Tudhope, D., Taylor, C., & Beynon-Davies, P. (1995). Classification and Hypermedia. 6th ASIS SIG/CR Classification Research Workshop, 173-192. doi:10.7152/acro.v6i1.12668

Classification and Hypermedia

Douglas Tudhope, Carl Taylor, Paul Beynon-Davies

Computer Studies Department
University of Glamorgan
Pontypridd
Mid Glamorgan CF37 1DL
Wales, UK.

fax: 01443-482715
e-mail: dstudhope@glam.ac.uk

Abstract

This paper discusses a research prototype demonstrating an architecture for a hypermedia system, in which the index space has semantic relationships between terms. Information items in a social history museum application are indexed according to spatial, temporal, and subject-based classifications. The spatial and temporal indexes are inter-linked. Measures of semantic closeness over the different classification dimensions allow access to information, which might be missed by retrieval tools that require exact matching of terms used in a request. Illustrative examples of hypermedia navigation tools that make use of semantic closeness are described. A method of integrating different index dimensions drawing on research in numerical taxonomy is discussed. The paper goes on to discuss possibilities for a richer set of semantic primitives, that could profitably draw on concepts in the classification research literature.

Introduction

The characteristic that distinguishes hypermedia from multimedia applications is that the user can interactively select information items to view, and can thus proceed through an information space in a 'non-linear' fashion (as opposed to presenting information in a serial manner to the user). Hypermedia applications tend to be differentiated from conventional database and information retrieval applications by an emphasis on movement or navigation through the information space; the current position (or previous results) are more important as a starting place for the next retrieval operation than is usual with database queries.

ISSN: 2324-9773

Tudhope, D., Taylor, C., & Beynon-Davies, P. (1995). Classification and Hypermedia. 6th ASIS SIG/CR Classification Research Workshop, 173-192. doi:10.7152/acro.v6i1.12668

Current research in both hypermedia and database systems however, tends to blur this distinction (eg, Agosti, et al, 1995). Like traditional database applications, hypermedia systems may also incorporate query engines. The results of a query are not simply passed back to the user, however, and instead, the query engine returns information with hypertext links for interactive browsing.

Most recent hypermedia research systems maintain a distinction between information components and link information, which is kept in a separate store. For example, the Dexter hypertext reference model (Halasz and Schwartz, 1994) defines a Storage Layer that holds a network of hypertext components and links, which is separate from the 'Within Component' layer. Frisse and Cousins (1989) characterise such hypermedia systems as consisting of a separate index space and document space (of actual information items). Some hypermedia architectures (including Dexter) permit link endpoints to be computed dynamically at runtime. This allows link destinations to be the result of a computation based on current session context, rather than being fixed in advance by the author, and offers the possibility of integrating queries with standard navigation. An early example of a Computer Aided Learning system which stressed computed links was Strathtutor (Mayes, Kibby and Watson, 1988), designed to facilitate 'learning-by-browsing'.

The University of Glamorgan hypermedia architecture is currently based entirely on computed links. All modes of navigation ultimately result in a query on a link store, which is independent of the node content. Accordingly, each information item is indexed and the resulting set of index terms drives the different navigation tools, as opposed to items being directly linked together. The first stage of research demonstrated that it was possible to implement standard hypermedia navigation techniques (browsing, graphical maps, simple query, guided tours) in this way (Tudhope, et al, 1994).

This paper reports on more recent work investigating more advanced navigation tools. Some advanced hypermedia architectures maintain a separation between index space and document space, but the index space is flat. The tools described here take advantage of the fact that the index space is not flat, but structured; semantic relationships exist between terms. Thus the index space can be seen as a classification system (or schema). However, since it is a hypermedia system, the emphasis is on methods of interactively navigating/querying the information space, that make use of reasoning over the relationships in the classification schema.

We have concentrated on the application area of museum hypermedia systems, in collaboration with the Pontypridd Historical and Cultural Centre (PHCC) and other local museums. The current media base is a collection of approximately 120 historical photographs of Pontypridd from the archives of the PHCC, as well as a small amount of other text and audio local history information. A series of research prototypes have been implemented, and demonstrated to curators and historians, but not subjected to serious usability testing. The first prototype was completely written in HyperCard, while the current system is a HyperCard / Lisp hybrid, with Lisp used for reasoning over the relationships in the schema, and HyperCard used for presenting the information and interacting with the user. We are commencing work on a more efficient version of the architecture.

The content of hypermedia museum exhibits is often an arrangement of existing material in the museum's archives. In many situations, the archive material will have been indexed using standard classifications or controlled vocabularies, not designed specifically for a particular collection. These classifications are typically created by 'expert' authorities. Thus, although they may conform to current standards, the terms employed and the ways in which they are organised may be unfamiliar to the general public. However the current trend

ISSN: 2324-9773

Tudhope, D., Taylor, C., & Beynon-Davies, P. (1995). Classification and Hypermedia. 6th ASIS SIG/CR Classification Research Workshop, 173-192. doi:10.7152/acro.v6i1.12668

towards integration of collection management systems with those designed for public access and exhibition (Besser, 1991) means that there may be a desire to use complex indexing systems as a basis for access by the general public. Such indexing systems need to support users who may not be precise in requests for information. They may not even have specific information they wish to retrieve from the system, but may just wish to explore. Thus, there is a need for access tools that use knowledge about the structure of the index space to assist a user in exploring public information systems, without requiring exact matching of terms in queries, such as the automated thesaurus tools suggested by Molholt and Petersen (1993).

Three Fundamental Index Categories

In our prototype, information items are indexed according to three fundamental categories, a subject index, a temporal index and a spatial index (referred to as the three dimensions of indexing). For the subject index we adopted the Social History and Industrial Classification (SHIC, 1983) which contains approximately seven hundred terms. SHIC is the result of collaboration between several UK museums, and is a hierarchical index (see Figure 1), with four first level facets, and five levels (we have implemented the first four). Photographs are classified at the lowest appropriate SHIC level, and can have multiple SHIC classifications. The spatial index models four editions of Ordnance Survey maps between 1880 to 1994 using a semantic (not coordinate) approach. The temporal schema ranges from 1755 (the bridging of the River Taff) to 1994. Space and time are first class categories and Lisp query processing works in all three dimensions.

Conventional graphical user interface hypermedia browsers exist in each dimension. A hierarchical menu system allows access to the SHIC hierarchy (Figure 1). Selecting a higher-level category brings forth narrower terms, and the user can view information items at any time. A 'media density' tool shows both the number of items indexed by a specific term, and the number of items if narrower terms are also included. A sliding bar mechanism provides an interactive timeline (Figure 2); this allows the user to select a period, either by dragging the Start/End markers on the timeline, or from a list of periods that contain the selected timespan. The period can be qualified by temporal operators, such as During, Pre, Post, Circa (in the Define Period panel). An A to Z arrangement of street names controls the spatial index (Figure 2). Choosing a letter from the A-Z, takes the user to the relevant part of the street index. The separate browsers can feed into a composite query window, allowing information requests to operate over all three dimensions, for example, Social Organisations in Pontypridd town centre during the Victorian era. At the implementation level, both simple and advanced navigation result in a query on the underlying link store, and early prototypes included a conventional query interface.

Examples of Advanced Navigation

Before discussing the structure of the index space and operations over it, we first consider the kind of navigation tools which might take advantage of the relationships in the classification schema. One application is in the case of an unsuccessful query or navigation route that leaves the user at a valid place in index space, but which has no associated information items. These situations may arise in museum information systems when there is an uneven distribution of information items across the classification space(s). There may be clusters of items at parts of a schema where a museum collection has a lot of coverage, and also sparse areas. This may pose particular problems for interactive browsing through multiple classification dimensions. The user may be unaware that information is available

ISSN: 2324-9773

Tudhope, D., Taylor, C., & Beynon-Davies, P. (1995). Classification and Hypermedia. 6th ASIS SIG/CR Classification Research Workshop, 173-192. doi:10.7152/acro.v6i1.12668

"nearby" in one of the dimensions. For example, the available set of historical photographs of Pontypridd tend to be located around the town centre and the sites of heavy industry.

For such situations, we have implemented a query generalisation tool, discussed elsewhere (Tudhope, et al, forthcoming). Briefly, on being dissatisfied with the results of an information request, the user may ask the system to 'Suggest alternatives' (this control panel can be seen in Figures 1 and 2). In fact, the tool is probably best suited to a composite, multi-dimensional request, where it takes into account which index dimensions have performed badly. For example, if a request concerned the SHIC and spatial dimensions, and no media items were returned for the SHIC term, but items were found for the street in question, then the SHIC term would be generalised. The underlying semantic net (relationships between terms) would be traversed, in order to see if nearby terms had associated information items. The slider bar in the 'Suggest alternatives' control gives a rough measure of the desired closeness, and corresponds to the semantic closeness measures discussed below. Of course, there is no guarantee that items suggested will actually be of interest, but the tool may provide a shortcut to items nearby in classification space, without the need for the user to systematically investigate terms in several index dimensions.

The main tool considered in this paper is 'navigation via similarity', an instance of a general technique which has been used in different applications in both information retrieval and hypermedia (Tudhope, et al, 1995). This method places less emphasis on the user being familiar with a classification system - having found one item of interest the user only has to ask for similar items. It was one of the methods for retrieving information identified by field studies of user-librarian interaction in the Danish BookHouse project (Pejtersen, 1989). The technique relies on a similarity measure, or coefficient, between sets of index terms. There may be terms from different dimensions of indexing and, as with SHIC, there may be multiple terms from a dimension. Different media items may have different numbers of terms. Of course, any similarity measure can only be in terms of the classification systems employed. Various similarity measures have been used in information retrieval (Salton, 1989). However, the measures have required exact matches of terms. Common measures used in hypermedia applications are variants of the cosine coefficient, involving set intersection between vectors of index terms with some normalisation mechanism. Matches occur when an index term is present in both vectors (see, eg, Tudhope, et al, 1995, Parunak, 1993).

Our measure takes account of relationships between index terms, and thus relaxes the requirement for exact equivalence of terms. We make use of a distance measure over the different index dimensions that returns a measure (in the range 0..1) of the degree of match. As an illustrative example from our prototype, consider two photographs and their index terms in the temporal and subject-index dimensions:

Media Item 22 'Graig Thistles' - a football team
SHIC: Men's-Costume, Sporting;
Date: 1918

Media Item 28 'Spaniards' - a musical group
SHIC: General-Costume, Social, Entertainment;
Date: 1926

ISSN: 2324-9773

Tudhope, D., Taylor, C., & Beynon-Davies, P. (1995). Classification and Hypermedia. 6th ASIS SIG/CR Classification Research Workshop, 173-192. doi:10.7152/acro.v6i1.12668

Both photographs (Figures 3-4) are of groups of people involved in some kind of leisure activity (sport and music), and with different gender mixes. The dates are close, considering the total time range in the schema (1755 to 1955). A similarity coefficient based on exact equivalence would return a completely null match. Although Men's and General are narrower classes of the broader term Costume, and Social and Sporting are sibling descendants of Organisation, this semantic proximity would not be reflected in the result. Entertainment is a more distant cousin to Organisation, but is still a subclass of the primary term Community Life. Costume belongs to a completely different facet to Community Life.

Figures 3-4 illustrate how this method can be used as a means of navigation, using our current prototype. Figure 3 shows that the user has previously navigated with the temporal browser to ask for information on the period 1900 to 1930. From the 33 media items available, one has been selected - the Graig Thistles, and its classification terms displayed. Wishing to view similar items, the user has selected the Media Similarity tool and set the Geographic dimension off but indicated closeness criteria with the slider bars for conceptual and temporal dimensions. Figure 4 shows the result - the Untouchables had the same SHIC index terms as the Thistles and are thus less interesting than the Spaniards. The calculation of similarity is based on the semantic closeness measures outlined below. An example involving the spatial dimension is discussed in Tudhope, et al (1995).

Semantic Modelling

Our initial set of primitive relationships has been derived from work on semantic modelling in database schema design (Hull and King,1987). This approach can be considered as 'more natural' than previous data models; the hierarchical, network, and relational models are more concerned with data storage formats than with the application's natural semantic units. Hull and King characterise a semantic model's main features as the explicit representation of domain objects and their relations, mechanisms for building complex types of objects, hierarchies of subtype/supertype relationships, and the ability to deduce new information in the schema from existing data. For example, commonly occurring attributes can be explicitly modelled once for parent classes, and then inherited by sub-classes lower down the hierarchy. Where required, sub-classes can specialise the original definition. Knowledge of the schema can also be exploited for integrity maintenance on deletion or modification of an instance of a data class.

We model relationships between the concept terms in the index space by lists of binary relations. Relationships are expressed as triples: subject, relationship type and object, and are held in a Binary Relational Store (Frost, 1982). For hypermedia architectures, employing a restricted set of relationships, as opposed to allowing an application author total freedom in constructing relationship types, allows the possibility of reasoning over the relationships to provide some intelligence in navigation or query processing. A set of primitive relationships, taken from the semantic modelling literature (and extended for the temporal and spatial dimensions) was adopted:

a) A Kind Of (AKO) - a hierarchical subclass-superclass relationship between classes of terms.

b) Part Of - a meronymic (aggregation) relationship modelling a part/whole relationship.

c) HasA - a definition of the attributes or properties of a class. For example, a temporal interval has a start date, end date and a name.

ISSN: 2324-9773

Tudhope, D., Taylor, C., & Beynon-Davies, P. (1995). Classification and Hypermedia. 6th ASIS SIG/CR Classification Research Workshop, 173-192. doi:10.7152/acro.v6i1.12668

d) IsA - although sometimes used to mean subtype/supertype, is strictly employed as relating an individual instance to a more general class. For example, Pontypridd Bridge IsA Footbridge (whereas Footbridge AKO Bridge).

The hierarchical SHIC subject index is modelled by four separate AKO hierarchies for the four first-level facets of SHIC. An example can be seen in Figure 1, where AKO relationships exist between narrower and broader terms. (A previous prototype on a local railway allowed more scope for combining AKO and IsA relationships in the schema.)

The temporal index is based on a unit of a year. The notion of a Temporal Interval captures both calendar periods such as Century and Decade, and more obviously socially-defined periods, such as the Victorian Era, the Depression, World War II. An implicit ordering is applied by the representation of years as integers. Temporal operators Pre, Post, During, and Circa can be applied to both kinds of period and are available to the user in the temporal browser. These operators can also be used to implicitly model temporal structural relationships, for example decades as belonging to centuries.

A geographic term consists of two categories. Firstly, it is part of an area. Area terms are grouped into strictly non-overlapping spatial containment hierarchies by the PartOf relationship. In addition a simple urban-geography term index is used for concepts like street, town, type of building - specific geographic terms are instances of street, etc. (SHIC was not seen to provide sufficient coverage of this sort.) In the current prototype items are mostly catalogued by street or building type. In order to provide scope for exploring spatial reasoning, a relationship, NextTo, was introduced to model a basic notion of spatial proximity, mostly street intersection.

A temporal-reference relationship exists to link a geographic term with a temporal interval. Each geographic term exists within a specific temporal period. This was to model the evolving meaning of geographic terms over time. The area denoted by a name may grow or shrink. Part of a town may be demolished; new streets or buildings may be built, or replace old areas. Names may change. Visitors to a museum might request information using terms that no longer apply to current administrative boundaries, Four editions of Ordnance Survey maps between 1880 and 1994 were used to derive the geographic terms for the schema, and their temporal periods. In all, the geographic classification holds 830 triples, containing 204 areas, 30 types of term, 32 temporal redefinitions, and 44 proximity relationships (over the town centre). The relationships, Redefines and Removal, model geographic change over time (Taylor, et al, 1994). Redefines provides a relationship between an old term and a new one that is replacing it. Removal is used in conjunction with Redefines to model the removal of a term through temporal inheritance. Operators exist to return changes or similarities for a geographic term between different time periods. The precise semantics of a geographical term is derived by tracing any Redefines or Removal relationships for the term. However, the interface does not currently permit the user to directly make use of this linkage.

Interlinking need not be confined to the temporal and spatial dimensions, although that is the extent of our current exploration. If we consider the full potential for interlinking, SHIC terms such as Transport have an obvious temporal dimension. On the other hand, some temporal periods such as World War II or the Depression are dependent on cultural or geographic context for their precise meaning. Nauta (1993) has discussed the changing meaning of artistic concepts over time, and how a hypertext system might allow a user to navigate among different uses of chiaroscuro, say, in different historical periods or areas. Users of the Art and Architecture Thesaurus are recommended to maintain older

ISSN: 2324-9773

Tudhope, D., Taylor, C., & Beynon-Davies, P. (1995). Classification and Hypermedia. 6th ASIS SIG/CR Classification Research Workshop, 173-192. doi:10.7152/acro.v6i1.12668

synonyms along with current (preferred) terms, although there is no explicit modelling of temporal linkage. As a compromise, archaic terms have to be searched via free text in narrative descriptions (Weinberg 1995).

Semantic Closeness

Flexible navigation/query tools embodying a degree of intelligence require some reasoning capability over the concept terms in the index space. If we are to relax exact matching of terms in a user's request for information (or current 'position' in hyperspace), then some notion of distance between concept terms is needed. This 'semantic closeness' underlies the navigation tools illustrated above. There are different implementations in the different dimensions of classification.

The simplest implementation is in the temporal dimension due to the numerical basis. This was refined by including a distance measure between temporal periods (Tudhope, et al, 1995). The semantics of distance between periods has to deal with factors including the relative length of the two periods, the amount of overlap if any, and the gap between the periods if no overlap. The measure implemented was parameterised so that the measure could be tailored to reflect the relative importance of these factors to a particular user. This measure would benefit from empirical usability evaluation. The current formula is

$$SemanticCloseness = W_1\left(\frac{MP}{IU}\right) + \frac{W_2}{1 + NMP/IU} + \frac{W_3}{1 + D/IU}$$

where:

- D = The distance between the temporal periods
- MP = The overlap (the matching portion) between periods
- NMP = The amount of non-overlap between the periods
- IU = The length of the period used as the basis for comparison.

The spatial distance measure in our prototype operates only over the NextTo relationship in the schema. Therefore it is based on a traversal of the shortest path between two geographic terms, with a weighting factor for each traversal in the path, and is a simpler case of the semantic closeness measure used for the subject index, which is based on the four primitive relations above and includes hierarchical relationships. This latter distance measure is still based on a traversal of the transitive relationships in the semantic net connecting the two classification terms, with a cost factor associated with each traversal between two directly connected terms. (We do not distinguish the direction of traversal in assigning the cost factor.) In an attempt to reflect the semantics of taxonomic hierarchies, we are experimenting with weighting the cost of a traversal between two neighbouring terms. One factor is the level in the hierarchy - cost is inversely proportion to level. Two siblings can be considered to have more in common at the bottom of a hierarchy than at the top. We are also experimenting with modifying the cost factor to reflect the type of relationship traversed. Intuitively, a term which is an instance of another class might appear closer than a term connected with a subclass or superclass relationship (AKO).

The total cost of a traversal over the classification is thresholded when the cost falls below zero, in which case the terms involved are considered far enough apart to be not close at all. Clearly, by varying the basic cost factor, this 'distance horizon' can be tailored to require either very close matches of query terms, or yield a very elastic measure. Criteria for choosing the degree of elasticity for a particular application need to be investigated, and

ISSN: 2324-9773

Tudhope, D., Taylor, C., & Beynon-Davies, P. (1995). Classification and Hypermedia. 6th ASIS SIG/CR Classification Research Workshop, 173-192. doi:10.7152/acro.v6i1.12668

whether it is useful to pass this control to the user. Again, given the possibility of weighting traversal costs differently, any empirical warrant for doing so needs to be investigated. Another subject for investigation is a better interface than the slider bars for indicating the desired semantic closeness in the different dimensions; the user needs feedback of the parameter in terms of the classification schema itself.

A recent experimental investigation of user perception of semantic distance asked users to express the relevance of a short text passage with an index term (and scope note) using a seven point sliding scale (Brooks, 1995). Index terms were taken from different levels of a hierarchical thesaurus in different trials on the same passage. There was only one descendant at each level down the hierarchy (no siblings). Perceptions of relevance were found to be inversely proportional to semantic distance, as measured by traversals (semantic steps) up/down the tree from the index term actually associated with the passage. However, the study also found that the rate of decline of relevance differed depending on the direction of traversal. Narrower terms, as the hierarchy was descended, were more quickly perceived as non-relevant than broader terms. Thus, although this study would appear to support the basing of our semantic closeness function on term traversal, and to provide some empirical justification for such measures, it also suggests that direction should be a factor in determining the cost of a traversal. However, the study was confined to AKO (in our terminology) relationships only, and did not consider the effect of sibling term descendants (only single parent-child relationships studied).

Rada, et al (1989) describe a metric to measure conceptual distance between sets of nodes in a hierarchical semantic net over IsA relationships (used in a general sense of broader/narrower), particularly with a view to the merging of (medical) thesauri. This metric was based on a shortest path traversal and did not distinguish direction. Experiments based on subjects assigning a measure of closeness to neighbouring pairs of terms did not support a significant contribution from direction; however this was based only on consideration of perceived conceptual closeness of terms and not any associated documents. Another experiment comparing the ranking of a set of documents with the index terms comprising a query by both subjects and the distance metric produced significant agreement. A recommendation of the paper was that extending the metric to non-hierarchical relationships should be treated with caution; in one experiment, introducing a non-hierarchical (causal) link between terms failed to produce agreement between the metric's and human subjects' measure of distance. The authors recommend that non-hierarchical relationships should only be traversed when there are indications that both document and query appear relevant to the relationship.

A further complication is introduced when implementing similarity measures between information items indexed by classification systems which allow an item to be classified by multiple terms. Then the semantic distance measure needs to compare two (possibly uneven) sets of classification terms. Rada et al's metric compares each term in one set with each term in the other set and returns the arithmetic mean. Our implementation sums the maxima achieved by each term with the members of the other set and normalises by the cardinality of the two sets. The reason for this is to take into account that different index terms may be contributed by different facets of a classification, or for a composite item like a photograph index terms may be related to different aspects of the item. To take the example of the two photographs described above:

ISSN: 2324-9773

Tudhope, D., Taylor, C., & Beynon-Davies, P. (1995). Classification and Hypermedia. 6th ASIS SIG/CR Classification Research Workshop, 173-192. doi:10.7152/acro.v6i1.12668

	Men's Costume	Sporting	Maxima
General-Costume	0.64	0.00	0.64
Social	0.00	0.64	0.64
Entertainment	0.00	0.19	0.19
Maxima	0.64	0.64	2.75

Set Similarity = 2.75 / (3 + 2) = 0.55

This method attempts where possible to compare like with like; the lack of match between General-Costume and Sporting does not detract from the high degree of match between the other terms. Rada, et al, discuss the possibility that distances are being calculated by their measure that are not meaningful (between quite different attributes, eg disease etiology compared to laboratory tests), and that a decomposition of index terms and different applications of the distance metric might be helpful. This would tend to support our approach above, where the maximum closeness values for the different index terms are summed. It would also support the unified similarity coefficient and different index dimensions described below; different facets of a classification could be assigned to different dimensions.

When navigating via similarity over multiple dimensions, we apply the (different) semantic closeness measures to each dimension and intersect the results. For large datasets, the number of comparisons may become a limiting factor. Our current prototype experiments with efficient implementation by applying a closeness threshold in each dimension, and only considers media items close to the original. This can result in media items being falsely rejected if they score very highly in other dimensions. More work is needed on such implementation issues.

Unified Similarity Coefficients

Recently we have been experimenting with a unified similarity coefficient that operates over multiple index dimensions of different data types in the same formula. This would mean that we would not have to take each dimension separately and intersect the results. This coefficient is adopted from the field of numerical taxonomy (Sneath and Sokal, 1973), where it is used for quite different purposes. Numerical Taxonomy measures taxonomic distance in biological classifications with a view to automatically deriving a classification of a set of operational taxonomic units, according to overall (phenetic) similarity. In our applications, we do not wish to derive a new classification (although that is an interesting possibility), but to integrate several extant classifications of a collection, in order to construct a method of navigation based on taxonomic distance. Thus we propose to turn numerical taxonomy on its head to construct navigation tools for a collection of information, which has previously been given classification schema.

We have implemented a version of Gower's (1971) general coefficient of similarity and are evaluating it. The general form of the coefficient S for two information items j and j is

ISSN: 2324-9773

Tudhope, D., Taylor, C., & Beynon-Davies, P. (1995). Classification and Hypermedia. 6th ASIS SIG/CR Classification Research Workshop, 173-192. doi:10.7152/acro.v6i1.12668

$$S_{ij} = \frac{\sum_{k=1}^{n} W_{ijk} S_{ijk}}{\sum_{k=1}^{n} W_{ijk}}$$

S_{ijk} is the similarity between the two items on the k_{th} dimension of indexing. W_{ijk} is a user-defined weight for each dimension, with the added criteria of being zero when no valid comparison is possible between two items on that dimension. Thus the coefficient handles missing data (for example, dates might not be available or indexing might be incomplete). Numerical taxonomists tended to view a priori weighting of characters with suspicion and Gower included it only as an option. However for our purposes, the weighting of dimensions allows users to assign the relative importance of different dimensions in their search. This coefficient is designed to work with multiple dimensions of indexing, where the different dimensions can have different data types. Each dimension returns an individual similarity measure between 0 and 1. Gower distinguishes between quantitative, dichotomous, alternative, and qualitative terms. Dichotomous refers to a single character (index term) that can be present or absent in the data (has two levels). When it is absent from both sets of terms then it does not count as a match, unlike alternative terms where two absences count as a match. Qualitative terms can have many levels, but they do not form an ordered set; there is no relationship between terms, and an exact match suffices. The match between two quantitative terms is calculated by

$$1 - \frac{|T_1 - T_2|}{\text{Range}}$$

We have adapted the coefficient to make use of semantic closeness measures when dealing with classifications where semantic relationships exist between terms, and also included the comparison between temporal periods. Thus the same coefficient is applicable to quantitative spatial or temporal metrics, binary indexing of nodes with keywords, and qualitative multistate or hierarchical classifications of information. Computational tractability remains to be explored; for large collections, comparing a given information item with every other item is expensive. It may be possible to produce a 'quick and dirty' option that achieves a faster solution, at the cost of some recall, by only considering a subset of items - this subset could be derived from partial results from the highly weighted dimensions. (Alternatively, a method of 'progressive refinement' could be used, where the user can terminate calculation of results at any time.)

Thus a unified similarity coefficient offers the possibility of incorporating:

- Multiple dimensions to the index space, where index terms in the different dimensions can be of different data types (as outlined above)
- User supplied weights to indicate the importance of the index dimension to this retrieval operation for the user
- Non-exact matches between terms in the query and terms classifying information items
- Missing terms in the classification of particular information items, for example when the date of a photograph is unknown, which will result in invalid comparisons for those items and terms

ISSN: 2324-9773

Tudhope, D., Taylor, C., & Beynon-Davies, P. (1995). Classification and Hypermedia. 6th ASIS SIG/CR Classification Research Workshop, 173-192. doi:10.7152/acro.v6i1.12668

- Multiple classifications in the same index dimension for an item, for example the multiple SHIC terms in the above example
- Hierarchical (or other structured) index dimensions, which allow the possibility of reasoning over relationships between terms.

Future Research Issues

For future research development, we wish to move towards larger data sets and to broaden the application area to include community information systems, as well as museum systems, in order to provide more scope for spatially indexed data. This will require a richer set of semantic modelling primitives, which should also allow reasoning mechanisms oriented to particular kinds of queries. We started by adopting a small number of general relationships from the semantic database literature; we wish to extend this and tailor it more closely to our application areas. In order to do this, we hope to also draw on semantic modelling work in information science. If we restrict ourselves to our current fundamental categories, we particularly wish to extend the subject-based and spatial indexing and to explore to greater depth the interlinking of dimensions.

Subject Indexes

Thesaurus relationships (eg Aitchison and Gilchrist, 1987) offer a rich set of semantic primitives applicable to a wide range of subject indexes. Many collection management standard classifications or indexes are based on an underlying thesaurus. A suitable base set for our purposes might include the following relationships. (Only one direction is listed since the other can be implicitly generated.)

Equivalence: indicates two terms are equivalent in the model. For example, it can be used to link Unpreferred terms to Preferred terms, and to provide synonyms.

Hierarchical: indicates general broader/narrower relationships (subclass-superclass). It is useful to distinguish different types of hierarchical relationships to allow more specialised query processing:

Generic: indicates the general taxonomic relationship, as used in biological (say) classification, which is equivalent to our use of AKO to model the SHIC taxonomy.

Whole/part: indicates a part/whole relationship, where the part is contained in, or possessed by, the whole. This coincides with our use of PartOf which was overloaded in the prototypes. It would be useful to more clearly distinguish different aspects relevant to our application area. Possible examples might include:

Geographic strict spatial containment:
(Trefforest IsGeographicallyPartOf Pontypridd)

Temporal containment:
So far this has been implicitly calculated in our system, but there may be a case for explicitly modelling some relationships between temporal periods to facilitate temporal reasoning:
(decade IsTemporallyPartOf century)

ISSN: 2324-9773

Tudhope, D., Taylor, C., & Beynon-Davies, P. (1995). Classification and Hypermedia. 6th ASIS SIG/CR Classification Research Workshop, 173-192. doi:10.7152/acro.v6i1.12668

Organisational containment:
In community information systems, it may be desirable to model relevant organisational structure, or tiers of responsibility:
(TownCouncil IsOrganisationallyPartOf CountyCouncil)

Instance: indicates that a specific term is a particular instance of a more general class - to take an example from our prototype:
(BrownLennoxChainworks IsA IndustrialPlant)
We considered IsA a quite separate relationship to AKO, although both are subject to inheritance, but it appears common to include Instance as a hierarchical thesaurus relationship. It is interesting to note that there has also been discussion in the database literature of precise distinctions of hierarchical and non-hierarchical semantic primitives with specific reference to database design, semantic nets and inheritance (eg, Brachman, 1983, Storey, 1993). Storey particularly discusses different kinds of part/whole relationships.

Poly-hierarchic: indicates that a term belongs to more than one broader category. It can apply to hierarchical and part/whole relationships. In our case, this particularly applies to overlapping spatial containment, where a geographic term such as a river may belong to several areas, or a colloquial district name overlaps more than one official subdivision. This would pose additional problems for semantic closeness traversal algorithms based on generalisation and specialisation. An example might be:
(RiverTaff IsPollyPartOf Pontypridd)
(RiverTaff IsPollyPartOf Trefforest)

Associative: indicates that a conceptual relationship between two terms exists, which is not an equivalence or hierarchic relation. Uses in our application might be whole/part relationships where a term might have constituent parts, but which might be considered more an association than a hierarchy. An example might be parts of a steelworks, or components of a railway network. Other examples might be causal relationships, or antonyms (may be less relevant to our applications).

An important employment of the Associative relationship for future development in our applications might be to express a (rough) association between two quite different classification systems. A term in one system would be associated with a corresponding term in another subject classification. This topic is discussed further below in regard to multiple classification systems.

Agosti et al (1995) and Rada et al (1993) have also discussed the three main thesaurus relationships as hypertext links.

Spatial Indexes

To extend the spatial modelling, we need to both provide a richer set of semantic primitives and incorporate coordinate-based Geographic Information System (GIS) functionality. If GIS functionality was incorporated then some semantic relationships between terms might be automatically derived from numerical attributes. Clementini and Felice (1994) discuss object-oriented spatial modelling, distinguishing instances from

ISSN: 2324-9773

Tudhope, D., Taylor, C., & Beynon-Davies, P. (1995). Classification and Hypermedia. 6th ASIS SIG/CR Classification Research Workshop, 173-192. doi:10.7152/acro.v6i1.12668

classes, and including both a complete and a partial containment part/whole relationship. Our very basic proximity relationship NextTo could be extended in an urban context by a more elaborate typification of intersections and/or the compass directions. Conventional commercial GIS functionality could yield quite sophisticated responses to coordinate-based spatial searches, although semantic spatial reasoning is still a research topic (Jones, et al, 1995). The meaning of relationships such as 'Near' varies with context, according to the type and size of objects compared, the purpose of the comparison, and whether a route-based distance metric is required. For example, a query might ask the system for towns 'near' Cardiff, and would thus involve the issue of whether Pontypridd (say) was 'near' Cardiff. Using explicit semantic relationships, or simple numeric calculations, would be problematic in this instance. (Is the query relating to pedestrians, cyclists, or car drivers?) A solution might require a procedural or rule-based response, rather than relying on an explicit model (for example see, Abdelmoty, et al, 1993). Hierarchical spatial relationships may inform calculation of proximity relationships.

Multiple Classifications

Work in the sociology of knowledge (eg, Barnes, 1994) has demonstrated the conventional nature of all classifications and located them as social and cultural artefacts. Classifications often involve a hierarchical ordering of terms and carry with them implicit views of what is important.

Feminist studies of science (Schiebinger, 1993) have shown how scientific classifications that we take for granted have their origins in specific historical and social settings and movements and have broader consequences. Schiebinger argues that the adoption of the concept 'mammal' by Linnaeus as a primary classification term for the animal kingdom was inter-linked with social debates at that time on breast feeding. As another example, the SHIC system employed in our social history prototype does not facilitate interrogation of the information based on gender or class. How a body of information is classified determines the kind of questions that can sensibly hope for an answer.

Hypermedia architectures with index spaces independent of the application content allow the possibility of multiple classifications of the same information. A given museum collection might be catalogued and also accessed by two quite different subject indexes, whether because of overlapping subject matter or because of evolving management of the collection. Another issue we are interested in is to incorporate both folk and 'expert' taxonomies. The SHIC system was designed for cataloguing purposes and we adopted it for public presentation. Museum professionals have pointed out that the depth of detail and academic terms employed render it problematic for public use.

Lay classification

Since we wish to facilitate public access, it would be attractive to attempt to provide parallel classifications that purport to be more 'lay-oriented' as opposed to 'expert' or 'scientific' classification systems. An example might be a simplified version of a complex classification, where a close parallel would hold. Links between terms in two parallel classification systems (eg lay to expert) could be another application of an associative relationship . Typically such a relationship would only be suggestive, as semantic equivalence would be unlikely, but it might provide pathways for either interactive browsers, or for automatic Guides to the system, with the capability of making suggestions to the user of nearby information according to the Guide's perspective (in a slightly different context, see Salomon, et al, 1989).

ISSN: 2324-9773

Tudhope, D., Taylor, C., & Beynon-Davies, P. (1995). Classification and Hypermedia. 6th ASIS SIG/CR Classification Research Workshop, 173-192. doi:10.7152/acro.v6i1.12668

Work in the field of cognitive anthropology is relevant to this endeavour. Here anthropologists are interested in discovering and describing how people organise and group the cognitive principles underlying culture and behaviour - in members' own terms. Although much of this work is directed towards study of very different cultures, some of the work on semantic groupings is of interest. Tyler (1969) lists taxonomies, trees and paradigms as basic observed semantic orderings. Taxonomies correspond to the hierarchic relationship described above, and trees are branching diagrams where only one character changes at each division (as in biological 'keys'). Paradigms are non-hierarchic orderings which cut across levels of taxonomic hierarchies by multiple intersections. For example, the concepts gender (male, female, neuter) and maturity (child, adolescent, adult) intersect with zoological classifications to yield terms such as mare, gelding, boar, colt, kitten, etc. Paradigmatic relationships might form an additional semantic primitive.

In fact, a similar type of relationship arises in some existing computer classifications. ICONCLASS, an iconographic classification system for the visual arts, is described as a hybrid classification system (Brandhorst, 1993, Grund, 1993) because it offers some extra features for users to create their own concepts. Among these is the Key number device which allows a user to define more detail about a concept in a hierarchy in a fashion that does not result in a combinatorial explosion of hierarchical terms. For example, an image of a saint could be drawn in front, back, side views and pointing up, down, etc. A person could be seen as old, as speaking, or expressing a particular emotion. These details are seen as additional to the primary organisation of the classification, and can themselves be unitary concepts or secondary hierarchies.

Concluding Remarks

Structured index spaces in hypermedia applications can benefit from research on classification. Reasoning over relationships between concepts in a classification system has the potential to assist hypermedia navigation, when a non-specialist user may be unfamiliar with how the information is arranged. This paper has described navigation tools, based on measures of semantic closeness, in a prototype implementation, and discussed future possibilities for a richer set of semantic modelling primitives and overlapping classification systems.

Acknowledgements

Assistance and historical information afforded by the Pontypridd Historical and Cultural Centre, and in particular Brian Davies and David Gwyer, is gratefully acknowledged. We would also like to thank Alice Grant from the Science Museum and Gerhard Jan Nauta from Leiden University for interesting discussions, and Phil Smith who provided helpful comments and suggestions in reviewing this paper.

Bibliography

Abdelmoty A., Williams M., Paton N. (1993). Deduction and Deductive Databases for Geographic Data Handling. Proc. Third International Symposium, SSD'93 Advances in Spatial Databases, (eds.) D. Abel and B. Ooi, Springer-Verlag, pp 443-464.

Agosti M., Melucci M, Crestani F. 1995. 'Automatic authoring and construction of hypermedia for information retrieval', Multimedia Systems, 3(1): 15-24.

Aitchison J., Gilchrist A. 1987. 'Thesaurus construction: a practical manual', Association for Information Management: London.

ISSN: 2324-9773

Tudhope, D., Taylor, C., & Beynon-Davies, P. (1995). Classification and Hypermedia. 6th ASIS SIG/CR Classification Research Workshop, 173-192. doi:10.7152/acro.v6i1.12668

Barnes B., Bloor D., Henry J. 1994. 'Scientific knowledge: a Sociological Analysis'. Athlone Press: London.

Besser H. 1991. 'User Interfaces for Museums', Visual Resources VII, 293-309.

Beynon-Davies P., Tudhope D., Taylor C., Jones C. 1994 'A Semantic Database Approach to Intelligent Hypermedia Systems'. Information and Software Technology. 36 (6), pp 323-329.

Brachman R.J. (1983). 'What Isa is and isn't: an analysis of taxonmomic links in semantic networks', Computer. 16(10). pp 30-36.

Brandhorst J. 1993. 'Quantifiability in Iconography', Knowledge Organisation 20(1), pp 12-19.

Brooks T. 1995. 'People, Words, and Perceptions: A Phenomenological Investigation of Textuality', Journal of the American Society for Information Science, 46(2), pp 103-115.

Clementini E., Felice P. 1994. 'Object-Oriented Modeling of Geographic Data', Journal of the American Society for Information Science, 45(9), pp 694-704.

Frisse M.E. and Cousins S.B. 1989. 'Information Retrieval from Hypertext: Update on the Dynamic Medical Handbook Project'. Proc. ACM Conference on Hypertext, Pittsburgh PA, pp 199-211.

Frost R.A. 1982. 'Binary-Relational Storage Structures'. The Computer Journal. 25(3), pp 358-367.

Gower, J. 1971., ' A General Coefficient of Similarity and Some of its Properties', Biometrics 27, December 1971, pp 857-74.

Grund A. 1993. 'ICONCLASS. On Subject Analysis of Iconographic Representations of Works of Art', Knowledge Organisation 20(1), pp 20-29.

Halasz F., Schwartz M. 1994, 'The Dexter Hypertext Reference Model', Communications of the ACM 37(2), pp 30-39.

Hull R., King R. 1987. 'Semantic Database Modelling: Survey, Applications and Research Issues', ACM Computing Surveys, 19(3), pp 201-260.

Jones C., Beynon-Davies P., Taylor C., Tudhope D. 1995 forthcoming. 'GIS, Hypermedia and Historical Information Access', Proc. 7'th International Conference of the MDA, Edinburgh.

Mayes J., Kibby M., Watson H. 1988. 'StrathTutor: the development and evaluation of a learning-by-browsing system on the Macintosh', Computers & Education, (12), pp 221-9.

Molholt P., Petersen T. 1993. The role of the 'Art and Architecture Thesaurus' in communicating about visual art, Knowledge Organisation 20(1), pp 30-34.

Nauta G. 1993. 'HYPERICONICS: Hypertext and the Social Construction of Information about the History of Artistic Notions', Knowledge Organisation 20(1), pp 35-46.

Pejtersen A. 1989. 'A library system for information retrieval based on a cognitive task analysis and supported by an icon-based interface', Proc. ACM Conference on Information Retrieval, pp 40-47.

Parunak, H. 1993. Hypercubes grow on trees and other observations (and other observations from the land of Hypersets). ACM Conference on Hypertext, Seattle. pp 73 - 81.

Rada R., Mili H., Bicknell E., Blettner M. 1989. 'Development and Application of a Metric on Semantic Nets', IEEE Transactions on Systems, Man and Cybernetics, 19(1), pp 17-30.

Rada R., Wang W., Birchall A. 1993. 'Retrieval Hierarchies in Hypertext', Information Processing and Management, 29(3), pp 359-371.

ISSN: 2324-9773

Tudhope, D., Taylor, C., & Beynon-Davies, P. (1995). Classification and Hypermedia. 6th ASIS SIG/CR Classification Research Workshop, 173-192. doi:10.7152/acro.v6i1.12668

Salomon G., Oren T., Kreitman K. 1989. 'Using Guides to Explore Multimedia Databases', Proc. 22nd Hawaii International Conference on System Sciences, pp 3-12.

Salton G. 1989. Automatic Text Processing. Addison-Wesley. Reading:MA.

Schiebinger L. 1993. 'Why mammals are called mammals: gender politics in eighteenth centrury natural history', American Historical Review, 98(2): 382-411.

SHIC Working Party. 1983. Social History and Industrial Classification: A Subject Classification for Museum Collections. (2 vols.) Published by Centre for English Cultural Tradition and Language, University of Sheffield, UK.

Sneath, P., Sokal, R.,1973, 'Numerical Taxonomy: The Principles and Practice of Numerical Classification.', W. H. Freeman and Company , San Francisco.

Storey V. 1993. 'Understanding Semantic Relationships', Very Large Databases Journal, 2(4), pp 455-488.

Taylor C., Tudhope D., Beynon-Davies P. 1994. 'Representation and Manipulation of Conceptual, Temporal and Geographical Knowledge in a Museum Hypermedia System'. Proc ACM European Conference on Hypermedia Technology. Edinburgh. pp 239-244.

Tudhope D., Beynon-Davies P., Taylor C., Jones C. 1994. 'Virtual Architecture based on a Binary Relational Model: A Museum Hypermedia Application', Hypermedia, 6(3), pp 174-192.

Tudhope D., Beynon-Davies P., Taylor C., 1995. 'Navigation via Similarity in Hypermedia and Information Retrieval', Proceedings of Conference on Hypertext, Information Retrieval, Multimedia. Konstanz, April, pp 203-218.

Tudhope D., Taylor C., Beynon-Davies P. (in press) 'Semantic navigation tools for museum hypermedia', CHART Journal of Computers and History of Art.

Tyler S. (ed.) 1969. Cognitive Anthropology. Holt, Rinehart and Winston: New York.

Weinberg B. 1995. In Depth Book Review of: Art and Architecture Thesaurus (2nd ed.), and Guide to Indexing and Cataloging With The Art and Architecture Thesaurus, (Eds. T. Petersen and P. Barnett), ', Journal of the American Society for Information Science, 46(2), pp 152-160.

ISSN: 2324-9773

Tudhope, D., Taylor, C., & Beynon-Davies, P. (1995). Classification and Hypermedia. 6th ASIS SIG/CR Classification Research Workshop, 173-192. doi:10.7152/acro.v6i1.12668

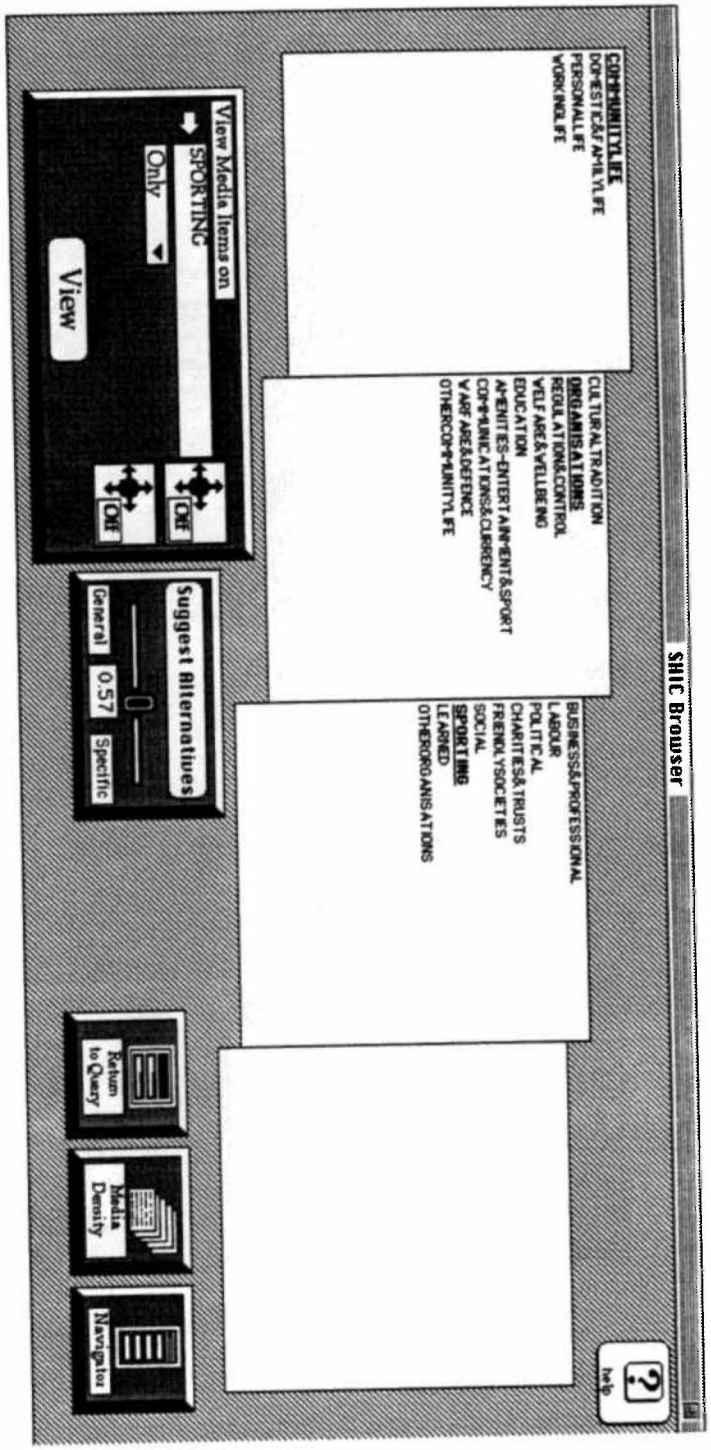

Figure 1: The SHIC Browser

ISSN: 2324-9773

Tudhope, D., Taylor, C., & Beynon-Davies, P. (1995). Classification and Hypermedia. 6th ASIS SIG/CR Classification Research Workshop, 173-192. doi:10.7152/acro.v6i1.12668

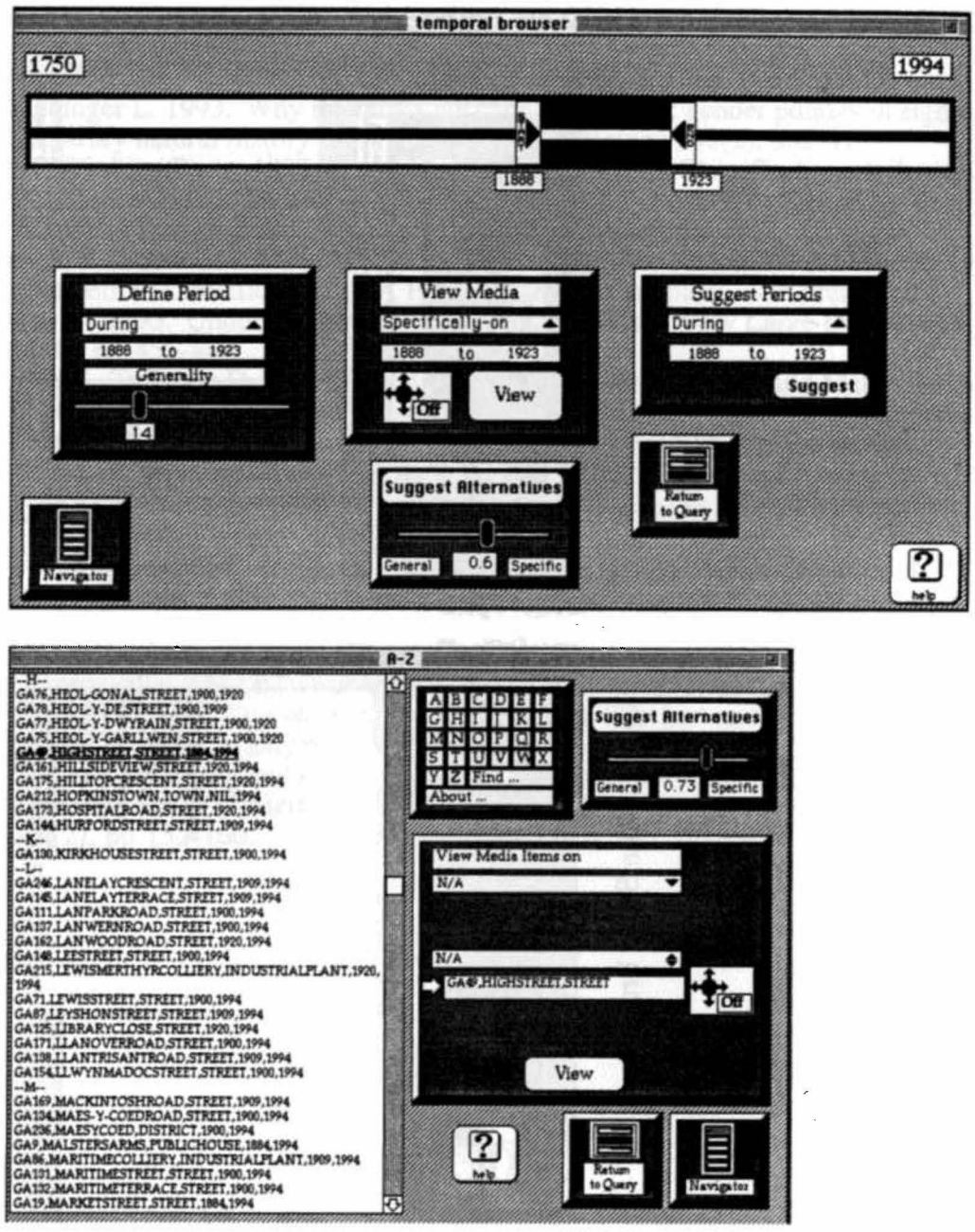

Figure 2: The Temporal and Geographical Browsers

ISSN: 2324-9773

Tudhope, D., Taylor, C., & Beynon-Davies, P. (1995). Classification and Hypermedia. 6th ASIS SIG/CR Classification Research Workshop, 173-192. doi:10.7152/acro.v6i1.12668

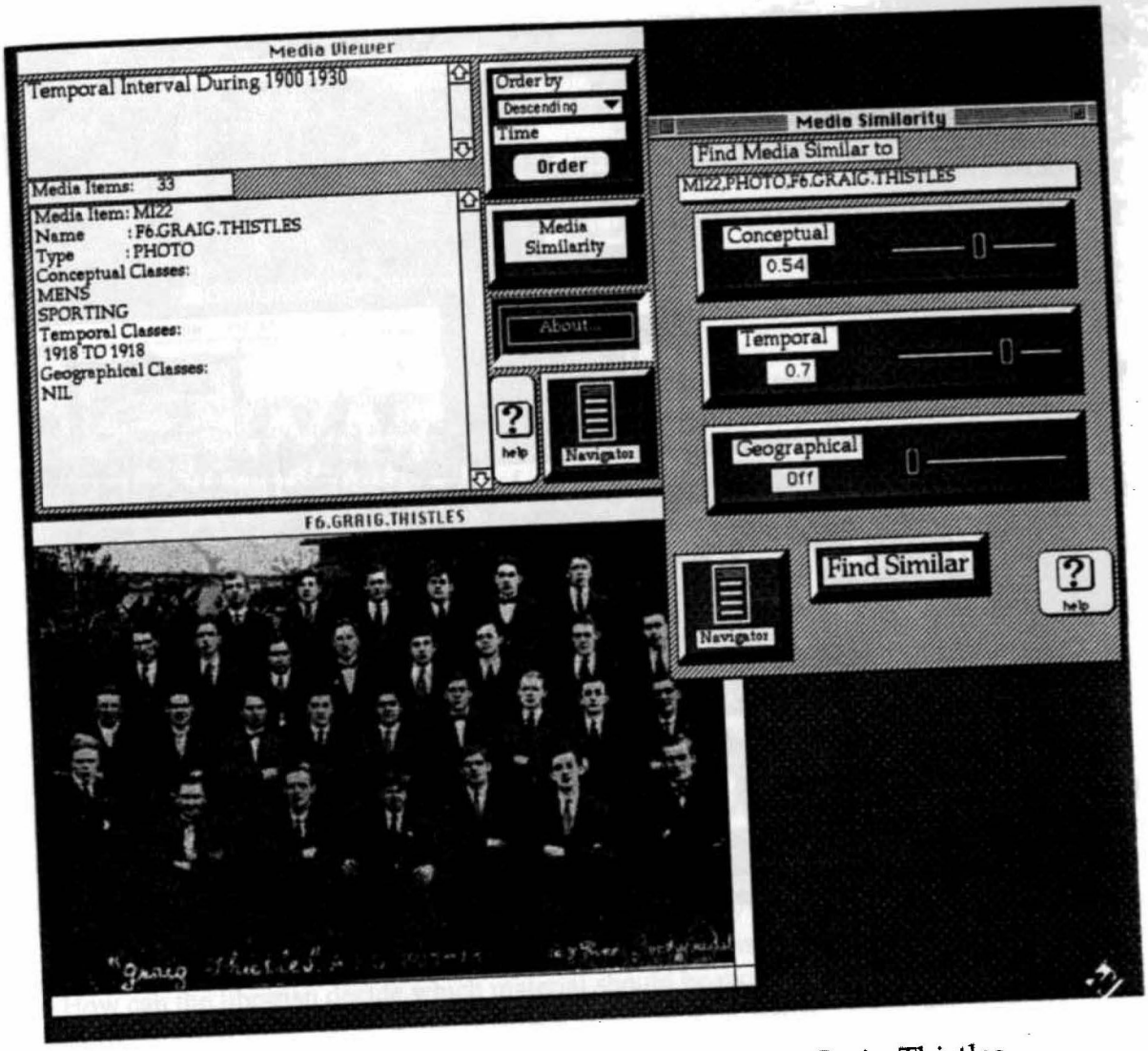

Figure 3: The Media Similarity Tool using Graig Thistles

ISSN: 2324-9773

Tudhope, D., Taylor, C., & Beynon-Davies, P. (1995). Classification and Hypermedia. 6th ASIS SIG/CR Classification Research Workshop, 173-192. doi:10.7152/acro.v6i1.12668

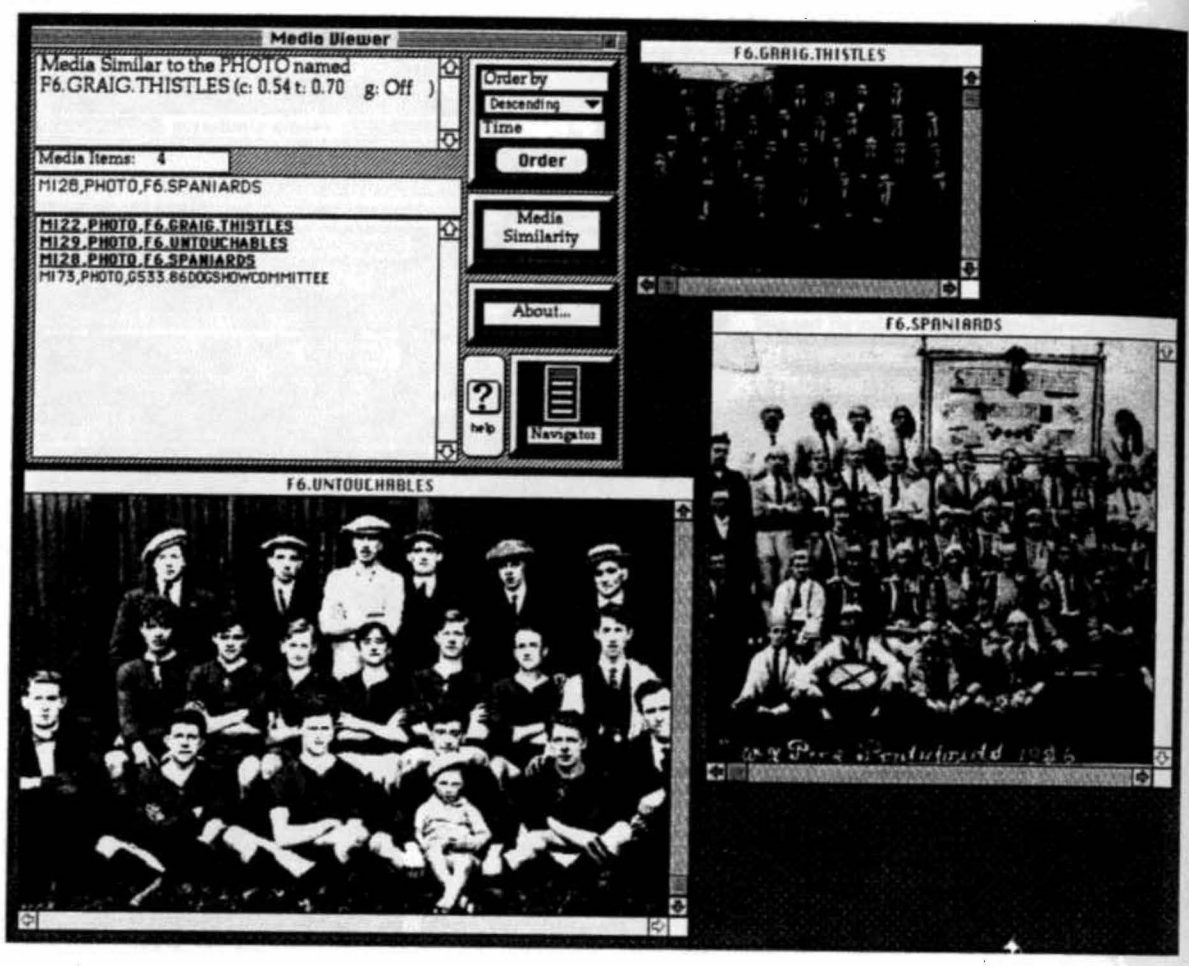

Figure 4: The Media Similar to Graig Thistles

ISSN: 2324-9773