (1956). On the shortest spanning subtree of a graph and the travelling salesman problem. *Proceedings of the American Mathematical Society, 7*, 48-50.

Kruskal, J. B. (1964). Non-metric multidimensional scaling: A numerical method. *Psychometrika, 29*(2), 115-129.

Larocca Neto, J., Santos, A., Kaestner, C., & Freitas, A. (2000). *Document clustering and text summarization.* Paper presented at the Fourth International Conference on Practical Applications of Knowledge Discovery and Data Mining (PADD-2000), London, UK, 41-55.

Leuski, A. (2001). *Interactive information organization: Techniques and evaluation.* Unpublished PhD dissertation. University Of Massachusetts, Amherst.

Lin, X., Soergel, D., & Marchionini, G. (1991, October 13-16). *A self-organizing semantic map for information retreival.* Paper presented at the Fourteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Chicago, IL USA, 262-269.

Lin, X. (1997). Map displays for information retrieval. *Journal for the American Society for Information Science, 48*(1), 40-54.

Lundquist, C., Grossman, D., & Ophir, F. (1997, November 10-14). *Improving relevance feedback in the vector space model.* Paper presented at the Sixth International Conference on Information and Knowledge Management, Las Vegas, USA, 16-23.

Marchionini, G. (1995). *Information seeking in electronic environments*. Cambridge: Cambridge University Press.

Markey, K., & Cochrane, P. (1981). *Online training a practice manual for ERIC data base searchers* (Vol. 2). Syracuse, New York: ERIC Clearinghouse on Information Resources.

Montello, D. R., Fabrikant, S., Ruocco, M., & Middleton, R. S. (2003). Testing the first Law of cognitive geography on point-display spatializations, *Spatial Information Theory: Foundations of Geographic Information Science* (Lecture Notes in Computer Science vol. 2825, pp. 316-331). Berlin: Springer-Verlag.

Muresan, G. (2002). *Using document clustering and language modelling in mediated information retrieval.* Unpublished PhD dissertation, Robert Gordon University, Aberdeen.

Muresan, G., & Harper, D. (2004). Topic modeling for mediated access to very large document collections. *Journal for the American Society for Information Science and Technology, 55*(10), 892-910.

Newby, G. (1998). *An information access model with a unified approach to data type, retrieval mechanism and information need.* Paper presented at the ASIS 1998 Annual Meeting, Pittsburgh, PA, 475-484.

North, S. C. (2002). *Drawing graphs with Neato*. Retrieved on 8th January, 2004, from shttp://www.research.att.com/sw/tools/graphviz/neatoguide.pdf

O'Day, V., & Jeffries, R. (1993, April 24-29). *Orienteering in an information landscape: how information seekers get from here to there.* Paper presented at the SIGCHI Conference on Human Factors in Computing Systems, Amsterdam, 438-445.

Olsen, K., Korfhagea, R. R., Sochatsa, K. M., Springa, M. B., & Williams, J. G. (1993). Visualization of a document collection: The vibe system. *Information Processing & Management, 29*(1), 69-81.

Ostler, T. (1999, July 14-16). *Information highlighting.* Paper presented at the 1999 International Conference on Information Visualization, London, UK, 528-534.

Over, P. (1997, November 19-21). *TREC-6 interactive report.* Paper presented at the Sixth Text REtrieval Conference (TREC-6), Gaithersburg, Maryland, 73-82.

Over, P. (1998, November 09-11). *TREC-7 interactive track report.* Paper presented at the Seventh Text REtrieval Conference (TREC-7), 65-72.

Pearce, C., & Miller, E. L. (1997). The TELLTALE dynamic hypertext environment: Approaches to scalability. In J. Mayfield & C. Nicholas (Eds.), *Advances in Intelligent Hypertext* (Vol. 1326, pp. 109-130): Springer-Verlag.

Pirolli, P., & Card, S. (1999). Information foraging. *Psychological Review, 106*(4), 643-675.

Porter, M. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130-137.

Prim, R. (1957). Shortest connection networks and some generalizations. *Bell System Technical Journal, 36*, 1389-1401.

Rorvig, M., & Fitzpatrick, S. (1998). Visualization and scaling of TREC topic document sets. *Information Processing & Management, 34*(2-3), 135-149.

Roussinov, D., & Chen, H. (2001). Information navigation on the web by clustering and summarizing query results. *Information Processing & Management, 37*(6), 789-816.

Salton, G., & McGill, M. (1983). *Introduction to modern information retrieval.* New York: McGraw-Hill Inc.

Salton, G., & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal for the American Society for Information Science, 41*, 288-297.

Salton, G. (1991, October 13-16). *The smart document retrieval project.* Paper presented at the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Chicago, Illinois, United States, 356-358.

Schvaneveldt, R., Durso, F., & Dearholt, D. (1989). Network structures in proximity data. In G. Bower (Ed.), *The Psychology of Learning and Motivation* (Vol. 24, pp. 249-284), Norwood, NJ: Academic Press.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika, 27*(2), 125-140.

Skupin, A. (2000, October). *From metaphor to method: Cartographic perspectives on information visualization.* Paper presented at the IEEE Symposium on Information Visualization 2000 (InfoVis 2000), Los Alamitos, CA, 91-97.

Skupin, A. (2002). A cartographic approach to visualizing conference abstracts. *IEEE Computer Graphics and Applications, 22*(1), 50-58.

Skupin, A., & Fabrikant, S. (2003). Spatialization methods: A cartographic research agenda for non-geographic information visualization. *Cartography and Geographic Information Science, Transitions in U.S. Cartography and Geographic Information Science, 30*(2), 95-119.

Stanney, K., & Salvendy, G. (1995). Information visualization: Assisting low-spatial individuals with information access tasks through the use of visual mediators. *Ergonomics, 38*(6), 1184-1198.

Sullivan, T., & Rorvig, M. (1998). *Converting MDS graphical images to precision/recall graphs*: Unpublished Technical Report, University of North Texas.

Swan, R., & Allan, J. (1998, August 24-28). *Aspect windows, 3-D visualizations and indirect comparisons of information retrieval systems.* Paper presented at the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, 173-181.

Takane, Y., Young, F. W., & De Leeuw, J. (1977). Non-metric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features. *Psychometrika, 42*, 593-600.

Taylor, R. (1968). Question-negotiation and information seeking in libraries. *College and Research Libraries, 29*, 178-194.

Tombros, A., & van Rijsbergen, C. (2001, October 5-10). *Query-sensitive similarity measures for the calculation of inter-document relationships.* Paper presented at the Tenth International Conference on Information and Knowledge Management (CIKM 01), Atlanta, GA, 17-24.

Toms, E. (1998). *Information exploration of the third kind: The concept of chance encounters.* A position paper for the CHI 98 Workshop on Innovation and Evaluation in Information Exploration.

Torgerson, W. S. (1952). Multidimensional scaling: 1. Theory and method." *Psychometrika, 17*, 401-419.

van Rijsbergen, C., & Sparck Jones, K. (1973). A test for the separation of relevant and non-relevant documents in experimental retrieval collections. *Journal of Documentation, 29*, 251-257.

van Rijsbergen, C. (1979). *Information retrieval.* London: Butterworths.

Vicente, K., & Willeges, R. (1988). Accomodating individual differences in searching a hierarchical file system. *International Journal of Man-Machine studies, 29*, 647-668.

Voorhees, E. (1985, June 5-7). *The cluster hypothesis revisited.* Paper presented at the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Montreal, Quebec, Canada, 188-196.

Voorhees, E., & Harman, D. (1996). *Overview of the fifth Text REtrieval Conference (TREC-5).* Paper presented at the Fifth Text REtrieval Conference (TREC-5), Gaithersburg, Maryland.

Voorhees, E., & Harman, D. (1997). *Overview of the sixth Text REtrieval Conference (TREC-6).* Paper presented at the Sixth Text REtrieval Conference (TREC-6), Gaithersburg, Maryland, 1-24.

Voorhees, E., & Harman, D. (1998). *Overview of the seventh Text REtrieval Conference (TREC-7).* Paper presented at the Seventh Text REtrieval Conference (TREC 7), Gaithersburg, Maryland, 1-24.

Westerman, S. J., & Cribbin, T. (2000). Mapping semantic information in virtual space: Dimensions, variance, and individual differences. *International Journal of Human-Computer Studies, 53*(5), 765-788.

Westerman, S. J., Collins, J., & Cribbin, T. (2005). Browsing a document collection represented in two- and three-dimensional virtual information spaces. *International Journal of Human-Computer Studies, 62*(6), 713-736.

Willett, P. (1988). Recent trends in hierarchical document clustering: A critical review. *Information Processing & Management, 24*(5), 577-597.

Williamson, C., & Shneiderman, B. (1992). Dynamic queries for information exploration: An implementation and evaluation. Paper presented at the SIGCHI Conference on Human factors in Computing Systems, 619-622.

Wise Jr, J., Thomas, J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., et al. (1995). *Visualising the non-visual: Spatial analysis and interaction with information from text documents.* Paper presented at the IEEE Symposium on Information Visualization (InfoVis '95), New York, 51-58.

Wise Jr, J. (1999). The ecological approach to text visualization. *Journal of the American Society for Information Science, 50*(13), 1224-1333.

Wu, M., Fuller, M., & Wilkinson, R. (2001). Using clustering and classification approaches in interactive retrieval. *Information Processing & Management, 37*(3), 459-485.

Xie, H. (2000). Shifts of interactive intentions and information seeking strategies. *Journal of the American Society for Information Science, 51*(9), 841-857.

Xie, H. (2002). Patterns between interactive intentions and information-seeking strategies. *Information Processing & Management, 38*(1), 55-77.

Xu, J., & Croft, B. (1996, August 18-22). *Query expansion using local context and global context analysis.* Paper presented at the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '96), Zurich, Switzerland, 4-11.

Xu, J., & Croft, B. (2000). Improving the effectiveness of informational retrieval with local context analysis. *ACM Transactions on Information Systems, 18*(1), 79-112.

Zamir, O., & Etzioni, O. (1998, August 24-28). *Web document clustering: a feasibility demonstration.* Paper presented at the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, 46-54.

PUBLICATION NOTE

A summary of publications, written or co-written by the author, that relate to the work presented in this dissertation:

Chen, C., & Cribbin, T. (2001, August 5-10). *Visualising and animating visual information foraging in context.* Paper presented at HCI International 2001, New Orleans, 1100-1104.

Chen, C., Cribbin, T., Kuljis, J., & Macredie, R. (2002). Footprints of information foragers: Behaviour semantics of visual exploration. *International Journal of Human-Computer Studies, 57*(2), 139-163.

Cribbin, T., & Chen, C. (2001, January 21-26). *Visual-spatial exploration of thematic spaces: A comparative study of three visualisation models.* Paper presented at Electronic Imaging 2001: Visual Data Exploration and Analysis VIII, San Jose, CA, 199-209.

Cribbin, T., & Chen, C. (2001, July 9-13). *Exploring cognitive issues in visual information retrieval.* Paper presented at the Eighth IFIP TC.13 Conference on Human-Computer Interaction, INTERACT 2001, Tokyo, Japan, 166-173.

Cribbin, T., & Chen, C. (2001, August 5-10). *A study of navigation strategies in spatial-semantic visualisations.* Paper presented at HCI International 2001, New Orleans, USA, 948-952.

Westerman, S. J., & Cribbin, T. (2000). Cognitive ability and information retrieval: When less is more. *Virtual Reality, 5*(1), 1-7.

Westerman, S. J., & Cribbin, T. (2000). Mapping semantic information in virtual space: Dimensions, variance, and individual differences. *International Journal of Human-Computer Studies, 53*(5), 765-788.

Westerman, S. J., Collins, J., & Cribbin, T. (2005). Browsing a document collection represented in two- and three-dimensional virtual information spaces. *International Journal of Human-Computer Studies, 62*(6), 713-736.

APPENDIX A: TOPICAL SCENARIO DESCRIPTIONS

1. Distinct aspect topic: Extinction

Description: The spotted owl episode in America highlighted U.S. efforts to prevent the extinction of wildlife species. What is not well known is the effort of other countries to prevent the demise of species native to their countries. What other countries have begun efforts to prevent such declines? A relevant item will specify the country, the involved species, and steps taken to save the species.

High recall query: "extinction OR endangered species"

Topical scenarios: Extinction

Aspect Number	Associated retrieved documents (numbered by rank similarity to the original query)	Aspect definition
1	22	Finland - saima ringed seal
2	11, 16	Brazil - golden lion tamarin
3	73, 36	Japan - Atlantic bluefish tuna, elephants
4	73	Int'l Commission for Conservation of Atlantic Tuna - Atl. blue tuna
5	112, 66	Kenya - elephants
6	90	Columbia - Andean condor
7	75, 99, 66, 14	South Africa - quagga, white rhino
8	78	Belize - jaguar, black howler monkey
9	31, 116, 96	Zimbabwe - rhino, elephants
10	8, 113, 70, 37	UK - capercaillie, tern, polecat, birds
11	16	Oman - Arabian oryx
12	None retrieved	EC - harp and ring seals
13	48	Spain - white-headed duck
14	None retrieved	Greece - elephants
15	123, 36, 14	Worldwide Fund for Nature - sea birds (long-tailed guillenot, shag, fulmar, little auk, Gr.

		North. Diver), elephants, panda, rhino, Bengal tiger, Barasingha deer.
16	34	Paraguay - teyu guazu iguana, cayman, boa constrictor
17	87	Poland - bison
18	16, 28	Indonesia - wild monkeys, chimps, Sumatra tiger
19	59, 3, 14	Cites - elephants
20	34	New Zealand - birds
21	87	Peru - vicuna
22	None retrieved	Canada - cod
23	13	India - tigers
24	18, 50	China - rhino, tiger
25	None retrieved	Romania - European mink
26	12	Zambia - elephant, black rhino

2. Overlapping aspect topic: Chunnel

Description: Impacts of the Chunnel - anticipated or actual - on the British economy and/or the life style of the British. Find as many DIFFERENT impacts of the sort described above as you can.

High recall query: "chunnel OR channel tunnel"

Topical scenarios: Chunnel 127 (retrieved documents 1-127), Chunnel 218 (retrieved documents 1-218).

Aspect Number	Associated retrieved documents (numbered by rank similarity to the original query)	Aspect definition
1	143, 39, 26, 116, 197, 138, 144, 160, 23, 165, 129	environmental impact
2	143, 9, 24, 64, 82, 75, 72, 36, 5, 19, 43	financing of high-speed rail line
3	133	cost of additional safety standards
4	67, 115, 106, 163, 1	merger (rationalization) of ferry companies
5	94, 203, 93, 98, 75, 86, 3, 19	location/relocation of industries because of Chunnel
6	203	loss of Japanese investors to Europe
7	203, 44, 2, 70, 197, 160, 97, 3, 6	changes in real estate market
8	186, 47, 85	increased quick visits to France
9	63	competition among railway companies
10	27, 184, 2, 16, 92, 38, 15, 29, 42, 48, 1, 12, 3, 56, 6	increased/improved rail freight/parcel/passenger service with Europe
11	89, 99, 30, 8, 39, 34, 2, 197, 25, 15, 183, 33, 86, 96, 3, 6, 90, 60	improved/harmed local economies because of rail lines
12	187, 30, 8, 93, 9, 198, 40, 26, 29	improved rail service (freight or passenger) within the UK
13	93, 131, 213, 75, 40, 147, 97, 183, 81, 165, 86, 96, 3, 6	changes in Keny economy/employment
14	131, 183, 96	increase chance for EC grants
15	10, 9, 24, 64, 82, 5, 37, 159, 101, 43	Chunnel impact on privitization of railroads

16	16, 163, 42, 12, 22	improvements in ferry/air services
17	31, 47	diseases entering UK
18	31, 47, 150, 57, 45	increased terrorism in UK
19	150	more drugs in UK
20	68, 163, 100, 97, 1, 22	changes in prices to cross Channel
21	21, 48	creation of British tunnel exports
22	87, 7	Brits driving on the Continent (esp. own cars)
23	85, 49, 7, 48, 81, 3	strengthen British links to EU and single market
24	85, 49, 45, 48, 11	improved British-French relationship
25	25	increased tourism anywhere on British island
26	45, 11	armed police in Britain
27	29	increased use of international shipments
28	3	removes psychological barrier of Channel

APPENDIX B: DOT LANGUAGE DEFINITION OF A MINIMUM SPANNING TREE

3. Example section of MST definition for the Extinction scenario

```
graph G {
19 -- 21 [weight=.933];
45 -- 66 [weight=.917];
15 -- 52 [weight=.854];
56 -- 65 [weight=.819];
27 -- 65 [weight=.812];
31 -- 112 [weight=.781];
41 -- 59 [weight=.755];
15 -- 56 [weight=.748];
33 -- 91 [weight=.735];
56 -- 107 [weight=.728];
10 -- 73 [weight=.728];
30 -- 65 [weight=.721];
21 -- 31 [weight=.721];
18 -- 50 [weight=.707];
50 -- 116 [weight=.693];
31 -- 59 [weight=.693];
49 -- 51 [weight=.686];
55 -- 88 [weight=.678];
1 -- 6 [weight=.678];
10 -- 26 [weight=.678];
:
:
64 -- 66 [weight=.436];
30 -- 35 [weight=.436];
24 -- 49 [weight=.436];
92 -- 126 [weight=.424];
90 -- 113 [weight=.424];
57 -- 110 [weight=.424];
23 -- 47 [weight=.424];
9 -- 115 [weight=.424];
1 -- 78 [weight=.412];
17 -- 40 [weight=.412];
1 -- 63 [weight=.412];
101 -- 126 [weight=.4];
92 -- 100 [weight=.4];
83 -- 97 [weight=.4];
62 -- 66 [weight=.4];
114 -- 122 [weight=.387];
95 -- 126 [weight=.387];
37 -- 70 [weight=.387];
46 -- 101 [weight=.374];
67 -- 95 [weight=.361];
}
```

APPENDIX C: DOCUMENT EXEMPLARS USED TO DEMONSTRATE LCD TERM APPLICATIONS

4. Document #3 (Extinction): Trade bans may save the whale, but not the elephant

BIO-DIVERSITY is the environmental lobby's latest buzz-word. Translated, it means the more species, the merrier. But this diversity appears to be under threat, at least according to statistics compiled by the World Conservation Monitoring Centre, which purport to show that species extinctions have risen rapidly over the past century. Humans cannot be blamed for the demise of all species, the extinction of the dinosaur being one obvious example. Let us accept, however, both that bio-diversity is worth preserving and that it is human beings who are responsible for the rise in extinctions in recent decades. What can be done to reverse the trend? The standard response, enshrined in numerous international conventions, is to ban economic exploitation of endangered species. Such a ban is the mechanism that the International Whaling Commission has used for over 40 years in its efforts to reverse the collapse in the number of blue and hump-back whales. A fortnight ago, at its 45th annual meeting, the IWC decided not to lift its ban on commercial whaling. The Convention on International Trade in Endangered Species (CITES) hopes that by banning ivory trade it can reverse the demise of the African elephant, whose numbers halved between 1979 and 1989, implying a loss of over 700,000 elephants. The ban was imposed in 1989 and reconfirmed a year ago, despite opposition from southern African governments. Do such trade bans work? Not always, argues Mr Timothy Swanson in the latest issue of Economic Policy. A ban on commercial fishing may be an effective way of protecting threatened oceanic species from excess farming, he argues, but halting trade in elephant products is not. Whales are threatened with extinction for three reasons: they breed slowly; they are cheap to catch relative to the market price for whale products; and access is available to anyone with a boat and the necessary expertise. If access to whale farming were controlled by quotas, their numbers could theoretically be stabilised. In practice, a ban on commercial whaling is a more effective way of reducing the economic return for fishermen and thus discouraging their capture. But the success of this policy for preserving the whales depends on the assumption that, left to their own devices, whales would breed freely and flourish. The same argument does not apply to elephants, which do not have the luxury of living in huge oceans. The survival of land species, especially such large and potentially destructive animals as elephants, depends on the willingness of humans to preserve their habitat. This depends on their economic return, compared to other land uses. It is because investing in elephants has not been sufficiently profitable, at least in the poorest African states, that elephants are threatened. While the proximate cause for the decline in the number of African elephants in recent years seems to be the availability of high-power weapons and the relatively lucrative ivory trade, elephants were killed in large numbers because government did not find it profitable to stop the poachers. In the 1980s, four countries alone - Tanzania, Zambia, Zaire and Sudan - lost 750,000 elephants. All spent less than Dollars 20 a year per square kilometre on park management. Zimbabwe, by contrast, spent Dollars 194 and saw its elephant stock rise by over 20,000. Little wonder that the higher spending governments of southern Africa are arguing for the ban on the ivory trade to be lifted. Banning trade reduces the incentive for African countries to keep poachers out of the parks or to preserve elephant-friendly habitats. If African elephants are to be saved, the economic return on elephant farming must be increased, rather than lowered, perhaps by granting export quotas to countries willing to invest in keeping the poachers out. Free trade in ivory may not be environmentally friendly, but neither is a trade ban. Timothy Swanson, 'Regulating endangered species', Economic Policy 16, April 1993. Cambridge University Press. -- INTERNATIONAL ECONOMIC INDICATORS: BALANCE OF PAYMENTS -- Trade figures are given in billions of European currency units (Ecu). The Ecu exchange rate shows the number of national currency units per Ecu. The nominal effective exchange rate is an index with 1985=100. -- UNITED STATES -- Visible Current Ecu Effective trade account exchange exchange Exports balance balance rate rate -- 1985 279.8 -174.2 -159.7 0.7623 100.0 1986 230.9 -140.6 -150.0 0.9836 80.2 1987 220.2 -131.8 -141.6 1.1541 70.3 1988 272.5 -100.2 -107.0 1.1833 66.0 1989 330.2 -99.3 -91.8 1.1017 69.4 1990 309.0 -79.3 -70.9 1.2745 65.1 1991 340.5 -53.5 -3.0 1.2391 64.5 1992 345.8 -64.1 -48.2 1.2957 62.9 2nd qtr. 1992 86.8 -16.9 -14.4 1.2717 63.6 3rd qtr. 1992 80.6 -17.7 -11.4 1.3831 60.1 4th qtr. 1992 91.5 -17.4 -17.4 1.2658 64.2 1st qtr. 1993 95.8 -21.8 1.1841 66.4 May 1992 28.4 -6.0 na 1.2676 63.8 June 29.2 -5.2 na 1.3039 62.3 July 27.3 -5.5 na 1.3693 60.5 August 25.9 -6.2 na 1.4014 59.8 September 27.3 -6.0 na 1.3786 60.2 October 29.4 -5.5 na 1.3210 62.1 November 30.5 -6.3 na 1.2372 65.1 December 31.6 -5.6 na 1.2391 65.3 January 1993 31.3 -6.4 na 1.1968 66.4 February 31.4 -6.7 na 1.1767 66.7 March 33.1 -8.7 na 1.1789 66.2 April na 1.2214 64.3 -- JAPAN -- Visible Current Ecu Effective trade account exchange exchange Exports balance balance rate rate -- 1985 230.8 76.0 64.5 180.50 100.0 1986 211.1 96.2 86.9 165.11 124.4 1987 197.3 86.1 75.5 166.58 133.2 1988 219.8 80.7 66.6 151.51 147.3 1989 245.3 70.5 52.4 151.87 141.9 1990 220.0 50.1 28.3 183.94 126.0 1991 247.4 83.1 62.9 166.44 137.0 1992 254.8 101.8 89.8 164.05 142.9 2nd qtr. 1992 63.9 26.1 23.1 165.60 139.9 3rd qtr. 1992 61.5 23.7 20.1 172.79 139.6 4th qtr. 1992 65.2 26.9 24.8 155.57

149.7 1st qtr. 1993 72.8 29.9 30.6 143.41 158.5 May 1992 21.1 9.6 8.8 165.57 139.7 June 21.3 8.3 6.3 165.32 141.7 July 20.5 8.1 6.9 172.22 139.2 August 19.9 7.4 5.9 177.11 137.0 September 21.1 8.2 7.2 169.05 142.5 October 21.3 8.9 7.7 159.93 148.2 November 22.1 9.1 9.3 153.22 150.3 December 21.7 8.8 7.8 153.57 150.7 January 1993 23.3 8.9 7.4 149.62 151.3 February 24.0 10.4 9.3 142.00 159.2 March 25.5 10.6 13.8 138.61 164.4 April 137.17 167.8 -- GERMANY -- Visible Current Ecu Effective trade account exchange exchange Exports balance balance rate rate -- 1985 242.8 33.4 21.7 2.2260 100.0 1986 248.6 53.4 40.3 2.1279 108.8 1987 254.3 56.8 39.8 2.0710 115.3 1988 272.6 61.6 42.9 2.0739 114.6 1989 310.2 65.3 52.3 2.0681 113.5 1990 323.9 51.8 37.2 2.0537 119.1 1991 327.4 11.2 -16.2 2.0480 117.7 1992 330.3 16.4 -19.9 2.0187 121.2 2nd qtr. 1992 81.1 3.6 -5.2 2.0511 118.7 3rd qtr. 1992 83.9 6.4 -6.4 2.0221 122.1 4th qtr. 1992 82.1 3.4 -4.1 1.9593 125.0 1st qtr. 1993 1.9348 125.6 May 1992 26.5 0.6 -2.1 2.0551 118.4 June 25.1 0.6 -2.1 2.0498 119.1 July 28.3 1.0 -3.8 2.0410 120.7 August 27.7 3.1 -0.7 2.0326 122.0 September 27.8 2.3 -1.8 1.9927 123.6 October 28.6 2.4 -1.3 1.9564 125.7 November 26.8 0.9 -0.3 1.9634 124.0 December 26.7 0.0 -2.5 1.9581 125.3 January 1993 25.8 -2.7 1.9327 125.3 February -2.7 1.9318 125.8 March 1.9399 125.7 April 1.9483 125.5 -- FRANCE -- Visible Current Ecu Effective trade account exchange exchange Exports balance balance rate rate -- 1985 133.4 -3.6 -0.2 6.7942 100.0 1986 127.1 0.0 3.0 6.7946 102.8 1987 128.3 -4.6 -3.6 6.9265 103.0 1988 141.9 -3.9 -3.4 7.0354 100.8 1989 162.9 -6.3 -3.6 7.0169 99.8 1990 170.1 -7.2 -7.2 6.9202 104.8 1991 175.4 -4.2 -4.7 6.9643 102.7 1992 182.4 4.3 2.1 6.8420 106.0 2nd qtr. 1992 46.2 1.5 0.9 6.9122 104.4 3rd qtr. 1992 45.2 0.9 0.0 6.8536 106.6 4th qtr. 1992 45.5 1.0 2.3 6.6529 109.3 1st qtr. 1993 6.5633 110.0 May 1992 15.0 0.59 1.38 6.9090 104.5 June 15.4 -0.16 -0.54 6.9001 104.9 July 15.5 0.87 -0.16 6.8872 106.0 August 14.2 -0.45 0.25 6.8944 106.3 September 15.6 0.49 -0.04 6.7792 107.6 October 15.1 0.11 0.99 6.6368 110.0 November 15.1 0.05 0.13 6.6426 109.0 December 15.3 0.85 1.14 6.6793 108.9 January 1993 13.7 0.48 0.69 6.5539 109.7 February 6.5442 110.3 March 6.5919 109.9 April 6.5875 110.5 -- ITALY -- Visible Current Ecu Effective trade account exchange exchange Exports balance balance rate rate -- 1985 103.7 -16.0 -5.4 1443.0 100.0 1986 99.4 -2.5 -1.4 1461.6 101.4 1987 100.7 -7.5 -2.1 1494.3 101.2 1988 108.3 -8.9 -8.0 1536.8 97.8 1989 127.8 -11.3 -17.0 1509.2 98.6 1990 133.6 -9.3 -18.0 1523.2 100.6 1991 137.0 -10.5 -28.9 1531.3 98.9 1992 137.9 -8.0 -11.0 1591.5 95.7 2nd qtr. 1992 35.8 -3.6 -2.9 1546.3 98.5 3rd qtr. 1992 32.9 0.5 -5.5 1564.6 98.2 4th qtr. 1992 34.9 0.0 0.0 1719.4 87.1 1st qtr. 1993 -4.9 1827.9 80.5 May 1992 11.5 -1.9 -0.9 1546.6 98.5 June 12.7 -0.5 -1.0 1550.3 98.5 July 13.9 0.8 -1.9 1546.2 99.5 August 7.7 1.1 -1.5 1543.4 100.1 September 11.3 -1.4 -2.0 1604.1 95.0 October 12.4 0.1 1.5 1723.8 87.3 November 10.8 -1.2 -0.9 1687.0 88.7 December 11.6 1.1 -0.6 1747.5 85.6 January 1993 9.7 0.4 -3.1 1784.9 82.5 February 0.6 1822.3 80.8 March -2.4 1876.4 78.5 April 1871.4 79.0 -- UNITED KINGDOM -- Visible Current Ecu Effective trade account exchange exchange Exports balance balance rate rate -- 1985 132.4 -5.7 4.7 0.5890 100.0 1986 108.3 -14.2 0.1 0.6708 91.6 1987 112.3 -16.4 -6.4 0.7047 90.1 1988 120.9 -32.3 -24.3 0.6643 95.5 1989 137.0 -36.7 -32.3 0.6728 92.6 1990 142.3 -26.3 -23.8 0.7150 91.3 1991 147.7 -14.7 -9.0 0.7002 91.7 1992 145.1 -18.7 -16.1 0.7359 88.4 2nd qtr. 1992 38.0 -4.5 -4.4 0.7034 92.3 3rd qtr. 1992 36.4 -4.5 -3.0 0.7261 90.9 4th qtr. 1992 34.3 -5.4 -4.6 0.8015 79.8 1st qtr. 1993 0.8017 78.5 May 1992 13.0 -1.2 -1.17 0.7000 92.8 June 12.5 -1.3 -1.30 0.7027 92.9 July 12.3 -1.6 -1.06 0.7137 92.5 August 12.3 -1.6 -1.09 0.7219 92.0 September 11.8 -1.3 -0.85 0.7428 88.2 October 11.5 -1.4 -1.19 0.7969 80.8 November 11.4 -1.7 -1.50 0.8100 78.3 December 11.5 -2.2 -1.93 0.7976 80.0 January 1993 0.7809 80.6 February 0.8179 76.8 March 0.8061 78.2 April 0.7894 80.5 -- All trade figures are seasonally adjusted, except for the Italian series and the German current account. Imports can be derived by subtracting the visible trade balance from exports. Export and import data are calculated on the FOB (free on board) basis, except for German and Italian imports which use the CIF method (including carriage, insurance and freight charges). German data up to and including June 1990, shown in italics, refer to the former West Germany. The nominal effective exchange rates are period averages of Bank of England trade-weighted indices. Data supplied by Datastream and WEFA from national government and central bank sources.

5. Document #31 (Extinction): Elephants in their sights: The arguments for lifting the ivory trade ban

A Zimbabwean villager had a blunt riposte to the world's 'elefriends' gathering in Kyoto this weekend, intent on maintaining a ban on ivory trade: 'Elephants eat people's food, and people are dying of hunger.' The question of whether to lift the ban will be among the most controversial issues this week at the triennial meeting of the Convention on International Trade in Endangered Species (Cites). As a test case for the effectiveness of trade measures in achieving environmental ends, it will provide important signals for action in defence of endangered animal and plant species. Although elephant populations have recovered in some areas, such as Zimbabwe, since the imposition of a ban on ivory trade in 1989, the species remains in danger. There is a heated debate over the extent to which the ban on trade has been responsible for the slim, localised recovery and whether extending the life of ban will sustain or undermine the future of the elephant. The danger facing the elephant is not in dispute. Africa's elephant population slumped from 1.2m to 600,000 between 1980 and 1988. Total trade in unworked ivory rose from about 200 tonnes a year in 1950 to about 1,000 tonnes a year in 1980, and remained at this level throughout the 1980s. The total of ivory exported between 1979 and 1988 accounted for more than 700,000 elephants. Since the imposition of the trade ban at the last Cites meeting in 1989, there has been progress. Demand in Europe and the US for ivory has virtually disappeared, according to customs statistics. Poaching has not been eradicated, but in certain countries (notably in southern Africa) success has been such that elephant herds now need to be culled. But can the trade ban be credited for these successes? And can they be maintained? Evidence derived from the ivory trade debate suggests that the ban is valuable as a source of publicity and has helped to reduce consumer demand for ivory products. As long as legal ivory cannot be distinguished from illegal ivory, a total ban also simplifies the international policing effort. But there are also concerns among conservation groups that success is only partly due to the ban and that illegal trade channels may expand and reverse the progress which has been achieved. Even the Worldwide Fund for Nature, a committed campaigner for maintaining the ban, concedes in a report published this month: 'These dramatic drops (in poaching) were brought about through increased law enforcement efforts.' African governments which are calling for a lifting of the ban base their case on the need to strike a balance between their rural communities and the local elephant population. The concern underpinning Zimbabwe's call for a resumption in trade is that the rising number of elephants, with their voracious appetites, are threatening the livelihood of the agricultural community. While they have no economic value, there is no incentive for villagers to tolerate them. The Zimbabwean government insists, therefore, that a controlled resumption of trading in ivory would provide villagers with an incentive to tolerate and protect local elephant populations. An alternative strategy is to promote Safari tourism. According to research by Dr Edward Barbier at the London Environmental Economics Centre, the annual value of ivory exports from Africa amounted to between Dollars 50m and Dollars 60m in the 1980s: 'Other values of the elephant, such as its importance to tourism earnings, may be considerably more significant,' he says. In a recent study of the economic value of elephants, colleagues at the Centre pointed out that in Kenya, earnings from viewing elephants came to about Dollars 25m a year - about 10 times the estimated value of poached ivory exports from Kenya. But despite the array of arguments mustered in favour of lifting the ban, such a policy poses clear dangers. Resumed trading would provide an avenue for poachers in countries where elephants remain under threat to 'launder' illegal ivory by mixing it with ivory from legal culls. Tests can now identify the DNA characteristics of individual pieces of ivory. It is therefore technically possible to identify poached ivory. Just how simply or effectively such tests could be administered is another matter. It is clear that no retail purchaser of ivory could tell the difference on a shop shelf, so oversight would need to be effective at source. Environmentalist groups, such as the Environmental Investigation Agency also emphasise the practical difficulties of monitoring the ivory at its source. They argue that corruption in large parts of Africa, and military conflict in Mozambique and elsewhere, as reasons for doubting whether DNA testing could be effective in preventing poached ivory from reaching world markets. Thousands of miles from the arguments in Kyoto, the elephant is unable to rest easily. Its security will not be guaranteed until demand in end-user countries has been staunched; until village communities in Africa can see some economic benefit from preserving this immensely disruptive pachyderm; and until the corruption and conflict on which poaching thrives have been brought under control.

6. Document #197 (Chunnel): Fury over French connection: The response to the Channel rail link route

You have to travel at least 50 miles north-west from Folkestone along the route of the planned high-speed rail link from the Channel tunnel to London before you meet people who are enthusiastic about the project. **At Stratford, in east London - not far from the link's final destination at St Pancras station - a young woman at a jobs agency is thrilled that the trains will stop just minutes from where she works. 'There will be a new shopping centre, new street furniture and green spaces -I think it will be the best thing that has happened to Stratford for a long time,' she says.** In contrast to this flash of enthusiasm a trail of anger, confusion and tragedy on small scale runs through the Kent countryside. For six years residents have been in limbo, not knowing when the link will be built, where it will be built and whether they will qualify for compensation. If anyone knows the meaning of planning blight it is Stuart Smith. Two-and-a-half years after moving from a house in Lenham Heath which was threatened by the original route of the link, he now faces the prospect of a second move. Revised proposals for the rail route, unveiled earlier this week by John MacGregor, the transport secretary, could bring trains within yards of The Mount, a Pounds 300,000 oak timber-framed farmhouse in the hamlet of Ram Lane near Ashford to which Mr Smith and his family moved in 1991. 'We got a sensible price from British Rail for our last house but there was no allowance made for the upset it caused,' he says. 'We had lived in that house for 22 years and you can't compensate for that.' Mr Smith is just one of thousands of home-owners who live near the 68-mile railway line, which is intended to speed sleek express trains at up to 140mph through the Garden of England. If the government can persuade private companies to invest at least half the Pounds 2.6bn cost of the project, Eurostar trains should be slicing through the Kent countryside by 2002. But before work starts on the line, Union Railways, the British Rail subsidiary working on the early stages of the project, hopes to have resolved the problems caused by years of planning blight. If the scale of resident's protests is maintained, Union Railways' negotiators are in for a tough time. While David and Ivy Hilliger, at Westnell Lane, are relieved that the new route will no longer run 30 yards from their back garden, they are not celebrating. 'I don't think Union Railway realises just how much it is affecting people's lives mentally. People in the village want to retire, but don't know whether they will qualify for compensation.' Mr MacGregor claimed that only 40 homes would be in the direct line of the route. But this small number is only arrived at because of the narrowness of the corridor which the government intends to 'safeguard' - that is, formally declare as the line of the route, a move which triggers the right to statutory compensation. A final decision on the corridor has yet to be taken by the government, but Union Railways says it is unlikely to be much wider than the 36 metres between the fences required to protect a twin railway track. This is in marked contrast to the 240-metre wide corridor declared by BR on its first route, abandoned in 1991, which would have run through south London into Waterloo. BR spent Pounds 140m buying homes along this corridor, acquiring practically the whole of the village of South Darenth and large swathes of Peckham. It has since been selling these properties off at a large loss. Union Railways says it has chosen a narrower corridor to keep costs down and reduce the area of blight. 'BR got its fingers burned last time but now Union Railways is being far too cautious,' says a Kent County Council official. In a recent study of the rail link project, the council called for a more generous compensation scheme for home-owners outside the 36-metre corridor. One problem facing residents is that there is no agreement on standards which should be applied to the disturbance that would be created by a fast railway line. The government and the local authorities involved are still discussing noise and vibration criteria. Kent County Council complains the present limits under which compensation is awarded are based on surveys of road noise carried out 20 years ago. Motorways create a background hum but fast trains cause a sudden rush of sound, it says. But such technical details are of little concern to Pat and David Henderson. Their three-bedroom semi-detached home on an estate at Pepper Hill near Gravesend will be just 24 feet above a planned tunnel. Pepper Hill and Ashford are the only two parts of the route which may be changed. The Hendersons were hoping to sell up, buy a smaller house and put some money in the bank. But they have seen their home plummet in value from nearly Pounds 100,000 to Pounds 60,000 in a few years. 'Estate agents say they won't even put us on their books,' says Mrs Henderson. Despite the uncertainties surrounding compensation, Union Railways says it is prepared to be more flexible than the law provides for. In theory, it cannot purchase properties compulsorily before the passage of the rail link bill through parliament, expected to take at least two years. But it says it will offer compensation as soon as an order safeguarding the route is published in the next few weeks. Compensation legislation allows it to offer market value plus up to 10 per cent. Home-owners in the direct line of the route will automatically be eligible for compensation but people living near but not on the line will have to apply to Union Railways. Estate agents will be asked to value properties. Only those 'very close' to the rail are likely to be bought while double glazing may be available for those living further away. Double glazing would not assuage the fears of Arthur Reeves, who runs a used car business next door to the Garden of England Mobile Home Park outside Harrietsham. Its pretty setting - old army ambulances and rusty Rolls Royces stand between the trees surrounding the house - is ruined by the drone of the M20 motorway which runs in front of the house. The Channel link, according to the latest plans, will run on his side of the motorway, compounding an already serious noise problem. 'They say that when the link is complete, there will be trains running every 10 minutes.' says Mr Reeves.

Union Railways insists that local residents' fears are exaggerated and that modern railways are built to such high standards that they will not create the noise and vibration many expect. But even if this turns out to be the case, the insecurity is causing unhappiness in Kent. 'There has been a lot of illness, and a lot of mental strain,' says Mrs Margaret Bottle, of Harrietsham, pointing out houses purchased by BR and now standing empty. 'It has been hanging over us for so long, and we can't get any sense out of Union Railway. They will not give us any straight answers. We won't know what any of this means to us, until the first train makes its first trip.

7. Document #75 (Chunnel): Blue Circle plans Pounds 500m Channel rail link station

BLUE CIRCLE Industries proposes to build a privately financed international railway station for the planned Channel tunnel link near Dartford in north-west Kent. The station, which could cost up to Pounds 500m, would occupy about 250 acres of chalk quarries and waste ground, part of a 2,500-acre site owned by Blue Circle, Britain's biggest cement producer. **The land is on the route of the Pounds 2.5bn rail link, to be financed jointly by the private and public sectors, which was announced by the government earlier this week**. Blue Circle's plans include hotels, a conference centre, offices and shops as well as international and domestic passenger terminals. The station could include an interchange with British Rail lines eastward to the Medway towns and westward through south-east London to the centre of the capital. The proposals are to be submitted shortly to ministers and **Union Railways**, the BR subsidiary responsible for developing the **high-speed** link, according to Mr Mark Pennington, Kent development manager for Blue Circle Properties. He will also seek meetings with banks and potential investment partners. Blue Circle said the cost of the station would be met out of proceeds from the commercial development. The company would provide the land as its contribution to the investment. It believes the project could be completed without any public finance. Blue Circle said it would begin local consultation shortly. The proposals were supported by Dartford District Council and by Mr Bob Dunn, MP for Dartford and chairman of the Conservative backbench transport committee, it said. Mr Pennington said the construction of a station providing a direct link to continental Europe could act as a catalyst for a much bigger development of the entire 2,500 acres owned by Blue Circle. This includes the Eastern Quarry which currently supplies the group's Northfleet cement works. Plans could eventually involve the construction of a new town with up to 12,500 homes, offices, shops, a conference centre, business and industrial parks, recreational and social amenities, creating some 34,000 jobs. The site is just east of the Dartford bridge and tunnel which carry the M25 across the River Thames. The development would be in line with plans to stimulate investment along the Thames. Mr Michael Howard, the environment secretary, this week established a 'task force' of civil servants to consider plans for redeveloping the corridor. He has not so far offered any contribution from the public sector.