Related Scientific Information

A study on user-defined relevance

Ulises Cerviño Beresi

A thesis submitted in partial fulfilment
of the requirements of
The Robert Gordon University
for the degree of Doctor of Philosophy

September 2011

Acknowledgments

Every thesis should have a section which acknowledges the contributions of those who directly, or indirectly, have helped create it. This thesis is no different and as such contains said section. It is very possible that this section is both incomplete and unjust in giving credit. I have tried my best so that this is not the case.

This thesis is dedicated to my wife Jorgelina, to whom I say thank you for your patience, understanding and endurance. However, and most importantly, thank you for trusting in me and joining me in this pursuit. Finishing this project would not have been possible without your support.

To my friends and family. In particular to my parents, Maria Rosa and Eduardo, who taught me the importance of approaching the world with a clear and educated mind, thinking for myself, questioning everything, and the importance of always keeping an open mind.

To my supervisory team, which should not go unnoticed as much of my success is due to them. To David, thanks for being direct. You taught me how to communicate both praise and criticism frontally. To Ian, thank you for trusting in both my abilities and judgement. When times looked bleak you provided me with encouragement and advice. I have to also admire your boldness: you agreed to supervise me after knowing me for about one hour (how silly was that?) To Mark, thank you for your sound advice. I will never forget your words: *"a good thesis is a finished one."* To Patrik, thank you for being there.

To both examiners of this thesis, Prof. Ian Allison and Prof. Birger Larsen. Whatever clarity and insights are contained in this work, they are due to your insistence. Not only your suggestions led to a much improved dissertation but your insights and questions

helped me realise the importance of communicating research in a clear and concise manner.

To Sandy, Thierry and Malcolm, who provided hours of amusement and sanity-preserving tea breaks. To Prof. Dawei Song, who supported and encouraged me endlessly; I would probably be still writing up without his support. Last, but certainly not least, to you, Yunhyong, without your constant challenges much of the quality of my work would be below par.

As some portions of this dissertation have been presented in various forums, each individual publication went through a review process which provided me with insightful comments from the reviewers. These were invaluable in improving the final dissertation. So, thank you anonymous reviewer.

Many others are to be thanked, however you are too many to be listed here. So, thank you (you know who you are).

To those who never stopped asking "are you done yet?"

Abstract

This dissertation presents an investigation into the manifestations of relevance observed in the context of related scientific information. The main motivation is to observe if researchers, in the context of knowledge discovery, use different criteria to judge the relevance of the information presented. Additionally, the effects that discipline and research experience background may have on these manifestations are investigated.

The scenario selected to carry out the observation is that of Literature Based Discovery (LBD). LBD is a trial-error interactive search strategy, developed by Swanson (1986a), which supports the finding and retrieving of complementary bodies of literature – sets of articles that are bibliographically non-interactive yet logically connected. Research scientists from three different disciplines and research experience backgrounds are observed while they interact with an LBD system built for the purposes of this study. Their cognitive processes and interactions are recorded and analysed. To aid in the analysis of the data, the concept of relevance criteria profiles is developed. Relevance criteria profiles are a technique to count and group the expressions of relevance criteria as observed during the search sessions. These offer the possibility of aggregating the observations into group profiles as well as the ability to measure the (dis)similarities that may arise in between profiles. As relevance criteria profiles provide a global view of the criteria used to judge relevance, a complementary visualisation technique is also developed. This technique displays the relevance judgement processes, as well as the interactions, in a sequential fashion allowing the researcher to perform temporal analyses on the session data.

The results show that researchers do use a variety of criteria when judging the relevance of information in the context of LBD. Moreover, individuals use these criteria in different frequencies; both discipline and research experience background seem to influence these frequencies however they may not be the only intervening factors. The observed interaction patterns suggest that researchers approach the problem in two stages: i) an initial more exploratory stage followed by ii) a more focused and engaged stage. The main contribution of this thesis is the observation of these manifestations of relevance together with the interaction patterns. The final recommendation offered is that the multi-dimensional nature of relevance in this context should be addressed when evaluating LBD systems. Additionally, it is acknowledged that certain interaction behaviours may also be used during the design and testing of such systems.

Contents

1 Introduction **2**

1.1 Aims of the Dissertation 5

1.2 Structure of the Dissertation 6

2 Literature Review **8**

2.1 Literature Based Discovery 10

2.2 LBD Search Models 11

2.2.1 Open Model — Hypotheses Generation 12

2.2.2 Closed Model — Hypotheses Validation 13

2.3 LBD Systems 14

2.3.1 The Models 14

2.3.2 Text Modelling 15

2.3.3 Evaluation Methods 18

2.4 On Relevance 23

2.5 Evaluation of IR Systems 26

2.6 Evaluation of LBD Systems 31

2.7 Summary 33

3 Study **39**

3.1 Method 40

3.1.1 First session 41

3.1.2 The offline processing 43

3.1.3 Second session 43

3.2 The System 45
3.2.1 First Session 46
3.2.2 The Offline Processing — An Automatic Open Search 47
3.2.3 Second Session 49
3.3 User groups 51
3.4 Collections 52
3.5 Search Tasks 53
3.6 Measurements 55
3.6.1 Written Data — forms and questionnaires 55
3.6.2 Verbal Data — Talk Aloud Protocols 57
3.7 Data Analysis 61
3.7.1 Interaction 62
3.7.2 Intent 62
3.7.3 Relevance Criteria 62
3.7.4 Relevance Criteria Profiles 64
3.7.5 Visualising Sessions 69

4 Results **73**

4.1 Participants 74
4.2 Collections 75
4.2.1 School of Computing 75
4.2.2 Information Management Group 76
4.2.3 School of Pharmacy 77
4.3 The Nature of the Data 77
4.4 Interaction 77
4.5 Intent 78
4.6 Interpretation of the Relevance Criteria 78
4.7 Relevance Criteria Profiles 84
4.7.1 Global Profile 85
4.7.2 School Profiles 85

4.7.3 Research Experience Profiles 88
4.7.4 Relevance Criteria - the Observations 88
4.8 Summary 108

5 Results II **112**
5.1 Profile Similarities 114
5.2 Relevance Judgement Processes 117
5.3 Interactions Revisited 130
5.3.1 Selecting Target Topics 130
5.3.2 Selecting Intermediate Topics 133
5.3.3 Assessing the Related Literature 137
5.4 Sessions Visualised 142
5.5 Summary 149

6 Discussion **154**
6.1 Verbal Protocols 155
6.2 Relevance Criteria 157
6.3 Measuring Profile Similarities 161
6.4 Relevance Judgement Processes 162
6.5 Interactions 165
6.6 Session Visualisation 167
6.7 Known Limitations 171
6.8 Summary 172

7 Summary and Conclusions **174**
7.1 The Three Research Questions 175
7.2 Designing and Evaluating LBD Systems 177

Appendices **179**

A Forms **180**

B Publications **189**

List of Tables

3.1 Research experience categories. 52

3.2 Codes used during the transcription step of the protocol 60

3.3 Encoding used to tag the utterances that express a relevance criterion . . . 65

4.1 Number of participants per school/group for which valid data was gathered 74

4.2 Queries issued to the search engine for constructing the collection for the Information Management Group 76

4.3 Number of occurrences for each criterion according to the global relevance criteria profile 85

4.4 Restrictions for variable j according to each grouping of participants 86

4.5 Number of occurrences for each criterion according to each school relevance criteria profile 87

4.6 Restrictions for variable j according to each grouping of participants 88

4.7 Number of occurrences for each criterion according to each research experience level relevance criteria profile 89

5.1 Average use (averaged across participants that expressed using them) of relevance judgement processes of complexity n 122

5.2 Frequency of each criterion as distributed across single criterion rule uses. . 124

List of Figures

2.1 The LBD models in action. The inner box depicts the open model in which a searcher is interested in forming hypotheses about the topics of interest. The outer box depicts the closed model; a model in which searchers either reject or find evidence for pursuing the verification of hypotheses. 13

3.1 Search task introduced during the first meeting with participants 43
3.2 Example search task introduced during the second meeting with participants 44
3.3 User interface of the system participants used during their first session . . . 46
3.4 An example document. On the top left the document identifier is displayed. 47
3.5 Visual representation of the open search algorithm. The first step is to model the topics contained in the user submitted documents and retrieve more documents about them (step a). The second step is to model the topics contained in the pooled documents and retrieve even more documents (step b). The final step is to model the topics contained in each document set retrieved in the previous step. 48
3.6 Topic tree. The A node is the participant's initial topic. The inner nodes, the B_i nodes, are the immediately related topics as described by both the literature and the topic extraction algorithm. The tree leaves are the indirectly related, to the initial topic A, C_{ij} topics. 49
3.7 First screen users see when doing the second part of the study. In the image the terms representing the initial topic are presented on the left panel whereas the initial C topics, also in the form of terms, are displayed. . 50

3.8 Navigation screen. The top panel (a) displays both the initial topic (listed as "Your topic") and the potentially related topic (listed as "Related topic"). The middle panel (b) lists the intermediate topics. The bottom panel (c) contains the supporting literatures. 51
3.9 Search task given to all research students that participated in the study . . 54
3.10 Search task given to all researchers that participated in the study 55
3.11 Search task given to all senior researchers that participated in the study . . 56
3.12 A typical relevance criteria profile. Frequencies are normalised, hence the y axis varies between 0 and 1. 66
3.13 An example with four relevance criteria and interactions plotted. 70
3.14 An example with four relevance criteria plotted. Interactions are further encoded and plotted accordingly. 70

4.1 Distribution of participants per research level and school 75
4.2 Relevance criteria profile of the *global* group. Values in the y axis vary between 0 and 1. 86
4.3 School profiles plotted together. 90
4.4 Research experience profiles plotted together. 90
4.5 The profiles of the schools, normalised within criteria, plotted together. . . 91
4.6 Research experience profiles, normalised within criteria, plotted together. . 92

5.1 JS Divergence scores between participants. Each cell represents a divergence score between two participants (rows/columns represent participants.) It must be noticed that the closer to 0 the score, the more similar the two profiles are. This is in line with traditional heatmap plots where red is used for higher activity. 115
5.2 Total number of participants that used, at least once, a relevance judgement process of complexity n 121
5.3 Average use of relevance judgement processes of length n. Bars represent a standard deviation. 121

5.4 Navigation screen. The top panel (a) displays both the initial topic (listed as "Your topic") and the potentially related topic (listed as "Related topic"). The middle panel (b) lists the intermediate topics. The bottom panel (c) contains the supporting literatures. 134

5.5 Proportion of interactions with the left (right) panel as the session progresses. The two curves mirror each other. For every interaction observed the proportion of one increase while the proportion of the other decrease, e.g. if an interaction with the left panel is accounted for at value 10 of the x axis and the curve for the interactions with the left panel increases accordingly while the curve for the interactions with the right panel decreases. 140

5.6 The anatomy of an anomalous search session. 143

5.7 Visual representation of the search session of participant 2 144

5.8 Visual representation of the search session of participant 19 147

5.9 Frequency of the relevance judgement processes by complexity. 148

5.10 Repeated expressions of *depth/scope/specificity* 149

6.1 Feedback form. Several criteria are present which, when combined, present a more informative judgement for why the video is *worth seeing*. 162

6.2 Participant 2 – First uses of relevance criteria 168

6.3 Time annotated graphs, in parallel for better comparison. 170

A.1 Form used to capture the demographics. 181

A.2 Form used during the first search session. 182

A.3 Research student simulated work task situation. 183

A.4 Researcher simulated work task situation. 184

A.5 Senior researcher simulated work task situation. 185

A.6 Form used to capture the documents selected as well as the topics they support. 186

A.7 Form used at the end of the second search session (page 1). 187

A.8 Form used at the end of the second search session (page 2). 188

Chapter 1

Introduction

As pools of information increase in size and complexity, the mechanisms to retrieve information from them must evolve accordingly. The discipline that deals with the issues involved in the design and testing of these mechanisms is called Information Retrieval (IR). Traditionally, the retrieval of information has been based on the matching of information objects to user requests. To perform the matches, keys are extracted, from both information objects and user requests, and they are compared. This procedure, however, places the burden of choosing the right keys on the users. Hence, users are faced with the problem of having to select the keys that *they* think are likelier to guide them to the *relevant* information.

A notion central to this matching procedure is that of *relevance*; IR is concerned with searching for and retrieving *relevant* information, not just any kind of information. However, this notion of relevance is elusive. At a basic level one could talk about relevance in a topical sense (also referred to as "aboutness"). Topical relevance is the simplest estimation of true relevance and the one usually embedded in the matching techniques of IR systems. However, *"...users can read into results a lot more than correspondence between noun phrases or some such in queries and objects, used primarily by systems for matching...and do find other information objects or other information relevant to their problem that is not retrieved by a system for a variety of reasons, for example not reflected in the query to start with"* (Saracevic 2006).

In a series of articles, Swanson provides examples and explores the possibility of deriving relevance relations between seemingly disconnected areas of medicine (Swanson 1986a, Swanson 1988b, Swanson 1990). Swanson (1986c) attributes the existence of these knowledge gaps to the natural fragmentation of science and defines these relevance relations as Undiscovered Public Knowledge (UPK) alluding to the fact that while the knowledge was implicit in the literature bodies, nobody had noticed the links before. The discovery of these hidden relationships –or the derivation of these relevance relations– led Swanson to systematically investigate this phenomenon resulting in an interactive trial-and-error search strategy for finding these logically connected but non-interactive (in terms of cross- and co-citations) literatures (Swanson 1986c, Swanson & Smalheiser 1997b). The proposed search strategy is composed of two stages where the first stage is *"...an exploratory process intended to stimulate human creativity in perceiving connections that identify logically-related pairs of literatures..."* and is followed by *"...a method for eliminating all pairs except those that are noninteractive"* (Swanson 1989). These two stages of the search process have been labelled the *open model* and *closed model* of discovery respectively (Weeber, Klein, de Jong-van den Berg & Vos 2001). Eventually, the art of retrieving these disjoint bodies of literature became known as Literature Based Discovery (LBD).

Researchers have taken on Swanson's message and started developing specialised Information Retrieval (IR) systems; systems that are designed and tuned to find these complementary literatures and provide information as to why they are potentially so. While traditional IR system are designed with a *best match strategy* in mind, in LBD systems this approach is only complementary to an initial step of *discovery* in which topics of interest and their relationships are modelled and assessed. Research in LBD has been mostly focused on developing new techniques. Although initially based mostly on co-occurrences of words, recent techniques have included information from specialised databases such as the Unified Medical Language System (UMLS) and the Medical Subject Headings (MeSH) (Lowe & Barnett 1994). LBD is a relatively young discipline, though the array of tried and available algorithms is quite varied.

Evaluation of LBD techniques typically involve identifying whether key concepts from

the Swanson discoveries were promoted through the results returned by the system. For example, the appearance of phrases such as "blood viscosity" and "platelet aggregation", which were important in connecting "Raynaud's disease" with "Fish oil" (Swanson 1986a) are considered to be indicators of good performance of an LBD system. This approach is useful to measure the performance at a system level (albeit in a limited fashion), however this approach leaves several questions unanswered that, in the case of LBD, might be central to measuring the success of a system. Effectively, the evaluation method used resembles that of the system-driven tradition of evaluation of IR systems; a tradition based on test collections. Tests collections –comprised of a collection of documents, a collection of requests, and a collection of relevance judgements– are used to assess the performance of systems in laboratory conditions, i.e. controlled settings in which the products of IR systems are evaluated (Harman 1997). In this tradition potential end users are rarely involved in the process, hence cognitive factors such as information seeking and interaction processes can not be taken into consideration.

Borlund (2003b) proposes a framework for evaluating Interactive IR (IIR) systems that takes into account user-centred issues. The framework is composed of three experimental components which allow the evaluation of IIR systems to take place under realistic settings while still retaining the control observed in the evaluations conducted using the system-driven method. Realism is attained by the involvement of potential end-users as test persons, while control is retained by the use of simulated work task descriptions; a core component of the framework proposed by Borlund. Simulated work task situations describe a situation in which needs for information are triggered on users; users are led to a cognitive state in which information needs arise and need to be satisfied before they can move on. Because potential end-users are involved, relevance can be treated as a dynamic and multidimensional concept. Furthermore, interaction and information seeking processes can be considered; all factors that may be central to the evaluation of LBD systems.

1.1 Aims of the Dissertation

LBD systems are inherently interactive since it is through interacting with the system, assessing proposed relationships and inspecting the literature that researchers propose, verify or reject hypotheses. Cognitive issues, therefore, such as information seeking and interaction processes, are factors that should be taken into account during the evaluation of such systems.

As shown by Swanson, topicality is a poor estimation of relevance in this context and it is likely that several factors, for instance the interactivity of the literatures, may affect the judgements made by the researchers. It is not entirely clear what the nature of relevance in the context of LBD is, how it can be observed (or measured) nor whether this can be done automatically. Additionally, the processes by which researchers make judgements have not been explicitly observed. The multidimensional and dynamic nature of relevance in this context is of particular interest in this work. Effectively, the core of this work lies in observing which criteria form part of this notion of relevance, the fluctuations across these criteria and whether researchers use them in varying proportions when assessing the appropriateness of the retrieved literature in an LBD scenario.

Hence, the main research question that this dissertation aims to answer is

> *What relevance criteria, if any, do researchers use when assessing the relevance of related, although potentially outside their research field, information?*

The observation of the relevance criteria, as used by researchers in an LBD context, is achieved by having the researchers interact with an LBD system while completing a knowledge discovery task. As LBD is mainly aimed at the scientific community, questions related to the users' background arise

> *Do researchers from different disciplines use different criteria and/or in different frequencies?*

Does research experience affect the relevance criteria, and their frequency of use, used in these relevance judgements?

To guide the evolution and improvement of LBD systems, and IR systems in general, researchers resort to their evaluation regarding their chosen performance measures. It is here where relevance plays an important role. Hence, the aforementioned questions gain importance and deserve attention.

To try and find answers to these questions, it is proposed in this dissertation to follow the guidelines and include the experimental components as proposed by Borlund & Ingwersen (2000) in the design of a user study. As it has been argued before, studies on the nature of relevance should be conducted with real end-users and in realistic settings (Schamber, Eisenberg & Nilan 1990). The context of the study is that of scientific discovery, hence the test persons are recruited from the scientific community; researchers are the end users of LBD systems. Researchers from three different disciplines, namely computing, information management and pharmacy, are invited to take part of the study.

1.2 Structure of the Dissertation

The rest of this dissertation is structured as follows. Chapter 2 provides a literature review of LBD: an introduction to the problems, a survey of the most representative techniques, and a description of the evaluation methods used are offered. Additionally both the system-driven tradition of IR research and Borlund's proposal for evaluating IIR systems are described here. In Chapter 3 details of the design and the materials used during the study including the characteristics of the user group, the collections searched and a description of the system used are offered. Relevance criteria profiles and session visualisation techniques are central to the analysis of the data gathered during the study. These analysis tools, developed during the course of this investigation, are described in Chapter 3. Chapters 4 and 5 present the data gathered during the study. In Chapter 4 an analysis of the relevance criteria observed using relevance criteria profiles is offered. Relevance criteria profiles are revisited in Chapter 5. Furthermore, the data is segmented to isolate and analyse the relevance judgement processes and the interactions with the system ob-

served are described and analysed. Chapter 6 includes a discussion of the implications of the results and a number of recommendations for future work. Chapter 7 concludes this dissertation with a summary of the research carried out regarding the research questions.

Chapter 2

Literature Review

As research fields become more specialised, academics tend to interact more with researchers and literature from their chosen speciality, and less with researchers and literature outside of their own specific area of interest. Consequently, the interaction between fields, through cross-referencing across fields and shared use of common literature, becomes reduced and related fields *detach* from one another. The result is relatively isolated (fragmented) and highly specialised bodies of literature, a phenomenon that has recently accelerated due to the increased rate of new publications available online (Swanson 1986c).

This detachment of research fields means that academics who share common interests and approaches but work in different fields can miss important connections. It is becoming increasingly challenging for most researchers, especially in established fields, to keep up to date with important developments in their own chosen speciality (Swanson 1988a). However, keeping track of useful new developments in allied fields is even more demanding, relying far more heavily on the inefficient processes of manual literature searching and browsing, chance discoveries through personal communications or selective manual dissemination of information. More often, cross-disciplinary connections have to be engineered through dedicated, but often small-scale, initiatives. Ashworth (1966) argues that librarians are in a unique position since much knowledge passes through their hands. A librarian should, then, be able to perceive which items of knowledge might be profitably combined even though he will not combine them himself or discover the full potential

behind said combination. Ashworth also suggests that *"this need for combination is so fundamental that library systems should be designed to meet the requirement, thus enabling librarians to play an active and vital part in future innovation"*. In Ashworth's proposed design one finds that, amongst others, systems should be able

- To demonstrate automatically when ideas are neighbours of each other, and therefore related
- To uncover, automatically, valid statistical correlations between apparently unrelated matters

The aim of research into Literature Based Discovery (LBD) is to help *discover* these connections between seemingly unrelated disciplines by mining publicly available academic literatures. This area of research is motivated by the findings of Don Swanson, who in the mid-80s discovered two disjoint literature bodies that were complementary i.e. when put together, they suggested an answer to a question which was not previously published. Swanson saw the potential in this procedure –combining knowledge from both literatures to form an answer– and started to systematically investigate it under the name of Undiscovered Public Knowledge (UPK), more recently known as Literature Based Discovery (Swanson 1986c).

To this day researchers keep on improving the techniques used to discover and retrieve these unrelated literatures. Techniques vary from system to system, however the search models implemented remain the same: the open model and the closed model. The open model of search is aimed at generating hypotheses and is an inherently exploratory model (Weeber et al. 2001). Users begin a search session with a topic in mind. The system is in charge of generating (suggesting) a set of directly related topics which in turn will be used to infer a new set of related topics. This last set of topics is interpreted as being indirectly related to the researcher's initial topic. Using these suggested topics, users can then start hypothesising about the potential relationships. The closed model, on the other hand, aids users in verifying the hypotheses generated using the open model. A typical search session begins with a hypothesis. This hypothesis is that the topics selected during the open model search session (the researcher's initial topic and an indirectly related topic)

are indeed related. A system implementing the closed model search is responsible for retrieving the intermediate topics and the relevant literatures. It is with these literatures that the researcher can start considering whether this hypothesis has merit or not.

To evaluate the performance of their systems, researchers have generally resorted to replicating Swanson's initial discoveries. To this extent, researchers model and extract the concepts contained in a subset of the relevant literature and observe, after filtering and ranking, how many of these modelled concepts correspond to those that Swanson discovered play a role in the discoveries made. We suggest that evaluating systems in such a way, albeit convenient and cheap, may not be the most appropriate method. For example, the users' context and interactions are disregarded. In a scenario in which background knowledge, for instance, and other contextual factors may greatly affect the outcome of the search sessions, an evaluation method that simply ignores them might not be the most appropriate.

In the following sections a description of the most representative work in the area of LBD is offered. The history of LBD is covered in Section 2.1 while the search models are described in Section 2.2. The techniques and evaluation methods are discussed in Section 2.3. In Section 2.5 a short overview of evaluation methods for Interactive Information Retrieval (IIR) is offered; it is our suggestion that these methods may be more appropriate when assessing LBD systems.

2.1 Literature Based Discovery

The most famous example of a successful discovery using LBD is that of the relation between dietary fish oil and Raynaud's disease (Swanson 1986a). Dietary fish oil has been shown to have several effects that improve blood circulation, e.g. reduction in blood lipids and platelet aggregation. Simultaneously, Raynaud's syndrome, being a peripheral circulary disorder, presents several symptoms that seem to be the negation of the effects achieved by dietary fish oil. Both arguments, if considered together, suggest that dietary fish oil might be beneficial for patients suffering of Raynaud's syndrome.

In his seminal article, Swanson (1986a) presented a set of 25 articles discussing the

beneficial effects of dietary fish oil on blood circulation and another set of 35 articles discussing the symptoms of Raynaud's syndrome that would be affected by these effects. A careful analysis of these two sets showed that they were not interacting with each other in the sense that no article in either set cited any article on the other set (cross-citation) and that only 4 articles cited at least one article from each set (co-citation). However, Swanson noticed and explained that out of the 4 articles which cited at least one article from each set, none did so in a way that related fish oil and Raynaud's syndrome. Contrasting the lack of interaction between the sets, the logical connections abounded.

In a subsequent article Swanson (1986c) presented a group of examples together with the general idea of Undiscovered Public Knowledge (UPK) of which the fish oil - Raynaud's syndrome was a particular instance. Swanson argued that the creative use of information searching strategies could lead to meaningful discoveries as these isolated, but complementary, bodies of literature were awaiting to be discovered. By working on the fish oil - Raynaud's syndrome example Swanson not only suggested that dietary fish oil might ameliorate or even prevent Raynaud's syndrome, he also suggested that, if his suggested link was indeed correct, logically related, but isolated, bodies of literature existed. At least in the field of medicine.

Swanson's discoveries were made on the medical domain. As it will be described in Section 2.3, this domain offers certain structural information and metadata that has been beneficial to enhancing the techniques used to not only replicate Swanson's original discoveries but also to make new discoveries. There has been a few examples of LBD being applied outside the field of medicine (Gordon, Lindsay & Fan 2002, Cory 1997) however, the question of whether LBD can be applied outside the field of medicine remains largely unanswered. Moreover, researchers frequently evaluate their systems by attempting to replicate Swanson's original discoveries which suggests that extrapolating the proposed techniques may be complex.

2.2 LBD Search Models

Weeber et al. (2001) suggested that most LBD systems are designed to implement one

(or both) of two search model. The open model, an exploratory model by nature, is characterised by the generation of a hypothesis at the end of the process. The closed model, on the other hand, aids the researcher in evaluating whether a hypothesis has sufficient merit so that it is properly tested, e.g. by conducting experimental research in a laboratory. Both models complement each other and are described next.

2.2.1 Open Model — Hypotheses Generation

The search process begins with a topic of a scientific problem or research question. This initial topic is usually referred to as "the starting topic" or "A-topic". For instance, a scientist may be interested in finding a novel treatment for Raynaud's syndrome. Next, the A topic (Raynaud's disease in the example) is used to query a database, Medline for instance, and the pertinent literature is retrieved. This literature is referred to as the "starting literature" or "A-literature". Important topics are then extracted and processed, according to techniques that vary from system to system. There may be human intervention in this extraction and processing step. The extracted topics are referred to as "intermediate topics" or "B-topics". If, for instance, the starting topic is that of our example (Raynaud's disease), the B-topics may then correspond to characteristics or symptoms of the syndrome, already tried treatments, etc. These B-topics are then used to query a database to obtain the pertinent literature; this literature is referred to as the "intermediate literature" or "B-literature". The B-literature is processed and topics are extracted. These topics are known as "target topics" or "C-topics". The searcher is then presented with this network of topics so that one or more target topics can be selected for further inspection.

This outlined process relies on the assumption of transitivity. If an A-topic is related to a B-topic, which in turn is related to a C-topic, then there is a chance that an indirect relationship between the A-topic and the C-topic exists. This simplification of the reasoning modelled by the process lends itself to implementations based on co-occurrence (mostly) which are reviewed in Section 2.3. And because the process is simplified, the result is an exponential growth in the number of potential relationships between any one A-topic and

any one C-topic. Hence, the two biggest challenges, in terms of building systems, posed by the open model are:

1. Modelling and extracting topics from documents

2. Modelling, finding and processing relationships between these topics

In Figure 2.1 (figure adapted from (Weeber, Klein, Aronson, Mork, de Jong-van den Berg & Vos 2000)) we can observe a graphical depiction of the open model.

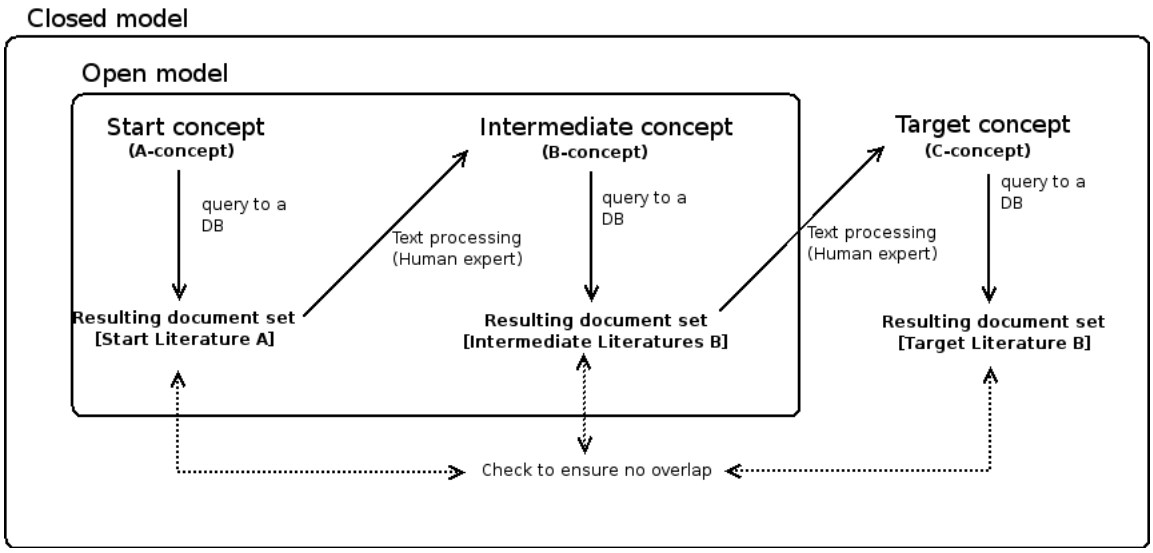

Figure 2.1: The LBD models in action. The inner box depicts the open model in which a searcher is interested in forming hypotheses about the topics of interest. The outer box depicts the closed model; a model in which searchers either reject or find evidence for pursuing the verification of hypotheses.

2.2.2 Closed Model — Hypotheses Validation

With a hypothesis in mind, a searcher embarks in using the closed model to assess whether it merit further investigation. The process begins with two topics, the A-topic and the C-topic. The database (or databases) is queried twice, once with the A-topic and once with the C-topic and both the A-literature and the C-literature are retrieved. These two literatures are pooled together and the resulting set is processed. The processing of the pool of literatures results in a set of intermediate B-topics that may, or may not, link

the two starting topics together. The searcher's job is then to investigate the B-topics extracted in relationship to the hypothesis under evaluation.

A Brief Note on the Models

Doyle (1961) points out that *"literature searchers value both the unexpected and the expected. When something unexpected is found, one thereby obtains information; when the expected is found, one obtains confirmation. However, when one formulates a search, the unexpected is hardly ever involved. Search requests are practically always constructed out of familiar combinations of terms."* Both models of LBD share a direct relationship with Doyle's observation. By using the open model of search, users will be confronted with the unexpected, with what is unknown to them yet related to their initial search. When using the open model, users should obtain information about the possibly related topics. As a complementary step, users could perform a search using the closed model. Results from this search may provide evidence supporting different types of relationships between the initial topics. At this stage users obtain confirmation (or refutation) on the relationship between the provided topics; that is, the hypothesis that the topics are linked.

2.3 LBD Systems

In this section a review of the most prominent examples of the techniques used to build LBD systems is offered. Both techniques used as well as the evaluation methods are covered.

2.3.1 The Models

Although Weeber et al. (2001) suggest that the closed model serves as complement to the open model, most research focuses in one of the two. The initial discovery made by Swanson (1986a), for instance, could be classified as a use of the closed model. After manually inspecting the literatures on Raynaud's disease and several articles discussing the benefits of dietary fish oil, Swanson suggested the link between them. Swanson (1988b) hypothesised next that migraine attacks may be linked, through eleven connections, to

magnesium deficiency. The procedure followed by Swanson, however, resembles the open model as it begins with the migraine literature (A-literature) and eventually reaches the literature on magnesium deficiency (C-literature). In later work, Swanson (1991) revisits the migraine-magnesium deficiency link by manually performing a closed model search in which he assumes that the relevant literature has been retrieved. Swanson analyses whether the initial topic can be logically linked to the target topic.

The procedure followed by authors whose research is focused on the open model, e.g. (Gordon & Lindsay 1996, Gordon & Dumais 1998, Lindsay & Gordon 1999, Pratt & Yetisgen-Yildiz 2003, Hristovski, Peterlin, Mitchell & Humphrey 2005), is such that it begins with a starting topic and the goal is to rank highly the known target topics. Authors who conduct their research following the closed model, e.g. (Weeber et al. 2001, Srinivasan 2004, Swanson 1988b, Smalheiser & Swanson 1996b, Smalheiser & Swanson 1996a, Swanson & Smalheiser 1997a), on the other hand, have a different goal. Their followed procedure begins with two topics for which their linking intermediate topics are known and try to rank these intermediate topics highly.

2.3.2 Text Modelling

While generally based on co-occurrence, techniques differ in how topics are modelled, how relationships are inferred and how these are filtered and ranked. Approaches based on co-occurrences and other statistical frequencies include Swanson's initial discoveries. After having manually inspected the literatures for co-occurring terms in the article titles, Swanson suggested two hypothesis:

1. That Raynaud's syndrome symptoms could be alleviated by means of a diet high in fish oil (Swanson 1986a, Swanson 1988b)

2. That magnesium defficiency may be related to migraine attacks (Swanson 1991)

Gordon & Lindsay (1996) investigate statistical approaches further, however they take a more principled analytical approach and use traditional IR weighting techniques such as *tf.idf* (Salton & Buckley 1988). Gordon & Lindsay (1996) model topics as either unigrams

or bi-grams (words or two-word phrases) and the statistics of these grams are used to filter and rank the modelled topics: those that do not meet a user-defined threshold are discarded (they are considered noise). While Swanson (1986a) analyses the co-occurrence of words in article titles, Gordon & Lindsay (1996) do so in the full text (whenever available) of the articles. Gordon and colleagues continue evaluating the appropriateness of term statistics for filtering and ranking purposes (Lindsay & Gordon 1999, Gordon et al. 2002).

Gordon & Dumais (1998) report on the application of a technique called Latent Semantic Indexing (LSI) (Deerwester, Dumais, Furnas, Landauer & Harshman 1990). The use of LSI in this context is to reveal hidden potential relationships amongst terms in text, as semantically similar terms lie close together in the LSI space. Unfortunately, no mention is made to how topics are filtered or discarded. Topics are ranked according to their proximity (using cosine distance) to the starting topic (e.g. "Raynaud's disease").

In addition to text statistics, other approaches have involved the use of external databases and metadata for topic modelling, filtering and ranking. The system implemented by Weeber et al. (2001) –an extension to their previously implemented system (Weeber et al. 2000)– uses the MetaMap algorithm (Aronson 2001). The MetaMap algorithm maps free-form text to the Unified Medical Language System (UMLS) medical concepts; concepts which are usually used as topics. The semantic information associated to these concepts is also usually used as filtering options. Algorithms that use MetaMap include LitLinker (Pratt & Yetisgen-Yildiz 2003) and the work conducted by Wren, Bekeredjian, Stewart, Shohet & Garner (2004). Pratt & Yetisgen-Yildiz (2003) initially map text to medical concepts using MetaMap. To uncover potential links, LitLinker uses association rule mining (ARM), an unsupervised machine learning algorithm for learning higher-order co-occurrences, e.g. tri-occurrences –a co-occurs with b and b co-occurs with c, hence a is considered to indirectly co-occur with c (R. Agrawal, Mannila, Srikant, Toivonen & Verkamo 1996). Topics are pruned according to several criteria. Firstly, topics that are considered to be too general (according to the UMLS) are pruned. Secondly, topics that appear in more than 10,000 titles are removed. Thirdly, the proximity of the topics to the initial A-topic is measured using the number of parents shared in the UMLS hi-

erarchy. Topics that are considered too close are also removed. Finally, UMLS semantic types are used to filter the remaining topics. Once pruning is over, similar concepts are grouped together to increase their statistics. Wren et al. (2004) model topics by mapping free form text to concepts using several databases amongst which we find UMLS and the Medical Subject Headings (MeSH) (Lowe & Barnett 1994). Relationships are identified by analysing the co-occurrence of topics within Medline records. To filter and rank the discovered relationships, they are compared to a random network model. The ranking of a connection depends on the ratio between the number of connections for any given link and the expected number of connections by chance. Those that exceed a defined threshold are ranked more highly.

Previous approaches rely on free form text to model and extract the relevant topics, however the algorithm proposed by Srinivasan (2004) relies only on Mesh (Lowe & Barnett 1994) and UMLS. The algorithm models topics as combinations of MeSH profiles where profiles are vectors of weighted MeSH terms. Topics are ranked according to the number of intermediate B-topics, and the strength of their association, that link the starting A-topic and the target C-topic.

A further specialisation is observed in the system developed by Hristovski et al. (2005) who include genetic information into their LBD system named BITOLA. This is an extension to their previous work (Hristovski, Stare, Peterlin & Dzeroski 2001, Hristovski, Peterlin, Mitchell & Humphrey 2003). Information about chromosome location of the starting and target topics is integrated so that disease-gene discoveries can be performed with their system. To this extent, despite the focus on the Medline literature, other sources of information such as LocusLink[1] and Human Genome Organisation (HUGO)[2] are included. Much like LitLinker (Pratt & Yetisgen-Yildiz 2003), to discover relationships, Hristovski et al. (2005) use association rule mining (R. Agrawal et al. 1996) between medical concepts.

[1] Now superseded by Entrez Gene – http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene
[2] http://www.hugo-international.org/

2.3.3 Evaluation Methods

Swanson's initial discoveries were later corroborated in subsequent articles in the biomedicine field (Weeber et al. 2001). Most researchers evaluate their systems using Swanson's discoveries as gold standards. The golden data is composed of golden topics, Swanson's reported A, B, and C topics. Although the details of each approach vary, to measure the performance of their systems, researchers follow a similar approach. The products of the systems are manually inspected to:

- Determine the number of golden topics present in the list of returned topics
- Determine the rankings of the golden topics

Gordon & Lindsay (1996) attempt to replicate the fish oil-Raynaud's disease discovery (Swanson 1986a). To do so, the authors perform their experiments with the original fish-oil documents dated between 1982 and 1986. The effectiveness of their system is measured by means of calculating Precision and Recall (Cleverdon, Mills & Keen 1966) however, the authors do not specify how they compare the topics retrieved by their system with the golden topics as to consider them either true or false positives. To assess further the ranked lists of topics extracted by their systems, human experts are involved and these suggest that "blood viscosity" is a salient topic. It remains unclear, however, why this topic was considered to be salient. Additionally, while it is true that its statistics placed this topic within the top ranked topics, other topics were also ranked highly and hence, if one was to set a threshold for picking potential candidates, these other topics could have been equally selected.

In a subsequent article, Gordon & Dumais (1998) evaluate the performance of LSI by comparing the results of this technique with those obtained through the use of term statistics of Gordon & Lindsay (1996). 136 LSI concepts which are found to be "near" to the concept "raynaud" (near in terms of being proportional to the cosine distance between two concepts) are compared to the union of 6 top-40 concepts from the results of (Gordon & Lindsay 1996). The authors find that the there is a 42% overlap between the two sets of concepts. More importantly however, is that the authors report that the top 10 LSI

concepts include 9 of the top 10 concepts of (Gordon & Lindsay 1996) and conclude that the approaches are, to a certain extent, uncovering the same relationships. The authors do not explicitly mention, however, how the correspondence between the LSI concepts and those of the previous work is made.

In an extension to (Gordon & Lindsay 1996), Lindsay & Gordon (1999) use tri-grams (three word phrases). The authors follow the evaluation procedure of (Gordon & Lindsay 1996) however the golden data is taken to be the Swanson's discovery that migraine might be related to a magnesium defficiency (Swanson 1988b). As the discovery model used is the open model, using "migraine" as the starting concept, the evaluation is done in two stages. During the first stage, the set of intermediate concepts is evaluated to see if any of the 11 connections reported by Swanson (1988b) as well as vasopasm (which the authors consider to also be related) are discovered. The authors report that *"10 of the 12 intermediate concepts linking migraine and magnesium were among the first few dozen items suggested by either token or record count analyses of the migraine literature"* (Lindsay & Gordon 1999). During the second stage the authors evaluate the results of further processing of all the 12 intermediate concepts[3] This processing involved extracting the top 500 topics from each intermediate literature, pooling them together and filtering to extract the top 50. These are in turn given to a medicine student who further groups and consolidates the list. The authors report that magnesium did not appear in the top ranks of any of the intermediate lists nor in the final list.

Weeber et al. (2001) also try and replicate Swanson's original discoveries (Swanson 1986a, Swanson 1988b). In the case of the fish oil-Raynaud's discovery, to build the corpus of documents to analyse, the authors run the query "Raynaud's disease" on Medline. The system then maps the raw text from the titles and the abstracts of the documents retrieved to UMLS concepts using MetaMap (Aronson 2001). The semantic information associated to these concepts is also used as filtering and ranking options. For the closed model, the authors examine the 68 top ranked intermediate metamap concepts and report that the

[3]The authors argue that since they are trying to replicate Swanson's findings, selecting only those for the experiment is reasonable. Additionally, it is suggested that if the purpose of the experiment was to generate new discoveries, all top intermediate topics in the ranking should be investigated further.

original B-topics of the fish oil-Raynaud's disease discovery are included in this list. When using the open model, the authors suggest that, while the original fish oil concepts were not ranked highly, other concepts related to fish oil were, and that researchers should be able to still recognise fish oil as a target concept from them. Again, it remains unclear how the golden topics from Swanson's discovery are matched against those produced by the system.

As the work by Gordon et al. (2002) is exploratory in nature –the authors set to explore the appropriateness of using LBD on the World Wide Web (WWW)– they do not attempt to replicate any of Swanson's discoveries. To conduct their experiments the authors focus on the open model. The process begins by retrieving[4] the top 50 documents on the topic "Genetic Algorithms". On these documents, several statistics are calculated such as term frequency and document frequency (the statistics calculated are explained in more detail in their previous work (Gordon & Lindsay 1996, Gordon & Dumais 1998, Lindsay & Gordon 1999)). The extracted concepts, represented as n-word phrases, are ranked and filtered according to the term statistics calculated on the pooled documents. A human expert proceeds then to select the 12 most salient topics from the list of ranked topics[5]. This selection process is subjective as it is explained that the judgement as to what a salient topic is largely depends on the expert's knowledge of the starting topic "Genetic Algorithms". Each of these topics is then used to query the search engine and the top 100 documents is retrieved from the WWW. Each document set is statistically analysed and the results are pooled together. Out of this new list of topics, the authors selected 42 topics as they *thought [they] might be important in relationship to genetic algorithms*(Gordon et al. 2002). To validate the results, an expert in the field is asked to generate a list of possible topics to be investigated in conjunction with "genetic algorithms". The authors report that they did not find any overlap between the list of the expert-generated topics and the topics selected from the output of their system.

Other discoveries made by Swanson were used as golden standard as well. For instance, the migraine-magnesium link is used as gold standard by Pratt & Yetisgen-Yildiz (2003).

[4]The search engine used is Altavista – http://www.altavista.com/

[5]No rationale is given as to why they chose 12 topics.

The experiment conducted by the authors is focused on the closed model of search. The system works by mapping the raw text of the documents to UMLS concept by means of the MetaMap algorithm (Aronson 2001). The resulting concepts are initially filtered in two steps:

1. Filtering too general concepts: the authors observe that many general concepts are usually present in the second level of the UMLS hierarchy of concepts and decide to prune concepts at this level.

2. Filtering too frequent concepts: as there are still too many general concepts present, the authors decide to filter out concepts with a document frequency greater than 10,000.

Once filtered, concepts are grouped together by folding upwards to parent concepts in an attempt to increase the statistics of the different concepts extracted. After starting an open model search with an A-topic representing "migraine", the authors observe that the target topic "magnesium" is ranked at position 11. The target topics are ranked according to how many in-links each topic has[6]. When conducting a closed model search using "migraine" and "magnesium" as start and target topics respectively, the authors report that 5 out of the 11 original links found by Swanson are suggested by the system.

Srinivasan (2004) test the effectiveness of relying on metadata exclusively as opposed to extracting concepts from the documents. Additionally, the authors' interest is to see whether the process can be fully automated and present the end-user with a final ranked list of C-topics, i.e. no human intervention is done during the ranking and filtering of intermediate B-topics. The topics used and returned by the system are composed of MeSH terms. These concepts are filtered according to the semantic types of the MeSH terms. Term statistics are then calculated within each semantic type. Concepts are then are represented as MeSH term weights, within the semantic types, as calculated for a particular document set. To test the performance of this approach, the authors perform experiments using both the open and the closed model. In the open model the end-user identifies the semantic types of interest and the system in turn returns a ranked list of

[6]In-links in this case is the number of intermediate topics related to the target topic.

topics to investigate further. The ranking is based on the number of connections from each topic to the initial topic as indicated by the user. These are the final target C-topics. In the closed model, the approach is similar however the user is in charge of providing both the starting A topic and the target C topic. The system returns a ranked list of intermediate B-topics filtered by the semantic types of both the A and the C topics. MeSH terms are ranked within B-topics hence each B-topic acts as a group of concepts to be analysed. The evaluation is performed by replicating both of Swanson's fish oil-Raynaud's disease and migraine-magnesium discoveries (Swanson 1986a, Swanson 1988b). After manually inspecting the returned list, the authors report that key MeSH terms were ranked within the top 10 ranked topics. The same conclusion is reached for the closed model experiment, however the authors express that comparing results with previous work is difficult as the ranking strategy for the closed model is incompatible with previous approaches. Additionally, the authors attempt to also replicate Swanson's indomethacin-Alzheimer link (Smalheiser & Swanson 1996a), Somatomedin C-Arginine link (Swanson 1990), and the Schizophrenia-Calcium independent Phospholipase A2 link (Smalheiser, Swanson & Ross 1998). The results for these experiments are reported to be consistent with those of the previous experiments.

Other approaches at evaluation included using artificial data sets (Van Der Eijk, Van Mulligen, Kors, Mons & Van Den Berg 2004) and conducting clinical trials on mice (Wren et al. 2004). Despite that Hristovski et al. (2005) did not evaluate their system, they reported that future evaluation was to be carried out and proposed a plan for doing so. The guidelines as proposed, suggest that in order to evaluate their system firstly valid disease-gene relationships are to be detected. Their plan is to observe whether the system can detect such relationships by only having access to the literature previous to the date of publication of the relationships.

The approaches, as described can be generalised and modelled by the principles of evaluation followed in the system-driven tradition of evaluation of IR system. As such, they suffer from the same drawbacks as the system-driven tradition. Some of these drawbacks are addressed by the user-centred tradition of evaluation and, more recently, a hybrid

approach which combines the two traditions. Before we discuss these models, however, we have to discuss the concept of relevance and its role in LBD. In the following section the reader is presented with a brief account of the pertinent, to this dissertation, research efforts on the notion of relevance. Once concluded, the evaluation models are described and discussed in the context of LBD.

2.4 On Relevance

Relevance[7] is an elusive concept which is intuitively understood by humans. However, due to the increase in size of the available information that we have to search today, we humans rely more and more on computers to retrieve relevant information for us. The contraposition and need for cooperation between computers and humans are what make the notion of relevance hard both to define and to investigate. IR systems have a notion of relevance embedded in their algorithms and retrieve and offer what may be relevant according to this notion. People, on the other hand, then go and assess relevance on their own. However, each version of the notion may have a different take on what relevance is (Saracevic 2006).

The concept of relevance in IR is not new as it is tightly tied to that of the evaluation of IR systems. Good systems are so because they return relevant information. The discussion on relevance has evolved over time and several interpretations and variations on the concept of relevance have emerged. Each variation has placed emphasis on different aspects of this notion. Amongst others we find the notion of Logical Relevance (Cooper 1971), Situational Relevance (Wilson 1973), Objective and Subjective Relevance (Swanson 1986b) and Psychological Relevance (Harter 1992). Each interpretation brings its own set of assumptions and rough edges and this only highlights the difficulty of the debate.

A commonly accepted interpretation is that of relevance as being a relation (Saracevic 2006). In its simplest form, this relation is taken to be a correspondence between a *document* and an *information need* as judged by a *person* (Saracevic 1975). And even

[7] This section offers a brief account of the most pertinent strands of research into the concept of relevance. The reader is, however, encouraged to read the review by Mizzaro (1998) and those by Saracevic (1975, 2006, 2007) for a fuller account of the research efforts in this area.

this simple interpretation comes with a variety of difficulties. The idea of *information need* is difficult to specify as the person is in an anomalous state of knowledge (ASK) (Belkin, Oddy & Brooks 1982). According to Belkin et al. (1982), this *non-specifiability of need* can be exhibited at two levels. Firstly, we have the cognitive level. At this level, the range of non-specifiability can vary between two extremes. At one end of the spectrum, we have people who know exactly what they need to solve their information needs. At the other end, we have people who are able to recognise they have the need for information but cannot express, or can only do so vaguely, this need. Secondly, in Belkin's hypothesis, we have the level of linguistics. To begin the interaction with the IR system, a need must be expressed as a *request* which in turn is translated to a *query*. The difficulty of linguistic non-specifiability resides in that people may not know how to best use the language to write queries to the system. This is aggravated as *documents* are usually created independently from *requests* and persons. Hence people are not aware of the underlying language included in the database and the documents. Finally, the *person*, when judging, brings with him all his background, knowledge, current state of mind, perspective and situation (Schamber et al. 1990).

Relevance judgements are traditionally used as proxy to measure relevance. These judgements are the verdicts that requesters emit on the information retrieved for their queries. Swanson (1977) proposed two frames of reference for relevance judgements:

1. Relevance is interpreted to be a correspondence between the document and the information need as seen by the requester

2. Relevance is interpreted to be a topical correspondence between the request and the document

While the first interpretation allows for more room for subjectivity from the requesters, the second one is more specific and specific to the point where judgements might be collected from third-party judges who might not necessarily be the original requesters. Two major studies on relevance judgements were conducted (Rees & Schultz 1967, Cuadra & Katter 1967). These studies suggested that there are about 40 variables, e.g. specificity

and difficulty of documents, that might affect relevance judgements, however not all variables were analysed. Even though, relevance, in both projects, was judged by third-party judges and in experimental conditions, it was still suggested that the reliability of human relevance judgement is questionable. That is to say that the desired stability and objectivity that researchers were after might not be achievable by using professional specialist judges.

By the late 1970s, the cognitive point of view in information sciences had gained enough impetus to influence most empirical studies causing a shift from a system-centric view, a view where relevance can be judged static and objectively, to a more user-centred view (Ingwersen 1988). The shift to a more user-centred approach made it possible for researcher to find, for instance, that psychological factors, such as cognitive states or affective feelings, influence the search behaviour (Kuhlthau 2004). Additionally, the nature of relevance began to be considered as being multidimensional and dynamic. Schamber et al. (1990) maintains that despite this dynamic and multidimensional nature of relevance, it is both systematic and measurable, hence its study can be done in a controlled and repeatable manner. This led researchers to argue that relevance studies should observe real users in natural settings holistically rather than study recruited judges under experimental conditions. Effectively, Schamber (1994) advocates that relevance studies should focus on real-life situations from the human information perspective.

During the last decades, several studies focused on observing the relevance criteria as used by various real end-users in real-world situations (Schamber 1991, Barry 1993, Barry 1994, Cool, Belkin, Frieder & Kantor 1993). In these studies, qualitative methods were applied to collect data. Users were not given a predefined set of criteria to perform relevance judgments, quite the contrary, criteria were derived directly from content analysis of verbal or written reports. This resulted in user-defined relevance criteria; criteria that is commonly used by real end-users during relevance judgement. During the analysis of the data from these studies, it was observed that most decisions were based on additional variables in document surrogates such as title, author, journal, and descriptor. After an in-depth analysis of the concept of topicality, Green (1995) points out that topicality is

only part of the subject contents of a document. One of the suggestions stemming from these studies is that topicality alone is not enough to make decisions and that other factors affect relevance judgments (Green 1995). A key difference between these studies and those by Rees & Schultz (1967) and Cuadra & Katter (1967) is that the variables found to affect relevance judgements were derived from real end-users on real-world tasks (as opposed to professional judges un laboratory conditions).

In the context of LBD, the notion of relevance might even be more complex. For one, there are additional moving pieces (as compared to the simple view of *request, document, person*) such as the start, intermediate and target topics. Additionally, the type of task to be solved is that of knowledge discovery which might be, in principle, quite heavy in terms of cognitive effort. The discussion on relevance in the context of LBD has been brief and mostly in the form of examples, e.g. (Swanson 1986c). However, the general consensus seems to be that topicality alone is not enough to derive relevance in this context (Swanson 1991). Effectively, the task of an LBD system is to first distill the topics that might compose a fruitful combination and second retrieve the pertinent literature. Pertinent, in this sense is not to be taken as being *on topic* since there may be several topics in play in any given combination, but rather *supportive*, amongst others, in the sense of supporting, or rejecting, a hypothesis.

2.5 Evaluation of IR Systems

Traditionally, most of the research in IR has been carried out following the principles of the system-driven tradition. The principles behind this tradition serve to design, and consequently test, better algorithms and systems. The evaluation of newly crafted algorithms follows the Cranfield tradition (Cleverdon et al. 1966); a tradition that is heavily based on the concept of test collections (Harman 1997). Test collections consist of:

1. A collection of document,
2. A collection of requests, and
3. A collection of relevance judgements

Before the collection of documents can be searched it is usually indexed; documents are converted to a suitable representation and stored in a database. In order to be able to search this new database, requests have to be transformed into queries: a representation that is suitable for matching against the contents of the database. The algorithm can, then, run the queries against the database. This procedure results in a ranked list whereby documents are ranked according to their likelihood of relevance (Belkin & Croft 1987). Because users are not involved in the evaluation of systems, it is commonly referred to as "laboratory IR", making emphasis on that the evaluation of system takes place in laboratory conditions (as opposed to real conditions) (Ingwersen & Järvelin 2007).

The quality of the results of any one algorithm is measured by assessing its ability to retrieve *relevant* documents. To measure this, relevance judgements are used as "ground truth". Relevance judgements link (relevant) documents to requests, hence the assessment of the products of the algorithms is possible. During evaluation, the recall and precision of a system are usually measured and averaged over requests (Cleverdon et al. 1966). These two metrics measure how many of all the known relevant documents have been retrieved (recall) and how many irrelevant documents have been included in the ranked list (precision). Because requests have been typically of a topical nature, and relevance judgements were made by experts in the topics, this type of relevance is usually referred to as "topical relevance", "topicality" and even "aboutness" (Saracevic 1996).

This model of evaluation fits the underlying structure of all of the approaches used to evaluate the performance of LBD systems described in the previous section: Swanson's original discoveries make the collection of relevance judgements (the B- and C-topics in the case of the open model and only the B-topics in the case of the closed model), a subset of Medline the collection of documents and Swanson's reported stating and target topics (the A- and C-topics) the collection of requests. However, these are not standard across research efforts. Each researcher that attempts to replicate any of Swanson's discoveries chooses which discoveries (collection of relevance judgements) to use, which subset of Medline to use (collection of documents) and which topics to include (collection of requests). When it comes to reporting performance, none but one group of researchers report standard

metrics such as Precision and Recall. Most researchers report subjectively whether the golden topics have been detected within the products of their system and the ranks of these topics. Additionally, the judgement as to whether the system has produced a true positive (how topics produced by a system are matched to those reported by Swanson) is subjective which makes the comparison of approaches very difficult. Further assumptions made make this model inappropriate to evaluate LBD systems. As it has been discussed earlier, LBD searches are inherently interactive and most researchers seem to agree that human intervention is required. However, all efforts either disregard the end-user by assuming that a researcher should be able to identify the salient topics or include humans in the loop in an ad-hoc fashion.

While the system-driven tradition of evaluating systems constitutes most of the literature of IR research, and accommodates the current approaches at evaluating LBD systems, the type of relevance assessment employed has been met with criticism, in particular, when dealing with interactive systems (Kekäläinen & Järvelin 2002).

Contrary to the laboratory IR tradition, in which the user is seldom involved, research in user-centred IR focuses on the behavioural and psychological aspects that affect how users interact with IR systems (Ingwersen 1993). It is considered to be a broader view of the problem. While laboratory IR considers the problem as a request-document matching one, research in the user-centred IR tradition sees it as a problem solving, and goal-oriented interactive task. It is by analysing and understanding these aspects that researchers aim to improve IR (as perceived by the research community) performance. The common research themes investigated include, for instance, the nature (Taylor 1967) and types of information needs that can be identified (Ingwersen 1996) as well as user defined relevance (see e.g. (Borlund 2003a, Barry & Schamber 1998, Saracevic 1996, Schamber et al. 1990)).

Initially, a user-centred approach might seem like an appropriate one to evaluate LBD system. However, this would mean completely disregarding the underlying systems and instead focusing on factors such as the perceptive power of different researchers, the appropriateness of workflows for particular discovery tasks, etc. Hence, applying such an approach would provide a partial picture of the performance of an LBD system, if at all.

This stems from the fact that the community involved in user-centred IR research, does not seem to concern itself with the details of the IR system with which the users under examination interacts. Comparable to the system-driven tradition, the user-centred IR approach considers systems as being constants and rarely linked to human beings (Ingwersen 1992). This view seems to be, in principle, as limited as that of the system-driven community though focused on a different aspect of the problem.

There are examples of research that attempt to integrate both viewpoints, however an evaluation of systems of this nature is particularly difficult. Because of their comprehensive approach two properties must be present in any evaluation framework of this nature:

1. Control: it is desirable that the control observed in experiments conducted in the system-driven tradition is retained as to achieve repeatability and the ability to compare with previous attempts

2. Realism: while control must be retained, the evaluation has to take place in a realistic, as possible, scenario since there must be an involvement of the end-users as to be able to observe/measure the cognitive phenomena of interest

Borlund (2000) who state that the user-centred approach is ideal for evaluating interactive IR systems, except for the lack of control and the expense incurred in during experiments (involving real users with real and evolving information needs). Motivated by the demand of evaluation methods that take into account both the system-driven and the user-centred traditions, Borlund (2003b) present an alternative evaluation framework for interactive IR (IIR). The aim of the framework is twofold: i) to allow for the controlled evaluation of IIR system in realistic scenarios considering multidimensional and dynamic information needs and relevance judgements and ii) to include in measures of system performance the non-binary nature of relevance judgements. Borlund's proposed framework is a reaction to the demands stemming from what has been termed as the three revolutions (Robertson & Hancock-Beaulieu 1992):

1. The relevance revolution: the fact that a request is not the same as an information need and that relevance should, therefore, be judged in relation to this need and not

a request. Additionally, there is an increasing acceptance that relevance is dynamic and multidimensional.

2. The cognitive revolution: an extension to the relevance judgement process in which informations needs are perceived to be multidimensional, dynamic and personal.

3. The interactive revolution: the fact that systems are interactive by nature (or are becoming more interactive) and that the system-driven approach at evaluation does not include interaction nor information seeking processes.

Borlund & Ingwersen (2000) acknowledge these revolutions and proposes three experimental components as part of their evaluation framework:

1. The involvement of potential users as test persons

2. The application of dynamic and individual information needs

3. The use of multidimensional and dynamic relevance judgements

The relevance revolution is acknowledged by involving potential users as test users. Relevance is then judged in relation to the user's information needs and the situation in which the needs arise. Further, the relevance of the information presented is also judged in an interactive and non-binary way, hence the concept of relevance is such that relevance is treated as being multidimensional and dynamic. By allowing end users to interact with the system the dynamic and multidimensional nature of information needs, and hence both the cognitive and interactive revolutions, are considered in the evaluation framework.

Realism is included in the evaluation framework by the involvement of end users in the experimental settings. This realism is reinforced, and control is achieved, by the use of simulated work task situations (Borlund & Ingwersen 1997). Simulated work task situations are a semantically open description of the context of a given work situation, i.e. they are a cover-story that provides context and describes a situation in which IR is needed. The simulated work task situation triggers the information needs in users. Realism is provided as simulated work situations lead users into a cognitive state which creates information needs; needs that need to be satisfied before the user can move on.

The experimental control is provided by the fact that simulated work situations remain the same across users (and possibly systems). Relevance judgements are made in relation to the information needs triggered by the simulated work situation, hence they provide comparable cognitive and performance data.

The approach proposed by Borlund & Ingwersen (2000) may be appropriate for conducting experiments in LBD, whether they are to test the performance of a give system or to explore some of the yet unexplored cognitive aspects involved in LBD. As the end-users of the system under test are involved in the evaluation loop the experimenter has then the option of gathering cognitive data. In the case of LBD, this means inviting researchers to take part of the proposed experiment and have them use the system under evaluation. Only then it becomes possible to effectively observe whether researchers are able to detect salient intermediate or target topics for further evaluation. Furthermore, which topics and why they are deemed salient becomes observable. This opens up the possibility of breaking free from having to replicate Swanson's discoveries onto a more general evaluation approach where discoveries are those that the researchers involved deem as such. By keeping the tasks stable across end-users (researchers) and potentially across systems, one attains control. Effectively, this provides anchors against which relevance judgements are made. These anchors are the artefacts that make it possible to compare results across both end-users and systems.

2.6 Evaluation of LBD Systems

Evaluating knowledge discovery systems is a complex task. One of the main obstacles is that if systems are successful they are, by definition, capturing new unproven knowledge (Pratt & Yetisgen-Yildiz 2003). Moreover, much of the success of a system may actually depend on the expertise and interpretation of the operator. A system will perform well as long as the operator is able to interpret the results and draw inferences from them. Systems do not make discoveries, people do.

Swanson's initial discoveries –fish oil-Raynaud's disease (Swanson 1986a) and migraine-magnesium (Swanson 1988b)– were validated by latter evidence as provided by medical

researchers (Gordon & Lindsay 1996). Finding supporting evidence by conducting clinical trials (in the case of the medical domain) is one form of discovery validation. Effectively, this is the approach that Wren et al. (2004) followed. However, this approach may not only be very expensive but also not be feasible at all in some cases.

As explained earlier, most researchers limited themselves to replicating Swanson's initial discoveries, while some either did not evaluate their systems or used human experts in an ad-hoc fashion for this task. Those that attempted to replicate Swanson's initial discoveries reported to have been successful at doing so. To measure the correctness of the results, most researchers resorted to measuring accuracy. The top suggested topics were usually inspected and compared to those reported by Swanson albeit in a rather subjective fashion. Researchers usually informally measured recall (Baeza-Yates & Ribeiro-Neto 1999) as they measured how many of the links reported by Swanson their systems had predicted (in any one position in the ranking). However, measuring how well the system replicates these discoveries is only evaluating the predictive power of it (albeit in a limited fashion). There are many factors in the performance of a system one could evaluate, and the ones one chooses to evaluate will depend on the ultimate goal of the system. For systems where the goal is to predict, as precisely as possible, the potential links between seemingly unrelated topics, measuring their accuracy/recall at predicting an already known piece of knowledge may be, at first sight, appropriate.

Systems aimed at aiding researchers make discoveries are more complex to evaluate as they are likely to be inherently interactive. Not only is their predictive power to be evaluated (as a minimum requirement, these systems should exhibit a level of predictive power) but also other factors such as their interface, usability and speed are to be evaluated as well. Two studies either evaluated or observed other factors such as interface usability and search habits (Smalheiser et al. 2006, Skeels et al. 2005).

Supporting evidence for Swanson's fish oil-Raynaud's discovery (the most replicated throughout the literature) has been provided, however there is no indication that the original links found are the only ones connecting the literatures. As newly crafted algorithms suggest potential links, there is no clear approach at measuring how "good" they

are not only in terms of their own merit but in comparison to Swanson's discoveries. Effectively, Weeber et al. (2001) report that even though their system did not rank highly the original fish oil related intermediate topics, other fish oil related topics were and that researchers should be able to infer the link based on those as well. The interpretation of these suggested links poses an additional problem. Most researchers assume, or suggest, that a researcher should be able to detect salient topics. However, this assumption is made regardless of the researcher's background experience, ability to use search engines, and information needs. It is also worth mentioning that while finding the hidden link between the active ingredients in fish oil and Raynaud's disease led Swanson to propose the hypothesis that the use of dietary fish oil might ameliorate the symptoms of Raynaud's disease, it also led Swanson to propose that there may be more hidden links awaiting for their discovery and to develop LBD. Hence, the interpretation of these suggested links is a factor to consider even for those systems in which the ultimate goal is to discover hidden connections.

Effectively, as reported by Smalheiser et al. (2006), unexpected uses of their application were observed, two of which were the visualisation of links between two topics (even if known)[8] and the browsing of literature detached from the searcher's research field (literature evaluation in a new context). Hence search tasks and outcomes, other than the discovery of hidden relationships, may be of value to searchers and should be evaluated. Furthermore, *relevance* in Swanson's discoveries could be interpreted as *causal relevance*, i.e. the topics A, B and C are deemed *relevant* precisely because "A affects B" and "B affects C", however relevance in these unexpected uses, and others, may be harder to define and assess.

2.7 Summary

Almost all approaches at discovering relationships rely firstly on co-occurrences, whether of n-grams or other entities, and secondly on a combination of filtering and ranking. Co-occurrences are measured on the basic building blocks chosen by the researcher, e.g.

[8] A use of LBD that had been previously suggested as a potential use by Gordon et al. (2002).

n-grams Lindsay & Gordon (1999). Filtering is then usually applied on the statistics of these building blocks. Infrequent items are filtered out on the assumption that if an item is too infrequent then it is likely to be noise. Too frequent items are also filtered out as they are considered to either be noise or too common to be of interest. Lastly, the remaining items are ranked before being presented to users. This ranking step varies according to both the researchers' idea of what characteristics of an item makes it potentially more profitable than others.

Evaluation approaches are varied, however they all can be modelled using the system-driven tradition of evaluation of IR systems. This includes a tendency towards performing this evaluation as automatically as possible and using the two original discoveries made by Swanson (Swanson 1986a, Swanson 1988b) as ground truth. Researchers then evaluate the performance of their systems by observing how many of the intermediate (or target) topics from these two discoveries are suggested by their system. Additionally, the ranks of these intermediate topics is measured and taken as a sign of (dis)favourable performance. These observations are subjective in two respects:

- The matching between the concepts produced by the systems and those reported by Swanson is done by the researchers and as such is not objective and
- The ranks of these concepts is also evaluated subjectively as to whether they are "reasonable" or not.

Observing how well a system has reproduced Swanson's discoveries poses some initial drawbacks. The dataset is composed of two single examples and is limited in size. Small datasets not only may prove to be ineffective in training models (Halevy, Norvig & Pereira 2009) but when it comes to evaluating systems, reproducing a reduced set of golden outcomes limits the conclusions that can be reached on the effectiveness of a technique. Additionally, over-tailoring a system to reproduce Swanson's findings may limit its generalisation power. An interesting approach at overcoming the data size limitation imposed by reproducing Swanson's discoveries is proposed by Van Der Eijk et al. (2004) whereby the golden dataset is crafted by following back the references of a set of articles detailing

the testing of a hypothesis. It is not entirely clear, however, how the dataset is to be constructed since starting and target topics should be defined as well as the intermediate topics.

The approach of evaluation followed by all researchers is semi-automatic in the sense that while not fully automated (researchers manually inspect the ranked lists of topics and take different measurements) the final end-users are never involved. Swanson (1991) argues that fully automating the discovery procedure is complex at best. The main issue is that neither the syntactic nor the semantic structures provide for the deductive chain of reasoning in which the author incurred. Additionally, Swanson (1991) suggests that human experts are to be involved to complement the potentially automated procedure of literature searching. Human experts, in this context, are to provide the background knowledge needed to interpret the statistical cues offered by the system as extracted from the literatures. This is contrary to the assumptions made during nearly all evaluation approaches that users should be able to make the discoveries once they have been presented with the appropriate combinations. This assumption is to be considered with care as it is here where different factors, for instance the users' background knowledge, come into play.

The notion of relevance is never discussed in the literature explicitly although it is suggested that *relevance* is present in all steps of the process. One must be careful though as to what one means by *relevance* since the word is heavily overloaded with meaning and sentiments. We refer to *relevance*, at this stage, to any interpretation that is inherently subjective and has been done by the researcher on behalf of the end user of a system. As such, this simplistic view on relevance encompasses the interpretation of extracted concepts from documents, the appropriateness of a combination of concepts as presented by a system, the saliency of a particular topic (or combination thereof) and the utility of a piece of information in relation to the completion of a goal.

During the text modelling phase of all systems, where concepts are extracted from natural language documents, is where the first decision is made. What is an appropriate representation of a concept extracted from text? The basic building blocks, such as

n-grams, are decided beforehand by the researcher since the system needs them to perform the desired processing and relationship-finding. This is a reasonable and necessary decision from a algorithmic point of view. However, it becomes less clear how reasonable each representation becomes when it is used during evaluation. For instance, to evaluate their system, Weeber et al. (2001) try and replicate both of Swanson's original discoveries (Swanson 1986a, Swanson 1988b). The authors suggest that, while the original fish oil concepts were not ranked highly, other concepts related to fish oil were and that researchers should be able to still recognise fish oil as a target concept from them. The main assumption behind this suggestion is that the presented combination should be interpretable, regardless of any other factor, by a researcher who would then go on to suggest the existence of a relationship. This is an assumption of relevance, however the interpretation has been made on behalf of the end users and this is where the process becomes less clear.

All throughout the literature the presence of a golden combination is taken to be a sufficient condition for the derivation of a relationship. When evaluating, researchers observe and count how many of the original intermediate topics have been found, ranked highly and hence proposed for further inspection by the system under evaluation. Firstly, it is assumed that the representation of the combination, be it n-grams or any other more complex structure, conveys enough information for it to be salient, i.e. the combination is readily interpretable without ambiguity. Secondly, it is assumed that the combination not only contains enough information but that the information present is of the right kind. For instance that the combination is easily mapped to a higher-level representation that corresponds to one of the original discoveries, e.g. "omega-3" is suggested by the system as potentially interesting and this is in turn mapped, by the evaluating researcher, to "fish-oil". Finally, the sole presence of a golden combination is taken as a positive sign of performance and it is inferred that if researchers were presented with it they should be able to derive the original relationship.

Additional assumptions are made on the context in which the discovery task is to be performed. Initially it is assumed that *any* researcher, regardless of contextual circum-

stances such as personal background and research experience, should be able to derive meaning from a golden combination once it has been found. This is to say that the leap from a golden combination to a meaningful discovery is granted once the right combination has been selected. Additionally, it is implied in this assumption that there is a single discovery to be made after a particular golden combination. Essentially, because the researchers know about Swanson's findings, they know what to look for, so when found they tend to conclude that any other researcher should be able to do so.

We argue that there are too many areas where the subjective judgement of the end-user play a crucial role and that these should be studied in more detail. However, it is not entirely clear how to proceed. On the one hand, one must retain control on how such study is performed. Control is necessary so that the results are comparable and repeatable. We analysed an alternative evaluation proposed by Borlund (2003b) in which not only performance data is obtained, but also cognitive data regarding the interaction and information seeking processes can be gathered. The method relies on the concept of simulated work task situations for experimental control and involves the use of potential end users as test persons to accommodate for realism. Simulated work task situations provide context and describe a situation which leads users into a cognitive state in which information needs arise and have to be satisfied before users can move on. Additionally, it was suggested that these properties of the framework make it appealing to conduct the experiments in LBD, whether they are designed to evaluate systems or explore the aforementioned cognitive aspects involved during an LBD discovery task.

In Chapter 3 the reader is offered a description of the design of a user study tailored towards finding an answer to the three main research questions of this dissertation. This study has been designed keeping the components of the framework proposed by Borlund (2003b) and hence:

1. Real end-users of the system are included: researchers were invited to take part of the study and use our system,

2. Real and dynamic information needs are applied: researchers were provided with a task that only provided a context in which they should formulate and execute their

searches, and

3. Multidimensional and dynamic relevance judgements are made: researchers judged the literature presented in any way they wished to and as many times as they deemed necessary (depending on how many times it was presented to them)

Chapter 3

Study

As discussed in previous chapters, relevance is not only dynamic but subjective. As such it is very hard to come up with an objective and universal measure of relevance. More so, in the context of LBD, relevance might be especially complex as it is derived from the combination of logically related literature. To better understand what people refer to as *relevance* and the reasons why it is derived in this context an observational study was conducted between the months of January and August of 2008. The study is described in the following sections. The purpose of this study was to observe the relevance criteria, as defined by Barry & Schamber (1998), used by participants when assessing the relevance of related literature. During the study data was gathered using a combination of feedback forms and talk aloud protocols (Ericsson & Simon 1993).

Harter (1992) suggests that only weak relevance, or hope for relevance, can be derived from reading surrogates of documents. This means that once a user is presented with surrogates of documents, as retrieved and generated by an IR system, the user can only hope that the document related to the presented surrogate is relevant. It is in this sense that relevance becomes *weak* (as opposed to full or strong relevance). In the context of LBD, the implication is that end-users of an LBD system would only derive full relevance of a suggested combination of topics after having read the documents that supported it. The focus of the study was on the closed model of search as it is at this stage where end-users try to either confirm or reject the potential of a hypothesis. As explained in

section 2.2.2, the closed model of search is aimed at aiding the user search for literature supporting (or refuting) a particular already formed hypothesis.

The original discoveries using LBD were restricted to the scientific community. In particular the scientific medical community. Even though attempts have been made to extrapolate the mechanisms and search patterns outside this community, e.g. the work done by Cory (1997) and by Gordon et al. (2002), much of the literature on LBD is in this domain. For this study, however, the scope was broaden by inviting participants from three different communities: the computer science community, represented by the School of Computing; the information management community, represented by the Information Management Group and the pharmaceutical community, represented by the School of Pharmacy. All these schools are part of the Robert Gordon University, located in Aberdeen, Scotland.

This chapter is structured as follows. In section 3.1 a description of the study and the two sessions that compose it is offered. The systems used in the study are described in section 3.2. Section 3.3 contains a description of the population that participated in the study. A description of the affiliations and categories, in terms of research experience, is offered. The collections used in the study and their particulars are discussed in section 3.4 while the search tasks and measurements taken are discussed in sections 3.5 and 3.6 respectively. Finally, in section 3.7, a description of the different types of data analysis that were performed on the gathered data is offered.

3.1 Method

The study consisted of two sessions with a time gap between them of no more than a week. This time gap was necessary as the system used by the participants during the study processed offline the results from the first session and this process took, on average, between 5 and 6 hours. The length of the time gap was chosen so that results from the first session would still be present in the participants's minds when doing the second session. The results of the offline processing, topics and potential relationships between them, were presented during the second session to be investigated by each participant.

3.1.1 First session

Participants were asked, before beginning the session, to read, agree to and sign a confidentiality agreement. The agreement stated that their data would be anonymised and that no personal references of any kind would be made on the write up of the study. Participants were assured that their data would remain secure and appropriately stored. Once the agreement was signed, participants were asked to fill in a form with information about their background such as research experience and confidence in using search engines. The form can be seen in appendix A. Instructions on how to operate the provided system were delivered next. The system used in this session is described in section 3.2.1. Once comfortable with the system participants were given the search task. The task consisted required that the participant searched for and found five documents that described or represented the participant's area of research. The search task given to the participants can be seen in Figure 3.1.

The goal of the first session was to gather data from the participants to initiate an automated open search. This initial open search is needed as only two paths can be followed to a closed search:

1. The user already has a relation in mind to investigate further or
2. The user has formed a relation in his mind after looking at the suggestions resulting from an open search.

The documents selected during the search session were interpreted as a representation of the participant's area of research and used to seed the automated open search as described in section 3.2.2. There was no time limit imposed on this session.

Representing an area of research

Representing an area of research might be a practically impossible task. Yet for this study it was needed that the participants provided such a representation in a form that either a person or a system could work with. An initial approximation of this representation could be a set of words. This approach would be one which participants could be familiar

with. Hence, one could have asked them to describe their area of research using a few words and then work with them. This would have provided, however, a limited view of their areas of research. This limitation is better worded in one of Swanson's works in which several postulates of impotence were put forward. It is the first postulate that states that *"an information need cannot be fully expressed as a search request that is independent of innumerable presuppositions of context"* (Swanson 1988a). Context which includes, amongst others, the participant's background knowledge and the database being searched. By providing a set of keywords participants would be hoping that:

- There are documents in the database that match the keywords and that
- The matched documents, if any, provide a good representation of their area of research.

Should any of these not be satisfied, participants would be left with an inadequate representation of their field of work; one that is lacking in information and generally incomplete or even erroneous (in respect to the contents of the database). Accepting the veracity of this postulate also rules out providing an interface that integrates the two search models of LBD. Moreover, the computationally expensive nature of the algorithm selected only aggravates the situation.

As an alternative approach participants were asked to search for and provide documents which represented or described their area of research. By interpreting these documents as a description of the participant's research area participants were freed from the burden of having to come up with terms that would, to a certain extent, describe their research. Following this approach meant that topics had to be automatically extracted from the documents. This also allowed for unsuspected but related vocabulary to creep in into the search and relationship discovery. Incidentally Swanson's original and current procedure works as described. Users of Swanson's system are asked to search PubMed for the source and destination literatures and then provide these to Swanson's system to work with[1].

[1]Visit the Arrowsmith website (`http://arrowsmith.psych.uic.edu/cgi-bin/arrowsmith_uic/start.cgi`) for more information.

```
Dear participant,

I'd like to ask you to search for documents that describe your
area of research, or an aspect of it. Whenever you think you have
found one, please write down the document ID (located at the
top of the viewing window) on the provided sheet. The purpose of
this search is so that the system under test can then suggest
topics that might be related to your area of research for you to
further investigate.
```

Figure 3.1: Search task introduced during the first meeting with participants

3.1.2 The offline processing

During the offline processing the documents found during the first session were used as the starting topic for an automated open search. The automated open search is described in section 3.2.2 and it results in a network of topics in which the entry topic (see Figure 3.6) represents the participant's area of research. This automated construction of the network results in a practically exponential explosion of relations. Asking the participants to investigate them all would render the study pointless. Instead the system ranks the final topics C_{ij} according to a simple algorithm (see 3.2) and presents the participants with the top 10 topics. During the second session participants were asked to investigate the potential relationships between their area of research and 3 out of these 10 topics.

3.1.3 Second session

At the beginning of their second session participants were given training on two aspects of the session: the system and the talk-aloud protocol. Once the instructions were delivered, participants were allowed to practise both using the system (on example but realistic data) and talking-aloud. This practice session had a time limit of 15 minutes. At the end of their practice, participants were given their search task.

Unlike the first session in which all participants were all given the same search task, participants were given a search task that corresponded to their expressed level of research experience. Information about the participants's research experienced was gathered at the beginning of the first session. An example search task can be seen in Figure 3.2. To

Dear participant,

Thank you very much for taking part of this experiment.
Now that you have done part 1 of it, it's time to do part 2.
In part 2, you are given a simulated work task. Briefly, a
simulated work task is a description of a task you should
perform. Below is the description:

Simulated work situation:

At a supervisory meeting you received constructive
criticism concerning the breadth of your literature
review. Even though the work you've been doing is
very good, your supervisor feels it is a bit too
specific/constrained. Your supervisor suggested you
look for connections between your research and other
research areas, e.g. other areas which have
techniques or ideas you might use in your research
or areas where your research might contribute. Your
supervisor suggested you identify these potential
areas as well as the pertinent literature so you can
discuss them together in your next meeting.

You will have a time limit of 1 hour to investigate 3 of the
possibly related topics you will see on the screen. You can
stop at any time though.
You will be required to think aloud as you investigate the
potential relations.

Figure 3.2: Example search task introduced during the second meeting with participants

complete the search task participants had an hour. During this hour they were asked to investigate exactly three of the ten potential relations presented. Participants were also asked to write down the document identifier whenever they thought they had found a document that they thought complied with the request. The form used during this second session to write down the document identifiers can be seen in Appendix A.

At the end of the session participants were asked to fill in a questionnaire which can be seen in Appendix A.7. Information gathered referred to, amongst others, the quality of the results obtained during the session as well as general comments about the whole procedure.

3.2 The System

In this section the details of the systems participants used in each session are described. Firstly the system used during the first session is described. In this session participants used a simple keyword driven search engine. This engine offered a search box and returned a ranked list of document surrogates. Each returned document surrogate had a hyperlink to access the full document. Depending on the affiliation of the participant, the system searched the appropriate database. Secondly the offline processing of the documents retrieved during the first session is described. During this offline processing topics were automatically extracted using a technique named Latent Dirichlet Allocation (Blei, Ng & Jordan 2003). These topics were used as queries to retrieve a new set of documents from the database. An extra set of topics were automatically extracted from this new set of documents. The process finishes by suggesting topics which are potentially related to the participant's area of research. Lastly the system used during the second session is described. This system presents and makes use of the topics discovered during the offline processing. Participants used the interface for navigating topics and retrieving the literature related to them.

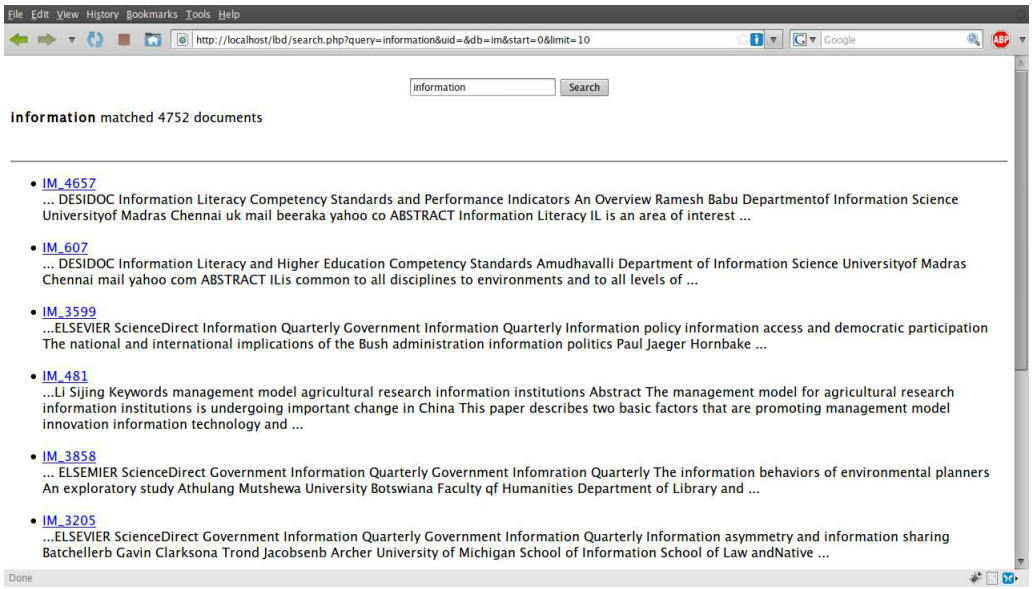

Figure 3.3: User interface of the system participants used during their first session

3.2.1 First Session

The search facilities provided during the first session are rather rudimentary. A screenshot of the user interface is presented in Figure 3.3. In the interface, after searching for "information" it can be seen that the query matched 4752 documents. The document surrogates of the initial 10 documents retrieved are listed below the horizontal line. The document surrogates were built as follows. The text in the hyperlink (in blue in Figure 3.3) is the document identifier — the internal code used when indexing the document. Initially this might seem like a poor choice for the text in the hyperlink as, for instance, the document title could have been used. Using document titles, however, was not feasible. As it is explained in section 3.4, the original documents were in the Portable Document Format (PDF). This format is particularly problematic when it comes to parsing and extracting particular portions of text such as titles. Given the different layouts of the documents, extracting titles was simply not possible. The document snippet below the hyperlink consists of the first 5 sentences of the document (a sentence was taken to be any character different from the newline character).

Clicking on a hyperlink would bring up a new window with the full contents of the document. An example document can be seen in Figure 3.4. On the top left corner the

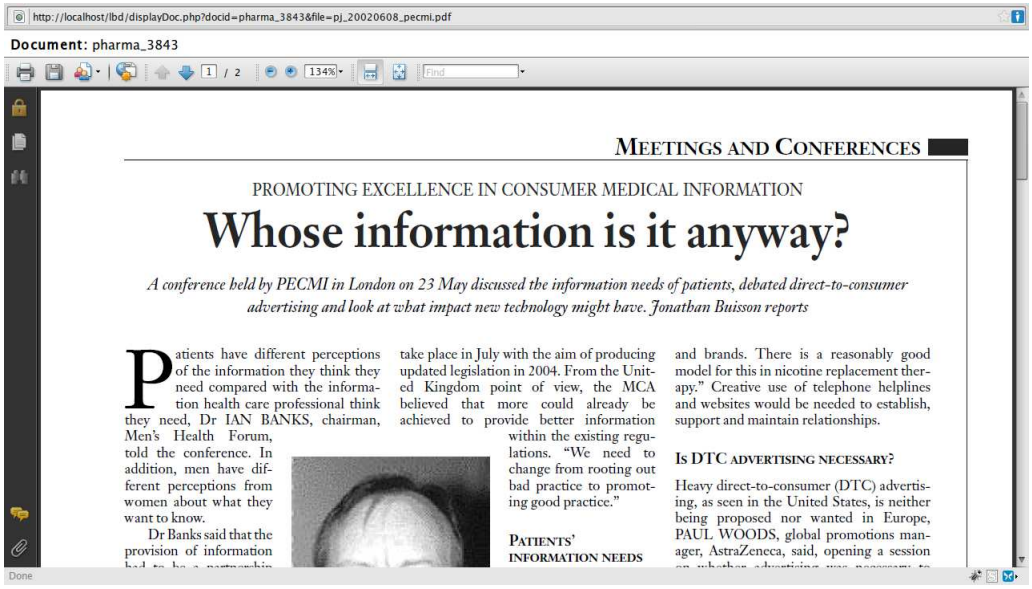

Figure 3.4: An example document. On the top left the document identifier is displayed.

document identifier is displayed again so that users could quickly identify it should they decide to keep the document.

The engine behind the interface was implemented using the Kullback-Liebler divergence retrieval model (Zhai & Lafferty 2004)[2]. In this retrieval model documents and queries are represented as statistical language models (Ponte & Croft 1998). A statistical language model is defined as a probability distribution over the vocabulary (the set of unique words in the collection of documents). The score of a document, then, can be defined as being proportional to the Kullback-Liebler divergence measure between the language model of the query and that of the document (Lafferty & Zhai 2001). For this system, the model was used as out of the box, i.e. no parameters were tuned. Potentially, this could have impacted on the length of the first session and participants might have taken longer than they would have had the parameters of the engine been tuned.

3.2.2 The Offline Processing — An Automatic Open Search

The initial step in the process is to automatically extract the topics contained in the documents selected during the first session. This is achieved by modelling the documents

[2]The actual system was written using The Lemur Toolkit software, Version 4.1 for Unix. Copyright ©2009 University of Massachusetts and Carnegie Mellon University.

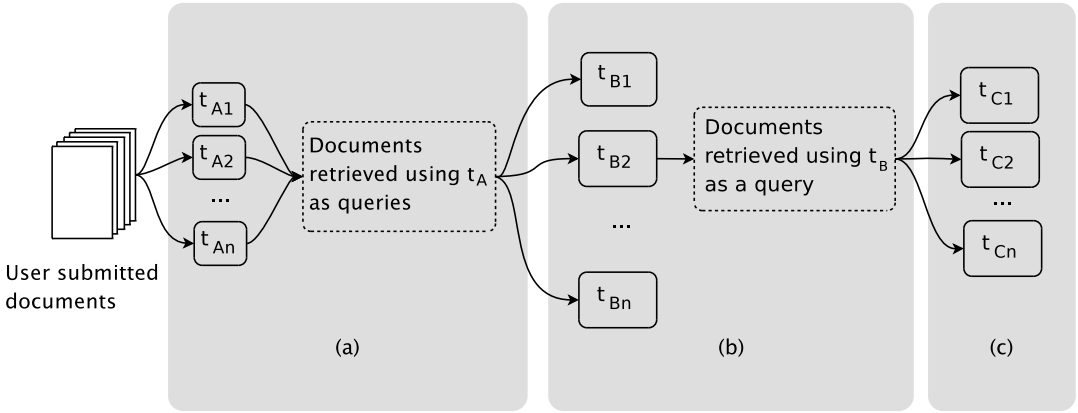

Figure 3.5: Visual representation of the open search algorithm. The first step is to model the topics contained in the user submitted documents and retrieve more documents about them (step a). The second step is to model the topics contained in the pooled documents and retrieve even more documents (step b). The final step is to model the topics contained in each document set retrieved in the previous step.

using a technique called Latent Dirichlet Allocation (Blei et al. 2003). In LDA each document is represented as a mixture of topics, where a topic is a probability distribution over words. We refer to the initial topics as topics t_A. These topics are each used as a query to retrieve more documents. To build a query representing each topic the top three words, according to their probability of being generated by the topic $P(w|t_A)$, are taken[3]. These queries are issued to the engine of the system used in session one and the top 50 documents are retrieved. This is depicted as step *a* in Figure 3.5. The assumption behind this first step is that topics discussed within a document share a relationship. Retrieving more documents using the top terms for the central topic should then increases the frequencies of occurrence of the related topics.

All documents retrieved in step *a* are pooled together. Topics are then automatically extracted again using LDA. This is effectively extracting the first layer of related topics in the open model search (referred to as *B topics* in the LBD literature). The process of building queries from topics and retrieving documents is repeated, however this time the retrieved documents are not pooled together. This is so that individual B-topic could be

[3]This is a rather ad-hoc procedure. A more principled approach would have been to take the actual topic —a probability distribution over words; a language model— and used directly as the query model as the engine was implemented using the Kullback-Leibler divergence model. This was not implemented for this study as the library used to build the system did not provide access to the low level language modelling framework and hence a custom language model could not be used to query the engine.

linked to their corresponding set of C-topics. This is depicted as step b in Figure 3.5.

On each set of retrieved documents a new round of automatic topic extraction is performed. These are the final t_C topics which are potentially related to the t_A topics through one or more of their directly related t_B topics (step c in Figure 3.5).

The result of this process can be represented as a tree. The root node of the tree is the starting topic as represented by the documents selected in the first session. The inner nodes of the tree are the immediately related topics t_B and the leaves are the potentially —indirectly— related topics t_C. This tree is depicted in Figure 3.6. Each edge represents a potential relationship between the nodes.

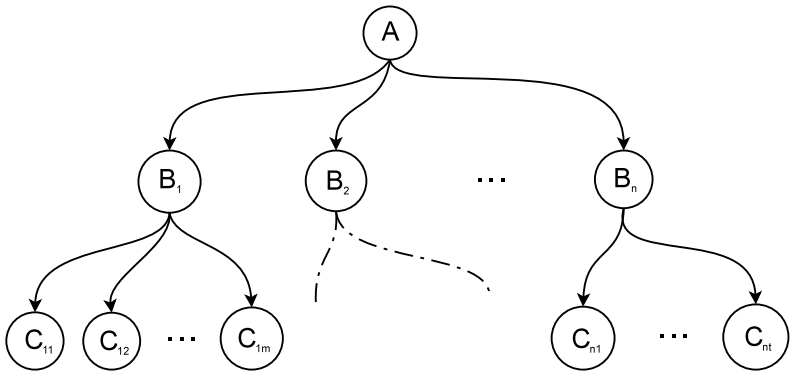

Figure 3.6: Topic tree. The A node is the participant's initial topic. The inner nodes, the B_i nodes, are the immediately related topics as described by both the literature and the topic extraction algorithm. The tree leaves are the indirectly related, to the initial topic A, C_{ij} topics.

3.2.3 Second Session

The system used during the session two consists of two parts: the entry screen and the navigation screen. The purpose of the entry screen is to provide an overview of the potentially related topics. On the left panel a representation of the initial topic, the participant's area of research, is provided. This representation is actually the query words used to retrieve more documents during step a of the offline processing. It is a set of keywords. On the right panel the 10 potentially related topics are listed. These topics are also represented as a set of keywords. These keywords are the three most important words for each potentially related topic. The number of terms selected for each C topic is

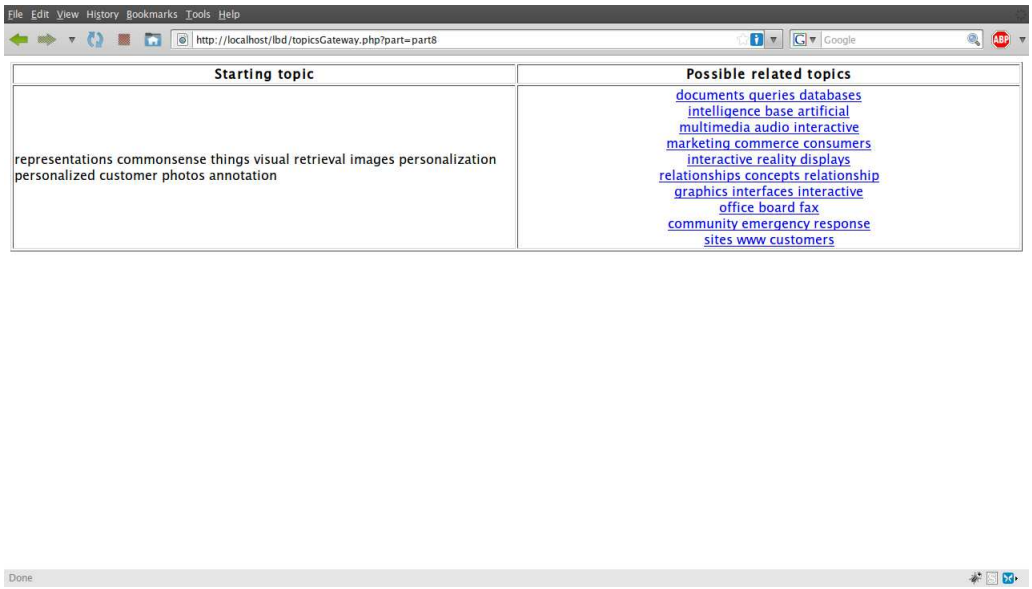

Figure 3.7: First screen users see when doing the second part of the study. In the image the terms representing the initial topic are presented on the left panel whereas the initial C topics, also in the form of terms, are displayed.

motivated by the observation that the average number of terms in user-submitted queries during web-searches is 3 (Spink, Wolfram, Jansen & Saracevic 2001). Although LBD might not particularly fit the context of web search, it was decided that this was still a reasonable approach. The importance of a word is measured by the probability that the topic will generate the word. The user interface of the entry screen can be seen in Figure 3.7.

On the entry screen the potentially related topics are actually hyperlinks. When a user clicks on any of these hyperlinks he navigates to the second screen; the navigation screen. On the navigation screen, the user can navigate and investigate the intermediate topics as well as the supporting literatures. The user interface of the navigation screen consists of three panels. On the top panel both the initial topic and the potentially related topic are listed again. This provides the context in which the intermediate topics are to be interpreted. In the middle panel the intermediate B topics are listed. Three columns of topics are presented to the user. Each entry is a hyperlink where the text is the top three words for the intermediate topic. Clicking on any intermediate topic performs a search for the literature supporting the topic. The lower panel, which is split into two,

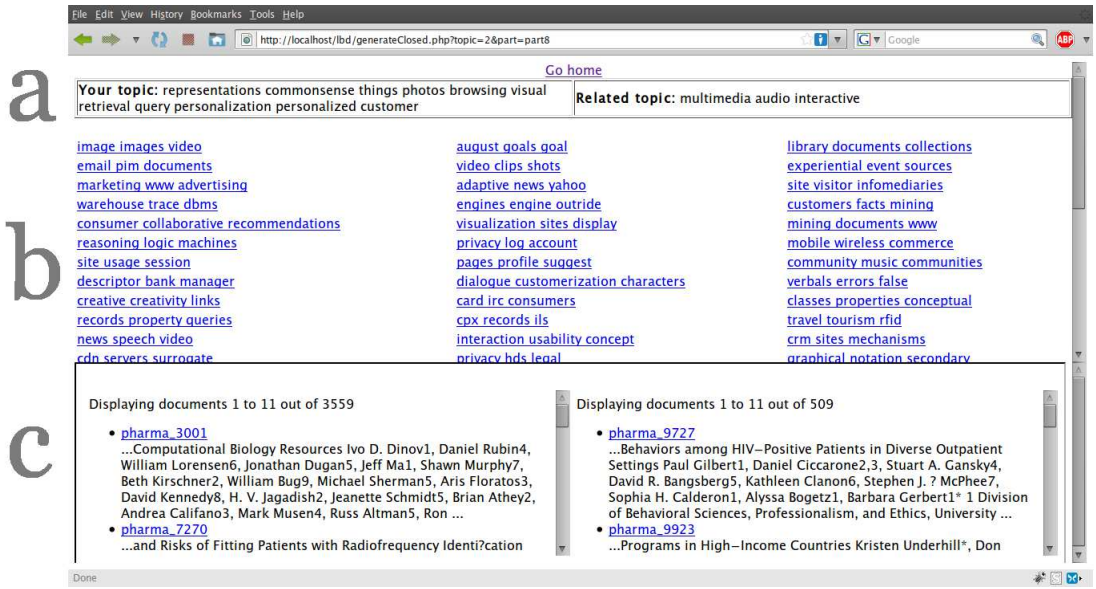

Figure 3.8: Navigation screen. The top panel (a) displays both the initial topic (listed as "Your topic") and the potentially related topic (listed as "Related topic"). The middle panel (b) lists the intermediate topics. The bottom panel (c) contains the supporting literatures.

is used to display the supporting literatures. On the left hand side of the lower panel the literature that supports the relationship $A \leftrightarrow B$ is listed. The literature supporting the relationship $B \leftrightarrow C$ is listed on the right hand side of the panel. The supporting literature was listed as a ranked list of document surrogates. The title of each surrogate is a hyperlink. Clicking on this hyperlink opened a new window with the full document contents (as seen in Figure 3.4). The document snippet consisted of the first five sentences in the document. A screenshot of this screen can be seen in Figure 3.8.

3.3 User groups

Participants from three different schools were invited to take part of the study. The three participating schools were the School of Computing, the School of Pharmacy and the Information Management Group (from the Business School[4]).

Potential participants were emailed with a request for participation. Prior to this the heads, of the respective schools were contacted and explained the purpose and mechanics

[4]Note that the name of the school, the Business School, is actually an umbrella name which groups many different schools, one of which is the Information Management Group.

Level of Experience	Description
Research Student	no differences were made on whether students were at their final or initial stages
Researcher/Lecturer	researchers with some years experience in research or lecturers who were still research active. This includes researchers working as either lecturers or post-docs.
Experienced Researcher	full time researchers with several years research experience. This includes senior lecturers and professors.

Table 3.1: Research experience categories.

of the study (this ruled them out immediately as potential participants), and asked for permission to contact the members of the school.

Participants were classified according to their expressed research experience into one of three categories as listed in table 3.1.

3.4 Collections

Inviting participants from three different schools meant that three different collections had to be created, one for each group, as no standard collections existed for this type of study. However different in content, all collections shared certain commonalities. First of all, all collections covered a variety of topics. Covering several different topics, always within a main theme, meant that participants were not restricted in their searches. Secondly, all collections contained a mixture of general public articles and scientific papers. This allowed for participants to have a good range of depth in the information they could find. Had the collections only contained scientific articles the second session would have had to be made longer than an hour.

All collections were indexed following the same procedure:

1. Converted documents from the Portable Document Format (PDF) to plain text: all collections harvested consisted of documents in the PDF format. The original documents were preserved for presentation purposes,

2. Removed too-frequent words that do not add information: all words (stop words) present in a simple list were removed from the documents and

3. Stemming: this standard step was not performed on the documents despite its potential detrimental effect on retrieval performance as the system estimates topics from document terms. Had the stemming step been performed the resulting estimated topics would contain the stemmed terms (instead of the original terms). We assumed that this would make the topics harder to interpret for the participants. Still this is an issue to investigate further[5].

The collections used during the study are further discussed in Chapter 4, Section 4.2.

3.5 Search Tasks

Simulated work task descriptions were written according to the three different levels of research experience as described in table 3.1. All tasks, however, were written with the intention of *pushing* participants outside their area of research and make them find (or create) relations with topics outside their own. This was masked as a request for a search for literature that aided the participant fulfil a particular task. An example task is presented in Figure 3.2.

Participants classified in the category of *research student* were given the search task depicted in Figure 3.9. This task suggested that, even though the work they had been carrying on was good, their supervisor had suggested they broaden their literary scope. Research students would have to search outside their own area of research and look for connections. Their ultimate goal was not only to make their supervisors happy (as we all did once) but also to enrich both their work and literature survey.

Participants in the category *researcher* were presented with the task depicted in Figure 3.10. The task suggested that they were immerse in the process of writing a grant proposal. As funding is vital to research activities, any help in getting a proposal accepted would be more than welcome. The search task mentions that a senior colleague had suggested the researcher write about the potential impact outside the main theme of the proposal. This would increase the chances of getting the proposal accepted. This sets the stage for

[5] An approach at solving this is the use of a dictionary for stemming and then doing a reverse lookup, however this may lead to situations where the stemmed word leads to at least two words, e.g.*walk* leads to at least *walked* and *walking*.

```
   Dear participant,

Thank you very much for taking part of this experiment.
Now that you have done part 1 of it, it's time to do part 2.
In part 2, you are given a simulated work task. Briefly, a
simulated work task is a description of a task you should
perform. Below is the description:

  Simulated work situation:

  At a supervisory meeting you received constructive
  criticism concerning the breadth of your literature
  review. Even though the work you've been doing is
  very good, your supervisor feels it is a bit too
  specific/constrained. Your supervisor suggested you
  look for connections between your research and other
  research areas, e.g. other areas which have
  techniques or ideas you might use in your research
  or areas where your research might contribute. Your
  supervisor suggested you identify these potential
  areas as well as the pertinent literature so you can
  discuss them together in your next meeting.

You will have a time limit of 1 hour to investigate 3 of the
possibly related topics you will see on the screen. You can
stop at any time though.
You will be required to think aloud as you investigate the
potential relations.
```

Figure 3.9: Search task given to all research students that participated in the study

the search for related literature.

Senior researchers were given the search task depicted in Figure 3.11. This task suggested that they had been invited to deliver a keynote speech at a very prestigious conference. To make their speech more appealing a colleague suggested they look for connections between their area of research and other research areas so that the speech was more focused on the grand-scheme of things rather than on the particulars of their research area. To do so, senior researchers would have to search for potentially related areas of research to mention in their speech.

```
   Dear participant,

Thank you very much for taking part of this experiment.
Now that you have done part 1 of it, it's time to do part 2.
In part 2, you are given a simulated work task. Briefly, a
simulated work task is a description of a task you should
perform. Below is the description:

  Simulated work situation:

  You are in the process of writing a grant proposal. A
  senior colleague as suggested you carefully write about
  the impact your proposed research might have on related
  fields or where related fields have ideas or approaches
  that you might exploit. Your colleague has advised that
  this might improve your chances of getting the proposal
  funded.

You will have a time limit of 1 hour to investigate 3 of the
possibly related topics you will see on the screen. You can
stop at any time though.
You will be required to think aloud as you investigate the
potential relations.
```

Figure 3.10: Search task given to all researchers that participated in the study

3.6 Measurements

Different measurements were made during the search sessions. Data was gathered in both written and verbal form. Data gathered included information on the searches performed and their results as well as background information on the participant. Data was not only anonymized but also kept securely to avoid both privacy breaches and tampering.

3.6.1 Written Data — forms and questionnaires

Participants had to provide information by answering different questions presented in three different forms. These questions aimed at understanding the quality of the documentation found (if any) amongst other things. Information regarding their background was also recorded.

At the beginning of their first session participants had to fill in the form depicted

Dear participant,

Thank you very much for taking part of this experiment.
Now that you have done part 1 of it, it's time to do part 2.
In part 2, you are given a simulated work task. Briefly, a simulated work task is a description of a task you should perform. Below is the description:

Simulated work situation:

You have been invited to deliver a keynote speech at a very prestigious conference in your research field. The organisers have kindly asked you to focus your speech on the future directions and implications of advances in your research field, especially on those fields outside your own. A senior colleague suggested that, in order to prepare your speech, you look for connections between your research and other areas of research, e.g. other areas which have techniques or ideas you might use in your research or areas where your research might contribute.

You will have a time limit of 1 hour to investigate 3 of the possibly related topics you will see on the screen. You can stop at any time though.
You will be required to think aloud as you investigate the potential relations.

Figure 3.11: Search task given to all senior researchers that participated in the study

in Figure A.1. This form is designed to capture as much information as possible on the participants's background. The background information collected includes,amongst others, the participants's professions, research fields and topics, their expressed confidence in evaluating literature and their preferred sources for literature.

During their second session participants had to record the documents they selected in a form. This form is designed not only to capture the document identifiers but also the suggested relationship supported. Participants had to write down the document identifier together with the intermediate topics which had retrieved it. Together with the initial topic and the indirectly related topic (topic C) the document identifier and the intermediate topic description provide a full picture of the interactions between them.

When the second session ends participants were presented with a final form. This form is designed to capture data on the search results of the closed model search system. The information gathered ranges from the topic variety present in the search results to the validity and intent regarding the connections suggested. This form is depicted in Figure A.6

3.6.2 Verbal Data — Talk Aloud Protocols

Talk aloud protocols are based on the idea that talking aloud while solving a task provides a view of the thoughts as the task solving process is ongoing. The assumption behind this idea is that people retain a small amount of information in a short term memory store. If you can tap into this memory store you can learn about the person's thought process in solving the task (Ericsson & Simon 1993). The information obtained could be used to improve not only the understanding on the problem-solving processes but also, for instance, to devise problem-solving computer algorithms. It was decided to use talk aloud protocols as they would provide a raw view of the relevance judgement process that users incur in when searching for literature.

Concurrent and Retrospective Reports

Two levels of reporting claim to be the closest reflection on the thinking process: concurrent reports and retrospective reports (Ericsson & Simon 1993). Retrospective reports refer to verbalisations where the thought process is no longer occurring. The person has to remember and summarise what they were thinking as they were solving the task. It can happen that the original contents of the memory store are changed as the person might leave out invalid reasoning steps, details that are deemed irrelevant and so on. Concurrent reports on the other hand happen while the thought process is ongoing. Reporting on the thought process as it happens provides with a much more raw view of the process, however it may also introduce irrelevant details (Green 1998). Additionally, the burden of verbalising thoughts concurrent to the action of solving a problem may overload the person's short-term memory and alter the normal process path (Ericsson & Simon 1993).

For this study it was decided to use concurrent reports as retrospective reports would not have provided the desired data granularity. Additionally, as participants were likely to judge the relevance of several documents during their second session a concurrent reporting would be more accurate than a retrospective one.

Talking aloud

The process of gathering verbal data relies on the participant talking aloud during the study. To ensure the quality of the verbal reports gathered, following two main guidelines is recommended (Ericsson & Simon 1993).

First, participants must be instructed to verbalise their thoughts. Considering that the nature of verbal protocol data can be influenced by the instructions received, these must be carefully and consistently worded (Ericsson & Simon 1993). In the instructions, participants should be encouraged to be precise in their talk-alouds and the researchers should reassure participants so that they are comfortable verbalising their thoughts using their own words (Pressley & Afflerbach 1995). Additionally, and to maximise reliability, all participants should receive the same instructions (Ericsson & Simon 1993). These procedures minimise interference with cognitive processing.

Second, to reduce both participant anxiety and misinterpretation of the instructions, participants should follow pre-session warm-up exercises. As the ability of different people to verbalise their thoughts varies, these warm-up exercises are also important to maximise the quality of the generated verbal protocol data (Pressley & Afflerbach 1995). These exercises usually require less than 15 minutes (van Someren, Barnard & Sandberg 1994) and result in several benefits. First, they ensure that participants have understood the instructions and that researchers and participants share the same understanding of the kind of data required (Ericsson & Simon 1993). Second, participant anxiety is reduced and so participants feel more comfortable in reporting their thinking (Ericsson & Simon 1993). To maximise reliability, and just like with instructions, all participants should not only follow the same warm-up exercises but these should also be designed as to be as similar as possible to the target task (Ericsson & Simon 1993, van Someren et al. 1994).

During the second session in the study, participants were instructed to verbalise all that came through their minds as they solved the task presented. These instructions were given verbally to each participant. Prior training was also provided. This consisted of a single warm-up session of up to 15 minutes navigating the system on example, but realistic, data. During the actual session, the verbal protocols were captured using digital audio recordings taken using a microphone connected to a PC.

Data Analysis

Reducing threats to data validity and reliability throughout the data analysis process can be achieved by following three suggestions found in the literature (Ericsson & Simon 1993, Pressley & Afflerbach 1995).

The first guideline refers to the transcription of the verbal data. When the recordings are transcribed, verbal data is to be transcribed verbatim, capturing as much verbal data as possible by including pauses, emphases, and indications of tone (Ericsson & Simon 1993). This additional information becomes secondary data sources as they assist the interpretation of concurrent verbal data (Pressley & Afflerbach 1995). The transcribed data is then to be segmented (divided into "utterances"). This step ensures that all the

Code	Description
...	pause or silence
[*read*] "text"	the participant is reading a document out loud; this appears usually in the form of a mechanical voice at a constant speed with the occasional mumbling
[*mumbles*] "..."	the participant is mumbling and the audio cannot be transcribed

Table 3.2: Codes used during the transcription step of the protocol

data is segmented in standard units for later encoding/analysis. Care must be taken, however, when defining the units (Ericsson & Simon 1993).

The second guideline suggests that a valid coding scheme that identifies major processes and patterns of knowledge in the data collected is to be designed. Special attention must be paid so that it facilitates cross-case analysis. The encoding of the data can be achieved with minimum threat to validity when the encoding scheme is developed from the data and, once developed, further data is encoded to check it (Ericsson & Simon 1993). However, there are advantages to building on existing encoding schemes. First, the method can be strengthened by applying and refining a common encoding scheme across data and second, the processes being studied can be further elaborated by the analysis of new data.

The third guideline pertains the assessment of the reliability of the encoding scheme and the encoding procedure (Pressley & Afflerbach 1995). The reliability of encoding schemes can be enhanced by the use of clearly defined codes, illustrated with examples (Rowe 1985) and the reliability of the coding procedure can be tested by using measures such as inter-rater agreement measures such as Kappa (van Someren et al. 1994) for determining the degree of agreement between independent coders in assigning codes to the utterances (Ericsson & Simon 1993, Pressley & Afflerbach 1995). Additionally, coders should practice using the encoding scheme until the codes are both familiar and applied consistently.

The recordings captured during the session were transcribed verbatim. The transcriptions were annotated using the tags listed in table 3.2. The annotated transcriptions were then split into "utterances". Utterances are defined as the minimum unit of speech that could be assigned a label from the encoding scheme. Often these minimal units were

bounded by breaths or pauses.

As coding schemes that suited the study were not readily available, to analyse the data gathered in the sessions a custom encoding scheme was developed. The scheme was designed so that three types of events could be categorised i) criteria used when assessing information, ii) any kind of interactions between the user and the system, and iii) the user's intents. Utterances were classified according to the following criteria:

- Interaction: any utterance that indicates the participant is performing an operation on/with the system or interacting with it, e.g. reading a document, clicking on a document surrogate, going back a page, etc.
- Intent: any mention of the participant's intentions regarding the obtained information or regarding their actions, e.g. using a retrieved document to impress their supervisor or initiating a search in the hopes of finding a particular type of information.
- Relevance Criteria: any mention of factors that may affect the participant's choices regarding whether they are to keep or not a document, e.g. if the user picks the document because it is a survey.

The encoders practised using the encoding on data gathered during a pilot study conducted to test the viability of the design of the study. The details of this pilot study are described in (Cerviño Beresi, Baillie & Ruthven 2008). Additionally, the reliability of the encoding was tested by measuring the overlap of the assigned codes by independent encoders on randomly sampled utterances.

3.7 Data Analysis

Once recordings are transcribed and segmented in utterances, they were labeled using the first level encoding described earlier. The utterances were then further analysed within each group.

3.7.1 Interaction

Utterances labeled with *interaction* were analysed to see if any search patterns emerged from the sessions. They were also analysed to confirm that the participants understood how to operate the systems.

3.7.2 Intent

Expressions of the participant's intents were observed and are presented in Section 4.5.

3.7.3 Relevance Criteria

Expressions of the relevance criteria used for either selecting or discarding documentation were the primary interest of this study. These expressions were classified further according to a second encoding scheme. The encoding scheme used was the one presented by Barry and Schamber in (Barry & Schamber 1998) which is briefly revisited in the following listing:

- Depth/Scope/Specificity: whether the information is in depth or focused, has enough detail or is specific to the user's needs. Also whether it provides a summary or overview or a sufficient variety or volume.
- Accuracy/Validity: whether the information found is accurate or valid.
- Clarity: whether the information is presented in a clear fashion. This includes well written documents and well as the presence of visual cues such as images.
- Currency: whether the information is current or is up to date.
- Tangibility: whether the information relates to tangible issues, hard data/facts are included or information provided was proven.
- Quality of Sources: whether the quality of the information can be derived from the quality of the sources of it. This includes authors as well as publications.
- Accessibility: whether there is some cost involved in obtaining the information.

- Availability of Information/Sources of Information: whether the information is available at that point in time.
- Verification: whether other information in the field, or the user, agrees with the presented information.
- Affectiveness: whether the user shows an affective or emotional response when presented the information.

According to Barry & Schamber (1998), *accessibility* refers to the cost or effort involved in obtaining the information. Effort, in their interpretation, refers to physical and not mental effort. If a document is available only through an interlibrary loan, then it would require physical effort from the user to obtain it. Cost involved refers to possible fees involved in obtaining such document. In this study documents were readily available and no fees were involved in obtaining them. Since the mental effort necessary to process the information is not interpreted to be a type of "effort", it was not expected that this criterion would be observed. *Availability* refers to physical availability of the document. Since documents were available at all times, this criterion was not expected to be observed. Despite these expectations all codes were included in the encoding scheme.

Extending the encoding scheme

Verifying the *validity* of a piece of information in a research field can be very hard to do for a newcomer. The task given to participants required them to branch out to potentially unknown areas of science placing them in the spot as newcomers. Considering this, *accuracy/validity* as a criterion, was not expected to be observed very often. What was expected, though, was that different forms of novelty would play an important role when users judged documents. Codes that would account for this were included. In the study done by Barry (1993) three types of information novelty are mentioned:

- Content novelty: whether the information is new to the user
- Source novelty: whether the source of the document is new to the user, e.g. an unknown author

- Document novelty: whether the document is new to the user

These codes were used to tag utterances that expressed if a document had been seen before (in the current session or not) and if the document or the information contained within it was known to the participant. The code *source novelty* was included to code utterances expressing, for instance, a known author writing in a different field or of a journal never read before.

As participants were asked to search for literature in potentially unknown areas of science it could well happen that the information found would be deemed non-relevant based on that they had actually not been able to understand it. In this situation participants could either silently reject the document or express their inability to correctly understand the information presented. To accommodate for this situation it was decided to include a code found in Barry's list of criteria denoted by the tag *ability to understand* (Barry 1994). According to Barry, utterances that denote "the user's judgement that he/she will be able to understand or follow the information presented" should be included in this category.

Participants's background knowledge, or experience, should also not be neglected as participants could possess prior experience that would enable them to make educated guesses much more easily which combined with the right information could lead to the creation of connections and in turn to positive relevance judgements. To encode the mentions of the use of background knowledge or information during the search session a code found in Barry's original listing was included: *Background Experience*. Barry (1994) states that this code is used to denote "the degree of knowledge with which the user approaches information, as indicated by mentions of background or experience". The encoding scheme used to tag the utterances found in the transcriptions is depicted in table 3.3.

3.7.4 Relevance Criteria Profiles

Once utterances have been coded they are grouped at the session level and counted, i.e. all mentions of a particular relevance criterion within the search session are added up and contribute to a single count for that session and that criterion. For any one participant

Tag	Description
Depth/Scope/Specificity	the extent to which information is in-depth or focused; is specific to the user's needs; has sufficient detail or depth; provides a summary, interpretation, or explanation; provides a sufficient variety or volume
Accuracy/Validity	information found is accurate or valid
Clarity	information is presented in a clear fashion
Currency	information is current or up to date
Tangibility	information relates to tangible issues
Quality of Sources	quality can be derived from the quality of the sources
Accessibility	the extent to which some effort is required to obtain information; some cost is required to obtain information
Availability	the extent to which information or sources of information are available
Verification	whether other information in the field, or the user, agrees with the presented information
Affectiveness	whether the user shows an affective or emotional response when presented the information
Background/Experience	degree of knowledge with which the user approaches information
Ability to Understand	user's judgement that he/she will be able to understand information presented
Content novelty	the extent to which the information presented is novel to the user
Source novelty	the extent to which a source of the document (i.e., author, journal) is novel to the user
Document novelty	the extent to which the document itself is novel to the user

Table 3.3: Encoding used to tag the utterances that express a relevance criterion

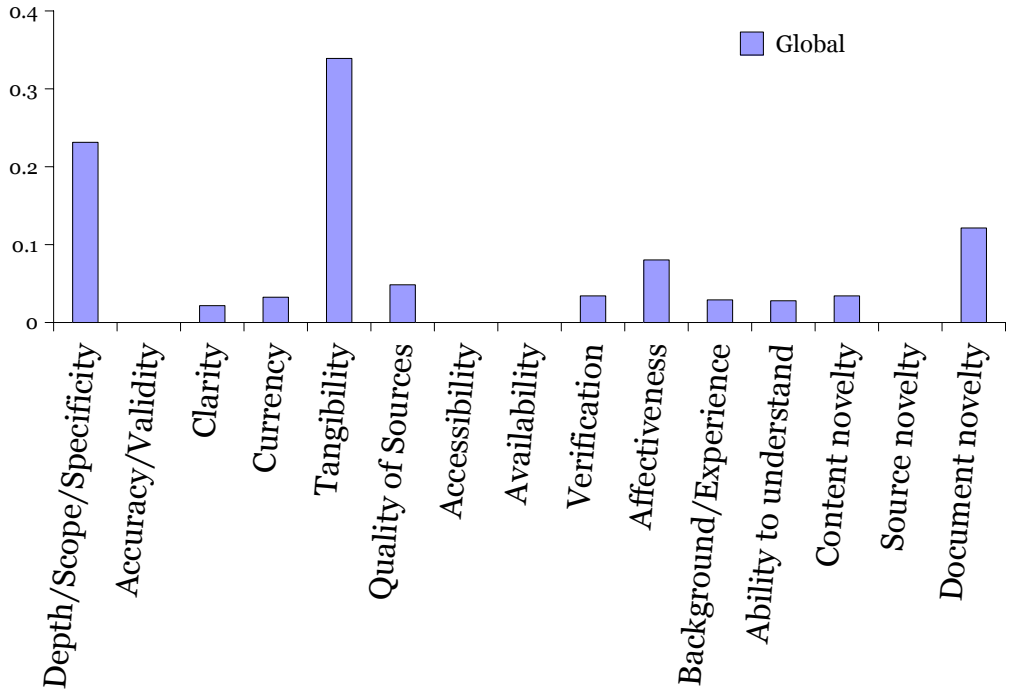

Figure 3.12: A typical relevance criteria profile. Frequencies are normalised, hence the y axis varies between 0 and 1.

there is what is defined as a "relevance criteria profile". A relevance criteria profile, simply put, is the grouping of the mentions of the relevance criteria during the search session. A typical relevance criteria profile, visualised as a chart, looks like Figure 3.12. These profiles provide a global view of the number of times, generally speaking, that each criterion has occurred during the search session for each participant.

Aggregating Profiles

Aggregating profiles, for instance if participants are grouped by their affiliation, does not require any special processing. Criterion counts are added together and the profile is normalised should any comparative analysis is to be performed.

Aggregating profiles is done by applying the following formula:

$$rc_i = \sum_j rc_{ij} \tag{3.1}$$

where rc_i is the count for criterion i in the new aggregated profile and rc_{ij} is the count

for criterion i for of the profile of participant j. The variable j is then restricted to the group for which the new aggregated profile is being calculated, e.g. $j = 1 \ldots 21$ such that p_j is a member of the School of Computing, where p_j is participant number j.

Normalising Profiles

Modeling the participants's preferences using relevance criteria profiles allows one to perform different types of analyses. Analysing a profile can be done with the profile as defined however comparative types of analyses need a normalising step before they can be performed. Two types of normalising can be performed and each allows a different type of analysis. On the one hand, normalising within a group (or individual session) is necessary when one wishes to investigate the relationships and relative weight of criteria within the group (or individual session). On the other hand, normalising within criteria is necessary when one wishes to investigate the relative weight across groups (or individual sessions).

To normalise within a group (or individual session) one applies the following formula:

$$rc'_i = \frac{rc_i}{\sum_{j=0}^{N} rc_j} \tag{3.2}$$

where rc'_i is the new, normalised, count for relevance criterion i, rc_i is the count for relevance criterion i in the relevance criteria profile of the group (or individual) and N is the total number of relevance criteria (in this study $N = 15$).

So that normalised profiles can be compared an extra normalisation step has to be applied. The result of this extra normalisation step is that criteria counts, in each profile, represent the proportional mentions across the profiles. To normalise across groups, and within each criterion, one applies the following formula:

$$rc'^{j}_i = \frac{rc^j_i}{\sum_{m=0}^{P} rc^m_i} \tag{3.3}$$

where rc'^{j}_i is the relative count of criterion i for profile j, rc^j_i is the actual count of criterion i in profile j and P is the number of profiles one wishes to compare.

Comparing Relevance Criteria Profiles

The Kullback-Leibler divergence (KL) is a natural measure function of a difference between a "true" probability distribution, p, and a target distribution q (Kullback & Leibler 1951). For discrete distributions, $p = \{p_1, \ldots, p_n\}$ and $q = \{q_1, \ldots, q_n\}$ the KL divergence measure is defined as:

$$D_{KL}(p||q) = \sum_{i=1}^{n} p_i \log_2 \frac{p_i}{q_i}$$

Although referred to as a metric, the KL divergence measure is not a true metric as it does not satisfy the triangle inequality. The KL divergence is also non-symmetric ($D_{KL}(p||q) \neq D_{KL}(q||p)$). The properties of the equation makes it non-negative and 0 if both distributions are equal ($p = q$). The smaller the divergence the more similar the two distributions are.

An alternative and symmetric measure of divergence is given by the λ divergence:

$$D_\lambda(p||q) = \lambda D_{KL}(p||m) + (1 - \lambda) D_{KL}(q||m)$$

where $m = \lambda p + (1 - \lambda)q$. A special case of the λ divergence is the Jensen-Shannon (JS) divergence (Lin 1991a). The JS divergence considers the KL divergence between p and q under the assumption that if they are similar to each other they should both be "close" to their average. Setting $\lambda = \frac{1}{2}$ results in the JS divergence:

$$D_{JS}(p||q) = \frac{1}{2} D_{KL}(p||m) + \frac{1}{2} D_{KL}(q||m) \tag{3.4}$$

where $m = \frac{1}{2}(p + q)$. As the JS divergence is based on the KL divergence, the smaller the divergence the more similar the two profiles are.

A discrete probability distribution $p(x)$ is a function that satisfies the following properties:

- The probability that x can take a specific value is $p(x)$, i.e.

  $P[X = x] = p(x) = p_x$

- $p(x)$ is non-negative for all x,
- The sum of $p(x)$ over all x equals to 1, i.e. $\sum_x p(x) = 1$

Normalised relevance criteria profiles satisfy all these properties so they can be interpreted as a discrete probability function. One can, hence, compare profiles using any of the above described divergence measures.

In particular the JS divergence was chosen, as it is symmetric[6], to compare the similarity between different profiles and to spot outliers in the data by comparing each individual profile with the global profile.

3.7.5 Visualising Sessions

Relevance criteria profiles provide a global view of the relevance criteria mentioned throughout a search session. This view however does not provide a view of the distribution of said criteria. A relevance criterion might not be evenly distributed; it could perhaps be that the distribution of its occurrences is skewed towards the beginning, or end, of the session. Another drawback of global profiles is that the sequence of occurrence of relevance criteria is lost. If during the session relevance criterion c_i was mentioned before c_j, this order is not considered as both occurrences contribute to their global count.

As a complement to global relevance profiles a technique for visualising search sessions was designed. Graphs resulting from applying this technique include information on the order of occurrence of the relevance criteria observed during a search session and the recorded interactions (if there were any).

Sequence is denoted by a time line. The time line only denotes an order in time and not any measure of it; equal spacing on the line does not mean equal time spans in the session. Relevance criteria ordering and grouping are represented as piles of coloured blocks. Each block represents the observation of a particular relevance criterion. Different criteria are assigned different colours. With relevance criteria piles relevance judgement processes are modelled. As long as relevance criteria are observed together one after the other with no other utterances of a different type in between, e.g. interactions, they are considered

[6]This is to say that $D_{JS}(p||q) = D_{JS}(q||p)$.

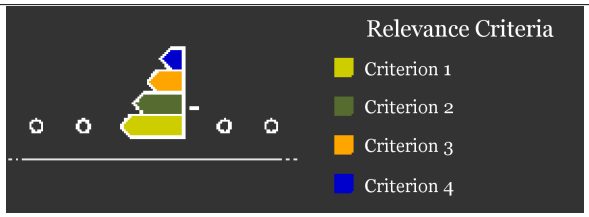

Figure 3.13: An example with four relevance criteria and interactions plotted.

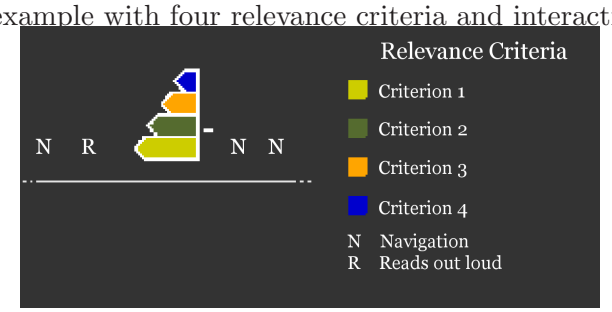

Figure 3.14: An example with four relevance criteria plotted. Interactions are further encoded and plotted accordingly.

to be part of the same relevance judgement process. Interactions are plotted in between relevance criteria piles.

Plotting Sessions

To plot a search session first the tagged utterances are grouped. For each group, the first relevance criterion in the sequence is plotted at the bottom of the pile, the second on top of it one unit to the right and so on. Blocks are made as long as need be so that the final shape of the pile resembles a staircase. The graph of the example sequence can be seen in Figure 3.13. In this graph there are two interactions on either side of the relevance pile which are plotted as circles.

There are assumptions behind the piles metaphor. First of all there is the assumption of aggregation. When a relevance criterion has been observed it is assumed that it will apply all the way until the user has made a final judgement. The length of each block in the graph symbolises this assumption. The application of criteria is done sequentially until the user is able to make a judgement about the relevance of the information. Each criterion contributes, either negatively or positively, to a final judgement. Negative contributions are represented as a minus sign next to the block in the graph (as seen in figures 3.13 and

3.14). One of the consequences, should this assumption hold true, is that the order in which criteria are used matters and that there might be a degree of relationship between relevance criteria. Users might follow a pattern when using relevance criteria. By using piles one can start analysing whether a user's relevance judgement process exhibits these dependencies between relevance criteria.

A second assumption is that relevance judgement processes can be isolated or delimited by the appearance of interactions. During the study it was observed that relevance judgements usually ended with the user navigating away from the document. This interaction can be preceeded by the explicit verbalisation of the relevance judgement, e.g. the user utters *"I don't like this document"*. A pile is then defined as occurrences of utterances that are not interactions. The shortcomings are obvious. First of all, depending on what the researcher considers to be an interaction, piles will (or will not) correspond to documents and their judgement processes as interactions are not necessarily all navigation interactions. Further encoding of interactions might alleviate this to a certain extent since the dynamics of the session might become more visible. For instance re-encoding the *interactions* plotted as circles in Figure 3.13 as *navigation* interactions and *read-out-loud* interactions results in Figure 3.14. Gathering click-through data and using it to better delimit the relevance judgement processes might also alleviate this situation.

Plotting sessions using this aforementioned technique allows a researcher to investigate the relative strength, or importance, of a relevance criterion within a relevance judgement process. In Figure 3.13 we see that one of the four criteria mentioned has a negative sign next to it ("Criterion 2"). This represents situations in which the user expressed a relevance criterion in a negative way, e.g. *"this is too old, it's from back in the 60's"*. In the example, Criterion 2 is negative yet the judgement process continues. This may suggest that the strength of Criterion 2, relative to the overall judgement process, is not as strong as to end it right there and then. The explanations can be varied, however the point is that researchers can direct their attention to further investigate these scenarios.

Choosing a Colour Sequence

According to Ware (1988) the effectiveness of using colours for coding is degraded as more categories are added. Ware recommends 12 colours which are normally used when labelling using colours. The first six colours, which also correspond to the basic colours in the colour opponent theory (Hurvich & Jameson 1957), are: white, black, red, green, yellow and blue. The remaining six colours are: pink, grey, brown, magenta, orange and purple.

Taking the colours as an ordered sequence of recommendations, it is suggested to use the number of occurrences of relevance criteria, in an aggregated profile, as indices to select an appropriate colour. The most occurring relevance criteria should then be assigned the first colour in the sequence, the second most occurring criterion the second colour in the sequence and so on. The rationale behind this procedure is that, since aggregated profiles are obtained by averaging across users, higher relevance criteria counts mean that users have mentioned the criterion, on average, more often hence the relevance criterion is likelier to be observed in any one search session. Choosing the most contrasting colours for the most commonly occurring relevance criteria should make easier the visual detection of the different criteria and the dynamics of the session.

Chapter 4

Results

In this chapter the data gathered during the study is described. Data dealing with the mentioned relevance criteria is presented and analysed in this chapter. We initially analyse the data using relevance criteria profiles (this technique is explained in Chapter 3, Sections 3.7.4). As discussed, relevance criteria profiles allow the analysis of the occurrence of relevance criteria at a global level. As such they provide a quick view of the occurrence of the relevance criteria on a per session basis while visualising one or more of these profiles as charts aids in uncovering the salient differences between individuals and groups.

Participants came from different schools and possessed different levels of research experience. Affiliation and research experience level lead to natural groupings of the participants. The following subsections describe the relevance criteria profiles of three groups, namely:

- **Global:** participants are not grouped. Statistics are calculated across all participants of the study regardless of their affiliation and research experience,
- **School:** participants are grouped by their affiliation. Statistics are calculated independently for each school (as listed in Table 4.1) regardless of all the other participants's particulars and
- **Research experience:** participants are grouped by their research experience level. Statistics are calculated for each research experience level (as described in Table 3.1)

independently of any other of the participants's particulars.

In Section 4.1 the user groups, their affiliations and research experience levels are described. The collections searched during the study are described next in Section 4.2. While a general overview of the data gathered is provided in Section 4.3, relevance profiles are discussed and their plottings are presented in Section 4.7. Section 4.8 concludes with a summary of the chapter.

4.1 Participants

A total number of 21 participants agreed to participate in the study. Out of these, 10 came from the School of Computing making this school the biggest school to take part of the study. Participants from the School of Computing were distributed as follows: 6 expressed that their research experience was that of a "research student", 2 that their research experience was that of a "researcher" and 2 that their research experience was that of a "senior researcher". The second largest school is the Information Management Group with 8 participants in total. Out of these 8 participants, 2 were research students, 4 were researchers and 2 were senior researchers. Only 3 people from the School of Pharmacy agreed to take part of the study, out of which 2 were research students and 1 was a researcher. No senior researchers from the School of Pharmacy accepted the invitation to take part of the study. The distribution of the participants according to their affiliation is displayed in Table 4.1 while the distribution of participants per research level (grouped by affiliation) in Figure 4.1.

School/Group	Participants
Computing	10
Information Management	8
Pharmacy	3
Total	21

Table 4.1: Number of participants per school/group for which valid data was gathered

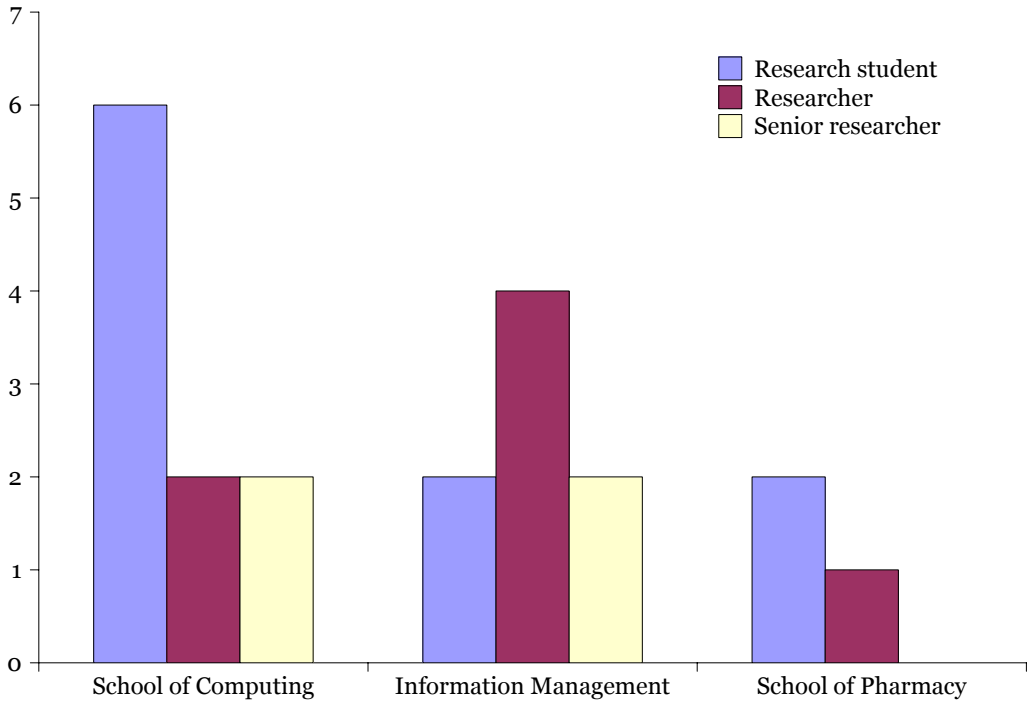

Figure 4.1: Distribution of participants per research level and school

4.2 Collections

4.2.1 School of Computing

The collection searched by participants coming from the School of Computing consisted of several volumes (up to volume 50) of the Communications of the Association for Computing Machinery (CACM)[1]. The collection contained 7028 articles covering several areas of Computer Science. Even though most recent articles in CACM are of a magazine type of an article, previous volumes contained scientific articles. Topics covered in this collection ranged from peer-to-peer (P2P) computing to software engineering theory and practice. The average document length is approximately 2676 terms with 85% of the documents containing between 0 and 5000 terms (after stopword removal). The longest document contains 34184 terms and the shortest only 79 terms. This collection was created by downloading all available documents from the CACM web site up to volume number 50.

[1] `http://www.acm.org/publications/cacm`

Query issued
Information Management
Content Management
Information
Information Retrieval
Information Systems
Knowledge Management
Profiles + content
Web

Table 4.2: Queries issued to the search engine for constructing the collection for the Information Management Group

4.2.2 Information Management Group

Participants from the Information Management Group searched a collection of several articles revolving around the topic *Information Management*. To create the collection, documents were searched for and retrieved from Library of Information Science Abstract[2](LISA). As the list of participants was known beforehand, queries that would reflect, as much as possible, each of the participants's research areas were crafted. To do so, each of the participants's *research interests* section from their home pages (whenever they were available) were visited and the most significant, but still generic, key words such as *Knowledge Management*, *Digital Libraries* were extracted. The full list of queries crafted can be seen in Table 4.2. The topics covered by the retrieved documents revolved around a main theme: information management. Amongst the topics covered, in particular, we find the web 2.0, law librarians and new trends in enterprise content management solutions.

A total of 1000 documents per query were manually downloaded. As participants came from a group with a common research theme there was a significant overlap in the queries constructed (e.g. the term *information* was present in several queries). This resulted in repeated documents when pooling all the document sets returned for each query. Duplicates were removed before the study. The collection contained a total of 4756 documents after de-duplication. The average document length is 3128 terms with 88% of the documents containing between 0 and 6256 terms.

[2]`http://www.csa.com/factsheets/lisa-set-c.php`

4.2.3 School of Pharmacy

The collection searched by participants from the School of Pharmacy contained documents coming from two different sources: the Public Library of Science (PLoS)[3] and The Pharmaceutical Journal Online (PJO)[4]. Articles published in PLoS, as well as in PJO, can be of two different natures: i) scientific articles or ii) magazine articles. Both types were included. The collection contained a total of 11426 documents. The topics covered by the documents were quite varied and ranged from tropical diseases (PLoS Neglected Tropical Diseases[5]) to articles on veterinary pharmacy editorials. The average document length is 5064 terms with 88% of the documents containing between 0 and 10128 terms.

To create this collection all available documents from both sources were downloaded and pooled together.

4.3 The Nature of the Data

A total of 1726 utterances were encoded as relevance criteria. An independent research encoded a total of 300 these (approximately 17%). We found that the overlap between our encoding and that of the researcher amounted to a total of 87%, i.e. 87% of the utterances had been assigned the same label by the two independent encodings (some utterances had more than one though).

4.4 Interaction

Interaction information was analysed mainly to see if participants had understood and knew how to use the system. Sessions would generally follow a pattern of interactions which could be laid out as follows. Participants, after being trained and having interacted briefly with the system, would initiate their searches by looking at the offered main topics as well as the description of their research topic. This was reflected in utterances like *"I'm going through the keywords first"* and *"the possible related topics for this is"*. Once

[3] `http://www.plos.org/`
[4] `http://www.pjonline.org/`
[5] `http://www.plosntds.org/home.action`

the participant had selected a topic to investigate further an initial quick search over the potential intermediate topics was usually done. Utterances like *"now I'm looking at the top B ones"* or utterances which denote that the participant was reading out loud the keywords of an intermediate B topic are examples of this. A click on an intermediate topic usually followed and the literature was retrieved. At this stage participants would be able to examine the actual literature connecting the two topics chosen. The session would progress with participants spending more or less time on a single intermediate topic, going back and forth between literatures. Unfortunately there were two exceptions where the participants were frustrated and abandoned the study prematurely. Interactions observed during the search sessions are analysed in more depth in Chapter 5, Section 5.3.

4.5 Intent

While participants usually verbalised their interactions with the system quite regularly, intentions were not expressed as often. Intentions generally referred to their search purposes such as "I need to find complementary". Mentions of other types of intentions such as the use of the information found (when found) were also observed ("I'm thinking that this will combine") however they were less frequent.

4.6 Interpretation of the Relevance Criteria

The encoding provided in Section 3.7.3 is a reinterpretation of the overlap of two other encodings as presented in (Barry 1993) and (Schamber 1991). Using this encoding for analysing the data is a sensible approach as

1. If it was found that the encoding applied to the data gathered during our study, more evidence for the generality of the encoding would be provided.

2. Evolving is expensive (time consuming), error prone, and generally difficult.

Since this type of investigation had never been done before in the context of LBD, it was decided that this was a reasonable approach.

In addition to interpreting most codes according to the interpretations and definitions provided by Barry & Schamber (1998), in this study a number of codes were used according to a personal reinterpretation. Moreover, as the Barry (1993) study is more in line with this study (when compared to that of Schamber (1991)), most interpretations are closer to those of Barry (1993) than to those of Schamber (1991). The interpretations of the encoding are offered next.

Depth/Scope/Specificity

In the definition provided by Barry & Schamber (1998) for *depth/scope/specificity* we see that utterances regarding to "whether the information is in depth or focused, has enough detail or is specific to the user's needs [...] it provides a summary or overview or a sufficient variety or volume." In this study, the code was interpreted as originally intended.

Accuracy/Validity

The code *accuracy/validity* is interpreted as to encode utterances referring to whether the information presented is accurate or valid. Even though the criterion refers to a personal judgement, information validity (or accuracy) does not depend on personal opinion nor personal agreement. A piece of information may be accurate, such as $2 + 2 = 4$, even if a person may disagree with it. This criterion refers to the act of the user judging the information to be valid (accurate). Originally, Barry referred to this criterion as *Objective Accuracy/Validity*.

Clarity

The differences in the definitions of *clarity* from the Barry (1993) study and the Schamber (1991) study are likely to be due to the different information objects studied in each study. In the cross-comparison study done by Barry & Schamber (1998), however, it is clarified that, in the broadest sense, what users are actually judging is whether the information is presented in a clear and understandable way. In this study, this is how the criterion was interpreted.

Currency

Barry's definition of *currency* (1993) agrees with that of Schamber (1991). According to this definition, *currency* refers to the extent to which users judge the information to be current, up to date, etc. In this study, expression of such nature were also coded as *currency*.

Tangibility

A code with which mentions of topicality (or aboutness) should be encoded is not included in the encoding used in this study. This stems from the assumptions in the Barry (1993) study and the Schamber (1991) study. The assumption behind both studies is that users judge relevance beyond topicality. In this respect, as the nature of both studies was to observe and list the relevance criteria observed, it seems that the participants of both studies did not mention topicality as a criterion explicitly. Regardless of the reasons for why the participants of both the Barry (1993) and the Schamber (1991) studies did not mention topicality, in this study participants did and so these mentions had to be labeled accordingly. Initially a code for this purpose could have been added. We believed, however, that a second interpretation of the code *tangibility* would be enough to accommodate mentions of the information being on topic (or about a topic). Consider the following excerpt from a search session:

> ‘‘...yeah, this is 2nd life, it’s about accessibility, it’s about different people’s abilities to use environments for Information Retrieval which is pretty bang on the topic I was looking for so, yeah ...’’

In this extract we can see three mentions of topicality. The first one is signaled by the utterance *“...this is [about] 2nd life...”*. The participant, in this case, is mentioning that he has recognised the overall theme of the document and what it talks about. The second and third mentions of topicality are *“...it’s about accessibility...”* and *“...it’s about different*

people's abilities to use environments for Information Retrieval...". The user refers, again, to the topics being discussed in the document. These mentions of topicality refer to the topics being discussed in the document. In this respect, it is interpreted that the document provides information about the topic. Utterances like *"it's about [topic]"* are interpreted as expressions of the document contents discussing the topic. It is in this sense that the document is providing information *about* the topic, so this type of utterances was coded as *tangibility.* The last utterance, *"...which is pretty bang on the topic I was looking for..."*, may seem as a mention of topicality however it actually is a mention of *specificity* and as such it is coded as *depth/scope/specificity.*

Quality of Sources

Quality of Sources, as interpreted in this study, refers to the different sources that participants could evaluate such as authors (or editors), affiliations or the publications in which the documents appeared. This interpretation is consistent with that of the criteria *Source Quality* and *Source Reputation/Visibility* in the Barry (1993) study. Utterances regarding the extent to which the quality of the information could be inferred either from personal experience with the source of the information or from the reputation (visibility) of the sources were encoded as *quality of sources.*

Accessibility

The differences between the definitions of *accessibility* provided by Barry (1993) and by Schamber (1991) seem to be a result of the information objects examined by the users in their studies. The interpretation of this criterion in this study refers to both the effort and cost involved in obtaining a document; these correspond to *Obtainability* and *Cost* as defined by Barry (1993). Mentions of the effort and/or the cost involved in obtaining the information, in this study, were coded as *accessibility.*

Availability

Barry (1993) defined *availability* on two levels: environmental and personal. Environmental availability refers to the extent to which the information presented is available in other documents within the environment. Personal availability refers to the extent to which the information presented was already possessed by the participant. In this study *availability* was interpreted as *environmental availability* only. Personal availability was interpreted to be part of a different code, namely *document novelty* as will be explained later on.

Verification

Despite its apparent similarity with *accuracy/validity*, *verification* refers to personal agreement with the information presented regardless of the validity (or accuracy) of the information. It may be the case that the information presented to the user is invalid (such as the statement $2 + 2 = 5$), however a person might still agree with it ("well, for large values of 2 that statement holds!"). Furthermore, the interpretation of the criterion refers to environmental agreement too; it refers to whether the information presented is agreed on by different sources of information. In the Barry (1993) study the code is actually *Subjective Accuracy/Validity*. Utterances referring to the extent to which the participants agreed with the information presented or to the extent to which the participants' point of view was supported by the information were coded with *verification*.

Affectiveness

Affectiveness, as defined by both Barry (1993) and Schamber (1991), refers to the extent to which the information provided participants with pleasure, enjoyment or entertainment. In this study, the interpretation of the code was extended to include expressions of raised (or diminished) interest.

Ability to Understand

The code *ability to understand*, according to Barry (1993), is used to code utterances that denote "the user's judgement that he/she will be able to understand or follow the

information presented". In this study the code has been interpreted accordingly.

Background Experience

Mentions of the use of background knowledge or information during the search session were encoded with *Background Experience*. Barry (1993) states that this code is used to denote "the degree of knowledge with which the user approaches information, as indicated by mentions of background or experience". In this study the code was interpreted as defined by Barry.

Content Novelty

Utterances expressing that the information contained within documents is known (or unknown) to the participant were coded as *content novelty*. Expressions indicating the extent to which the information is novel to the participant, and consistent with the definition and interpretation provided by Barry (1993), were encoded as *content novelty*.

Source Novelty

The code *source novelty* refers to the extent to which the source of the information (for instance the author) was novel to the user. In this study the interpretation was extended to also include mentions, for instance, of a known author writing in a different field or on an unexpected journal. It is not only interpreted to refer to the extent to which the sources are novel but also to the extent to which the relationship between the information and the sources is novel or unexpected.

Document Novelty

Utterances expressing that a document had been seen before (in the current session or not) were coded as *document novelty*. An expression classified as *document novelty* can express that, for instance, a document was not known by the participant prior to finding it. As such, this means that the document is novel but also that the document is not available in his personal collection. Personal availability was covered by Barry (1993) under the

code *availability* (explained earlier) however in this study it is covered as part of *document novelty*.

4.7 Relevance Criteria Profiles

Analysing relevance criteria information was performed at a global level using relevance criteria profiles. Relevance criteria profiles allowed the analysis of frequencies at an individual and group level. Comparative types of analyses are also possible on relevance criteria profiles, though a normalisation step is required beforehand. Two approaches were followed when plotting profiles: plotting profiles individually and plotting profiles together. Plotting profiles individually aids the interpretation together provides a quick overview of the differences, or similarities, between them.

In Section 4.7.1 an account for the most mentioned relevance criteria, according to a global relevance criteria profile, is provided. Profiles for the groups listed in Section 4 are also calculated and plotted together. In sections 4.7.2 and 4.7.3, the school and research experience profiles are presented and briefly analysed. An account of the observed relevance criteria, together with plottings of the profiles, is offered in Section 4.7.4. To produce these plottings one of two approaches, depending on the desired analysis, were followed: i) profiles were plotted together as is or ii) a second level of normalisation was performed before plotting the profiles together. Plotting profiles together as they are is done in the attempt to uncover salient (dis) similarities in terms of within-profile proportions. By plotting two, or more, profiles together the salient criteria, within each profile, can be immediately visualised. Re-normalising the profiles, before plotting, is needed to uncover a different type of pattern: the proportional mentions within criteria. By re-normalising and then plotting the profiles together we can observe how each criterion is distributed across profiles.

Because verbal reports are only a subset of the thought processes that occurred during the search session, the results presented next can only be interpreted as indicative and never as conclusive. In addition, as there were differences in verbalisations (volume and coverage) from user to user, the absence of mentions of a particular criterion does not

Criterion	Number of Occurrences	Percentage
Depth/Scope/Specificity	406	23.06%
Accuracy/Validity	0	0.0%
Clarity	38	2.15%
Currency	57	3.23%
Tangibility	595	33.80%
Quality of Sources	85	4.82%
Accessibility	0	0.0%
Availability	0	0.0%
Verification	60	3.40%
Affectiveness	141	8.01%
Background/Experience	51	2.89%
Ability to understand	49	2.78%
Content novelty	60	3.40%
Source novelty	0	0.0%
Document novelty	213	12.10%

Table 4.3: Number of occurrences for each criterion according to the global relevance criteria profile

mean that the participant never considered it during the relevance judgement process. Moreover, as the volume of the verbalisations varied from participant to participant, certain normalisations had to be performed in order to be able to compare the relevance criteria profiles.

4.7.1 Global Profile

The global relevance profile was obtained by applying Formula 3.1 and restricting j to all participants, i.e. $j = 1 \ldots 21$. In Figure 4.2 we can see that, overall, criteria dealing with the *tangibility* and with the *depth/scope/specificity* of the information were the two most common. *Document novelty* and *affectiveness* follow in third and fourth place respectively. These four criteria account for a 76.9% of the total number of observations (1355 occurrences). The list of all counts per criterion can be seen in Table 4.3 are also depicted in Figure 4.2.

4.7.2 School Profiles

The distribution of participants according to their affiliation can be seen in Table 4.1. The profiles of the three schools were obtained by applying formula 3.1 with the variable

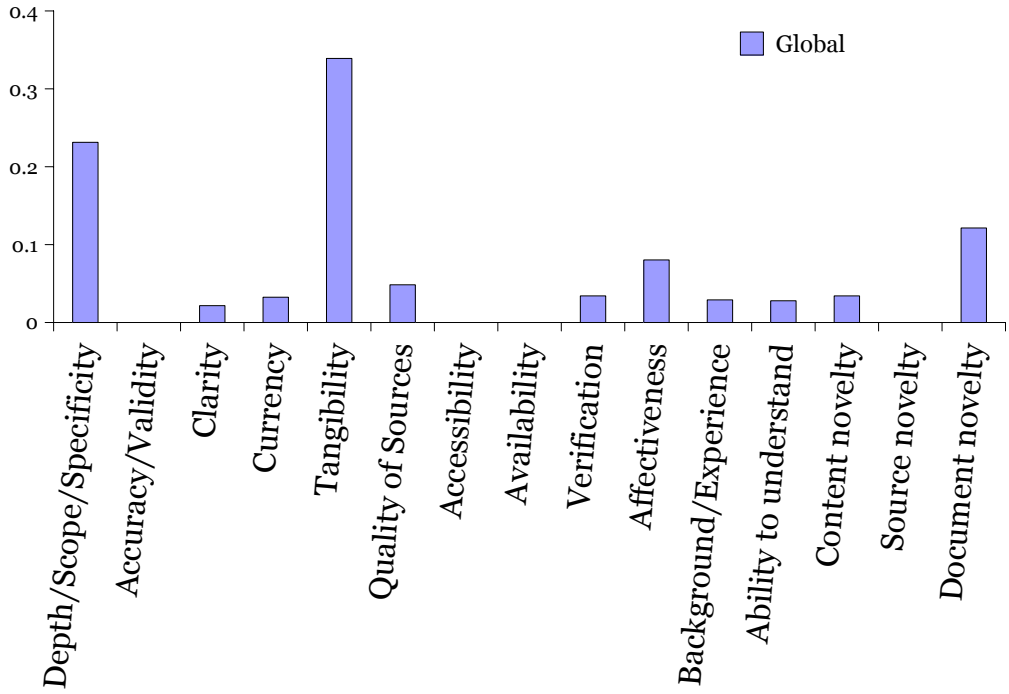

Figure 4.2: Relevance criteria profile of the *global* group. Values in the y axis vary between 0 and 1.

j restricted as listed in Table 4.7.2.

Group	Restriction
School of computing	$j = 1 \ldots 21$ such that p_j is a member of the School of Computing
Information Management Group	$j = 1 \ldots 21$ such that p_j is a member of the Information Management Group
School of Pharmacy	$j = 1 \ldots 21$ such that p_j is a member of the School of Pharmacy

Table 4.4: Restrictions for variable j according to each grouping of participants

Ten participants from the School of Computing took part of the study. This is the most represented school in the study. The most mentioned criteria, by members of the School of Computing, are *tangibility* and *depth/scope/specificity*. *Tangibility* was mentioned 356 times (about 40.6%) while *depth/scope/specificity* 127 times (about 14.4%). Their third most mentioned criterion is *document novelty* which has been mentioned 98 times (about 11%). This may suggest that members of the School of Computing prefer tangible data

over, for instance, voluminous information. Eight participants were affiliated to the Information Management Group making it the second most represented school in the study. Members of this group mentioned *depth/scope/specificity* 231 times (about 30.3%) and *tangibility* 215 times (about 28.2%). *Document novelty* was mentioned 103 times (about 11.7%). Unlike members of the School of Computing, who seem to have a preference for tangible data, members of the Information Management Group seem to be more interested in other properties of the information such as its volume and its specificity. This preference, however, is not as marked as that of the members of the School of Computing. Only 3 participants from the School of Pharmacy accepted the invitation and took part of the study. Members of this school also seem to exhibit the same preferences as members of the Information Management Group as they have mentioned *depth/scope/specificity* 48 times (about 38.7%) and *tangibility* 24 times (about 19.3%). The criterion *document novelty* was mentioned by members of this school 12 times (about 11%).

	Computing	Information Management	Pharmacy
Depth/Scope/Specificity	127	231	48
Accuracy/Validity	0	0	0
Clarity	23	13	2
Currency	19	33	5
Tangibility	356	215	24
Quality of Sources	51	32	2
Accessibility	0	0	0
Availability	0	0	0
Verification	14	40	6
Affectiveness	91	42	8
Background/Experience	26	18	7
Ability to understand	30	13	6
Content novelty	39	17	4
Source novelty	0	0	0
Document novelty	98	103	12
Total	876	760	124

Table 4.5: Number of occurrences for each criterion according to each school relevance criteria profile

As explained in Section 4.6, utterances encoded as *tangibility* may include mentions of *topicality*, so care must be taken when comparing the mentions of *tangibility* with those of any other criterion. This will be examined in more depth in the following subsections.

4.7.3 Research Experience Profiles

The distribution of participants per research level is depicted in Figure 4.1. The profiles for these three groups were obtained by applying formula 3.1 using the restrictions listed in Table 4.6. The distribution of the utterances according to each profile is depicted in Table 4.7

A total number of 10 participants expressed that their research experience level was on par with that of a research student. This is the largest group. The second largest group is the group of participants that were classified as researchers. This group consisted of 7 participants. The smallest group, with 4 participant, is that of the senior researchers.

Group	Restriction
Research Student	$j = 1 \ldots 21$ such that p_j has expressed that they are a research student
Researcher	$j = 1 \ldots 21$ such that p_j has expressed that they are a researcher
Senior Researcher	$j = 1 \ldots 21$ such that p_j has expressed that they are a senior researcher

Table 4.6: Restrictions for variable j according to each grouping of participants

Regardless of research experience level, the two most mentioned criteria were, in order, *tangibility* and *depth/scope/specificity*. Students mentioned *tangibility* 249 times (about 32%) and *depth/scope/specificity* 166 times (about 22%), researchers mentioned *tangibility* 218 times (about 39%) and *depth/scope/specificity* 150 times (about 26.8%) and senior researchers mentioned *tangibility* 128 times (about 29.7%) and *depth/scope/specificity* 90 times (about 20.5%). As with the school profiles, it must be noticed that there may be mentions of "aboutness" or "topicality" that have been encoded as *tangibility* and are driving the counts up. This phenomenon will be analysed further in the next subsections.

4.7.4 Relevance Criteria - the Observations

Plotting the profiles together may help visualise the (dis) similarities between uses of criteria per school (or research experience level) more clearly. Different individuals had varying degrees of verbosity which resulted in different numbers of utterances coded. To make the

	Student	Researcher	Sr.Researcher
Depth/Scope/Specificity	166	150	90
Accuracy/Validity	0	0	0
Clarity	14	15	9
Currency	10	5	42
Tangibility	249	218	128
Quality of Sources	48	18	19
Accessibility	0	0	0
Availability	0	0	0
Verification	18	20	22
Affectiveness	68	20	53
Background/Experience	25	14	12
Ability to understand	30	17	2
Content novelty	37	16	7
Source novelty	0	0	0
Document novelty	92	66	55
Total	757	562	441

Table 4.7: Number of occurrences for each criterion according to each research experience level relevance criteria profile

comparison of profiles possible, the criteria counts had to be converted to proportions by normalising the profiles. The normalisation step was done applying Formula 3.2. A graphical depiction of the normalised school profiles can be seen in Figure 4.3 while the normalised research experience level profiles are depicted in Figure 4.4.

In both figures we can observe that the two most mentioned criteria are *depth/scope/specificity* and *tangibility*. In the case of the school profiles we can see that the School of Computing mentioned *tangibility* more often than *depth/scope/specificity* while the other two schools mentioned *depth/scope/specificity* more often. This preference is not as clear in the case of the Information Management Group however the data in Table 4.5 confirms that indeed they mentioned *depth/scope/specificity* more times than they mentioned *tangibility*. Students, researchers and senior researchers all mentioned *tangibility* the most. Their second most mentioned criterion is *depth/scope/specificity*.

Plotting relevance criteria with a single level of normalisation is useful in observing top mentioned criteria across groups and doing basic comparisons, however, adding an extra normalising step, coupled with combined plotting, helps reveal even more patterns. Applying Formula 3.3 on the already normalised profiles results in normalised proportions

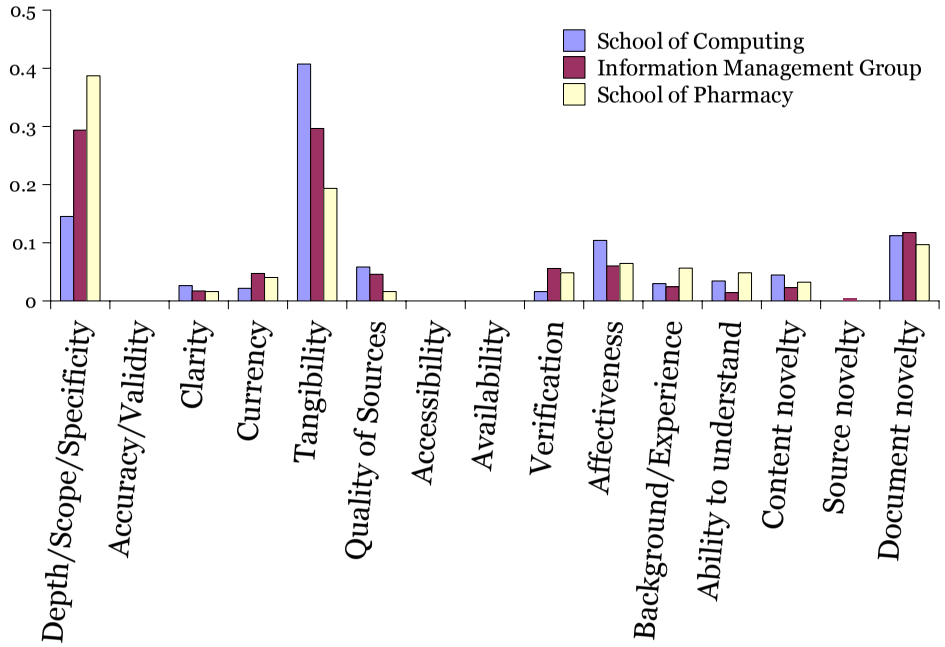

Figure 4.3: School profiles plotted together.

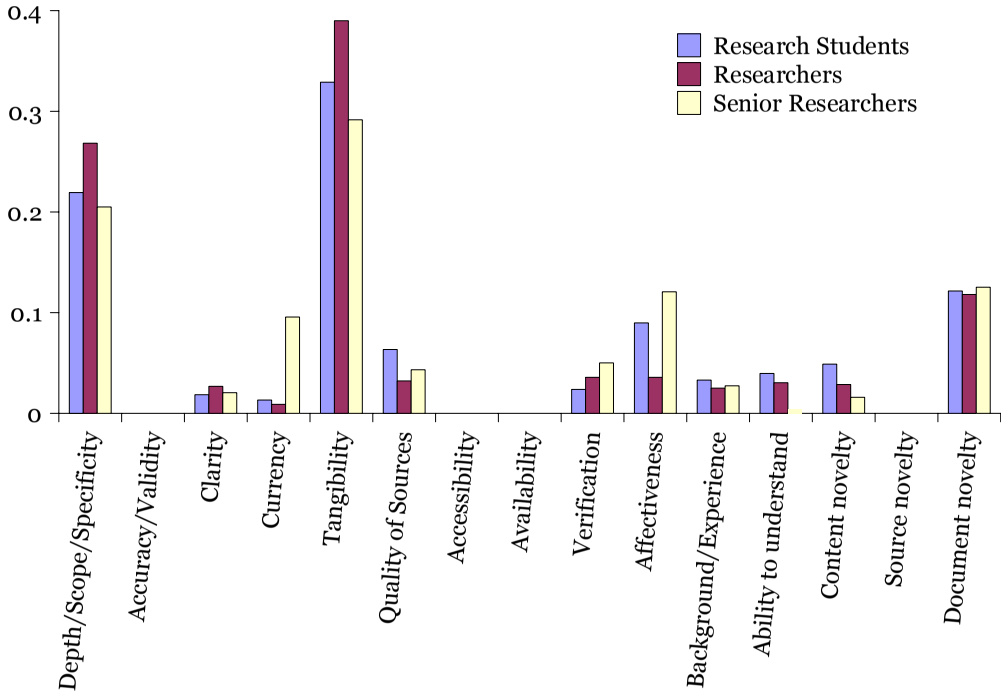

Figure 4.4: Research experience profiles plotted together.

per criteria, i.e. for each criterion the proportions are normalised resulting in a distribution over groups of the proportional mentions of the criterion. As the resulting proportions of

each criterion now sum up to 1 (100%), we can observe, when plotting these profiles, how much different schools, in respect to the others, have used each criterion.

Figures 4.5 and 4.6 depict the re-normalised school profiles and the re-normalised research experience level profiles respectively. These two figures will be used as guides in the next subsection when discussing each criterion in more depth.

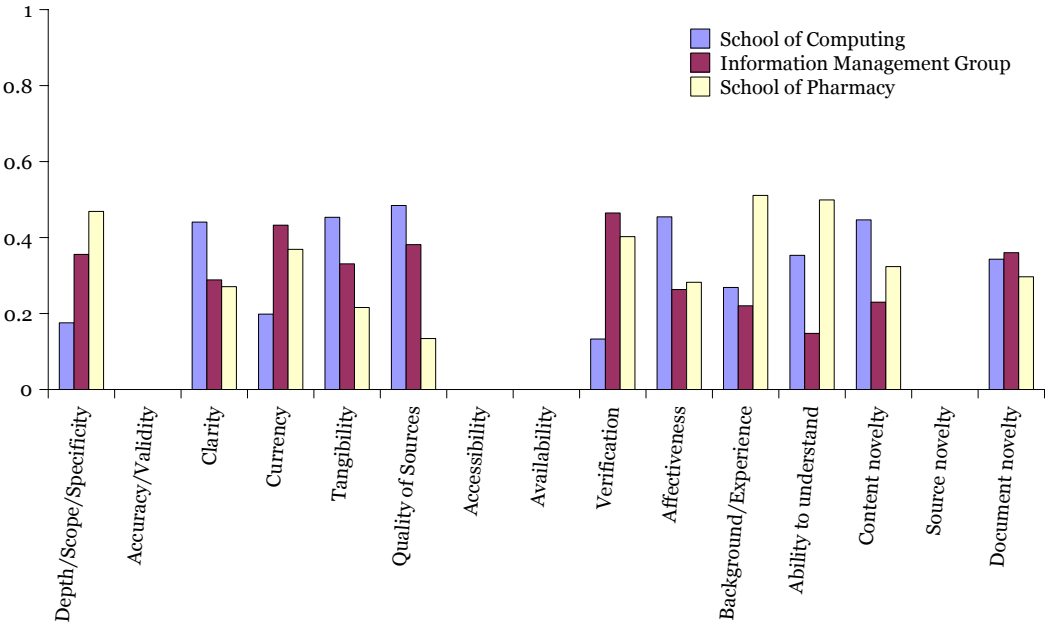

Figure 4.5: The profiles of the schools, normalised within criteria, plotted together.

Depth/Scope/Specificity

At a global level, *depth/scope/specificity* was observed to occur a total of 406 times (23.07%). This criterion deals not only with scope, but also with specificity, volume, detail and even genre of the information contained in the document. Reasonably so, participants were interested in these properties of the information obtained. As utterances that express that the document refers to a topic specific to the user's needs were also coded as *depth/scope/specificity* a number of references to topicality may be included in the counts of this code. Examples of the utterances coded as *depth/scope/specificity* include:

- "general summary"
- "detailed enough"

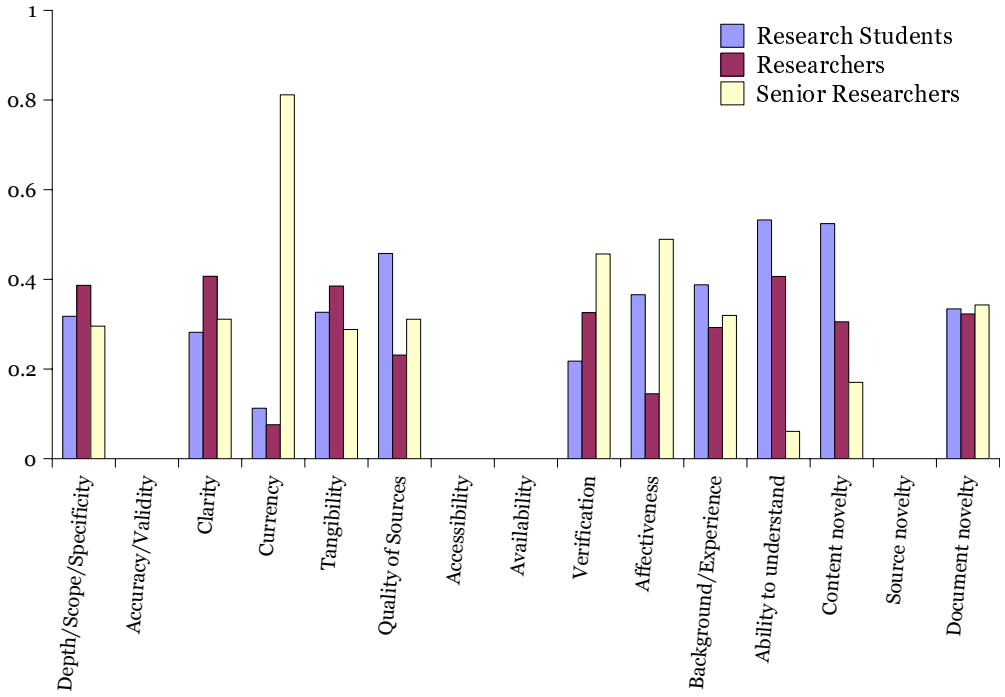

Figure 4.6: Research experience profiles, normalised within criteria, plotted together.

- "lots of information"

Out of these 406 occurrences, 61 (15%) were references to *exemplary documents*. According to Blair & Kimbrough (2002), *"exemplary documents are those documents that describe or exhibit the intellectual structure of a particular field of interest"*. Vocabulary varies significantly across research fields in science. One function these exemplary documents perform is to provide a definition of the words included in these vocabularies. This is reflected in mentions such as *"...lot's of jargon I don't recognise..."*[6] and *"...[provides] definitions of acronyms..."*[7]. By mapping the structure of the field, exemplary documents also provide a context in which the vocabulary is to be interpreted. An example of such documents in the scientific community is the *survey article*. Survey articles summarise up to a point in time the most important advances and issues to be treated in a field, include a list of references to follow up and possibly a list of important academics and institutions. Participants of this study referred to this type of exemplary documents in ways such as *"...an overview of the key papers..."* and *"...an overview of data collection techniques..."*.

[6]This utterance could also be classified as *clarity* or *ability to understand*.
[7]This utterance could also be encoded as *tangibility*

There is some evidence that suggests that an exemplary document of this type may, for instance, ease the entrance of a newcomer to the world of research in that field. It would ease this entrance by not only providing an overview of the field itself but also of the pertinent vocabulary and major players in it. It would be reasonable to observe users preferring a survey article to the latest article on a specific topic when getting acquainted with the field being investigated. This must be kept in mind as exemplary documents can be of most use when their topicality has been assessed to be outwith the participant's own field of work. Effectively, this has been mentioned by a participant who negatively expressed that a document on his own research was *"high level"*. As the participant continued the session, a positive expression of *"high level"* was observed again, however this time the participant was referring to a document on a topic, different but related to the participant's own field of research. Preferences for exemplary documents included, but were not limited to:

- "general summary"
- "gentle introduction"
- "good overview"

A possible answer for why this type of documents are preferred by users when entering a new field may be because these documents may have a high ratio of information obtained vs. processing effort (both concepts introduced in Harter's theory of psychological relevance (Harter 1992)). Perhaps by providing this roadmap to the field, together with the associated jargon, survey articles offer plenty information in exchange for little mental processing effort. This would afford users a quick judgement to whether or not it would be beneficial to go deeper into the field and search for possible connections.

For two schools, namely the Information Management Group and the School of Pharmacy, this criterion was their most mentioned; each school mentioned it 231 and 48 times respectively. Members of the School of Computing mentioned *depth/scope/specificity* 127 times. In Figure 4.5 we observe that, proportionally speaking, the school that mentioned *depth/scope/specificity* the most is the School of Pharmacy, followed by the Information

Management Group and the School of Computing. It seems reasonable that, while members of the School of Computing have used this criterion to a certain extent, they have verbalised other criteria, such as *tangibility* for instance, more often.

The distribution of the mentions of *depth/scope/specificity* across research profiles is as follows: students mentioned it 166, researchers 150 and senior researchers 90 times. This makes this criterion the second most mentioned criterion by any one research experience profile. In figures 4.4 and 4.6 we can also observe that it was researchers who, proportionally speaking, mentioned *depth/scope/specificity* the most (when compared to students and senior researchers).

Accuracy/Validity

Judging the *accuracy/validity* of the information presented when entering new fields of research may be very hard to do, if not impossible. It was not expected to be observed very often, and in fact it was not observed at all. A potential explanation is that users, as they had entered new territories, took for granted the *accuracy/validity* of the information as the information came from documents that had been published in different, and sometimes well known, publications. In a sense, it may be that there was an implicit use of *quality of sources*.

Clarity

The code *clarity* refers to the extent to which the information was presented in a clear and understandable way and it has been observed 38 times (about 2.1%). This is a criterion that might have an effect on the mentions of *ability to understand*, as it may happen that because the information is not presented in a clear fashion, the user expresses his (or her) inability to understand it and vice-versa. Because the collections consisted of published documents which have gone through a reviewing process, it seems reasonable to have observed this criterion less than most the other criteria as the peer-reviewing process is supposed to guarantee, amongst others, a certain level of clarity. However, the expressions of the criterion alone (its counts) only indicate its presence. It may have happened that

the information was in general very clear so that participants mentioned *clarity* only when the information had been presented in an outstandingly clear and understandable way (or the total opposite) and that generally the clarity of the information is silently ignored. Mentions of *clarity* included:

- "the title is quite explicit"
- "[topic] is a clear one"
- "[it is] hard to read"

Currency

Mentions coded as *currency* accounted for a 3.2% of all relevance criteria mentions (57 utterances coded). *Currency* refers to the extent to which the information was judged to be current or up to date. It is not entirely clear what role of this criterion plays in this context as both "old" as well as "new" information could be potentially very relevant, i.e. regardless of the date published, related information would remain being related. However, users might prefer more current information as the chances of making "new" discoveries may increase by incorporating current information. Users of this study mentioned *currency*, for instance, in the following ways:

- "outdated...yeah, 1985"
- "ancient but relevant"
- "it's up to date"

Senior researchers seem to be the most interested in current information. As depicted in Figure 4.6, most mentions of this criterion came from senior researchers. This phenomenon could have been influenced by the search task given to senior researchers: while students had to complement their literature review and researchers had to write a proposal, senior researchers had to gather information for a keynote speech. Perhaps, the activities undertaken by both researchers and students allowed them more room when it

came to the currency of the information. Despite the preference usually given to current information when conducting a literature review, students still have to demonstrate an understanding of the research field they are involved in. A similar line of reasoning applies to researchers in the process of writing a research proposal and while special attention is usually paid to the state of the art and very current information, potentially outdated information serves as a complement in this task. Preparing a keynote speech, where the organisers have asked you to *"focus your speech on the future directions and implications of advances in your research field, especially on those fields outside your own"*, however, might be more restrictive regarding the currency of the information used for this. Perhaps it was this that made senior researchers concentrate on very current information and hence the higher proportion observed in terms of mentions. This suggests that bias might have been inadvertedly introduced while crafting the simulated work tasks. More specifically, by making a difference, in an attempt to make the tasks more realistic in regard to the research experience levels, in the underlying tasks that participants had to perform. Each task has an implicit time[8] constraint included which was not detected beforehand.

In terms of the school profiles, members of the School of Computing were the least interested in the currency of the information; most mentions (about 81%) came from member of the Information Management Group and the School of Pharmacy.

Tangibility

Recalling the explanation of the code *tangibility* from section 3.7.3 it is not surprising that this was the most common criterion used by the participants (it appeared 595 times representing a 33.8% of all the criteria mentioned). Tangibility refers to the document's contents, the actual explicit information contained within. However, some of the utterances coded with *tangibility* actually refer to the topics discussed. Out of the 595 utterances encoded as *tangibility*, 257 (43.2%) correspond to mentions of topicality. Recoding these utterances as *topicality* would result in *tangibility* becoming the second most mentioned criterion followed by *topicality* with 338 and 257 counts respectively (*depth/scope/specificity* would become the most mentioned criterion with 406 mentions).

[8]Time not in the sense of having a restricted amount of time to perform a task but in the sense of date.

Participants of this study mentioned the topicality of the information presented less often than, for instance, the volume of it. There are at least two potential explanations for this phenomenon. On the one hand, it may be that topicality is less important than, for instance, mentions that are to be encoded as *depth/scope/specificity*. When presented with a short document, for instance, a participant could express that *"it's too short"* and disregard it. In this situation, regardless of the topics being discussed, the information would have been deemed irrelevant by the user simply by the expression of it not being enough[9]. On the other hand, it could be that topicality is a requisite *sine-qua-non* relevance cannot be judged. When presented with information, it could be that a user firstly assesses whether it is on topic or not and expresses this like *"it's about [topic]"*. Once the information has been deemed on topic, the judgement can proceed and the user applies other criteria. Any of these two potential explanations could be the cause for the higher counts of both *tangibility* and *depth/scope/specificity* compared to *topicality*.

Examples of the utterances coded with *tangibility*:

- "[the document is] talking about"
- "it does illustrate"
- "there's some interesting facts in this one"

Members of the School of Computing mentioned *tangibility* more often than any of the other criteria. Out of the 356 mentions of *tangibility*, 135 (37% of all mentions encoded as *tangibility*) are actually mentions of *topicality*. Differentiating between mentions of actual *tangibility*, according to the Barry & Schamber (1998) interpretation, and *topicality* does not affect the ranking in terms of mentions of criteria: members of the School of Computing still mentioned *tangibility* more often than any other criterion. The revised ranking of criteria would have, if mentions of topicality were to be encoded as *topicality*, *tangibility* as the most mentioned criterion with 221 mentions followed by *topicality* with 135 mentions. In third place we would find *depth/scope/specificity* with 127 mentions.

[9]We must remember that this analysis is solely based on what users verbalised. Silent rejections or use of criteria, as they cannot be detected, cannot be taken into account when analysing and proposing potential hypotheses.

This preference for hard data, as exhibited by members of the School of Computing, is not entirely surprising given the nature of the discipline. As the code *tangibility* refers to the actual contents of documents, this suggest that even when users referred to the contents of the document, topicality was not the only factor affecting the judgements. This can be observed in the individual transcriptions as participants from this school referred to this hard data in ways such as *"..that's an application..."* and *"...there's some interesting facts in this one..."*. One must be careful, however, when interpreting this observation as not all mentions of *tangibility* were positive. Indeed, at least one member of the School of Computing referred to *tangibility* negatively as *"...too many formulae and stuff..."*.

As observed before, in the case of the profile of the Information Management Group and the profile of the School of Pharmacy, mentions of *depth/scope/specificity* outnumbered those of *tangibility*. Re-coding utterances encoded as *tangibility* that refer to "aboutness" or "topicality" makes the difference even larger. In the case of the mentions from members of the Information Management Group, out of the 215 utterances encoded as *tangibility*, 115 (53.5%) were actual mentions of topicality and, in the case of the mentions from members of the School of Pharmacy, out of 24 utterances encoded as *tangibility*, 7 were actual mentions of *topicality*. Taking this into account means that the members of the Information Management Group and members of the School of Pharmacy have a marked preference for *depth/scope/specificity* to the point where mentions of topicality outnumber those of *tangibility*. Despite that interest was shown by members of these schools in tangible data (expressed, for instance, as "...they achieved a 50% repose rate...") most did not regard hard data as an affecting factor in terms of relevance of the information presented.

All three research profiles mentioned *tangibility* the most, however, if we take into account that *topicality* accounts for a 43% of the mentions coded as *tangibility* this situation changes. As observed before for the school profiles, re-encoding mentions of *topicality* with their own code results in a new ranking where *depth/scope/specificity* is the most mentioned criterion regardless of research experience level. The second most mentioned criterion, however, in this new ranking depends on research experience. Students and re-

searchers mentioned *tangibility* mostly while senior researchers mentioned *topicality*. Potentially, this could be influenced by the task each group had to try to complete. Students were asked to complement their literature review. The nature of literature reviews is to provide an overview of the landscape of a field of practise and as such they are required to include not only the subtopics that might be found within the field but also more tangible information such as previous results and techniques. Perhaps this is what motivated students to mention *tangibility* more often than *topicality*. A similar explanation applies to researchers who were asked to write a funding proposal. In the case of senior researchers, however, the situation changes. Senior researchers mentioned *topicality* more often than *tangibility* and this might also may have been influenced by the task they had to complete. Senior researchers were requested to write a keynote speech. It may be that preparing a keynote speech actually does not require tangible data but only information that is on topic as keynote speeches are usually about the future of an area. When suggesting what the future may bring, only related topics and potential interactions are described but not in great detail.

Quality of Sources

Participants of the study resorted to the reputation of the authors, their affiliation and/or the reputation of the publications as an indicator of the quality of the information. Mentions of this usage were encoded as *quality of sources* and were observed 85 times (4.4%). In the context of entering new, and possibly unknown, fields of research, resorting to this criterion seems like a sensible approach. However, evaluating the credentials of the sources of the information may not be an easy, and even feasible, task. Perhaps it was this that made the participants refer mostly to generic qualities such as position and not to more specific factors such as names and familiarity with the authors's work. Some examples of these expressions are:

- "I see the name of"
- "never heard of him"

- "he's guest editor"

While most mentions referred to generic attributes like position in an organisation, there were more specific mentions of names and expressions of personal relationships with the authors of the documents such as *"...she's a friend of mine..."* and *"...Carol's review again..."*.

Mentions of *quality of sources* came mostly from students and senior researchers. It may be the case that students, as they are beginning their career as researchers, are more impressionable by positions and affiliation and have mentioned these often. This could potentially explain the high proportion of mentions coming from students. In addition senior researchers, as they are established in their field, are familiar with who are the major players in the field and their work, and potentially share a personal relationship with them, may have mentioned the *quality of sources* in a less generic way and referred to people and names instead of positions and affiliations. While members of the School of Computing and members of the Information Management Group mentioned the *quality of sources* of the information, members of the School of Pharmacy almost did not express their interest in this criterion. Potentially this could be due to an inadequacy of the collection searched by members of this school. It could be that because all, or most of, the articles and even the journals are not well known by members of the school, they silently disregarded the reputation of the authors and the publications.

Accessibility

Accessibility was not observed at all. This is likely to be due to the settings in which the study took place. Obtaining any document from the collection did not involve any sort of fees nor effort (except the effort of clicking on the provided hyperlink). This could potentially explain why there were no mentions of *accessibility*, however, it may have still happened that participants had referred to the potential cost of obtaining documents cited within the documents being inspected in ways such as *"...getting that might be expensive..."*.

Availability

Availability was also not observed at all. As the collections were all crawled beforehand, all the documents were available when requested. This, however, does not mean that there could have been mentions of *availability*. For instance, a participant could have expressed interest in reading a document cited in the document he was examining and then mention its unavailability, however this did not happen.

Verification

Utterances coded as *verification* accounted for 3.4% (60 occurrences) of all the coded utterances. The code *verification* was used to tag utterances from participants expressing that the information presented was, for instance, supported by other information within the field. Mentions of personal agreement with the information, or the support of a user's point of view, were also coded as *verification*. Participants were placed in the spot as newcomers by the very search task they had to solve so confirming, or rejecting, that a piece of information was supported by other information within the field seems unlikely (environmental agreement). Effectively, most mentions of *verification* referred to personal verification (personal agreement). Examples of utterances coded as *verification* include:

- "[reads out loud — history matters] yes it does!"
- "goes back to my [topic] thoughts"
- "I'm thinking about [topic X] but this one is looking all the way through [topic Y]"

Verification was mostly mentioned by senior researchers, followed by researchers and students. While senior researchers are quite interested in obtaining information that is verifiable, research students seem to be more relaxed and accepting. *Verification*, as interpreted, refers to the extent to which the information supports the user's point of view or is agreed on by the user. It is a subjective criterion in the sense that it does not depend on the accuracy or validity of the information (as reflected by the code *accuracy/validity*). It may be that the the level of agreement is related to the research experience level as

students do not mention this criterion as much as researchers and senior researchers do. Perhaps the more experienced, in terms of research, a person is, the more views and beliefs he (or she) has. Having more views, or more established views, would mean that the (dis) agreements with the information presented, in terms of these views, are likely to happen more often. This observation does not contradict what can be observed in Figure 4.5: the school that least mentioned *verification* is the School of Computing and 60% the members of this school were classified as students.

Affectiveness

Affectiveness was observed 141 times (7.5%). Utterances coded with the tag *affectiveness* included expressions of surprise, rejection and disregard amongst others. Even though *affectiveness* was analysed in isolation[10] it is possible that affectivenes (to the information) has an effect on the user's search experience. A person constantly expressing negative affectiveness towards the information retrieved by the system may develop a level of animosity to the point where future judgements are negative regardless of the appropriateness of the information presented. Under such circumstances, the user may even choose to end the search session prematurely. Some examples of the utterances coded as *affectiveness* are:

- "I like [topic]"
- "a turn off in terms of my research"
- "got all excited there!"

Members of the School of Computing seem to be, according to their expressions, relatively more affective than members of any of the other schools (Figure 4.5). Moreover, the least affective group, according to Figure 4.6, is that of researchers.

[10] Mentions of *affectiveness* were treated in isolation in the sense that the potential accumulative aspect was not considered. Utterances were observed, coded and counted for each session and independently of each other. To fully investigate whether *affectiveness* accumulates throughout a search session one could perhaps look at the distribution of the mentions of *affectiveness* over time during the search session and then see if the relevance judgements are affected by it.

Ability to Understand

Ability to understand was mentioned 49 times (about 2.7%). The code is used to encode utterances that refer to the extent that users express that they will be able to understand or follow the information presented. As such, this criterion may be related to *background experience*. It seems reasonable to relate them as it may be that the availability of (or lack of) *background knowledge* could result in users expressing their (in) ability to understand the information presented. Hence, it is be reasonable to observe that most mentions are produced by the less experienced people in the group, e.g. students. This might suggest that experience does affect its occurrence but also that there might be a relationship between the two codes. Participants mentioned their (in) ability to understand in ways such as:

- "not sure how these things connect"
- "I see what happened here"

Background Experience

Background experience mentions were observed 51 times (about 2.9%). These mentions referred to the extent to which the background knowledge of experienced was used during the session to judge the relevance of the information presented. Initially it could be suggested that this code is similar to *content novelty*, however, they differ as *background experience* refers to uses of background experience while judging the information presented and *content novelty* refers to whether the information had been encountered before. This criterion was observed in ways such as:

- "I am familiar with [topic]"
- "I have researched before"
- "[I have] done enough of that [in the past]"

In Figure 4.5 a correspondence between the mentions of *background experience* and *ability to understand* can be observed. The distributions of the mentions, at first sight,

seem similar to each other. Members of the School of Pharmacy mentioned both criteria more often than members of the other two schools. This could be a coincidence however it could also be that both criteria are related (as suggested earlier). For instance, a lack of background knowledge could result in an inability to understand the information presented. Exactly the opposite interaction could also explain the occurrences: due to the presence of background knowledge, an expression of an ability to understand the information presented could be observed. This pattern in usage can only be confirmed (or rejected) by analysing the individual search sessions in more depth.

Content Novelty

Verification and content novelty were both observed an equal number of times; utterances coded as *content novelty* accounted for 3.4% (60 occurrences) of all coded utterances. Examples of expressions of *content novelty* being used as a relevance judgement criterion include:

- "same topics coming up as before"
- "didn't know [topic] had an impact on that"
- "nothing that is really new to me"

Content novelty refers to the extent to which the information contents are novel to the user. As it can be seen in Figure 4.6, expressions of content novelty came mainly from research students. It may be that the research experience level exhibited by participants, as well as their background knowledge, have an effect on how novel they will find the information presented.

Content novelty was mostly mentioned by members of the School of Computing. This phenomenon could also be attributed to the choice of collections. It must be noticed, however, that the sign of the mentions is not depicted in the figures. Members of the School of Computing could either be familiar with the contents of the collection or the contents could be totally new. Should most content in the collection be new to the members of this school, one should expect a high proportion of positive mentions of *content novelty*

and vice-versa. Relevance criteria counts alone only indicate the presence (or absence) of a criterion being used to judge the relevance of the information presented and not whether their contribution to the relevance judgement is either positive or negative.

Background experience, *ability to understand* and *content novelty* are three criteria that, according to the figure, have been all mentioned primarily by students. Even though whether the mentions are either positive or negative is not displayed on the graph, it is sensible to expect students to mention these criteria more often than researchers and senior researchers. Since students are supposed to have a limited experience in both their domain as well as other domains, it is expected that mentions of *background experience* are negative in the sense that students judge the information irrelevant as they lack the background knowledge. Higher mentions of *ability to understand* and *content novelty* by students could be explained by students being less experienced than researchers and senior researchers (the same argument applies to researchers). Being students less experienced could result in higher, and negative, mentions of *ability to understand* (students express their inability to understand) and also in higher, though positive, mentions of *content novelty* (being less experienced could also mean less familiar with research topics outside their own).

Source Novelty

Source novelty refers to whether the source of the information is novel to the user. This criterion has not been observed during this study. There are at least two potential explanations for this. On the one hand, as users entered new field of research it may have been that they did not know any of the sources of the information (the authors, the journals, etc.) and they did not express this. When everything is new, perhaps, one perhaps cannot say this all the time or even at all. On the other hand, it may be that they knew all the sources of the information, but that is very unlikely. There is also the situation that some utterances, for instance *"I don't know who he is"*, were coded as *quality of sources* only when they could have also been encoded as *source novelty*. This is due to the subjective nature of the encoding and analysis and, while a researcher could have encoded these ut-

terances as *source novelty*, another research could have encoded these utterances as *quality of sources*.

Document Novelty

Utterances regarding the novelty of the documents were observed 213 times (10.1%). According to almost all authors, e.g. (Weeber et al. 2001, Gordon et al. 2002, Pratt & Yetisgen-Yildiz 2003) novelty is an important factor to consider. Systems are actually tailored to prefer novel information, according to the definition of novelty embedded in the system, and rank them higher. Its importance is deemed so to the point where non-novel connections, documents, examples or information are considered to be of little importance. As Gordon et al. (2002), amongst others, mention, novelty is highly subjective. Novelty is also dependent on the context in which is observed. At least three scopes of novelty can be suggested: *community*, *personal* and *task*. Novelty at the community scope means that documents are novel to a whole community of practice. Novel, in this context, means that the document has not been discussed nor cited in any of the documents produced by the community. This type of novelty may be observed, for instance, when an member of a community of practice is made aware, by personal communication for example, of a document that might benefit his work and, in turn, the community itself. Validating the novelty in this scope, however, might not be feasible in practice, and can only be partial, as it would involve examining all publications produced by the community and their references.

When documents are already known to a community of practice, they might still happen to be novel to an individual. In situations like this, the novelty is purely personal. While doing a literature search for his (or her) Ph.D., a research student might come across a document he had not come across before. From his point of view, this is a novel document. It could well be, however, that the document in question is widely known and hence the novelty is not considered to be in the scope of the community.

Novelty in the scope of a task refers to documents which are novel in the context of the task being solved but not necessarily in either the personal scope nor the scope of a

community. Re-finding a document in a different search task is and example of novelty in the *task scope*. This could result in the searcher gaining new insight, due to the influence of the context, on the information contained in the document. Novelty in this scope is heavily related to personal novelty but is not necessarily tied to novelty in the scope of a community.

It is very hard to determine the novelty scope from the mentions of *document novelty* observed in this study. However, the polarity of such mentions can be determined as the criterion was often mentioned in either a positive or negative way. Intuitively, within this context of knowledge discovery, one would expect a correlation between positive mentions of novelty and a positive influence towards relevance judgements however this is not entirely true. Some participants actually interpreted being presented with the same document over and over ("negative" novelty) as a sign of relevance while others took that as a sign of poor system performance. Negative mentions of *document novelty* followed a pattern like *"I've seen this before, therefore I'm not interested in it"*. Examples of such negative mentions are:

- "I've seen already"
- "it is that damn article again!"
- "our old friend"

Initially this is the most expected behaviour in these settings: already known information (or documents) has little to offer. However, a reversed pattern was also observed. Some participants followed a pattern like *"[because] I've seen this before, I will take it"*. This is observable in expressions such as:

- "always getting that article so it must be relevant"
- "again here we have [title]"
- "it was this [title] again"

This reoccurrence of the same documents was interpreted, by a group of users, as a reinforcing positive sign, regarding the relevance of the documents. Documents being re-retrieved for different intermediate topics were judged to be relevant as they kept *"cropping up everywhere"*. Perhaps it was the different context, as provided by different intermediate topics, what made the re-occurrence of known documents to be judged as a positive sign and not a negative sign. If this was the case, it could then be derived that the novelty happened at the task level as explained before.

4.8 Summary

We began this chapter with an account of the collections used during the study, the characteristics of the user groups and our interpretation of the encoding used to label the transcribed verbal protocols. This interpretation, although subjective to a certain extent, was aligned to the original interpretation as described in (Barry & Schamber 1998). When the segmented transcriptions were encoded, we observed a total of 1726 relevance criteria mentions. A random sample of 300 utterances previously encoded as relevance criteria was re-encoded by an independent researcher and the overlap between assignments was found to be 87%, i.e. out of the 300 utterances, 261 shared at least one label with our original encoding. This suggests that, although there were differences, the interpretation and the act of encoding is stable.

The types of relevance criteria observed were analysed at three different levels:

- Global: no breakdown by either affiliation nor research experience. All relevance criteria occurrences were analysed in together.
- Affiliation: relevance criteria were grouped by affiliation.
- Research experience: relevance criteria were grouped by research experience.

Initially, profiles were described both quantitative and qualitatively and the top two relevance criteria were reported. These were consistent across groupings: *tangibility* and *depth/scope/specificity* and the two most mentioned criteria.

Relevance criteria observations were normalised at the criterion level. This normalisation step allows the observation of how each criterion had been mentioned across groupings. These new slices of the data were visualised to aid their analysis. For each criterion, the occurrences at the three levels were analysed and examples of these occurrences were provided.

While discussing *depth/scope/specificity*, we observed that a 15% of its occurrences referred to *exemplary documents* (Blair & Kimbrough 2002). We argued that this was a reasonable observation since these type of documents could well be used by newcomers as an entry point to a research field of potential interest. Additionally, it was suggested that this preference might indeed relate to the theory of psychological relevance (Harter 1992) as the information to processing effort ratio of exemplary documents would be expected to be high. It must be noted, however, that this type of documents would only be of use once their topicality had been assessed.

An unexpected introduction of potential bias was detected while discussing the criterion *currency*. It was noted that as the research experience level increased, the mentions of this criterion would increase. It was suggested that this might be an undesired effect of how the simulated work tasks were crafted. Because all efforts strived to make these more realistic, they were tied to the research experience level by creating different narratives and tasks for each of the levels. It is possible that each task has an implicit association to the concept of time, in the sense of dates, that made participants mention the criterion *currency* in different proportions. This suggests that the differences in mentions might be an artefact of the bias potentially introduced as opposed to being a real phenomenon.

Mentions encoded as *tangibility* were further dissected. This was motivated by the realisation that several of these were mentions of *topicality* while encoding the utterances. *tangibility* was the most mentioned criterion, however relabelling utterances as *topicality* meant that *depth/scope/specificity* would become the most mentioned criterion. Two explanations as to why this could be happening were offered. On one hand, it might be that there are situations where relevance can be judged immediately and independently of topicality, e.g. if a document is too short, it does not matter whether it is on topic or

not; if it's too short, it's too short. On the other hand, it might be that topicality is a pre-requisite so that relevance can be judged and that most participants made an implicit assessment of topicality even before starting to verbalise other criteria.

quality of sources mentions were done mostly by research students and senior researchers. It was further observed that students had referred to generic attributes such as an author's position in an institution while senior researchers referred to the quality of sources in more specific ways such as author's names and the reputation of some publication houses. When it comes to verifying the information presented (utterances encoded as *verification*) we observed that the more experienced as a researcher a participant was, the more mentions of this criterion would occur. It was posited that as researchers progress in their careers and gain experience, the likely it is that (dis)agreement will arise with certain works. This is due to the strengthening (weakening) of current and past beliefs.

A correspondence between mentions of *ability to understand* and *background knowledge* was suggested as the distributions of these two criteria were indeed very similar. It seems reasonable to suggest that if a person lacks the relevant background knowledge, their ability to understand a piece of information would be diminished.

Lastly, the criterion *document novelty* was analysed. The different scopes in which novelty could occur were discussed and it was pointed out that deriving this context from the utterances would be a very difficult, if not impossible, task. Additionally, the polarity of the mentions was briefly discussed and examples of these and how *negative* mentions of *document novelty* sometimes resulted in a positive judgement were given.

Overall, both a quantitative and qualitative description of the observed relevance criteria were given and attempts to provide explanations for each of the observed frequencies were made. Behaviour was inspected to see if any of its types, as represented by the frequencies of mentions of a particular criterion, would relate to either the research experience levels or to the affiliations or both. It was discovered, however, that this might not be straightforward. The explanations as to why criteria were mentioned in the observed frequencies varied, however none clearly related the frequencies to either type of grouping. Additionally, while analysing the frequencies of the criterion *currency*, it was discussed

how the way the simulated work tasks were crafted might have introduced bias.

Chapter 5

Results II

In this chapter relevance criteria profiles are revisited in Section 5.1. Here these profiles, and their similarities, are analysed to see if there are any relationship between participants' affiliation and their mentions of relevance criteria. As profiles can be interpreted as a discrete probability distribution, we can compare them using divergence measures. A comparative analysis of their divergences suggests that there might be three naturally emerging clusters. This analysis is aided by the use of heatmaps and the Jensen-Shannon divergence (Lin 1991a).

Isolated mentions of relevance criteria give partial insight into the cognitive processes that were present during the search session. However, it may be wished to analyse how these mentions interact together and how these groups of mentions are used to judge the presented information. In Section 5.2 a technique for isolating and analysing these groups of relevance criteria mentions is presented. In this technique the encoded transcriptions are segmented with the objective of isolating the so-called relevance judgement processes. These processes are defined as uses of relevance criteria as delimited by user interactions. Relevance judgement processes are interesting in two respects:

1. Complexity: one can quantify the complexity of a relevance judgement process as the number of relevance criteria included in the process and one can use this as a proxy for the actual complexity of evaluating a particular piece of information.

2. Selection rules: one can relate the number of criteria in each relevance judgement

process to the six rules of selection presented by Wang & White (1999). These rules can then be interpreted in the context of LBD and related to user intentions.

Additionally to our analysis of the relevance criteria, judgement processes, and selection rules, the interactions with the system in which the participants incurred are described and analysed in Section 5.3. The description and analysis is broken down into three subsections, each which correspond to one of the three main activities in a closed model search in LBD:

1. The target topic selection process.
2. The intermediate topic selection process.
3. The literature inspection and selection process.

After analysing both relevance criteria profiles and user interactions in isolation, both of these ideas are analysed in conjunction. Relevance criteria profiles, complemented with user interactions, provide a fuller picture than either alone. Carrying out such analysis, however, can be difficult and time-consuming. Having a bird's eye view of the search session which included both these notions –the relevance criteria and the user interactions– would help detect, for instance, the segments of the session which might be of particular interest to us. Unfortunately, there are currently no available plotting techniques that combine these two entities into a single picture[1]. In Section 5.4 a custom visual representation of a search session is included. This visualisation includes the relevance judgement processes, the user interactions and the search session evolution. This visual representation provides a holistic view of the search session as it progressed and aids in the analysis of the complexity and interactions between the relevance criteria observed and the user interactions with the system.

Section 5.5 summarises and concludes the chapter.

[1] To the best of our knowledge.

5.1 Profile Similarities

Individual and aggregated relevance criteria profiles provide a global view of the most commonly mentioned criteria for that particular session or group of sessions. Aggregated profiles, however provide a view for arbitrary groups of profiles, e.g. profiles grouped by the affiliation of the participants. An alternative view may be provided by grouping profiles by measuring their similarities to each other. Since relevance criteria profiles can be interpreted as a discrete probability distribution (Section 3.7.4) they can be compared using standard divergence measures such as Kullback-Leibler or Jensen-Shannon (JS) (Lin 1991a).

The JS-divergence was chosen to measure the similarity between relevance criteria profiles. The profiles were first normalised and then the JS-divergence value was calculated for every possible pair of profiles. This is depicted as heat maps in Figure 5.1.

In each heat map, the value in cell (i, j) corresponds to the JS-divergence value between the profiles of participants i and j. The matrices are symmetric as the JS-divergence is a symmetric measure, i.e. the value in cell (i, j) is equal to that in cell (j, i). Rows and columns are ordered by date in which the participant took part of the study. This leads to the participants being ordered by school. Index values from 1 to 10 represent the School of Computing, from 11 to 18 the Information Management Group and from 19 to 21 the School of Pharmacy.

In the figure we find four heat maps. The matrices in each map are all equal and the only difference between maps is the number of colours used as palette for the JS-divergence values. In all maps, the redder the colour of the cells the less divergent the two profiles are. In Figure 5.1 (a) only two colours have been used. In this map we can observe that the profile in row/column 6 has a high divergence with almost all the other profiles in the map. The divergence between the profile and most others (with the exception of two) is above 0.4[2]. This suggests that the participant represented by the profile in row 6 is an outlier. In Figure 5.1 (b) three colours have been used in the palette and we begin to better

[2]JS-divergence values closer to 0 mean that the two distributions compared are less divergent and vice-versa.

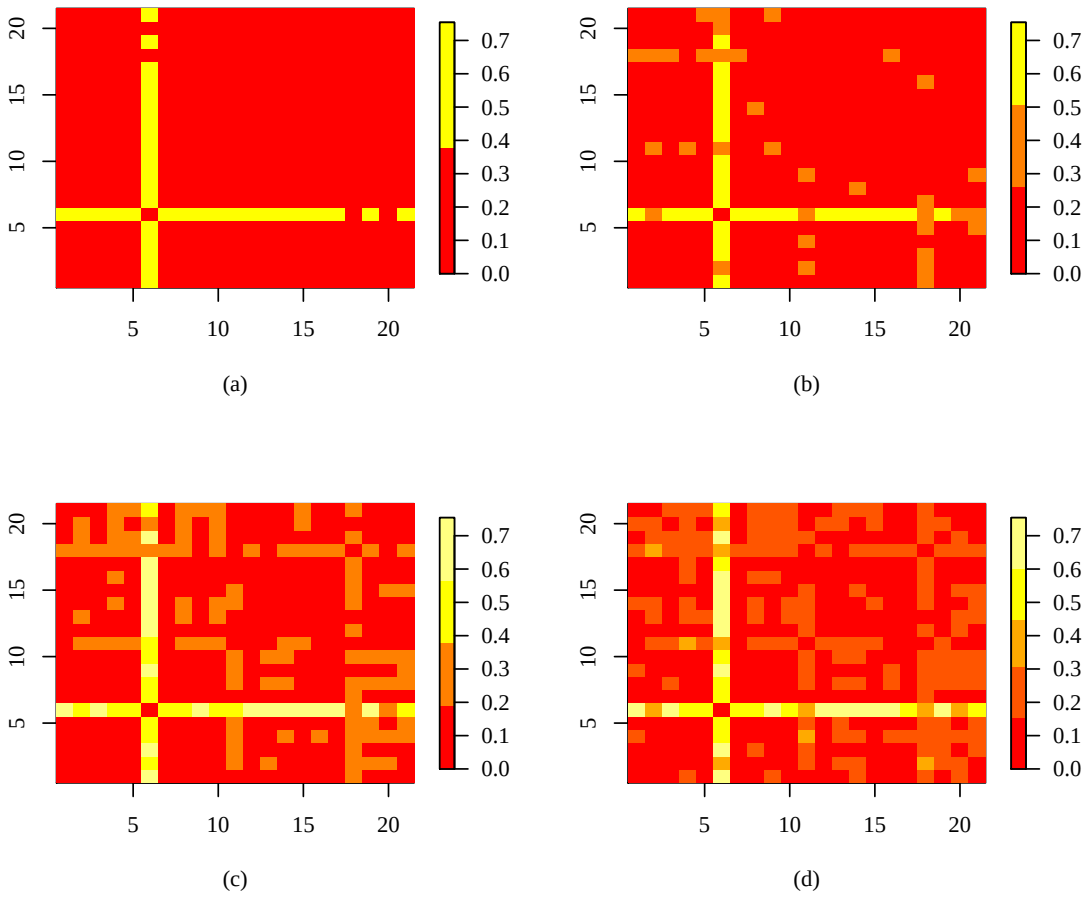

Figure 5.1: JS Divergence scores between participants. Each cell represents a divergence score between two participants (rows/columns represent participants.) It must be noticed that the closer to 0 the score, the more similar the two profiles are. This is in line with traditional heatmap plots where red is used for higher activity.

observe the divergences between profiles. Profiles in rows 11 and 18 divergence mostly with profiles of members of the School of Computing (rows/columns 1-10). In the third heat map, (c) in the figure, four colours have been used in the palette. The divergences become more noticeable and we can observe that the profile in row 18 is actually divergent with most other profiles (with the exception of some profiles of members of the Information Management Group). We can also observe that the profiles of the participants from the School of Computing exhibit a level of convergence and this shows as a red block on the bottom left of the map. Moreover, their profiles do not diverge with most of the profiles

of members of the Information Management Group (with the exception of profiles in rows 11 and 18 which seem to diverge with almost all profiles). In the last heat map, Figure 5.1 (d), five colours were used. In this figure the divergences become even clearer. The profile in row 18 diverges with practically every other profile but with two. One of these two profiles is that in row 11 which also seems to diverge with most other profiles. In the figure we can also observe that the profiles of the participants of the School of Computing remain convergent and that they diverge more with the profiles of the members of the School of Pharmacy than with those of the Information Management Group. The profile in row 17 seems to be very similar to almost every other profile with the exception of two: profiles in rows 18 and 4. There seems to be a group of profiles that are convergent with almost every other profile. These profiles are those in rows 1,2,3,7 (members of the School of Computing) and 12 and 17 (members of the Information Management Group). That these profiles are convergent with most other profiles could be due to that the participants represented by these profiles follow a globally shared behaviour in using relevance criteria to judge the relevance of the information presented, however before confirming/rejecting this suggestion, a closer inspection to the search sessions should be conducted.

Analysing the JS-divergence reveals three emerging clusters, however these do not correspond to any of the groups analysed in the previous chapter; neither cluster wholly corresponds to either schools groups nor research experience groups. The three clusters found through the analysis of the heat maps are: i) that of the potential outliers –profiles in rows 6, 11 and 18– ii) that of the potential representatives of the whole group –profiles in rows 1, 2, 3, 7, 12 and 17– and iii) the rest. The profile in row 6 is divergent with practically every other profile. This suggests that the participant, in terms of his/her associated relevance criteria profile, may be an outlier. The participant's profile is only close to two other profiles and they may also be considered outliers. Effectively, the profiles in rows 11 and 18 do not seem to be convergent with almost any other profile. An outlier is a sample that is numerically distant from the rest of the data. As such, outliers may be an indication of measurement errors. It has been suggested that the profiles in rows 6, 11 and 18 may be outliers, therefore checking whether there have been errors in measuring

during the search sessions of the participants represented by these profiles, can only be done, by closely inspecting these and analysing their evolution. However, outliers can also indicate the areas in which a theory might not be valid. It may be the case that the participants in question behave in a way that does not correspond with that of the rest of the group. This should also be checked by analysing their search sessions in more detail.

Some initial suggestions can be proposed from this simple divergence analysis. Firstly, visualising the JS-divergence in this fashion helped detect that some search sessions may either be anomalous or at least different enough, in terms of the relevance criteria mentions, so that they merit a closer inspection. Secondly, the profiles of the members of the School of Computing seem to exhibit a certain level of convergence, however, this level of convergence exhibited could be due to the way the profiles were ordered in the matrix. For instance, swapping places between participants 11 and 12 would have revealed a bigger red-ish block of convergent profiles however, participant 12 is affiliated with the Information Management Group. Before suggesting that members of any one predefined group behave in a certain way, or use relevance criteria following a shared pattern, a deeper inspection of the search sessions should be conducted. Thirdly, there seems to be a clustering of profiles that are convergent with almost every other profile. This group of profiles could be signalling that the participants, whom these profiles represent, may be using a group shared pattern in terms of usage of relevance criteria when judging information.

Analysis of the JS-divergence between relevance criteria profiles is only useful to indicate which participants (or groups of) may be behaving in a particular way. As such, this analysis may only be useful to detect these individuals and so further inspect their search session use of relevance criteria and interactions with the system.

5.2 Relevance Judgement Processes

In the study conducted by Wang & White (1999), participants were asked to select, from the results of searches conducted by librarians, which documentation they would use for their projects. One of the observations resulting from the analysis of the selection process is that users applied a set of decision rules when selecting this documentation. The selection

process, as described, consisted of six rules:

1. Single criterion decision: if the user detects a single salient unwanted aspect in the information, it is immediately discarded. This rule represents the principle of least effort.

2. Multiple criteria decision: if users cannot reach a judgement after applying the single criterion rule, they apply several criteria until a judgement is reached.

3. Dominance rule: users select documents such that they excel in at least one criterion and are no worse in any of the other criteria, e.g. two documents which provide the same information, however one of them is more current than the other.

4. Scarcity rule: when information is scarce, users tend to be more lax regarding the criteria used to judge the information.

5. Abundance rule: when users have found enough information, they tend to stop accepting more information even if it would be deemed relevant under different circumstances.

6. Chain rule: when users have detected that they are on a chain, or vein, of information they tend to make a collective information on the set, e.g. because the previous document, deemed relevant, is on this chain, a new document on the same chain is likely to be considered relevant.

In this study, it was observed that the participants applied a subset of these rules in varying proportions. This suggests that the rules found in the study conducted by Wang & White (1999) are also applicable to the context of LBD making them more general. To estimate the frequency with which these rules were used, the following procedure was applied. Firstly, search sessions were segmented to obtain the relevance judgement processes as described in Section 3.7.5. Secondly, the length of each of these sequences was measured; the number of utterances encoded as a relevance criterion within each sequence is counted. Lastly, once the length of each sequence is measured, sequences of length n

were counted per participant, i.e. for each session how many sessions or length 1, 2,...n there are.

The complexity of a relevance judgement process is defined as the number of criteria used in it. This definition stems from the assumption that the more criteria is mentioned within any one process, even if one criterion is mentioned many times, the more complex the process is. It was observed that participants expressed applying relevance criteria in sequence when evaluating the presented literature, and that the more criteria that was applied the more time-consuming and difficult the judgement was. Hesitations, together with mentions of criteria and backtracking, were an indication of this type of behaviour. Although this definition rests on a limited number of observations, Wang & White (1999) found that *"participants often apply a salient criterion to reject a document. Participants tend not to scan all aspects of a document in decision-making"* suggesting that the complexity of a decision may be related to the number of relevance criteria used to reach said decision.

The complexity of a relevance judgement process makes it possible to classify a process into one of the two selection rules:

- Processes of complexity 1: these include the mention of a single relevance criterion and hence map to the *elimination* rule of (Wang & White 1999). In this case this rule is referred to as the single-criterion rule
- Processes of complexity > 1: these include mentions of multiple relevance criteria and hence map directly to the multiple criteria rule of (Wang & White 1999)

A total of 589 relevance judgement processes (of any complexity) were counted. Out of these, 215 (36.5%) are of complexity 1 and 374 (63.5%) of complexity 2 and larger. The bars in Figure 5.2 denote the total number of participants (y axis) that used a sequence of n criteria (x axis) to assess the relevance of the information presented at least once. In the figure we can see that all participants applied, at least once, a sequence of one criterion to judge the information presented. This corresponds to the rule of single criterion decision described by Wang & White (1999). We can also see that most participants (at least 14 participants) used up to 6 criteria in any one relevance judgement processes. More complex

relevance judgement processes are used by fewer participants. Processes of complexity larger than 7 were used by, at the most, 7 participants.

In Figure 5.3 we see the average use (y axis) of relevance judgement processes of complexity n (x axis). In the figure we see that sequences of complexity 1 (single criterion rule) were used, on average, about 10.6 times per session. The more complex the relevance judgement process becomes, the less it is used. As it can be seen in Table 5.1 (and also in Figure 5.3), the average number of uses decreases as the complexity of the process increases (with the few exceptions) and is always lower than the average use of relevance judgement processes of complexity 1. The multiple criteria selection rule corresponds to using relevance judgement processes of complexity 2 or above. After aggregating these processes of complexity 2 and above, on average, they were used 3.5 times per session. This suggests that participants wanted to quickly judge the information and keep the information flow dynamic. Quickly assessing the information presented, possibly with aims of a quick dismissal, means that they could spend more time assessing more information and obtain a broader coverage of the information space. This behaviour may have been encouraged by the settings in which the study took place: search sessions had a time limit of 1 hour.

The single criterion decision rule, as described by Wang & White (1999), suggests that this rule is mostly applied to quickly dismiss information based on salient unwanted features. During this study, two types of use of the single criterion rule were observed:

- Filter out: in concordance with the original description of the rule, participants detected salient unwanted features and quickly dismissed the information.
- Eager acceptance: contrary to the original mention of the rule, participants detected a salient feature that made them consider the presented information relevant automatically.

The frequency with which these two uses were observed can be estimated as follows. Firstly, for each use of the single criterion rule the criterion used in it was counted. The polarity of the expression was also taken into account. A positive mention of *currency*, for

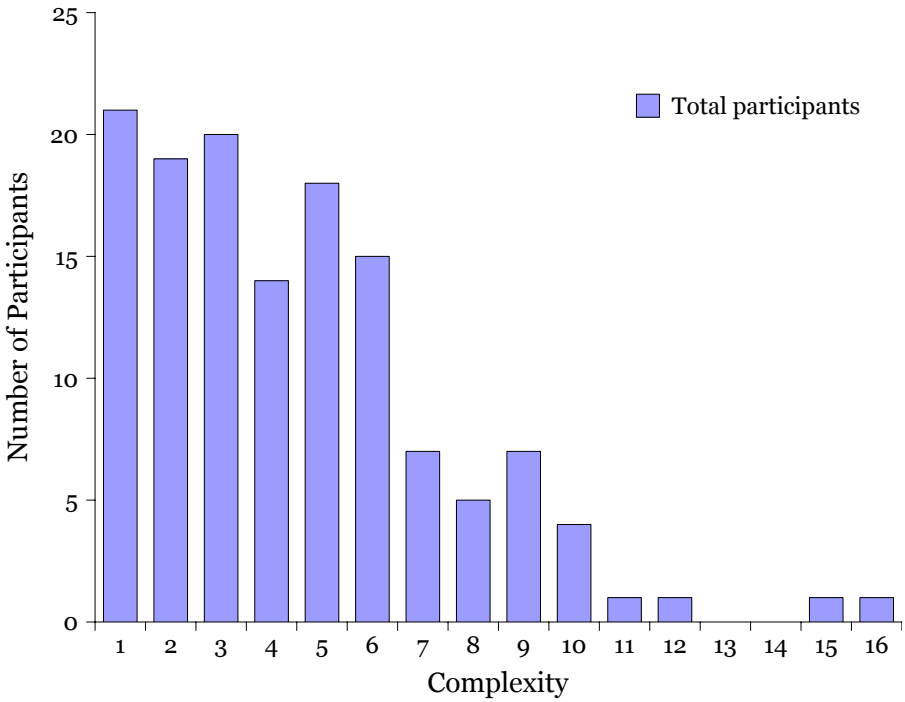

Figure 5.2: Total number of participants that used, at least once, a relevance judgement process of complexity n

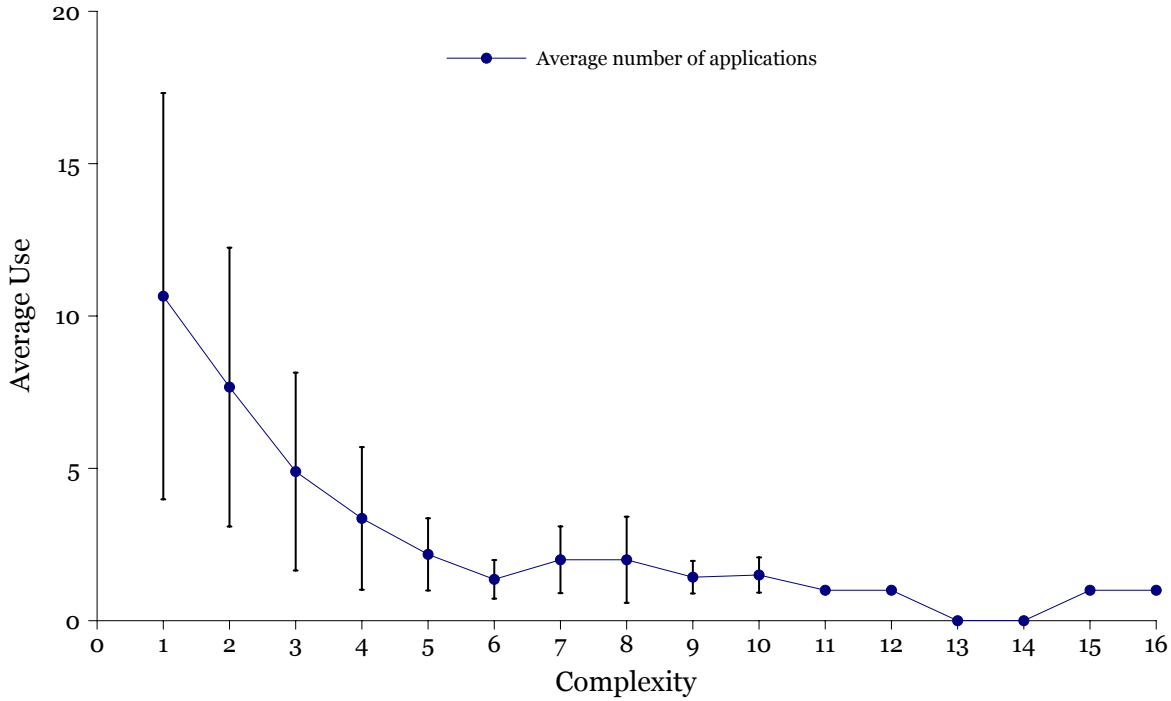

Figure 5.3: Average use of relevance judgement processes of length n. Bars represent a standard deviation.

Complexity	Average use (times)	Deviation
1	10.65	6.6
2	7.6	4.5
3	4.9	3.2
4	3.35	2.3
5	2.17	1.1
6	1.35	0.6
7	2	1
8	2	1.4
9	1.42	0.5
10	1.5	0.5
11	1	-
12	1	-
13	-	-
14	-	-
15	1	-
16	1	-

Table 5.1: Average use (averaged across participants that expressed using them) of relevance judgement processes of complexity n

instance, was considered different from a negative expression of the same criterion. This expresses our assumption that negative mentions of criteria correspond to uses of the rule to filter out irrelevant information and that positive mentions correspond to uses of the rule to eagerly accept the relevance of the presented information. Finally, the counts were then converted to proportions, i.e. the vector of counts was normalised.

The assessment of the polarity of any one utterance was done by analysing the type of words used in the utterance itself. In the few cases where the language itself was not enough to determine the polarity, the tone of the voice of the participant and the preceding utterances were taken into account. Consider the following example. A participant mentions that *"...[the document] is too old..."*. This utterance is classified as *currency* and its polarity deemed negative. The negative polarity is inferred from the use of *"too old"* in the utterance. This expression suggests that the participant deemed the information to not fulfil a specific criterion: that the information is current or up to date. *Currency* is used as a criterion, but in a negative fashion and will probably influence so that the final relevance judgement of the information presented is negative. Consider now the polarity of utterances such as *"...it's from 2006..."*. The language used in the utterance indicates

that the person is referring to the date of publishing and the potential *currency* of the information, however it does not offer any indications regarding the polarity of the expression. In cases like this, one can resort to the audio recordings to assess the tone of the person's voice and also consider the influence of the preceding utterances. Consider these two potential scenarios in which the polarity of the utterance is to be inferred. Both begin in the same fashion: as soon as the user is presented with the document, the relevance judgement process begins and the use of criteria mentions start. Suppose that the first mention of a criterion is negative and that it is to be encoded with *depth/scope/specificity*, e.g. *"...it's only 2 pages long..."*. As expressed, the user is already starting to lean towards a negative judgement. Should the next utterance be *"...[and] it's from 2006..."*, then its polarity would be deemed negative. This stems from the use of the word *and* to connect the two mentions of criteria suggesting that they share the same polarity. On the contrary, should the next utterance be *"...[but] it's from 2006..."*, then its polarity would be deemed positive. In this case, what makes the polarity to be deemed positive, instead of negative as in the first example, is that the expression is contraposed by the appearance of the word *but* which signals an opposite polarity to that of the first utterance (negative). The preceeding utterance and its polarity are used as a reference point against which the polarity of the following utterance is judged.

The counts depicted in Table 5.2 show that criteria mentions are almost evenly distributed across polarity; out of a total of 215 criteria mentions, 114 correspond to positive mentions while 101 are negative mentions. All criteria was used –either positively, negatively or both– at least once in a single criterion rule.

Because the verbal data gathered from participants did not always correspond to actual relevance judgements of documents, a portion of the uses of the single criterion rule were observed in a different context. Positive uses of this rule were used mostly for assessing the *potential* relevance of the information. That is, participants expressed using a criterion in a positive fashion to decide whether the information could be relevant. The relevance of the information would then be decided, possibly by using more than one relevance criterion, once it had been assessed more thoroughly. Negative uses, on the other

Criterion	Positive		Negative	
Depth/scope/specificity	9	4.2%	17	8.0%
Accuracy/Validity	-	-	-	-
Clarity	3	1.0%	-	-
Currency	3	1.0%	6	3.0%
Tangibility	**55**	**25.6%**	16	7.5%
Quality of Sources	7	3.3%	-	-%
Accessibility	-	-	-	-
Availability	-	-	-	-
Verification	2	1.0%	3	1.0%
Affectiveness	13	6.0%	5	2.0%
Background knowledge	3	1.0%	2	1.0&
Ability to understand	1	-	9	4.2%
Content novelty	3	1.0%	2	1.0%
Source Novelty	-	-	-	-
Document novelty	16	7.5%	**41**	**19.0%**
Total	114	53%	101	47%

Table 5.2: Frequency of each criterion as distributed across single criterion rule uses.

hand, were always used to immediately dismiss the information and hence corresponded to negative judgements of relevance.

Filtering out irrelevant information was mostly done on the grounds that the documents were not novel, e.g. a document had been re-retrieved. Participants mentioned *document novelty* in a negative fashion 41 times (about 19%) when using a single criterion rule:

> *"old ants! ah the little buddies ... that's the document I already have now so I'm not going to read that ..."*

> *"... yes, I've seen it before ... oh, not again, no, still not want to see that, hmm ..."*

The second most used criterion, for filtering out irrelevant information was *depth/scope/specificity*. Document length, in particular, was considered by the participants as an important factor when assessing the relevance of the information:

> *"... no it's very short, I'll put it back ..."*

> *"... I'm gonna put it back because it's very brief and a bit journalistic ..."*

Positive relevance judgements made using the single criterion rule had *tangibility* as the most used criterion. This may suggest that a group of participants found hard data a good indicator of the relevance of the information and when this criterion was met they were quick to accept the information as relevant. However, one must remember that mentions of topicality were also encoded as *tangibility*:

> *"... oh yes it's about simulations, interactive kind of thing, I'll write that one down as well ..."*

> *"hmm, yeah, that could be an interesting application, all right, oh I need to write this stuff at the top don't I?"*[3]

As observed earlier, *document novelty*, when used negatively, seems to be an indicator of irrelevance. This observation, coupled with that *document novelty* is the second most used criterion in positive relevance judgements suggests that the correlation between relevance judgements and the polarity of *document novelty* may be high. *Document novelty* was mentioned in positive judgements as:

> *"... again that one has already been identified as high up which is really emphasising to me that I probably should read it first, and it probably is a major one in this I'm kinda liking this one ..."*

> *"... I think maybe I've seen before, again it gives me a lot of theoretical underpinning it has a lot of really nice, well not really nice, mathematical stuff anyway and yeah I think that's probably the one I would take ..."*

Using the segmentation and counting process described we can estimate the frequencies with which participants applied the single and multiple criteria rules. Estimating the frequencies of the other rules is a much more subjective task, hence no quantitative data is provided in the analysis of these rules.

[3] In the study, participants were asked to write down the document identifier whenever they thought they wanted to keep it for later reference, i.e. they considered it *relevant*.

The dominance rule was mentioned by participants. The use of this rule suggests that participants assessed the relevance of documents not only in relation to the topics being inspected but also in relation to the previously assessed documents:

> *"...that's the kind of paper that I'm looking for, it's probably the most appropriate that I have found, more than previous ones ... "*

> *"... this must be one of the best ones I've found so far ... "*

A reversed version of the dominance rule was also observed:

> *"... I'll put it as relevant but it's not as relevant as the others ..."*

This mention suggests that the participant deemed a document relevant, though when compared to the previously assessed documentation, it was not "as relevant". This could be due to two potential reasons: a reversed dominance rule, which means that the document is considered relevant despite it being worse (in some aspects) than the previously found documents or that this is an expression of usage of the scarcity rule. It could have been the case that the participant had found a few other documents before the one assessed, however, the number of documents found was not enough. If the scarcity rule was applied and, even though the document might have been "less relevant" than the others, the participant would still decide to keep the document.

Expressions of use of the chain rule were also observed:

> *"... can't help feeling that this one should be a rich vein ..."*

> *"... that [topic] was quite a rich one so ... I got quite a lot out of that one ..."*

and one mention was even coupled with a suggestion for a desired feature of the system:

> *"... this is something that I want, I'm not going to read it because the title says it all ... now what I really desperately want is a little box at the bottom that says "find lots of other things like this" ..."*

That the participant requested a feature that retrieved "more like this" suggests that the participant suspected that the document might have been the first example of a set of documents in the same vein of information and that they might have all been interesting. One could, therefore, consider this expression as a use of the chain rule.

In addition to the six rules presented by Wang & White (1999), participants of this study also applied the following rules:

1. The reoccurrence rule: participants selected, for further inspection, reoccurring documents.

2. The concordance rule: participants interpreted as a signal of potential relevance when a document appeared high on the ranked list on both the left and the right panel.

These two rules were applied to assess the *potential* relevance of the documents (defined as weak relevance by Harter (1992)). This means that participants used these rules to decide whether they would click on a link to obtain the full documents and assess them further. The reoccurrence rule refers to documents being re-retrieved during the session. As such, it seems that the number of times the document would reoccur was interpreted as a signal of relevance:

> *"... it's been presented to me for every single damn search query I input in this thing so something tells me it might be relevant ..."*

> *"... again this one, this one is cropping up everywhere this document, it's about chips ..."... simulation model" okay that's definitely something that I will have a look ..."*

> *"... in this case the Schneider thing has come up again and again, we'll have a look at it just to see ... yeah, well okay we'll pick that just because it's interesting to me ... just because it has popped under my nose enough ..."*

> *"... it's [the document] popped up a few times so I'll take it ..."*

This rule was also affected, or so it seems, by the intermediate topic being evaluated:

> *"... yeah I've seen this one before, I think there is some overlap between the topics which I guess it's a good thing ... I've seen that before as well ... already have it ... but I'll put it again ..."*

However, in this case, it may have been that the document was relevant regardless of which intermediate topic had retrieved it (the relevance of the document could be considered invariant to a certain extent) or the document was relevant because a new interpretation had been derived due to the context provided by the intermediate topic. Unless participants expressed it, one cannot assess whether it was one case or the other. The reoccurrence rule seems to contradict special cases of the single criterion rule. It was observed that the most used criterion, in single criterion negative relevance judgements, was *document novelty*. So how could reoccurring documents be selected for further inspection when there was a high proportion of documents automatically dismissed based on that they were not novel? Based on observations, it is suggested that the reoccurrence rule depends not only on documents reoccurring but also on the frequency of the reoccurrence. A document reoccurring for the first time may be initially dismissed automatically on the grounds on "having seen it before", however should the same document reoccur for the nth time, then it may have been selected for further inspection because it had "cropped up everywhere." Moreover, for a document d to reappear frequently, it must be retrieved, and ranked highly, on several of the intermediate topics inspected by participants. That a document frequently reappears in different contexts, such as the contexts provided by the different intermediate topics, seems to have prompted participants to select them for further inspection.

The concordance rule refers to documents that appear ranked highly on both the right and the left panel. Participants inspected the top ranked documents on both the left and right panel to find coincidences:

> *"... So, again I've looked at the top four of the results I've been produced and I'm noticing that they are quite different ..."*

These coincidences were interpreted by these participants as an indication of the potential relevance of the agreeing documents:

> *".. again the top one on both has matched so again I'm thinking that that's probably a good place to start ..."*

> *"... and again here at the top, the top ... here I'm thinking that the top two have again agreed ..."*

And some participants were even puzzled when they did not see this concordance:

> *"... So, the first time that the two documents listing haven't agreed at all so I'm thinking that I will really have to think about how I'm gonna tackle this topic ..."*

This interpretation of the search results as presented on both panels is plausible. Documents on the left panel are supposed to be discussing the relationship between the participants's research area and the intermediate topic while the documents on the right panel should be discussing the relationship between the intermediate topic and the target topic. The coincidence rule seems reasonable as it suggests that documents highly ranked on both panels may be likelier to explain both sides of the relationship and as such they may provide more information in a single place.

Even though listed as two separate rules, participants applied the two rules combined in a single step:

> *"two wildly different results come out of the search on either side, our old friend the first article I picked has actually come out at the top again which again makes me think that this must be some article, must be really really good, and yeah I'm kind of thinking that I'm always getting that article and that I should really just take the hint and go and read that particular one ..."*

5.3 Interactions Revisited

Participants interacted with the system in different ways, however patterns were observed. These patterns depended on two factors: the information presented by the system at any one stage and how this presentation was done, i.e. the user interface. A typical search session can be divided into three main stages:

1. The selection of target topics
2. The selection of intermediate topics
3. The assessment of the related literatures

5.3.1 Selecting Target Topics

At the beginning of the search session, participants were presented with their research topic and a list of ten possibly related topics (target topics). Participants were asked to investigate, one at a time, three out of these ten target topics. Initially, participants went over the list of the topics, starting with their own, and tried to assimilate them. Participants usually began by trying to assimilate their own area of research first:

> *"Ok, right, so, I'm looking at the starting topic first ... evolutionary ehw means nothing to me, heuristics, p2p peer yeah ok let's see where that came from ... computation ... very general ... genetic constraints ... genocop was a particular kind of optimisation software ants and food, so these are my starting topics and they may well relate to some grant sort of thing that I was doing..."*

> *"Ok, so let's see what keywords I've got to start with ... induction maybe, expert yes, training possibly, CBR retrieval definitely ... belief evidence ... mining definitely ... costumer ..."*

> *"... the ones immediately jump off the page are off the screen are archives cultural heritage probably human if we take human in the broadest sense, user studies would come into that ..."*

Participants may have deemed this step necessary due to how the initial topics were presented. Each topic was represented as a bag of words. Because these bags of words had no structure, participants had to interpret what the represented topic might have been. After inspecting their initial topics, participants set out to decide which related topic they would investigate further. The selection of topics, at this stage, was based on two main factors (as expressed by participants):

1. Whether the topics *"jumped out"*, i.e. they stood out by either being *"obvious choices"* or strange enough combinations of words such that they arose the participant's curiosity.

2. Whether relationships between their area of research and the presented topics could be inferred at this stage.

Participants initially scanned the list of related topics in search for something that was salient enough that would make a particular topic stand out from the list. As such, they tried to assimilate them and make sense out of the bags of words they had been presented with. This is an area in which the system could be improved. A more intuitive representation of the topics may make the selection making process easier by leaving more cognitive energy to be used for finding connections instead of interpreting bags of words. Once participants had a topic in mind, they started forming initial potential explanations to why/how the topics were related:

> *"hmm related to this maybe, maybe possible applications for ... [...] ... integrations transaction, what the heck is that? ... [...] ... oh heck! Difficult to find any obvious meaning from the keywords for the topics, my guess health care seems a very clear one so let's start with health care ... "*

> *"... possible related topics ... yes, ok, I'm looking down these to see if anything jumps out at me as particularly interesting or that I think might me an application of my research ... ok, I'm going to start with "mathematical computation logic" I do expect that to relate to my research in terms of graph theory or some*

sort of computational theory ... relating to evolutionary algorithms so the idea would that I would be able to discovery something about that ... so I'll select that one ..."

"the possible related topics, hmm ... I supposed if I'm looking at it from a technological perspective which if I'm thinking outside the box as my senior colleague has told me, then "retrieval classification" and "evaluation", "evaluation" particularly ... I like the next one which is "indigenous" I don't like the "Africa" but the "indigenous" bit is something that I quite like and that would actually tie in with the Australian people the aboriginal ones ... [...] ... but dare I say in curious combinations, so indigenous is interesting, "Africa" is not, indigenous is interesting aboriginals is not specifically ..."

When participants clicked on a topic, they were presented with a second screen in which they were given the opportunity to investigate the intermediate topics that completed the potential relationships between their area of research and the selected topic. Participants entered this stage, usually, with a preconception of the type of relationship they were after. This may have affected how they evaluated the documents as these preconceptions may have imposed a certain structure and even biased their expectations:

"ah it's interesting what it decided to come up with "computational java infrastructure" ... I would've thought it would've been about just general optimisation of code and such like ..."

"I'm thinking that my initial thing is that "teachers ..." has come up which I'm finding a bit strange, I'm finding also a bit strange that some of the other ones that have come up are quite application-based maybe that's the "applied" coming out; our old friend from number one task is back ... so I'm thinking that I'm possibly going to struggle with my initial thoughts ..."

"... it's floating up for a few things and I'll take it although it's not directly relevant to the current ... it's not what I'd expect to find in this current thing ..."

"... I think this is interesting that this document would be related to this topic, this doesn't make any sense to me ..."

Almost all participants devoted an equal amount of time to the inspection of each related topic, i.e. they spent about one third of the allocated time to each topic. This may be because participants were offered to be prompted every 15 minutes before the session began and almost all participants accepted the offer with the exception of one participant who asked to be prompted every 20 minutes. Only one participant spent almost the entire session inspecting a single related topic. When prompted at minute 45, the participant realised that there was not much time left in the session and decided to end it.

5.3.2 Selecting Intermediate Topics

The second screen offered three different panels. The top panel was vertically split into two panels. The top left panel contained the representation, as a bag of words, of the participant's research area. The top right panel contained the bag of words representing the selected related topic. These two topics were static in the sense that participants could not interact with them by either modifying them, selecting a new topic, etc. In the middle panel, a list of intermediate topics was offered. These intermediate topics were also represented as a bags of words. In the bottom panel, also split vertically into two panels –left and right–, the supporting literatures were displayed. Initially, as participants had not selected any intermediate topic, each panel in the bottom listed the retrieved literature for the participants's research area and the selected related topic. The bottom left panel contained the literature retrieved for the participant's research area and the bottom right panel the literature retrieved for the related topic. At this stage two common behaviours were observed (Figure 5.4):

- Participants directed their attention directly to the literature panel.
- Participants directed their attention directly to the middle panel and the intermediate topics.

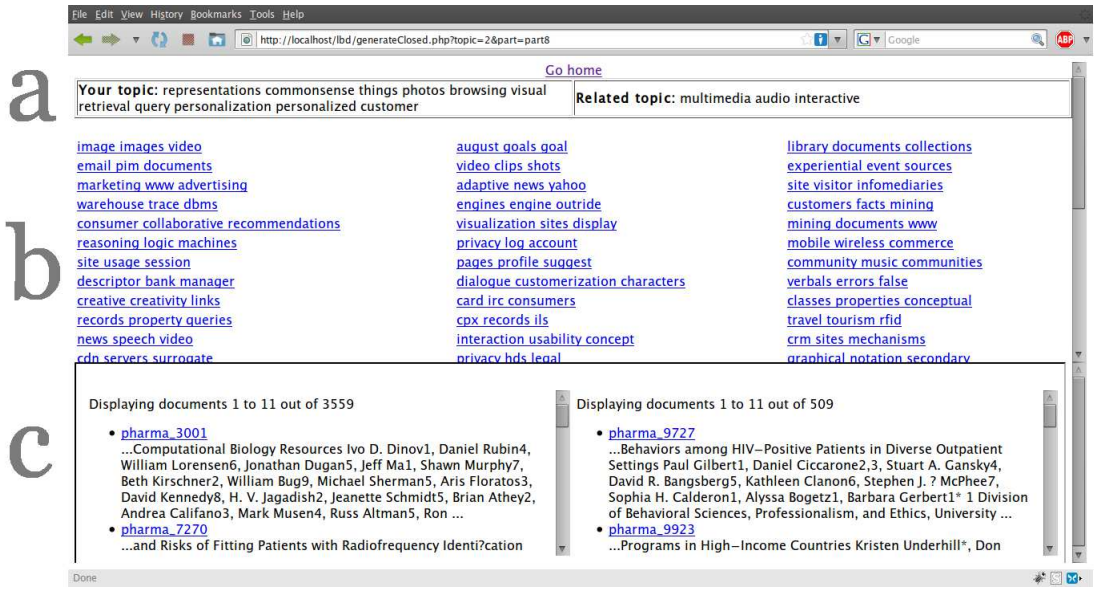

Figure 5.4: Navigation screen. The top panel (a) displays both the initial topic (listed as "Your topic") and the potentially related topic (listed as "Related topic"). The middle panel (b) lists the intermediate topics. The bottom panel (c) contains the supporting literatures.

When participants started analysing the literatures first, they found that most of the documents listed on the left panel had already been seen. This is due to the fact that the representation of their research area, the bag of words, had been extracted from their initial set of documents retrieved during their first session and hence were likely to re-retrieve and rank these documents at the top of the list. To some participants, this did not seem to be an impediment for selecting these documents again. Perhaps the new context, as provided by the related and intermediate topics, shed a new light into the information contained in the documents:

> *"I'm intrigued by [reads] "record culture institutions" ... I'm drawn to this one ... but I think it's one of the ones I took last week ... it is ... and that very much interest me ... even though it's under [topic] is giving examples of the preservation of collective community memory in different media really ... even though it's of the USA there are lots of examples that are probably transferable ... so I'd be interested in that ... and that's from the left ..."*

The purpose of the participants's research topic and selected related topic was to provide a context in which the intermediate topics were to be evaluated. When participants selected the intermediate topics, they also applied the same rules as for the target topics:

> *"okay, let's explore some subtopics and see if some of these relate to health care and CBR problem solving ... I don't think most of these do but ... face video participants ... hmm ... let's see this one ... I guess I should look at both sides rather than just ... to see what's related to ..."*

> *"ok so I'm now looking down the list of things that we have there ... ok, so I see the one that says combinatorial complexity which is quite close to evolutionary algorithms or I suppose that before I click on that ... okay I will click on that combinatorial complexity travelling and now I've got a bunch of things that relate specifically to that ..."*

In this case, the participants mentioned their use of the rule of connection making for choosing the intermediate topic selected. In this context, however, the rule of connection making may have been harder to apply than before as there is less freedom, i.e. while initially participants had to speculate about what the potential relationships between their areas of research and the suggested target topics may be, now they would have to make sense of their speculations in such way that they included some of (if any) the intermediate topics presented. This may explain why there were cases in which only the rule of saliency was used to select the intermediate topic:

> *"... this is where we get our related topics ... I see we've got various that have their countries of origin because I remember that some of the papers where in Uganda and Ghana and so on ... Nigeria gets a mention ... some mention nurses which clearly relates to the medic professions papers, clinical trials, again seems to be related to the medical papers ... health women ... South Africa, Nigeria again, oncology ... right ... so if again I try and find one that seems more general and generic as opposed to one that has a very specific geographic or sectoral kind of focus ... I'll try "viewed abstracts abstract" ..."*

"... [reads] "routing street women" I want to look at that just because it seems a bit odd ..."

"... I'm very tempted by [reads] "chatterbot terrorism" because it's such a weird combination ..."

"... there are not so many intermediate topics that are jumping out to me this time ... "

"... I've kinda been through the ones here that kinda stick out, the ones that jumped out to me ..."

Once an intermediate topic was selected, each panel would list the retrieved documentation for:

- left panel: the participant's area of research combined with the selected intermediate topic.
- right panel: the selected intermediate topic combined with the target topic.

Once the literatures had been retrieved and listed on both panels, participants proceeded to investigate them, however, to decide when it was time to finish their current inspection, by either selecting a new intermediate topic or a new target topic, participants applied one of two different rules:

1. Satisfaction: participants were satisfied with the information gathered through the inspection of the documents retrieved by the intermediate topic.
2. Frustration/boredom: participants showed signs of frustration (or boredom) as the information obtained from the selected intermediate topic was not satisfactory/enough/etc.

Examples of uses of the rule of satisfaction follow:

"... the feeling at this point is that I have a round selection, I have news pieces which are going to give me examples of sites that I can go to and look at actual practice and I've got theory and I've got actual practice ..."

The rule of frustration or boredom was also applied to end current searches. Some participants decided to select a new intermediate topic when they either could not find any relevant documentation or simply got bored by the information retrieved by the selected intermediate topic:

> *"... I'm going to do what a good librarian should not do and say that I'm now bored with "information retrieval classification" and go back and do another topic ..."*

> *"...find example of not finding anything or boredom..."*

5.3.3 Assessing the Related Literature

The purpose of the bottom panel was to present the retrieved literature for each side of the relationship. The left panel displayed the literature for the combination of the participants's research topic and the selected intermediate topic (if any). The right panel displayed the literature for the selected related topic combined with the selected intermediate topic (if any).

Initially, participants considered mainly documents closely related to their research topic. This was observed as the documents on the left panel were inspected more frequently than the documentation on the right panel. Participants also expressed this verbally. The literature on the right panel started being considered once participants had exhausted that in the left panel:

> *"... it's interesting because I'm instinctively drawn to the left column first rather than the right column, I don't know why that should be, maybe because that leads my topic field closer to me than those in these fields ..."*

> *"... I'm scrolling through the ones on the left hand side ... [...] ... still on the left hand side, scrolling down ... I'm gonna have a look and see how many others are of interest in this area ... [...] ... searching though the ones on the left for terms that broadly fit the brief that I've been given well, related to the initial terms ..."*

"... so I'm more drawn to the left side because it's more related to my keywords, which is why I'm getting all the 2nd life stuff on the left hand side ..."

Participants seemed more comfortable, initially, with staying *"close"* to their initial topic (their research area). Perhaps they felt that they would be more competent at evaluating the literature if it was closely related to their research area. There were, however, exceptions to this behaviour as some participants set out to explore new territories from the beginning. One participant, for instance, expressed at the beginning of the session (right after selecting an intermediate topic):

"... so I'm motivated to go to the right hand side first because I feel that that's more relevant to the direction I'm trying to go in so I'm gonna look at more carefully ... 'cause I don't really care whether the left hand side is that relevant to what I'm interested in just yet ..."

As the sessions progressed, however, participants realised that diverging into other research areas may be beneficial if they were to fulfil the task they had been assigned and started evaluating the documents on the right panel more often. Documents listed on the left panel became *"ironically too close"* to their research topic so they started diverging and analysing documents on the right panel (assuming that these were "closer" to the target topic). This was observed in two ways: i) mentions of examination of the right panel became more frequent and ii) some participants expressed this explicitly:

"... looking at the documents on the left hand side it's very close to ... but in terms of the brief, future relations, it's actually ironically too close to the initial area and probably isn't looking at relations with other fields ..."

"... I'm on the right hand side ... which is really I suppose where I should be as I'm realising I should be looking for impact of KM on other fields ..."

An example of the increase in frequency in mentions of interactions with the right panel follows. Below there is the transcription of an entire search session that depicts this

diverging behaviour. Irrelevant parts have been omitted, e.g. mentions of use of relevance criteria, interactions that are not related to the right/left panels, etc.

> *"[...] ... I'm scrolling through the ones on the left hand side ... [...] ... still on the left hand side, scrolling down ... [...] ... searching though the ones on the left ... [...] ... I'll look at the ones on the right hand side ... [...] ... still looking at the right hand side ... [...] ... looking at the documents on the left hand side it's very close to ... [...] ... I'll look at the documents on the right hand side now ... [...] ... pick one of the left again ... [...] ... moving on to the right ... [...] still haven't looked at the ones on the left hand side ... [...] ... looking at the left hand side ... [...] ... take one on the right hand side ... [...] ... theres another article on the right hand side ... [...] ... picking up more on the right hand side ... [...] ... again an article from the left ... [...] ... looking at the right hand side ... [...] ... okay pulled up one on the left ... [...] ... okay still scrolling through the left ... [...] ... scrolling on the ones of the right ... [...] ... and that's from the right ... [...] ... I picked from the left hand side ... [...] ... I'm on the right hand side ... [...] ... I'll have a look at the next 10 on the right hand side ... [...] ... I'm not really seeing anything on the left so I'll just focus on the right ... [...] ... scrolling through the right ..."*

During the session, the participant verbalised a total of 23 interactions with either panel. Out of these, 10 were interactions with the left panel and 13 were interactions with the right panel. Overall, the proportions seem to suggest that both panels are equally relevant in terms of user interactions, however it is how these interactions were distributed, as the session progressed, what depicts the previously mentioned behaviour. Encoding the interactions in the transcription using the code L, for interactions with the left panel, and the code R, for those with the right panel, results in an encoded stream of interactions as follows: L L L R R L R L R L R R R L R L L R R L R R R. Visualising these interactions can be done as follows. First, each interaction is considered to occur at a point in time and in an ordered fashion. At any one point in time one of the two types of interactions can occur: either the participant expresses to be interacting

with the left panel or with the right panel but not both at the same time. As such, these interactions are considered to be mutually exclusive. At each step, all previous interactions are considered to be the total number of interactions observed, up to that point, and the proportion of interactions corresponding to each panel is calculated. At any point in time, the proportions sum to 1 (100%). Figure 5.3.3 depicts the sequence of interactions corresponding to the transcription.

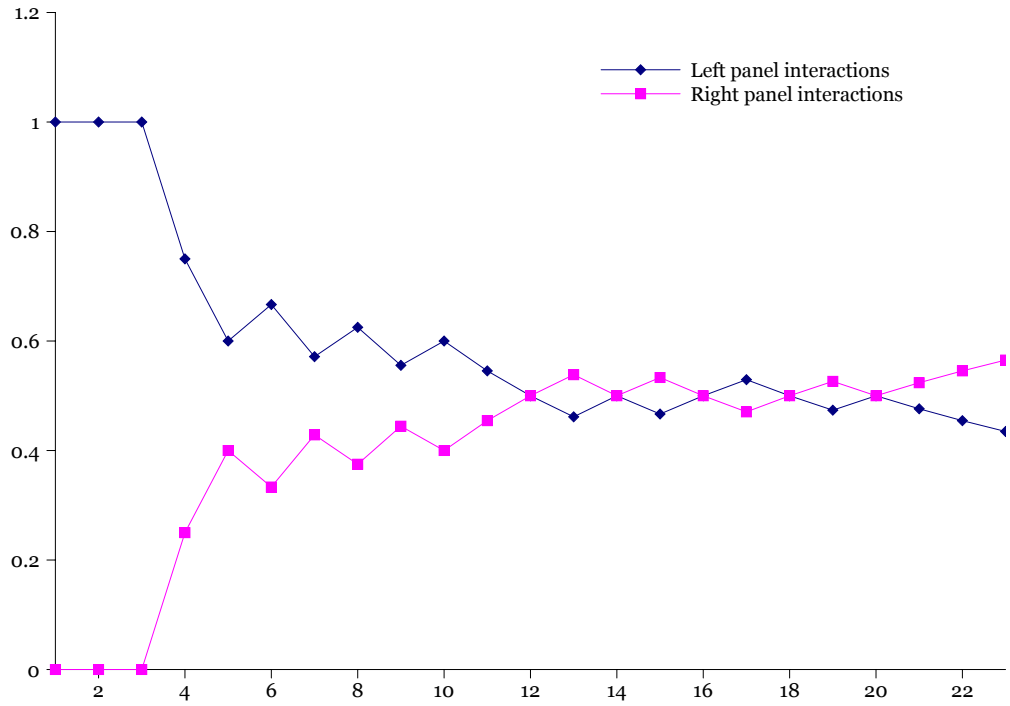

Figure 5.5: Proportion of interactions with the left (right) panel as the session progresses. The two curves mirror each other. For every interaction observed the proportion of one increase while the proportion of the other decrease, e.g. if an interaction with the left panel is accounted for at value 10 of the x axis and the curve for the interactions with the left panel increases accordingly while the curve for the interactions with the right panel decreases.

The proportions for the interactions with the left panel seem to follow a decreasing trend while those of the interactions with the right panel seem to follow an increasing trend. At the beginning of the sequence of interactions (closer to the beginning of the search session) the proportion of interactions with the left panel is larger than the proportion of interactions with the right panel. However, as the session progresses, both proportions tend towards the centre, i.e. interactions with the left panel stabilise while

interactions with the right panel seem to become more frequent. At this point, the participant seems to be interacting in equal proportions with both panels. As the session reaches its end, however, the proportion of interactions with the right panel follows an increasing trend as interactions with this panel become more and more frequent. At the same time, interactions with the left panel begin decreasing in frequency. Despite this potential behaviour, not all participants found the right panel a "richer source" of information:

> *"...my feeling is that there's more useful on the left side than on the right side so maybe I could pick another one ..."*

> *"... so I'm just going to look at the left hand side because that was more successful in the previous search ... I will look at the right hand side in case it is better this time... left hand side ... not finding much interest on the right hand side ..."*

These examples uncover some of the deficiencies with the analysis of the interactions with the left and the right panel performed on the participant's transcription. This type of analysis is very difficult to perform on verbal data. Firstly, verbalisation of these interactions may not be available. Participants were asked to say out loud "anything that went through their minds" during the search session, however, that does not guarantee that they will verbalise all interactions with the system. Secondly, assuming that these verbalisations are available, they may not be properly aligned with time. Verbalisations of interactions with the left and the right panel, for instance, may be observed for the first time after half the search session has elapsed and hence the temporal analysis is rendered invalid. Finally, the verbalisations may be ambiguous. Even if the verbalisations are observed and properly aligned with the progress of the search session, they may still be ambiguous which makes the analysis highly subjective. Some participants verbalised the interactions with the panels in ways such as *"...on the other panel..."* and *"...I will now investigate the other side..."*. These utterances cannot be encoded with either L nor R unless one backtracks until a non ambiguous interaction utterance is found.

Alternatively to verbal data one could resort to analysing the user-click logs or eye-tracking information to get a more reliable account of the interactions with the panels.

The technique for plotting the interactions remains unchanged for either type of data assuming that it was collected appropriately. For instance, if one was to analyse user logs one would have to record clicks for each panel and have them be distinguishable but also to record scrolling actions on both panels as it could be the case that the surrogates are briefly inspected but that the user does not click on any document. Unfortunately, none of these data were available hence an example of how such analysis could be carried out was provided using verbal data.

5.4 Sessions Visualised

In Section 5.1, it was suggested that some relevance criteria profiles may be considered outliers. One such profile was that on row 6. In this section, the corresponding search session is visualised and analysed in more depth. The search session is firstly segmented as described in Section 3.7.5 and then plotted as described in Section 3.7.5. This procedure is also applied to the search sessions of rows 2 and 19 in the divergence matrix (see Figure 5.1) which correspond to participants 2 and 19 respectively.

The Anatomy of an Anomalous Session – participant 7

The result of the segmentation and visualisation process for the profile of participant 7 (in row 6 in the JS-divergence matrix) is presented in Figure 5.6. At first sight it can be seen that the participant spent almost all of the session reading out loud. This could reflect a misunderstanding in the instructions for the study. The participant may have interpreted the request of *talking out loud* as a request for the participant to *read out loud*. We can also observe that the participant did not mention many relevance criteria nor did so very frequently. This explains the high divergence value between the participant's profile and the other profiles. Because the participant may have misinterpreted the instructions and spent most of the session reading out loud, fewer expressions of relevance criteria may have been observed. However, it could also be that the participant did not find any documents that were even remotely interesting and hence silently (in the sense of mentioning relevance criteria) dismissed all of them.

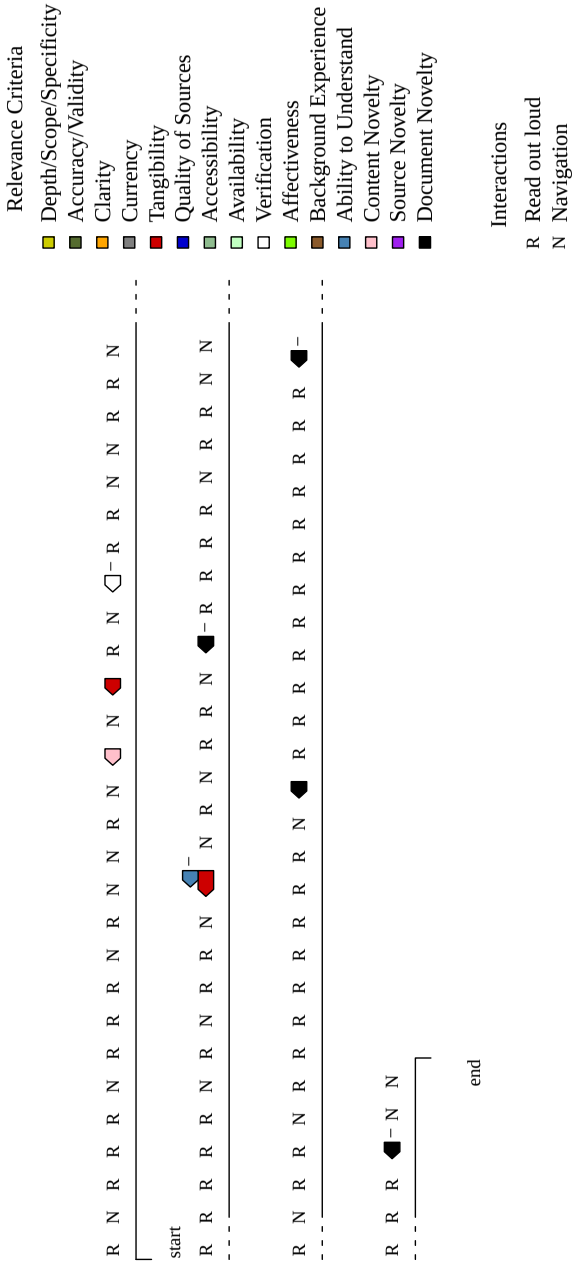

Figure 5.6: The anatomy of an anomalous search session.

Participants 2 and 19

Participant 2 is a research student from the School of Computing. Applying the segmentation and visualisation technique to the participant's search session results in a graph as

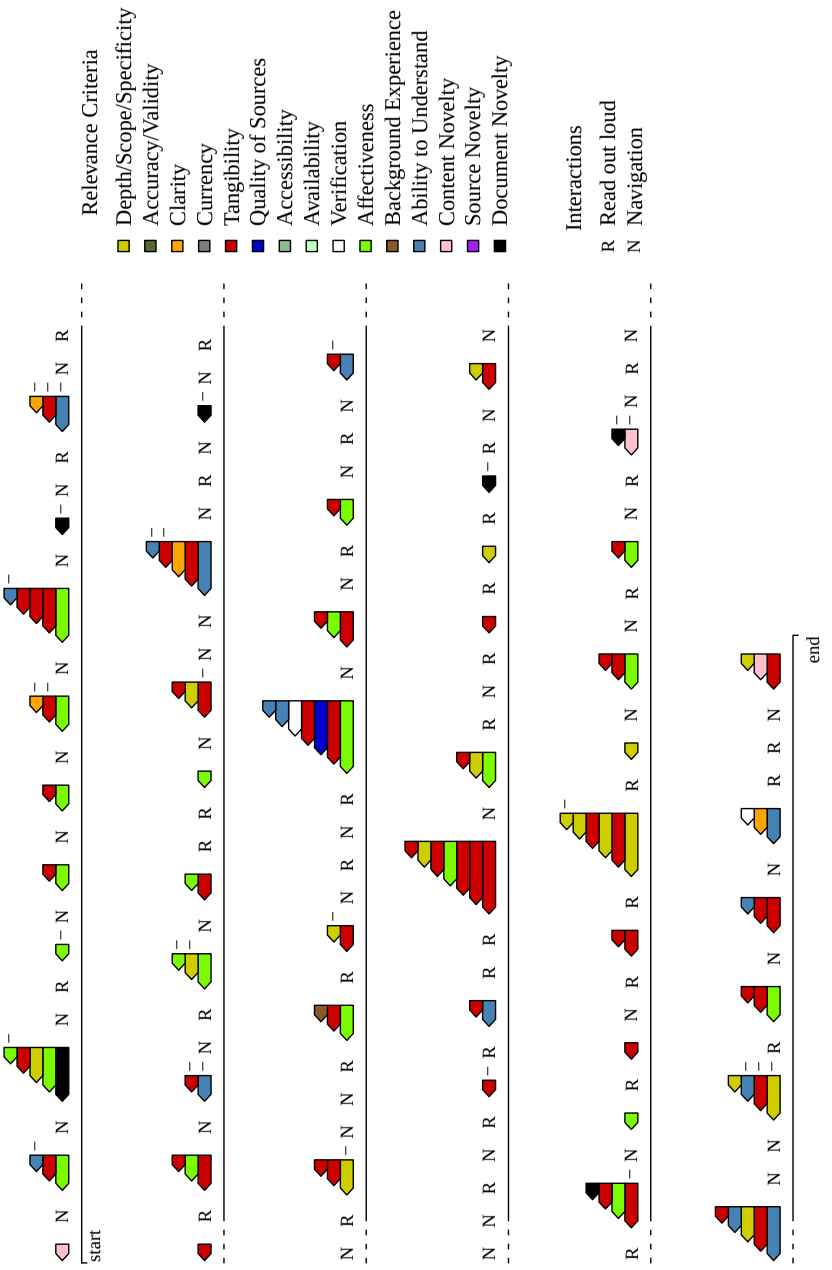

Figure 5.7: Visual representation of the search session of participant 2

depicted in Figure 5.7. In the figure we can observe that the user was engaged during the session. If we interpret the number of expressions of affectiveness as a measure of engagement, then we can observe that the participant is engaged, and remains so throughout the session from the beginning of it. These affective responses, encoded as *affectiveness*, are represented as blocks coloured in light green. Effectively, out of 49 relevance judgement processes (depicted as coloured piles in the graph) 22 (about 45%) contain at least one expression of *affectiveness*. Affective responses seem to be, however, more frequent at the beginning of the session than closer to the end of the session. Perhaps the participant begins to express less emotions (or have less emotional responses) as the session progresses and he becomes more familiar with the underlying collection.

The participant seems to engage in simple to semi-complex relevance judgement processes very often. The interweaving of piles and interactions (including acts of reading out loud) is frequent. This may suggest a more "careful" approach at searching for relevant information. A frequent alternation between interactions and uses of relevance criteria may be due to the participant constantly analysing the presented information looking for cues to derive its relevance. As such, it may be a sign of the participant's experience in finding these cues. A person relatively inexperienced in finding these cues may have to sequentially assess each information piece in more detail. This would be translated to stacks of 2 or 3 relevance criteria blocks. It may also be that the participant is wary and does not want to filter out potentially relevant information too quickly. Hence, the participant assesses in more detail (than average) each piece of information. As the participant expressed: *"... hmmm ... I'm usually crap at selecting things for my literature review, I either go for everything or select hardly anything ..."* which suggests that the participant will use a more careful strategy.

Tangibility, which includes topicality, seems to play an important role during the participant's search session. Out of the 49 relevance judgement processes, 37 (75.5%) include at least one utterance encoded as *tangibility*. This complements the global view presented by the relevance criteria profile (see Figure 4.3) which showed that *tangibility* was a commonly used criterion by participants from the School of Computing. During the par-

ticipant's session, *tangibility* not only was a commonly used criterion, but also one that was present in most relevance judgement processes. Moreover, the criterion is present in relevance judgement processes of different complexities covering almost the full range.

Participant 19 is a senior researcher from the Information Management Group. The participant's search session is depicted in Figure 5.8. Contrary to the interaction behaviour exhibited by participant 2, participant 19 seems to navigate more diligently. Whenever the participant considers to have found a promising source of information, however, the relevance judgement processes are rich both in the number of uses of relevance criteria and in their variety. On average, the relevance judgement processes in which the participant engaged seems to be more complex than those of participant 2. Figure 5.9 contains a bar chart depicting the frequency of the relevance judgement processes in which both participants incurred. Participant 2 seems to mostly engage in processes of complexity 1, 2 and 3 with some occasions in which more complex processes are used. Participant 19, on the other hand, seems to make use, on average, of more complex processes. Even though simple processes (of complexity 1) are used frequently –possibly for quickly filtering irrelevant information– the remaining processes are more evenly "spread out" and more complex processes are more frequent.

In the figure, we can observe that *tangibility* is not as prominent a criterion as it is for participant 2. Effectively, out of 41 relevance judgement processes, 19 (about 46%) contain at least one use of *tangibility* as relevance criteria. *Depth/scope/specificity*, on the other hand, appears at least once in 27 (about 65%) relevance judgement processes. As depicted in Figure 4.3, members of the Information Management Group mentioned in near-equal proportions the criteria *tangibility* and *depth/scope/specificity*. Participant 19, however, seems to unbalance this proportion in favour of *depth/scope/specificity*.

In both sessions we see that some criteria are repeated within relevance judgement processes. *Tangibility*, for instance, is mentioned up to 5 times within one relevance judgement process (participant 2). This is, however, reasonable. The code *tangibility*, as it was interpreted in this study, includes mentions of topicality. Furthermore, several different expressions of references to hard data are to be encoded as tangibility. Expressions

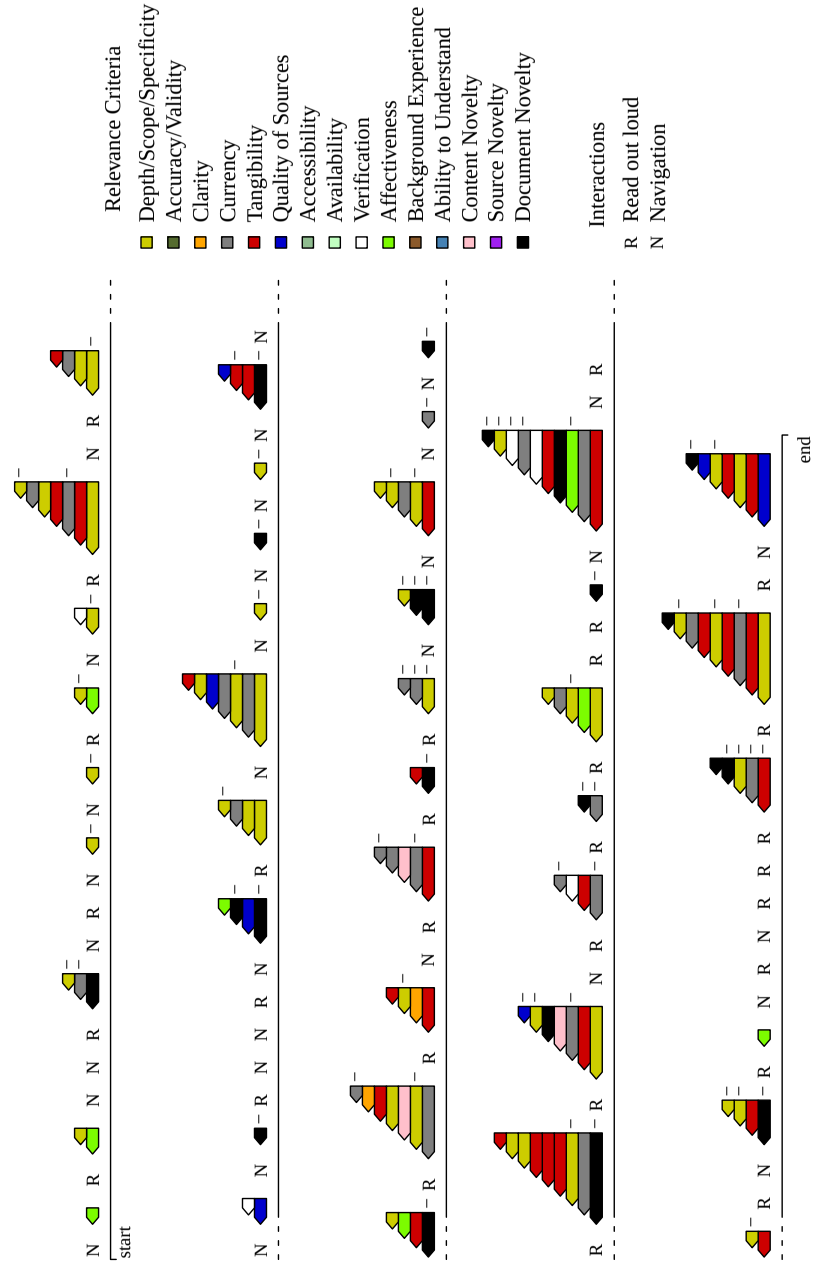

Figure 5.8: Visual representation of the search session of participant 19

like *"... [a] neural network, ah you know what that could almost be an application if a neural network can do it you would be able to evolve it as well ..."* and *"... it'd be*

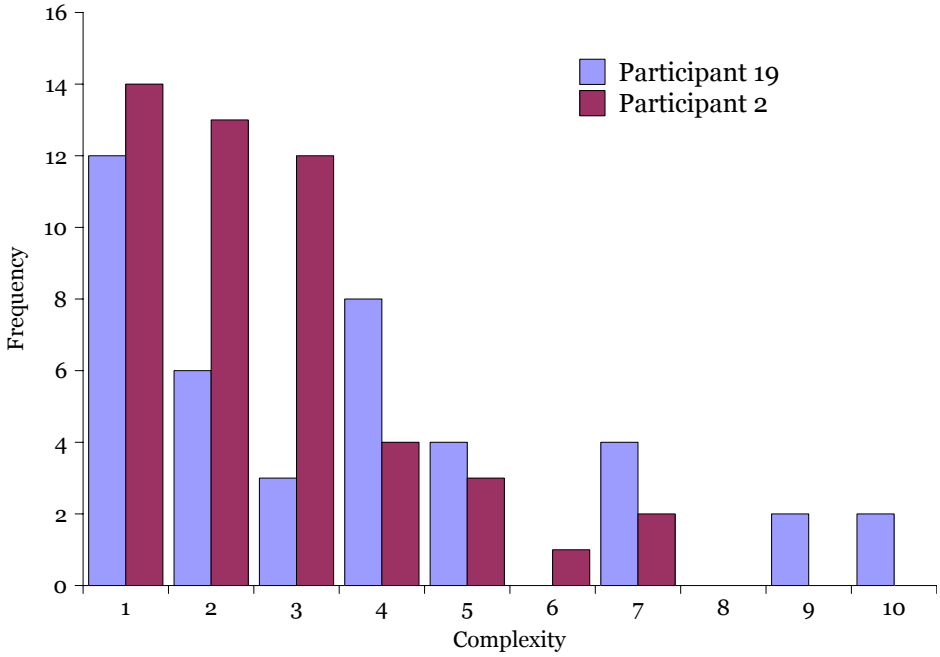

Figure 5.9: Frequency of the relevance judgement processes by complexity.

one to look at to get references from this ...” are both encoded as *tangibility*, however they both refer to different types of tangible information being analysed. One refers to the details of an implementation of a technique (a neural network), and in some sense it could also be encoded as *depth/scope/specificity*, while the other expression refers to the references to be extracted from the document (which could also be encoded as *intent*). *Depth/scope/specificity* is another such code. It was mentioned up to 4 times within any one relevance judgement process (participant 2). As a criterion that encompasses mentions of different properties of the information being assessed (its depth, its scope, its specificity with respect to the user's information needs, etc.), *depth/scope/specificity* is likely to be repeated within relevance judgement processes. Consider these expressions from participant 19:

> *“... [this] is really what I'm interested in and again is really relevant to the brief which is find new technologies or technologies used in a new way for knowledge management and sharing ... looks quite current, november 07, looks a wee bit anecdotal and it's very short so I'll put it back ...”*

These expressions correspond to Figure 5.10. As it can be observed, there are repeti-

tions of *depth/scope/specificity* and even with opposite polarity. This is due to that there are expressions referring to the specificity with respect to the participant's information needs (*"it's really relevant to the brief"*) and to the volume of the information (*"it's very short"*). This repetition of mentions of the criterion *depth/scope/specificity* was observed frequently for participant 19 (and other members of the Information Management Group) but not for participant 2 (and the remaining members of the School of Computing) while repetitions of mentions of *tangibility* where more frequently observed during the session of participant 2 (and remaining members of the School of Computing).

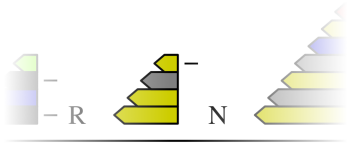

Figure 5.10: Repeated expressions of *depth/scope/specificity*

5.5 Summary

In this chapter the relevance criteria profiles and the user interactions at different levels were analysed. Initially, the similarity of the relevance criteria profiles was examined using a divergence measure. The Jensen-Shannon divergence (Lin 1991b) was measured between pairs of relevance criteria profiles and these similarities were plotted as a heatmap. On the heatmap we observed that there may be three naturally emerging clusters. However, it was noted that the emerging clusters did not map either on to the participating schools nor the research experience levels of the participants. This may be due to the fact that there are no common behaviours, in terms of relevance criteria inter-relationships, that are common to either the schools or the research experience levels. Alternatively, it might have been that the analysis technique is not entirely appropriate and that other methods would reveal that there are indeed clusters that correspond to these predefined groups.

Relevance judgement processes were analysed next. It was described how these processes are obtained through a segmentation process of the search sessions. This segmen-

tation processes relies on expressions of interactions and uses these as delimiters. The segmentation process rests on the assumption that a judgement process begins and ends with user interactions. This implies that relevance judgement processes are not necessarily related to document relevance judgements. The complexity of a relevance judgement process was defined as the number of criteria mentioned within the process. The complexity of a process was used to map it to either the single-criterion rule (referred to as *elimination* due to its most common usage) and the multiple-criteria rule as presented by Wang & White (1999). It was observed that participants usually engaged in simpler processes, however there were observations of more complex processes (of complexity up to 16 criteria). Single-criterion rules were used, in accordance to (Wang & White 1999), to quickly reject unwanted (or potentially irrelevant) documents. However, we also observed the opposite behaviour. The single-criterion rule was also used to quickly accept and judge as relevant the information presented. This was discussed through the analysis of the polarity of the criteria involved in such uses of the rule. Negative mentions of the criterion used in the rule were mapped to rejections of the information presented while positive mentions of the criterion to the potential acceptance of it. Overall, all mentions were evenly distributed across polarity, however some criteria leaned more towards one of the two. Rejections were mostly done based on the basis of the documents not being novel, i.e. 19.0% of all mentions of a single criterion were negative mentions of *document novelty*. Most positive mentions of a single criterion (25.6%) were mentions of *tangibility*. It was observed that the second most mentioned positive criterion is also *document novelty*, which suggests that the polarity of *document novelty* may be highly correlated with the actual relevance judgement of the documents. Additionally, a discussion, through examples and qualitative data, of the occurrences of the other 4 rules listed in (Wang & White 1999) was offered.

It was also described and analysed, with examples, the user interaction patterns. These patterns were described at three stages:

1. During the selection of target topics

2. During the selection of intermediate topics

3. During the assessment of the related literatures

The way topics were represented –as bags of words– may have had an effect on the interactions and how participants decided which ones to inspect further. The selection process was based mainly on two approaches:

1. Standing out: if the topic immediately stood out, it would be selected for further inspection. This included "obvious" topics or topics which would arouse the participant's curiosity.

2. Inferred relationship: if a relationship between the starting and destination topic could be inferred, the destination topic would be selected for further inspection.

The selection of intermediate topics was based on similar patterns, however when it came to inferring relationships, the degrees of freedom were more restricted (when compared to the top level topic selection) since there was more context in which the selected intermediate topic should be placed.

The interactions during the literature selection process were also examined. A qualitative example in which a participant slowly digressed from interacting frequently with the left panel to interacting more frequently with the right panel was offered. To exemplify this behaviour, a simple approach for plotting these interactions was used. Interaction were modelled as a sequence of L and R events (signifying an interaction with the Left and Right panel respectively) and interpreting each of these as contributing to a total number of interactions at any given point in time. Verbal data from an example session was used. Plotting the proportions for this session then resulted in mirroring curves showing the digression the participant had incurred in. It was suggested that verbal data might not be the best data for this kind of analysis and that user-clicks or eye-tracking data might be more appropriate. The technique presented, however, remains data-agnostic and could be equally applied to either of those two data sources.

During the analysis of the divergence measures between profiles it was suggested that one of the participant's profiles was very different from the other profiles. When we visualised the sessions of three participants, it was confirmed that said participant was

indeed an outlier. Further, it was suggested that the way in which the participant behaved may have been due to a misinterpretation of the instructions delivered at the beginning of the session. This was reflected in the visualisation of the participant's session as being mostly read-alouds with very few mentions of relevance criteria. The other two sessions were briefly analysed, and it was observed that some commonly occurring criteria were also distributed across the search sessions (as opposed to being concentrated on some portion of them). One of the participants, participant 2, was engaged all throughout the search session and to have mostly used relevance judgement processes of moderate complexity. This behaviour suggests that the participant approached the task with caution as to not filter out any potentially relevant information. Participant 19, on the other hand, exhibited a more diligent approach at judging information. Although the participant used the single-criterion rule quite often (perhaps for quickly filtering out irrelevant information), he also incurred in processes of increased complexity and variety.

The custom visualisation tool developed in this dissertation has been useful in:

1. Outlier detection: as suggested, looking at the plot of the search session quickly confirmed that the session was drastically different from the other two

2. Criteria distribution throughout the session: it was possible to observe, in a glance, the spread of different relevance criteria across the search sessions. This is useful in complementing the global analysis of relevance criteria profiles which provide with a total number (or proportion) of mentions of criteria for the entire session

3. Relevance judgement processes complexity: we could observe, again in a glance, how processes of different complexities are used across the search session which gives an indicator of the type of search behaviour that a participant may be exhibiting

Session visualisation has a number of drawbacks however. Firstly, visualisations are not comparable across sessions. This stems from the fact that they are not time-annotated, hence one cannot analyse, for instance, at which point in time (during the session) each participant mentions their first relevance criteria. This drawback comes as the encoding of the verbal utterances was not time-annotated either. Had they been so, the informa-

tion would be available and it could be included in the session visualisation. Secondly, relevance judgements are not visualised. This is related to the way relevance judgements processes are isolated. Relying on user-interactions as delimiters for relevance judgement processes does not guarantee that they will be aligned with document judgements hence these cannot be included in the visualisation. Coupling verbal analysis with the analysis of logs to discover the user interactions might alleviate this situation. Finally, a scheme for picking colours for labelling the different relevance criteria was recommended, however if the number of observed criteria increases the number of different colours must do so accordingly.

Chapter 6

Discussion

Carrying out an experiment designed as explained in Chapter 3 allowed for the observation of the relevance criteria used when researchers, of different disciplines and research experience levels, judged the relevance of information related to their area of research. In Chapter 4 the observations were presented in the form of relevance criteria profiles; a technique for grouping these observations developed in Section 3.7.4 which allows for different visualisations of the data as well as comparisons between the observations. These profiles showed that the participating researchers used different criteria in different occasions and frequencies.

In Chapter 5 the relevance profiles were compared against each other using a divergence measure. It was observed that even though there may be three naturally emerging groups, none conformed to the imposed groups: discipline or research experience level. This suggests that additional factors may be affecting the uses of relevance criteria. Relevance judgement processes were isolated and the global trends and rules used analysed. It was reported that while all participants used a single criterion for judging the relevance of the information presented often, more complex processes were also common however less frequently used. The chapter finalised with a description of the common interaction patterns when selecting target topics, intermediate topics and the supporting literature. Additionally, it was suggested that participants followed an interaction pattern during their search sessions. An example was provided in which this behaviour could be observed.

Participants usually began by analysing the literature that was deemed to be "closer" to their research experience to eventually drift to the literature concerning the selected related topic.

Overall, the approach taken seems appropriate for conducting, at least, observational studies on the cognitive processes involved in LBD searches.

6.1 Verbal Protocols

The design of the study included the use of verbal protocols for gathering data regarding the cognitive processes involved in the assessment of the relationships suggested by the system. It was decided that concurrent reports, as opposed to retrospective reports, would be used as they would provide a raw view of these processes. An alternative, although costlier, approach would be to complement concurrent reports with retrospective reports as this might not only improve the reliability of the data gathered but also provide new insights into the cognitive processes being observed (Taylor & Dionne 2000, Ericsson & Simon 1993). Both convergent as well as divergent information contained in complementary reports are of use to the analysis of the data. Convergent information offers an opportunity for validation as well as elaboration on behalf of the researcher, while divergent information may indicate where the complex relationships that are part of the cognitive process under study lie.

During the description of the recommended guidelines for gathering data using verbal protocols (Section 3.6.2) it was suggested that all participants should receive the same instructions in order to maximise reliability. This guideline was followed as during this study all participants received the same set of instructions, however these instructions were verbally communicated. To maximise reliability and repeatability while minimising threats such as the natural fluctuations in language, for instance, that are likely to be present when communicating instructions verbally, the instructions should have been given in writing to participants. This has the added benefit of ensuring that participants understand what it is required from them. Although great care was paid when making sure participants had understood the instructions, during the analysis of the data gathered during the study, it

was suggested that some participant might have misunderstood the instructions since all the data gathered during their sessions consisted of reading out loud data points. Providing the instructions in writing, allowing for reading and re-reading time, and allowing for questions to be asked afterwards would have minimised the risk of misunderstandings and improved reliability, and hence it is the preferred approach when delivering instructions to participants.

After having been instructed, participants performed a warm-up exercise on example, but realistic, data. The exercise, however, was not identical to the task they would have to perform afterwards. The exercise allowed for maximum freedom in navigating the system. The goal of the exercise was twofold: i) to make sure participants would be familiar and comfortable with the system as its user interface diverged from those of traditional search engines and ii) to make sure participants had understood the instructions and were comfortable verbalising their thoughts as they used the system. The design of future studies, however, should include a warm-up exercise that mimics, as much as possible, the actual task to be solved. This is to augment the chances that participants not only have understood the instructions and are familiar with the system in question, but also so that they are comfortable providing verbal data as they *solve a similar task* as opposed to simply freely navigating data using the new system.

The reliability of the encoding procedure was assessed by means of having an independent encoder encoding a random sample of utterances and then measuring the overlap between codes. Although this overlap was found to be significant, this is an ad-hoc procedure in that it is an intuitive approach but that its reliability was not validated. Approaches better founded and grounded on literature should have been followed instead. For example, using inter-rater agreement measures such as Kappa (van Someren et al. 1994) might be more appropriate. Future studies should take this issue into account seriously as using standard measures allows not only to obtain a measure of reliability but also to judge its standard. Additionally, having more than a single independent encoder will reduce the threat of subjectivity and personal interpretation in the encoding of the transcribed data.

As a final note on the gathering and analysis of verbal data, it must be pointed out

that although a predefined encoding was used, a custom encoding could have been derived from the data as this is the suggested approach in the literature. However, using a predefined encoding increases the stability of the existing encoding as well as provides both refinements and new interpretations of the codes. Much like complementing concurrent reports with retrospective reports, a similar approach could be followed in future experiments where independent encoders evolve their own encoding from the data in parallel to encoders labelling utterances using a pre-defined encoding scheme (if a suitable encoding is readily available). Although measuring the agreement between encodings might prove to be a very difficult task, doing so will increase the likelihood that any one code is objective and describes the cognitive process under study (or part of).

6.2 Relevance Criteria

Researchers do use different relevance criteria when judging the relevance of the presented information. In this case, information related to their area of research. The most frequently used criteria were *tangibility* and *depth/scope/specificity*. *Tangibility* is the criterion dealing with whether the information is tangible; utterances coded with *tangibility* were observed a total number of 595 times. Mentions of topicality accounted for almost half of its mentions: topicality was observed 257 times – a 43.2% of all mentions encoded with *tangibility*. If we take this into account, we then have to conclude that it is *depth/scope/specificity* the most used criterion, followed by *tangibility* and *topicality*.

It seems sensible that topicality was not the most used criterion. When assessing the relevance of related information, it is the potential relationship between the presented information and the starting topic that is assessed and not its topicality. At least not initially. Keeping in mind that assessing the relevance of related information, if we accept that it is the nature of the potential relationship that is being assessed, is hard, it is likely that the participants were actually judging the potential relevance and not the actual relevance of the presented information. In this respect, they may have judged the relevance in a shallow fashion, e.g. “this looks like it has potential, I’ll save it for later.” Especially considering the time constraints imposed. Hence, interpreting properties of the

information such as document length, the genre of the document, and the specificity of the information as signals of potential relevance seems sensible. Longer documents may offer *more* relevant information, general overview documents may be *easier* to understand and may provide a *broader* picture of the field. An interesting case of the mentions encoded as *depth/scope/specificity* is that of those corresponding to mentions of *exemplary documents*. These documents *"...describe or exhibit the intellectual structure of a particular field of interest. In so doing, they provide both an indexing vocabulary for that area and, more importantly, a narrative context in which the indexing terms have a clearer meaning"* (Blair & Kimbrough 2002). An example of exemplary documents is the *survey article* which broadly covers an area of research and as such offers, amongst others, a list of the major trends in it, the state of the art techniques as well as the most influential authors. Based on the observation that these documents were chosen across research experience levels and affiliation group, it is not unreasonable to suggest that these are worth special attention. Perhaps these documents, as they offer an overview in a single document, are a good entry point for researchers to a new field of research. Moreover, considering the amount of information usually contained in these documents, it may be that the rewards obtained once read, when compared to other documents in the field, is large, i.e. the ratio $\frac{information\ obtained}{effort\ needed\ to\ obtain\ it}$ is large (Harter 1992).

Members from the School of Computing exhibited a preference for tangible information as utterances encoded with *tangibility* were their most frequently expressed. Members of the other two schools on the other hand, namely the Information Management Group and the School of Pharmacy, mentioned *depth/scope/specificity* more often than any other criteria. Given the nature of the discipline, it is not unreasonable to observe that members of the School of Computing had a preference for *tangible* information (even after discriminating between *tangibility* and *topicality*). However, the reasons behind the difference in behaviour of the group, regarding the global trends, are not clear. It may be that members of the different disciplines make inferences at different levels of abstraction. While members of the School of Computing may be making connections at very detailed levels, and hence they require tangible information for doing so, members of the other two disciplines

may be making them at more abstract levels, hence the requirements of depth and scope of the information.

When the data was grouped according to the research experience level expressed by the researcher, the picture changed. Regardless of experience level, all participants mentioned *tangibility* the most followed by *depth/scope/specificity*. However, when re-encoding mentions of *tangibility* that are actual mentions of *topicality*, the preferences actually depend on the research experience level of the researcher. Students and Researchers preferred *tangibility* over *topicality* while Senior Researchers did not. *Depth/scope/specificity* was the most mentioned criterion regardless of research experience. This behaviour may have been provoked by the search tasks assigned to each group. While research students and researchers had to complete their literature review and a research proposal respectively, senior researchers had to deliver a keynote speech at a conference. It may be that when writing a keynote speech, tangible data is not as important as topicality since in keynote speeches more speculative ideas are communicated without many details.

Implications

The implications of these observations are clear: relevance, in the context of LBD, is multi-dimensional. Moreover, its manifestations depend on several factors amongst which we find discipline and research experience. Despite the subjective an highly personal nature of these manifestations, potentially global trends were observed which can be used either during the design or testing of LBD systems.

Operational estimations of the two most observed criteria[1] may be embedded in systems in an attempt to increase their performance in returning relevant information. If, and only if, we can measure *tangibility*, for instance by looking at the number of tables in a document, and *depth/scope/specificity*, for instance by looking at the number of pages in a document (document length has been mentioned frequently as a relevance criteria), we may then embed these measurements in ranking algorithms. Two approaches may then be followed. If one knows beforehand the target audience of the system one may decide

[1] *topicality* is excluded as this is already operationalised, to a certain extent, in the underlying *best match* algorithms

to favour one criterion over the other. For instance, if one knows that the systems is to be tailored towards the Computer Science scientific community then *tangible* information may be preferred. Systems tailored, on the other hand, to Information Management and Pharmacy would favour documents for which the *depth/scope/specificity* score is high. However, things are not this simple. When researchers where categorised by their research experience level, the preferences, as exemplified by the criteria frequencies, are different to those of the school categories. This complicates matters as now, if we had embedded in our system the measurements proposed before, it is not too clear how one should favour one over the other. Moreover, detecting when we are in presence of a research student, a researcher, or a senior researcher may be very difficult to accomplish. A compromise may then be reached by taking into account both criteria in equal proportions and favour documents which exhibit both properties over those that only exhibit one or none.

Simulated work task situations are a powerful tool if used correctly, however their crafting should be approached with great care. Moreover, as suggested in this dissertation, comparing results obtained with different variants of these task descriptions may not be as straightforward as initially considered. Effectively, we suspect we might have introduced bias while crafting the simulated work task situations. Especially in terms of the relevance criterion *currency*. When we designed the tasks we decided that we wanted to have one per research experience level and to word it so that participants would relate better to them. Each task required participants to search outwith their area of expertise. However, when we asked senior researchers to find information to use for their keynote speech we might have asked them, implicitly, to favour more recent information. This implicit requirement may be not present, for instance, in the request for finding information to complete a literature review for a doctoral dissertation (the task given to research students). The proportional observations of the relevance criterion *currency*, then, might have been artificially influenced by the wording of the tasks. This implies that the data obtained might not be comparable across tasks, i.e. research experience levels.

6.3 Measuring Profile Similarities

Relevance profiles were compared using a divergence measure and the similarities analysed. Three naturally emerging clusters were observed: those that do not conform to group trends, those that do, and the rest. Unfortunately no cluster corresponded to any of the imposed groupings, i.e. neither cluster corresponded to the school disciplines nor the research experience levels. This suggests that neither discipline nor research experience level alone are driving the behaviours in terms of relevance criteria used. However, being able to measure the similarity in terms of judgement behaviour may be a useful tool for detecting naturally emerging groups which can then be traced to groups in terms of population features, e.g. people more proficient in use of search engines use the criterion *topicality* less often as they assume the results are already topically relevant.

Implications

Measuring (dis)similarities between relevance criteria profiles may result in improved user modelling and collaborative features of IR systems. Assuming that relevance criteria profiles can be estimated (built) using cheap approaches, i.e. cheap when compared to verbal reports, then being able to compare them may be a good approach, for instance, for improving recommender systems. Multi-dimensional relevance criteria profiles would allow recommender systems to escape the binary ratings usually used to build the co-occurrence matrices upon which they based their recommendations. Additionally, individual profiles may be used when modelling user preferences in IR systems targeted at personalised search. As more detailed information about the nature of the user judgements is available, better estimations of future relevance may be achieved.

Individual profiles can be compared to other individual profiles, and what is more, they can be aggregated to build group profiles. These two mechanisms, that of comparison and that of aggregation, form the basis for building group profiles and testing which individuals are to be a part of such groups. Hence, they may have an impact in how communities are modelled, for instance for collaborative search or social network analysis. Additionally, because the profiles define clear relevance criteria, they may serve as guide-

Figure 6.1: Feedback form. Several criteria are present which, when combined, present a more informative judgement for why the video is *worth seeing*.

lines for designing feedback systems in such scenarios. Effectively, one such example can be found in the rating system implemented by the videolectures website[2] which includes criteria such as *tangibility* (valuable and informative), *clarity* (well presented) and *ability to understand* (easily understandable). For all the criteria considered see Figure 6.1

6.4 Relevance Judgement Processes

Relevance judgement processes were defined as a set of relevance criteria used to judge the information presented in between interactions with the system. Their complexity was defined as the number of criteria used during the process. It was observed that processes of complexity 1 were used by all participants, and that there were two common scenarios for their use:

- To quickly filter out irrelevant information
- To quickly accept information that is potentially very relevant

More complex processes were also observed, however as processes became more complex their use decreased on average. The polarity of the judgements was evenly distributed

[2] `http://videolectures.net/`

for 1-criterion processes which makes it difficult to decide, initially, how these processes are likelier to be used. In the case of negative uses, i.e. to filter out irrelevant information, the most used criterion was *document novelty*. This suggests that the criterion is an important one in this context of LBD. Despite the many observations of re-occurrence being interpreted as a sign of relevance, the negative uses of *document novelty* is larger. Positive uses of the 1-criterion rule has as their most frequently mentioned criterion *tangibility*. This observation seems to correspond with what was suggested: that once the potential profitability of the relationship had been established, *tangibility* would be used when devising details of the relationship and that *topicality* may eventually act as a proxy criterion for relevance.

Expression of uses of the dominance rule (and a reversed version) and the chain rule (Wang & White 1999) were also observed. The dominance and chain rule are related to a certain extent. While the dominance rule relates to comparative relevance judgements in which users judge the relevance of the new information in regards to the previously judged information, the chain rule refers to the cases in which users detect to be on a chain of information and make a collective judgement on the set. These two rules are related in the sense that relevance judgements are dependent on the previously judged documents and as such suggest that the relevance of any one document is not invariant to the order and set within which it has been evaluated, i.e. the context of the relevance judgement is important.

In addition to the rules defined by Wang & White (1999), two new rules were observed:

1. The re-occurrence rule

2. The concordance rule

Sometimes the negative impact of re-occurring documents was overridden by the frequency of this re-occurrence. Frequently re-occurring documents were interpreted as being relevant. In a sense, the users expressing this, exhibited a certain trust that the system knew better than them as it was constantly suggesting the same document as being relevant. The concordance rule is related to the re-occurrence rule in the sense that when

both the left and right panel showed a concordance on the top ranked documents, then these would be selected as potentially relevant and inspected further. When the panel contents did not match, users were forced to analyse the document surrogates in more detail before deciding which documents to analyse further.

Implications

The high frequency of the 1-criterion processes observed suggests that users were often interested in quick judgements. Quick judgements, perhaps, would allow users to cover more of the search space, so in a sense it may be that they consider that knowledge discovery is a recall oriented task. This seems to be contradictory with the approaches taken by most LBD systems which attempt to filter out as much information as possible (when modelling, filtering and ranking the topics) before presenting the results to the user. If users actually consider it to be a recall oriented task, filtering out information is a potential risk. Hence, it may be more appropriate to build systems so that they present as much information as sensibly possible while offering an interactive user interface appropriate for quick judging of the information. Moreover, the observations of the used rules suggests that there may be two stages to the closed-model search: the first stage is recall oriented, hence systems should support users to cover as much ground as possible before aiding them (second stage – precision oriented) to focus on the potentially relevant pockets of information.

Document novelty has been frequently used in 1-criterion judgements, and the polarity of the use may have a strong correlation with that of the relevance judgement. Hence *document novelty* may be embedded the user interface of LBD systems. Including information about document novelty in the user interface is already done, to a certain extent, in modern browsers. Browsers change the colour of already visited links, letting users know that they have already seen the document linked. However, this use can be extended to include the number of times it has been read, the number of times it has been retrieved but not necessarily read, and the context in which these events have happened. For instance, next to links leading to already seen documents one could include the number of times it has

been read and when hovering over the link the user interface could inform users in which contexts (intermediate topics) this document has been read. Alternatively, already seen documents could be hidden altogether from users, offering, however, the option to display them should users wished to do so. In addition, this suggestion would accommodate for the observations of the re-occurrence rule, as either hiding re-occurring documents until their n^{th} re-occurrence has been observed or displaying the number of re-occurrences next to their link would inform users of these events. As expressed by a participant, *"...I would say maybe it would be nice to have on the paper, to have which topics it's related to ... because if you say one [document] is relevant, then if it's involving some of the other topics which also appear to be relevant then you may want to look at them, you may not have noticed that because you get so many ... so I think that if you find a key paper you could have "I want more like this" then you may want to know, out of all these, because they re-appear, which ones that is appearing in ... 'cause it's often the case you find something from a totally different angle and you want more of that ... and ok, you do it by following the references at the end of the paper or visiting the author's website but if you have something like this you could actually tell me "ok this paper is relevant, tell me all the topics in here, the suggested ones, in which ones this paper appears" because that gives you an idea of what else to look for ... 'cause some of these ones have some rectangle topological, now there may be things in there but I don't know ... 'cause the keywords may not be the ones I would've picked ..."*

The concordance rule has been observed, and it has been suggested that this rule aided users in starting the evaluation of the retrieved literature. Using colour codes, for instance, to provide visual cues to aid users detect which documents are present in both panels may be an improvement to user interfaces of LBD systems.

Lastly, the observation of uses of the dominance and chain rule suggests that the relevance of the retrieved documents cannot be assessed independently of their context.

6.5 Interactions

Interactions were observed and patterns analysed at three different stages:

1. During the selection of target topics

2. During the selection of intermediate topics

3. During the assessment of the related literatures

How topics are presented seems to have had an effect in how the selections were made. Because the system used during the study modelled topics as tri-grams (bags of three words), participants had to add an interpretation step into the topic selection stage. Selections were guided by two factors: i) surprise and/or ii) ease of inference of the relationships. The surprise factor included strange combinations of words for which the participants could not make sense. The ease of inference of the potential relationships suggests that participants entered the evaluation of the literature phase with a preconception of what they were looking for.

Interactions with the system also included those with the literature panels, i.e. the left and right panels displaying the related literatures. Participants initially inspected the literature on the left panel more often than that in the right panel, suggesting that they felt inclined to analyse the literature "closer" to their area of research first. Eventually, the interactions diverged towards the literature in the right panel. As participants felt more comfortable with the literature outside their research area, these interactions increased in frequency.

Implications

It was observed that the interactions with the panels containing the literatures followed a pattern of slow drifting towards the literature pertinent to the related topic. It may be that, as initially users focused on the literature closer to their research area, the option to hide either panel may be a useful one. By being able to hide either panel, users could fully concentrate on the literature presented. Moreover, offering colour cues to aid users identify which portions of the document surrogates correspond to which topic may aid the transition to the literature retrieved by the related topic.

6.6 Session Visualisation

Three search sessions were visualised and analysed. One session was confirmed to be an outlier as the corresponding graph showed that the participant had spent the entire search session reading documents out loud. The other two sessions were deemed to be more representative of what a "normal" session looked like. From the visualisations it was suggested that participant 2 exhibited a generally affective behaviour as *affectiveness* was a frequently expressed criterion. Furthermore, the participant was "careful" when judging the relevance of the information presented. This may correspond to the participant's research experience level (research student). *Tangibility* was not only globally, in the participant's relevance criteria profile, frequent but also used in most relevance judgement processes. Participant 19, on the other hand, mentioned *depth/scope/specificity* more often than any other criteria. The criterion was also present in most relevance judgement processes; processes which were mostly of low complexity suggesting that the participant engaged in a highly interactive session. Repetition of uses of criteria within relevance judgement processes were observed in both sessions. These may be due to the many manifestations of any one relevance criterion, e.g. mentions of document length, information breadth, and the amount of information contained are all to be encoded as *depth/scope/specificity*.

Implications

The technique developed in Section 3.7.5 results in graphs that offer sequential information needed to analyse whether criteria are used in different portions of the search session and are a good complement to relevance criteria profiles. However, as explained in Section 3.7.5, there are drawbacks to the visualisation technique used to display the search sessions. The reasons behind these shortcomings are not all particular to the technique itself as they include incomplete taxonomies (of both interactions and relevance criteria), the assumptions behind the segmentation technique (that relevance judgement processes are delimited by interactions) and the partial gathering of data collection during the study. Using the beginning of the search session of participant 2, depicted in Figure 6.2, we exemplify these shortcomings and propose solutions to them.

Figure 6.2: Participant 2 – First uses of relevance criteria

The session begins with an expression of recognition of the contents presented (pink box denoting an expression of *content novelty*). The novelty of the contents analysed is in relation to the participant remembering that in the set of documents retrieved during the first session *"... there was something about networking or something [...]"* and that *"[...] that's where the other keywords come from ..."*. The participant then interacts with the system. This interaction, even though encoded as *N*, has the user selecting a top related topic from the presented list at the beginning of the session. This is expressed as *"... I will try the top "logic speed processor" ..."*. After this interaction, the user engages in a relevance judgement process of complexity 3. This process begins with an affective expression; the participant utters that *"... [the topic] could be interesting ..."*. The list of intermediate topics is assessed next; the participant expresses that there are *"... things to do with circuits and ... logic ..."*. The judgement process is ended by an expression that *"... actually there's a few things I've no idea what they are to do with anything, but never mind ..."* (encoded as *ability to understand*.)

As described, the participant is using relevance criteria to assess the relevance of the information presented by the system; in this case, the related and intermediate topics. In the graph, no distinction is made on whether the relevance criteria (and hence the relevance judgement processes) correspond to the evaluation of topics, document contents, document surrogates, etc. Everything is information and all information is assessed. Because no distinction is made, certain types of analysis may become more difficult to perform on the plot as is. Should one want to analyse the relevance judgements on the contents of the documents inspected, for instance, one would have to resort to the transcription of the verbal reports to decide whether the relevance judgement process corresponds to the contents of a document or to other information objects. To overcome this limitation, the graph should be annotated to indicate what type of information is being assessed. To

this extent, and to annotate the graph accordingly, extra information should be collected during the search sessions. Click information, for instance, can indicate when the user has opened (or closed) a document. Hence the graph can be annotated to indicate whether the information being judged is coming from a document or some other part of the system. Distinguishing which information is being assessed is important as the type of relevance judged may differ from information object to information object. When participants evaluate the relevance of of the related (or intermediate) topics, for instance, they are actually evaluating the potential relevance of the documents that will be retrieved when the topic is submitted as a query. As expressed by a participant: *"... (reads/mumbles) "video games..." that might be relevant, that's right hand side, which is presumably meant to lead me into new areas which might be of interest ..."* Harter (1992) referred to this potential relevance as *weak relevance*.

Relevance judgements (either binary –positive or negative– or graded) can also be included into the graph. Depending on how the researcher wishes to do it, a mechanism for the user to provide feedback on whether the judgement has been positive or negative can be built into the system. Judgements can then be incorporated into the graphs. For instance, a button to "print" documents that the user wishes to keep for later reference could be added to a document viewing window. Sequences of the form *open document*→*print*→*close document*, may be interpreted as positive judgements, i.e. the user wants to print the document for later reference. Sequences with the *print document* step missing, may then be interpreted as implicit negative relevance judgements.

The taxonomy used in this study to encode interaction utterances is simple: an interaction with the system is either a *Navigation* interaction or the act of *Reading out loud*. However, as exemplified, interactions, as expressed by the participants, can be more varied. Discriminating these, and assigning those that wish to be analysed further their own code, is necessary as otherwise the analysis of the interactions using the graph may become too difficult. In this respect, the shortcoming is not one particular of the visualisation technique itself but of the study design. Extending the encoding should suffice to make the analysis of interactions patterns possible. Reliably recording interactions is

also of extreme importance. Resorting to verbal reports of these interactions is far from optimal. As described in Section 5.3.3, interactions with the left and right panel were not always verbally reported. Interaction data with these panels, for instance, should be extracted from the interaction logs.

Another shortcoming is that of alignment. Because the graph includes information about order only, sessions (or portions of) cannot be directly compared, i.e. one cannot compare the first quarter of one participant's session with that of another. To be able to do so, the timeline should be annotated with time and the stacks and interactions placed accordingly. Having time annotated sessions, would mean one could draw the graphs in parallel and compare, for instance, the session lengths or session subsets.

Figure 6.3 depicts a mock up of the first 3 minutes of two search sessions. The sessions have been drawn in parallel and aligned on the time axis for better comparison. In addition, information about document relevance judgements was added and the taxonomy with which interactions were encoded extended.

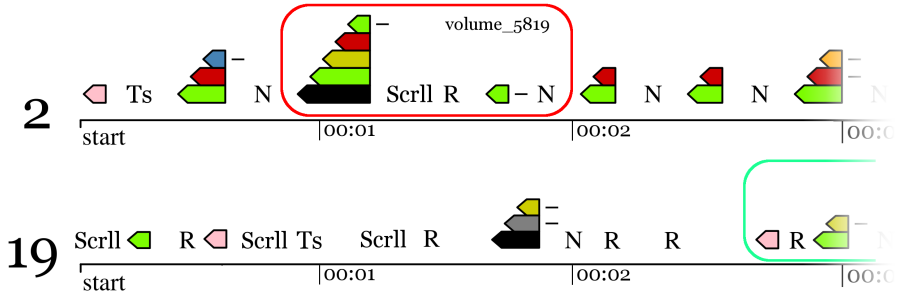

Figure 6.3: Time annotated graphs, in parallel for better comparison.

In the figure we see that participant 2 (top graph) immediately engages in assessing the information presented. We also know that this information is not coming from a document; this narrows down the possibilities to the related or the intermediate topics (as it is the beginning of the session). Participant 19, on the other hand, starts navigating the collection and only selects a topic around minute 1, point at which participant 2 begins judging the first document of the session. The red box around the stacks of criteria and interactions indicates that the participant found the document irrelevant. The document identifier is also present on the top right of the box. Participant 2 judges the document

to be irrelevant around minute 2 and the participant continues searching for information. Even by this time, participant 19 has not judged a single document yet. The first relevance judgement of a document by participant 19 comes around 2:35 and we know the participant has found it relevant because of the green box around the relevance criteria stacks.

The taxonomy of interactions is extended to include actions of scrolling (whether the document contents or the screen) and actions of selections of topics. These are encoded as *Scrll* and *Ts*. The full taxonomy allows for the encoding of scrolling of contents (*Scrll*), selection of topics (either related or intermediate – *Ts*), acts of reading out loud (*R*), and, as a general placeholder, navigations (*N*).

6.7 Known Limitations

Results from this study are to be interpreted with care. Firstly, results from the talk-aloud reports are to be interpreted as indicative rather than definitive. Verbal reports indicate only a subset of the cognitive processes in which the researchers incurred, hence they provide a partial view of reality. Furthermore, as there were variations in terms of volume and coverage of the observed utterances, absences of mentions of criteria cannot be interpreted as not being considered during relevance judgements by participants. Secondly, the interaction analysis was also performed on the verbal reports. The taxonomy used to encode interactions utterances is limited to two events: read out loud and navigation, being *navigation* a placeholder for any interaction event that is not the act of reading out loud. Clearly this limited the analysis and a more varied and comprehensive taxonomy should be used in future studies. However there is also the issue of ambiguity. Certain interaction utterances depended on the interaction history, e.g. *"...I'll check out the other panel..."*, and this issue cannot be resolved with an extended taxonomy. Hence, interactions should actually be analysed from log data, for instance, gathered through hooking mechanisms embedded in the user interface of the system under test. Thirdly, the user groups, although varied, were small and skewed. Too little data was gathered to reach any significant conclusions. Moreover, the skewness of the distribution of the participants made the comparison analyses difficult and the results very tentative. Finally, the time imposition

on the search sessions may have influenced the outcome of the study. Researchers under pressure may have modified their behaviour in terms of both relevance judgements and interactions to comply with the time limits imposed.

6.8 Summary

In this chapter the implications of the results presented in chapters 4 and 5 were summarised and discussed. Firstly, it was suggested that relevance in the context of LBD is multi-dimensional. The direct implication of this is that relevance cannot be treated as a binary notion, whether in an system-driven evaluation method or in a user-centred evaluation method. Additionally, there were criteria that were very frequent and, if operationalised correctly, could be embedded in systems in an attempt to improve their performance. Conversely, they could be used as indicators of system performance during evaluation experiments. Secondly, it was suggested that relevance criteria profiles and the ability to compare may be useful tools in improving or extending collaborative features of IR systems. Having more detailed information regarding why certain judgements have been reached may prove useful in recommender systems which are traditionally based on binary co-occurrences of events. Thirdly, the observation that 1-criterion relevance judgement processes are the most frequently used may be due to that users may have been interested in covering as much information as possible and to do so they attempted to quickly judge the presented information and continue browsing. Although not rare, more complex processes were observed in inversely proportional frequency. This would suggest that the users may have approached the solution of the task in two stages: an initial recall-oriented stage in which they aimed at quickly judging the information presented followed by a precision-oriented stage in which they more meticulously analysed the retrieved information. Systems may then be designed to aid this searching activity by supporting the two stages.

Because *document novelty* was frequently used in 1-criterion judgements, regardless of polarity, it may be a good candidate for embedding an operational version of the criterion in ranking algorithms or used as a performance indicator when conducting LBD system

evaluations. Although obvious in the context of LBD, it was suggested that the relevance of documents depends on the context in which they are judged. This observation stemmed from the observations of uses of the dominance and the chain rule.

Other observations resulted in suggestions for improving LBD user interfaces. It was suggested that the number of times a document has been read, retrieved, and the context, for instance which topics lead to the documents, in which these events happened should be provided. Regarding the concordance of the left and right panels, perhaps visual cues indicating the matching documents would provide an indication of this concordance and would support users in planning their approach to assessing the literature bodies. Finally, to accommodate for the drifting behaviour observed in the interactions between panels, hiding either panel would allow users to focus on one literature at a time. Complementing this feature with visual cues indicating topic (in terms of the starting, intermediate and related topic) matches both in the document surrogates and the document reader window may provide extra context to users when reading the documentation.

Chapter 7

Summary and Conclusions

During the preceding chapters we described the steps followed in the investigation of the manifestations of relevance in the context of LBD. In Chapter 2 the reader was introduced to the problem of LBD. Initially we described the original discoveries made by Swanson. These discoveries were made manually, however they resulted in a recommended set of steps to follow when attempting to make new discoveries. The common LBD search models were described next. The open model was described as an exploratory model aimed at stimulating the researchers's intellect and suggest potentially fruitful relationships. The closed model was described as a filtering model in which the researchers approach the search task with a potential relationship in mind and set out to find if there is information that supports it. Next, the most representative works in LBD were described and analysed. While most techniques rely, initially, on word co-occurrence statistics, it was seen that newer approaches use specialised metadata for topic modelling, filtering and ranking. A broad overview of research in IR was offered next. The system-driven tradition of evaluation was described and the concept of a test collection explained. Following, Borlund's approach Borlund & Ingwersen (2000) was suggested as a potential set of guidelines for designing and conducting experiments to evaluate LBD systems.

In Chapter 3 we described the design of the study and the components involved. The system and the search sessions were described first. The population was described next; a group of research scientists coming from three different disciplines and three different

research experience backgrounds. For each discipline, a collection was crawled and indexed. Three simulated work task situations were created: one for research students, one for researchers and one for senior researchers. The method for gathering the data, both in written and verbal form, was explained; it was the verbal reports which allowed for the observation of the multi-dimensional nature of relevance. How data was to be analysed was explained next. The encoding protocol and the relevance criteria profiles were explained. Further, the technique for sessions segmentation and visualisation was developed in this chapter.

In Chapters 4 and 5 the results were presented. Researchers use a variety of criteria and in different frequencies when evaluating the relevance of information in an LBD context. For instance, we observed that participants from the School of Computing frequently mentioned *tangibility* while participants from the Information Management group mentioned *depth/scope/specificity* more often. Moreover, they engage, initially, in quick judgements with aims, potentially, of covering as much as possible of the presented information space. When interacting with the system, they seem to try and stay "close" to their area of research initially while they are still learning about the related domains, however they eventually diverge towards the related topic.

The results were summarised and the implications discussed in Chapter 6. Several suggestions regarding the inclusion of operational versions of the criteria observed into both LBD user interfaces and ranking algorithms were made. These included, amongst others, adding visual cues denoting *document novelty* and *concordance* between literature panels and embedding operational estimations of *tangibility* and *depth/scope/specificity* into ranking algorithms.

7.1 The Three Research Questions

The main objective of the dissertation is to investigate the relevance criteria used by research scientists as they solve an LBD type of search. To do so, a study, designed as described in Chapter 3 to take these components into account, was conducted. The main outcome of the study is the observation of the multidimensional and dynamic nature of

relevance and the interaction patterns in which users incurred. In Chapter 1 we presented the research question that motivated this dissertation:

> *What relevance criteria, if any, do researchers use when assessing the relevance of related, although potentially outside their research field, information?*

Researchers do use different relevance criteria, and in different frequencies, when evaluating the relevance of the information in an LBD context. We observed researchers as they interacted with an LBD system and gathered information about their mental processes as they judged the relevance of the information presented. After analysing this information we were able to conclude that researchers use a variety of criteria and in different proportions. The list of criteria used was presented in Chapter 3, Section 3.7.3. Because relevance is dynamic and multi-dimensional in the context of LBD, evaluation methods have to take this into consideration.

Two additional questions, related to the main question investigated in this dissertation, were introduced:

> *Do researchers from different disciplines use different criteria and/or in different frequencies?*

> *Does research experience affect the relevance criteria, and their frequency of use, used in these relevance judgements?*

Tentatively, the answer is "yes" to both questions. However the real answer may be more complex than that. Although researchers from different disciplines used the same array of criteria, the frequencies of the uses are different. The relevance criteria profiles for each discipline showed that different criteria were expressed with more frequency, suggesting that researchers from different disciplines exhibit certain preferences when evaluating the presented information. There is also evidence that research experience also affected the frequencies of these uses. The research experience background relevance criteria profiles showed differences in terms of frequencies of expression of criteria. However, because

there may be naturally emerging clusters in terms of relevance criteria profile similarities and none correspond to either discipline nor research experience background, we suggest that other unobserved factors are affecting these frequencies.

7.2 Designing and Evaluating LBD Systems

The results of the study led to suggestions for both the design and testing of LBD systems. It is our belief that LBD systems should not be tested using methods that resemble the system-driven tradition of IR research as the cognitive processes that are involved during LBD searches have an effect on to how relevance is judged in this scenario. We described and analysed how previous attempts at evaluating LBD systems approach the problem and pointed out that there are several places where researchers have taken decisions on behalf of the end-users, e.g. how to map the concepts, as represented by their system, to the golden topics from Swanson's original discoveries. This makes the results obtained with this method hard to compare as different researchers might take different decisions on behalf of the end-users. Additionally, we observed that certain rules affect relevance judgements, e.g. the re-occurrence rule, and that these depend heavily on the user interactions with the system, hence interaction patterns are to be considered to be part of the evaluation method.

Several by-products resulted from our investigation. As we used some of the experimental components of the framework proposed by Borlund (2003b), we have an initial attempt at modelling simulated work tasks in the context of LBD. We have discovered that it is possible, although the crafting must be approached with utmost care as to not introduce bias inadvertently. Although we paid a great deal of attention to the crafting of the simulated work tasks, we might have introduced bias regarding the currency of the information to be judged in at least one of the tasks. Had we not decided to investigate whether different groups expressed criteria in different proportions, we might have used a single work task situation and hence retained more control over the experimental settings.

We have also developed three tools for the analysis of data gathered in future studies of this nature. Firstly, we developed the concept of relevance criteria profiles. These

represent the criteria expressed by a participant or group during a time window (in our case an entire search session) and provide a global view of the proportions include therein. These profiles can be compared using standard divergence measures. These measures can eventually be used as basis for analysis, e.g. the divergence measures between profiles can be used as input values for the affinity propagation clustering (Frey & Dueck 2007). In our case we resorted to plotting these divergence values as a heatmap and manually inspecting these to see if any natural clusters emerged. Secondly, we developed the notion of relevance judgement process. These are smaller groups of relevance criteria as delimited by interactions. These were used during our data analysis to investigate the different uses of selection rules as defined by Wang & White (1999). Additionally, these processes can be used to analyse the behaviour of a user in terms of judging information along the dimensions of variety of criteria used in any one process, average complexity of the processes and more. Finally, we developed and provided two examples of a custom session visualisation tools. Using this tool we confirmed our observation that a participant in our study was an outlier in terms of the data gathered during his session. This was made more explicit when visualising the session as it was clear that the participant had spent most of the session reading out loud and that very few relevance criteria had been mentioned. Additionally, we presented an example of how the session visualisation tool can be used to further analyse the behaviour of participants in terms of their relevance criteria used. We plotted two example sessions and each was analysed in terms of relevance judgement process complexity, process variety, cadence of the judgements and most commonly preferred criteria.

Appendices

Appendix A

Forms

Participant _____

1) Profession: _____

2) Position: _____

3) Research field: _____

4) Research topic: _____

5) On a scale from 1 to 5, 1 being strongly disagree and 5 strongly agree:

1. I am confident in my knowledge of papers in my research field	1	2	3	4	5
2. I am confident in my knowledge of research out with my research field	1	2	3	4	5

6) Please state your preferred methodology for searching for literature (you can choose more than one):

1. Online search engine, e.g. Google, Yahoo, etc.
2. University library
3. Ask librarians
4. Personal references, e.g. colleague, friend, etc.
5. Journals
6. Conference proceedings
7. Other (please state): _____

7) On a scale from 1 to 5, 1 being strongly disagree and 5 strongly agree:

1. I am proficient using search engines	1	2	3	4	5
2. I have recently used a search engine for literature searches	1	2	3	4	5

Figure A.1: Form used to capture the demographics.

Dear participant,

I'd like to ask you to search for documents that describe your area of research, or an aspect of it. Whenever you think you have found one, please write down the document ID (located at the top of the viewing window) on the provided sheet. The purpose of this search is so that the system under test can then suggest topics that might be related to your area of research for you to further investigate.

1)

2)

3)

4)

5)

Figure A.2: Form used during the first search session.

Dear participant,

Thank you very much for taking part of this experiment. Now that you have done part 1 of it, it's time to do part 2. In part 2, you are given a *simulated work task*. Briefly, a simulated work task is a description of a task you should perform. Below is the description:

Simulated work situation 1:

At a supervisory meeting you received constructive criticism concerning the breadth of your literature review . Even though the work you've been doing is very good, your supervisor feels it is a bit too specific/constrained. Your supervisor suggested you look for connections between your research and other research areas, e.g. other areas which have techniques or ideas you might use in your research or areas where your research might contribute. Your supervisor suggested you identify these potential areas as well as the pertinent literature so you can discuss them together in your next meeting.

Indicative request:

Find, for instance, about a technique employed/developed in another area that might share commonalities with (or could be applied to) your research, e.g. an algorithm that could be adapted, a process that could be abstracted/refined, etc.

You will have a time limit of 1 hour to investigate 3 of the possibly related topics you will see on the screen. You can stop at any time though.

You will be required to think aloud as you investigate the potential relations.

Figure A.3: Research student simulated work task situation.

Dear participant,

Thank you very much for taking part of this experiment. Now that you have done part 1 of it, it's time to do part 2. In part 2, you are given a *simulated work task*. Briefly, a simulated work task is a description of a task you should perform. Below is the description:

Simulated work situation:

You are in the process of writing a grant proposal. A senior colleague as suggested you carefully write about the impact your proposed research might have on related fields or where related fields have ideas or approaches that you might exploit. Your colleague has advised that this might improve your chances of getting the proposal funded.

Indicative request:

Find, for instance, about a technique employed/developed in another area that might share commonalities with (or could be applied to) your research, e.g. an algorithm that could be adapted, a process that could be abstracted/refined, etc.

You will have a time limit of 1 hour to investigate 3 of the possibly related topics you will see on the screen. You can stop at any time though.

You will be required to think aloud as you investigate the potential relations.

Figure A.4: Researcher simulated work task situation.

Dear participant,

Thank you very much for taking part of this experiment. Now that you have done part 1 of it, it's time to do part 2. In part 2, you are given a *simulated work task*. Briefly, a simulated work task is a description of a task you should perform. Below is the description:

Simulated work situation:

You have been invited to deliver a keynote speech at a very prestigious conference in your research field. The organisers have kindly asked you to focus your speech on the future directions and implications of advances in your research field, especially on those fields outside your own. A senior colleague suggested that, in order to prepare your speech, you look for connections between your research and other areas of research, e.g. other areas which have techniques or ideas you might use in your research or areas where your research might contribute.

Indicative request:

Find, for instance, about a technique employed/developed in another area that might share commonalities with (or could be applied to) your research, e.g. an algorithm that could be adapted, a process that could be abstracted/refined, etc.

You will have a time limit of 1 hour to investigate 3 of the possibly related topics you will see on the screen. You can stop at any time though.

You will be required to think aloud as you investigate the potential relations.

Figure A.5: Senior researcher simulated work task situation.

Participant _____

Starting topic:
Description:

Topic:
Description:

Documents selected	Topic supported

Figure A.6: Form used to capture the documents selected as well as the topics they support.

Participant _____

1) You have completed the final session of the experiment. Please answer the following questions to help us understand how your search experience was.

How varied were the suggested topics?	Very varied (e.g. every topic was different from each other)	Varied (e.g. most topics were unique but some overlapped with each other)	Somewhat varied (e.g. there were several topics that overlapped with each other)	Not varied enough (e.g. most topics overlapped with each other)	Not varied at all (e.g. they were all the same topic)
How valid were the suggested relations between the suggested topics and your research area?	Very valid (e.g. all connection we valid and supported by the literature)	Valid (e.g. most connections were valid and supported by the literature)	Somewhat valid (e.g. some connections were valid and supported by the literature)	Somewhat not valid (e.g. most connections weren't valid or weren't supported by the literature)	Invalid (e.g. I didn't find any valid connections)
Would you pursue any of the connections found?	Absolutely (e.g. I would pursue all the connections found)	Yes (e.g. Yes, I would pursue most of the connections found)	Maybe (e.g. I would perhaps pursue some of the connections found)	Not really (e.g. I'm not sure if I would pursue most connections found)	No (e.g. I wouldn't pursue any of the connections found)
How much knowledge have you gained from the connections?	Lots of knowledge (e.g. I have understood the relation between my research field and all the other topics suggested)	Plenty knowledge (e.g. I have understood the relation between my research field and most of the suggested topics)	Some knowledge (e.g. I have understood the relation between my research field and some of the suggested topics)	Not much knowledge (e.g. I have understood the relation between my research field and a few of the suggested topics)	None (e.g. I haven't understood any of the relations between my research field and any of the suggested topics)
How much knowledge have you gained from the documents found?	Lots of knowledge (e.g. I have understood the relation between my research field and all the other topics suggested)	Plenty knowledge (e.g. I have understood the relation between my research field and most of the suggested topics)	Some knowledge (e.g. I have understood the relation between my research field and some of the suggested topics)	Not much knowledge (e.g. I have understood the relation between my research field and a few of the suggested topics)	None (e.g. I haven't understood any of the relations between my research field and any of the suggested topics)
How novel were the documents you selected?	Very novel (e.g. I didn't know of any of the documents found)	Novel (e.g. most documents were new to me)	Somewhat novel (e.g. some document were new to me)	Not very novel (e.g. most documents were not new to me)	Not novel (e.g. I already knew about those documents)

Figure A.7: Form used at the end of the second search session (page 1).

Were the documents you selected from your own area of research?	Absolutely (e.g. all documents selected are from my own research field)	Yes (e.g. most documents are from my own research field)	Maybe (e.g. some documents are from my own research field)	Not really (e.g. a few documents are from my own research field but most are not)	No (e.g. no document is from my own research field)

2) Do you have any final comments?

Figure A.8: Form used at the end of the second search session (page 2).

Appendix B

Publications

Portions of this thesis have been published in different forums (ordered chronologically):

- Ulises Cerviño Beresi, Yunhyong Kim, Ian Ruthven and Dawei Song, **Why did you pick that? Visualising relevance criteria in exploratory search** *to appear in International Journal on Digital Libraries, Special Issue on ECDL 2010*

- Ulises Cerviño Beresi, Yunhyong Kim, Mark Baillie, Ian Ruthven and Dawei Song, **Relevance in Technicolor** in *Research and Advanced Technology for Digital Libraries (ECDL 2010).* Glasgow, UK, September 2010. Springer Verlag Lecture Notes in Computer Science.,

- Ulises Cerviño Beresi, Yunhyong Kim, Mark Baillie, Ian Ruthven and Dawei Song, **Colouring the Dimensions of Relevance**, in 32^{nd} *European Conference on Information Retrieval (ECIR 2010).* Milton Keynes, UK, March 2010. Springer Verlag Lecture Notes in Computer Science.,

- Ulises Cerviño Beresi, Mark Baillie and Ian Ruthven, **Towards the Evaluation of Literature Based Discovery**, in *Workshop on novel methodologies for evaluation in information retrieval , ECIR'08*

- Ulises Cerviño Beresi, Narrowing gaps in Science, in *BCS IRSG Symposium: Future Directions in Information Access 2007 (FDIA) held in conjunction with ESSIR 2007*

Bibliography

Aronson, A. (2001). Effective mapping of biomedical text to the umls metathesaurus: the metamap program, *Proceedings of AMIA Symposium* pp. 17–21.

Ashworth, W. (1966). Librarianship and other disciplines, *Midlands Branch A.G.M.* .

Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*, 1st edn, Addison Wesley.

Barry, C. (1993). *The identification of user criteria of relevance and document characteristics: Beyond the topical approach to information retrieval*, PhD thesis.

Barry, C. (1994). User-defined relevance criteria: an exploratory study, *Journal of the American Society for Information Science* **45**(3): 149–159.

Barry, C. & Schamber, L. (1998). Users'criteria for relevance evaluation: A cross-situational comparison, *Information Processing and Management* **34**(2-3): 219–236.

Belkin, N. & Croft, W. (1987). Retrieval techniques, *Annual review of information science and technology* **22**: 109–145.

Belkin, N., Oddy, R. & Brooks, H. (1982). Anomalous states of knowledge as a basis for information retrieval, *Journal of documentation* **38**(2): 61–71.

Blair, D. & Kimbrough, S. (2002). Exemplary documents: a foundation for information retrieval design, *Information Processing and Management* **38**(3): 363–379.

Blei, D., Ng, A. & Jordan, M. (2003). Latent dirichlet allocation, *Journal of machine Learning Research* pp. 993–1022.

Borlund, P. (2000). *Evaluation of interactive information retrieval systems*, PhD thesis.

Borlund, P. (2003a). The concept of relevance in IR, *Journal of the American Society for Information Science and Technology* **54**(10): 913–925.

Borlund, P. (2003b). The IIR evaluation model: a framework for evaluation of interactive information retrieval systems, *Information Research* **8**(3): 8–3.

Borlund, P. & Ingwersen, P. (1997). The development of a method for the evaluation of interactive information retrieval systems, *JOURNAL OF DOCUMENTATION* **53**: 225–250.

Borlund, P. & Ingwersen, P. (2000). Experimental components for the evaluation of interactive information retrieval systems, *Journal of Documentation* **56**(1): 71–90.

Cerviño Beresi, U., Baillie, M. & Ruthven, I. (2008). Towards the evaluation of literature based discovery, *Workshop on novel methodologies for evaluation in information retrieval* .

Cleverdon, C. W., Mills, J. & Keen, E. (1966). Factors Determining the Performance of Indexing Systems, Vol. 1: Design, Vol. 2: Test Results, *Aslib Cranfield Research Project, Cranfield, England* .

Cool, C., Belkin, N., Frieder, O. & Kantor, P. (1993). Characteristics of text affecting relevance judgments, *NATIONAL ONLINE MEETING*, Vol. 14, pp. 77–77.

Cooper, W. (1971). A definition of relevance for information retrieval* 1, *Information storage and retrieval* **7**(1): 19–37.

Cory, K. (1997). Discovering Hidden Analogies in an Online Humanities Database, *Computers and the Humanities* **31**(1): 1–12.

Cuadra, C. & Katter, R. (1967). Experimental studies of relevance judgments: Final report. vol. i: Project summary (nsf report no. tm-3520/001/00). santa monica, *CA: System Development Corp* .

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.

Doyle, L. (1961). Semantic road maps for literature searchers, *Journal of the ACM* **8**(4): 553–578.

Ericsson, K. & Simon, H. (1993). *Protocol analysis: verbal reports as data*, MIT Press Cambridge, MA.

Frey, B. & Dueck, D. (2007). Clustering by passing messages between data points, *science* **315**(5814): 972.

Gordon, M. & Dumais, S. (1998). Using latent semantic indexing for literature based discovery, *Journal of the American Society for Information Science and Technology* **49**(8): 674–685.

Gordon, M. & Lindsay, R. (1996). Toward discovery support systems: a replication, re-examination, and extension of swanson's work on literature-based discovery of a connection between raynaud's and fish oil, *Journal of the American Society for Information Science and Technology* **47**(2): 116–128.

Gordon, M., Lindsay, R. & Fan, W. (2002). Literature-based discovery on the world wide web, *ACM Transactions on Internet Technology* **2**(4): 261–275.

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook*, Vol. 5, Cambridge University Press.

Green, R. (1995). Topical relevance relationships. i. why topic matching fails, *Journal of the American Society for Information Science* **46**(9): 646–653.

Halevy, A., Norvig, P. & Pereira, F. (2009). The unreasonable effectiveness of data, *IEEE Intelligent Systems* **24**: 8–12.

Harman, D. K. (1997). The trec conferences, pp. 247–256.

Harter, S. (1992). Psychological relevance and information science, *Journal of the American Society for Information Science and Technology* **43**(9): 602–615.

Hristovski, D., Peterlin, B., Mitchell, J. & Humphrey, S. (2003). Improving literature based discovery support by genetic knowledge integration, *Studies in Health Technology and Informatics* **95**: 68–73.

Hristovski, D., Peterlin, B., Mitchell, J. & Humphrey, S. (2005). Using literature-based discovery to identify disease candidate genes, *International Journal of Medical Informatics* **74**(2-4): 289–298.

Hristovski, D., Stare, J., Peterlin, B. & Dzeroski, S. (2001). Supporting discovery in medicine by association rule mining in Medline and UMLS, *Studies in Health Technology and Informatics* **10**(Pt 2): 1344–8.

Hurvich, L. M. & Jameson, D. (1957). An opponent-process theory of color vision, *Psychological Review* **64**: 384–404.

Ingwersen, P. (1988). Towards a new research paradigm in information retrieval, *Knowledge engineering: expert systems and information retrieval*, Taylor Graham Publishing, pp. 150–168.

Ingwersen, P. (1992). *Information retrieval interaction*, London, Taylor Graham Publishing.

Ingwersen, P. (1993). Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory, *Journal of documentation* **52**(1): 3–50.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory, *Journal of documentation* **52**: 3–50.

Ingwersen, P. & Järvelin, K. (2007). On the holistic cognitive theory for information retrieval, *Studies in Theory of Information Retrieval. Budapest: Foundation for Information Society* **2007**: 135–147.

Kekäläinen, J. & Järvelin, K. (2002). Evaluating information retrieval systems under the challenges of interaction and multidimensional dynamic relevance, *Emerging frameworks and methods: CoLIS 4: proceedings of the Fourth International Conference on Con-*

ceptions of Library and Information Science, Seattle, WA, USA, July 21-25, 2002, Libraries Unltd Inc, p. 253.

Kuhlthau, C. (2004). Seeking meaning: A process approach to library and information services, *Westport, CT* .

Kullback, S. & Leibler, R. A. (1951). On information and sufficiency, *Annals of Mathematical Statistics* **22**: 79–86.

Lafferty, J. & Zhai, C. (2001). Document language models, query models, and risk minimization for information retrieval, *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* pp. 111–119.

Lin, J. (1991a). Divergence measures based on the shannon entropy, *IEEE Transactions on Information theory* **37**: 145–151.

Lin, J. (1991b). Divergence measures based on the Shannon entropy, *Information Theory, IEEE Transactions on* **37**(1): 145–151.

Lindsay, R. & Gordon, M. (1999). Literature-based discovery by lexical statistics, *Journal of the American Society for Information Science and Technology* **50**(7): 574–587.

Lowe, H. & Barnett, G. (1994). Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches., *JAMA: the journal of the American Medical Association* **271**(14): 1103.

Mizzaro, S. (1998). How many relevances in information retrieval?, *Interacting with computers* **10**(3): 303–320.

Ponte, J. & Croft, W. (1998). A language modeling approach to information retrieval, *Research and Development in Information Retrieval*, pp. 275–281.

Pratt, W. & Yetisgen-Yildiz, M. (2003). Litlinker: capturing connections across the biomedical literature, *K-CAP '03: Proceedings of the 2nd international conference on Knowledge capture*, ACM, New York, NY, USA, pp. 105–112.

Pressley, M. & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading.*, Lawrence Erlbaum Associates, Inc.

R. Agrawal, H., Mannila, Srikant, R., Toivonen, H. & Verkamo, A. (1996). Fast discovery of association rules, *Advances in knowledge discovery and data mining* **12**: 307–328.

Rees, A. & Schultz, D. (1967). A field experimental approach to the study of relevance assessments in relation to document searching. final report to the national science foundation. volume ii, appendices.

Robertson, S. E. & Hancock-Beaulieu, M. M. (1992). On the evaluation of ir systems, *Information Processing Management* **28**(4): 457–466.

Rowe, H. (1985). *Problem solving and intelligence*, Lawrence Erlbaum Associates.

Salton, G. & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval* 1, *Information Processing & Management* **24**(5): 513–523.

Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion in information science, *Journal of the American Society for Information Science* **26**(6): 321–343.

Saracevic, T. (1996). Relevance reconsidered, *Proceedings of the 2nd Conference on Conceptions of Library and Information Science (COLIS 2), Copenhagen, Denmark* pp. 201–218.

Saracevic, T. (2006). Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Part II, *Advances in Librarianship* **30**: 69.

Saracevic, T. (2007). Relevance: A review of the literature and a framework for thinking on the notion in information science. part iii: Behavior and effects of relevance, *Journal of the American Society for Information Science and Technology* **58**(13): 2126–2144.

Schamber, L. (1991). *Users'Criteria for Evaluation in Multimedia Information Seeking and Use Situations*, PhD thesis, Syracuse University.

Schamber, L. (1994). Relevance and information behavior., *Annual review of information science and technology (ARIST)* **29**: 3–48.

Schamber, L., Eisenberg, M. & Nilan, M. (1990). A re-examination of relevance: toward a dynamic, situational definition, *Information Processing & Management* **26**(6): 755–776.

Skeels, M. M., Henning, K., Yildiz, M. Y. & Pratt, W. (2005). Interaction design for literature-based discovery, *CHI '05: extended abstracts on Human factors in computing systems*, ACM, New York, NY, USA, pp. 1785–1788.

Smalheiser, N. & Swanson, D. (1996a). Indomethacin and Alzheimer's disease, *Neurology* **46**(2).

Smalheiser, N. & Swanson, D. (1996b). Linking estrogen to Alzheimer's disease: an informatics approach, *Neurology* **47**(3): 809.

Smalheiser, N., Swanson, D. & Ross, B. (1998). Calcium-independent phospholipase A2 and schizophrenia, *Archives of General Psychiatry* **55**(8): 752.

Smalheiser, N., Torvik, V., Bischoff-Grethe, A., Burhans, L., Gabriel, M., Homayouni, R., Martone, A. K. M., Perkins, G., Price, D., Talk, A. & West, R. (2006). Collaborative development of the arrowsmith two node search interface designed for laboratory investigators, *Journal of Biomedical Discovery and Collaboration* **1**(1): 8.
URL: *http://www.j-biomed-discovery.com/content/1/1/8*

Spink, A., Wolfram, D., Jansen, M. & Saracevic, T. (2001). Searching the web: The public and their queries, *Journal of the American Society for Information Science and Technology* **52**(3): 226–234.

Srinivasan, P. (2004). Text mining: generating hypotheses from MEDLINE, *Journal of the American Society for Information Science and Technology* **55**(5): 396–413.

Swanson, D. (1977). Information retrieval as a trial-and-error process, *Library Quarterly* **47**(2): 128–148.

Swanson, D. (1986a). Fish oil, Raynaud's syndrome and undiscovered public knowlege, *Perspectives in Biology and Medicine* **30**: 7–18.

Swanson, D. (1986b). Subjective versus objective relevance in bibliographic retrieval systems, *The Library Quarterly* **56**(4): 389–398.

Swanson, D. (1986c). Undiscovered public knowledge, *The Library quarterly(Chicago, IL)* **56**(2): 103–118.

Swanson, D. (1988a). Historical note: Information retrieval and the future of an illusion, *Journal of the American Society for Information Science* **39**(2): 92–98.

Swanson, D. (1988b). Migraine and Magnesium: eleven neglected connections, *Perspectives in Biology and Medicine* **31**: 526–557.

Swanson, D. (1989). A second example of mutually isolated medical literatures related by implicit, unnoticed connections, *Journal of the American Society for Information Science and Technology* **40**(6): 432–435.

Swanson, D. (1990). Somatomedin C and arginine: Implicit connections between mutually-isolated literatures, *Perspectives in Biology and Medicine* **33**: 157–186.

Swanson, D. (1991). Complementary structures in disjoint science literatures, *SIGIR '91: Proceedings of the 14th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, ACM Press, New York, NY, USA, pp. 280–289.

Swanson, D. R. & Smalheiser, N. R. (1997a). An interactive system for finding complementary literatures: a stimulus to scientific discovery, *Artificial Intelligence* **91**(2): 183–203.

Swanson, D. & Smalheiser, N. (1997b). An Interactive System for Finding Complementary Literatures: A Stimulus to Scientific Discovery, *Artificial Intelligence* **91**(2): 183–203.

Taylor, K. L. & Dionne, J.-P. (2000). Accessing problem-solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing the university of manitoba university of ottawa, *Journal of Educational Psychology* .

Taylor, R. (1967). Question-Negotiation and Information-Seeking in Libraries.

Van Der Eijk, C., Van Mulligen, E., Kors, J., Mons, B. & Van Den Berg, J. (2004). Constructing an associative concept space for literature-based discovery, *Journal of the American Society for Information Science* **55**(5): 436–444.

van Someren, M., Barnard, Y. & Sandberg, J. (1994). *The think aloud method: a practical guide to modelling cognitive processes*, Academic Press.

Wang, P. & White, M. D. (1999). A cognitive model of document use during a research project. study ii. decisions at the reading and citing stages, *Journal of the American Society of Information Sciences* **50**(2): 98–114.

Ware, C. (1988). Color sequences for univariate maps: Theory, experiments and principles, *IEEE Computer Graphics and Applications* **8**(5): 41–49.

Weeber, M., Klein, H., Aronson, A., Mork, J., de Jong-van den Berg, L. & Vos, R. (2000). Text-based discovery in biomedicine: the architecture of the DAD-system, *Proceedings of the AMIA Symposium* **20**: 903–7.

Weeber, M., Klein, H., de Jong-van den Berg, L. & Vos, R. (2001). Using Concepts in Literature-Based Discovery: Simulating Swanson's Raynaud–Fish Oil and Migraine–Magnesium Discoveries, *Journal of the American Society for Information Science and Technology* **52**(7): 548–557.

Wilson, P. (1973). Situational relevance, *Information storage and retrieval* **9**(8): 457–471.

Wren, J., Bekeredjian, R., Stewart, J., Shohet, R. & Garner, H. (2004). Knowledge discovery by automated identification and ranking of implicit relationships, *Bioinformatics* **20**(3): 389–398.

Zhai, C. & Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval, *ACM Transactions on Information Systems* **22**(2): 179–214.