Taxonomic Distance
Classification and Navigation

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

Hypermedia museum exhibits will often be indexed acording to different criteria. This paper discusses an approach to hypermedia design for museum applications that attempts to capture underlying semantic relationships between conceptual terms in a separate index space. A distance measure over the index space makes possible high level navigation tools to assist users unfamiliar with the organisation of the information. This semantic closeness function allows flexible information requests with non-exact matches of terms. A prototype implementation of a social history exhibit is described, where information is indexed by temporal, spatial and subject-based terms. The spatial and temporal dimensions are inter-linked. The paper describes navigation tools intended to allow easier public access, and discusses how they might be extended drawing on work in numerical taxonomy.

Introduction

This paper describes work at the University of Glamorgan, on providing flexible access to hypermedia applications. We have concentrated on museum hypermedia systems, in collaboration with the Pontypridd Historical and Cultural Centre (PHCC). Work on museum hypermedia at the University of Glamorgan began with a project commissioned by PHCC for a map-based historical exhibit on old Pontypridd, which is currently on display in the museum. Subsequently a research group formed to investigate the possibilities of encoding semantic knowledge of the application domain in a Semantic Hypermedia Architecture or SHA, in order to provide assistance to the user in retrieving information from the system.

The current media base is a collection of approximately 120 historical photographs of Pontypridd from the archives of the PHCC, as well as some textual and oral histories. There are three dimensions of indexing. The geographical schema models four editions of Ordnance Survey maps between 1880 to 1994. The conceptual schema represents the hierarchical Social History and Industrial Classification (SHIC 1983) to the fourth level and contains approximately seven hundred terms. The temporal schema ranges from 1755 (the bridging of the River Taff) to 1994. The current prototype system is a hybrid HyperCard / Lisp system. Lisp supports reasoning over the relationships in the schema, and HyperCard is used for the user interfaces and media presentation.

In contrast to most commercial hypermedia systems, we separate link information from the actual content of media items; there is a separate index space. If hypertext links are embedded in the information content, it makes it more difficult to rearrange material in an application, and it hinders attempts to provide advanced navigation tools. The SHA is based on a semantic modelling perspective (Hull and King 1974); it aims to provide a set of binary relationships between terms in the index space, to model the wider context of terms in the application domain. Relationships include notions of subclass/superclass, instances, part/whole, and spatial and temporal versions of these (Tudhope et al 1994, Jones et al 1995). All relationships are held in a Binary Relational Store (Frost 1982), and thus at the lowest level all hypermedia navigation is implemented by a query on the store. Thus links are computed at run time, allowing the possibility of more flexible types of navigation tools.

Recent years have seen a fusion of ideas from hypermedia and information retrieval, seeking to combine query and navigational (browsing) access to information (eg, Agosti et al 1995). For example, the user could be allowed to take the results of a query and browse in a hypertext fashion. Many standard information retrieval domains (very large text-based collections of documents) rely on automatic generation of index terms with queries formed by boolean combinations of terms or free-text forms of query (Salton 1989). In contrast, hypermedia applications have tended to contain smaller amounts of information with links entered 'manually' by the author.

The content of hypermedia museum exhibits is often not so much specially authored, but is arranged from existing material in the museum's archives or other sources. Alternatively, the items in an exhibit may be based around an existing collection of archive material, with interpretations specially written for the new system. The coverage of such collections then may depend upon the museum's previous acquisition policy and the vagaries of chance. 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 towards integration of collection management systems with those designed for public access and exhibition ('a re-thinking of the traditional separation between systems for handling exhibition and other user-oriented information from those that handle collection management information' - Besser, 1991) means that there may be a requirement to use complex controlled vocabulary indexing systems as a basis for information retrieval by the general public. For example, we have made use of SHIC which was designed originally as a collection management tool (Taylor et al 1993). This means that you cannot expect public users to be precise in their requests for information (even assuming they have a specific purpose in using the system). There is a need for access methods that facilitate a guided form of exploration and that do not require exact matching of terms in queries. Molholt and Petersen (1993) see automated thesaurus tools as having potential to offer clues or suggestions that guide users along a pathway between what they want and how the information is organised. Hogan et al (1991) discuss related issues, with reference to content-based (non-verbal) addressing of image databases, in order to facilitate access.

Multi-dimensional indexing

An exhibit may be indexed using several dimensions or classification spaces. For example, an exhibit on the biology of an area might contain specimens indexed by the locations where they were found, the dates collected, one or more scientific classifications for botany or zoology, and perhaps indexes reflecting more popular interest. In the prototype social history exhibit, we classified material by a subject index (SHIC), and also by spatial and temporal dimensions. We decided to make space and time first class entities in the SHA (Jones at al 1995), reflecting the museum orientation. Thus hierarchical and adjacency relationships exist for both spatial and temporal terms, and the Lisp code in the query processor is able to traverse the relationships in the index spaces to provide imprecise matching of the terms in a user's information request with the actual terms used to index particular media items. The foregrounding of geographical and chronological terms is supported by a project (Bates et al 1993) to study search terminology used by humanities scholars. The study also commented on the need of humanities scholars to combine terms from a wide variety of categories in online searching. With the SHA, we have attempted to provide navigation tools that allow users to request information, with different levels of generalisation ('give') permitted in each dimension.

The Subject Index, SHIC, is the result of collaboration between several UK museums, whose joint working party produced a hierarchical index, with four first level facets, 'Personal Life', 'Domestic and Family Life', 'Community Life' and 'Working Life'. SHIC provides for multiple classifications of an item which is crucial for classifying media such as photographs where one scene can be seen in several different ways. Although we here use SHIC other classification systems are obviously applicable.

The temporal classification schema stores information relating artefacts to some existence in time. The schema consists of year time point identifiers that constitute the relevant temporal span. The schema also allows us to model periods as the aggregation of years - decades, centuries, or cultural notions like the Victorian Era, or the Depression. Temporal operators, such as Before, After, Circa, Overlaps act on both kinds of period.

With the geographical schema, the intention of the project was to generate a model capable of supporting the evolutionary nature of geographic entities (Beynon-Davies et al 1994). The geography of an area does not stay constant over time; new towns or streets are built and names are all susceptible to change. Thus the geographical and temporal index spaces in the SHA are linked. The basic geographical unit is the street, and street intersection is modelled as a relationship in the schema. (In the future, we also intend to include spatial coordinate data.) Information from four editions of Ordnance Survey maps between 1880 and 1994 are modelled; geographical terms have existence in a temporal period. Relationships exist to express a change in name of a geographical term, and allow for the removal or addition of attributes to a current definition. A user can request changes that have taken places in a given area within a timeframe.

Each of the index dimensions supports standard browsing activity; the conceptual dimension can be navigated by a hierarchical browser; the temporal dimension through a sliding bar mechanism; and the geographical dimension with an A to Z arrangement of street names. Information requests do not need to be limited to one particular dimension, but can have terms of reference in all of the three dimensions. For example, we could retrieve information on: Social Organisations associated with Pontypridd town centre during the Victorian era.

Higher Level Navigation

Modelling the structure of the different index dimensions, allows us to take advantage of the relationships in the classification schema to support higher level navigation tools than simple browsing. A problem in public access to information arranged according to standard subject indexes is that this may result in an uneven distribution of viewable information items over the different classification dimensions. There may be clusters of items at some parts of the schema, and sparse areas in other places. In such cases there may be additional problems for navigation through the information. Usually navigation takes place via the classification concepts (index terms), so that an unsuccessful strategy may result in few information items available for viewing; the user may be unaware that much more information is available "nearby". In our prototype, the set of historical photographs of Pontypridd from the PHCC archive are generally located around the town centre and the sites of heavy industry surrounding the town, and the subject classification employed had many sparse areas.

Consider an example from our current prototype, where a composite request for information using the SHIC term Retail Distribution and the geographical term High Street failed to yield any results (Tudhope et al forthcoming). The user is able to ask the system to generalise the unsuccessful request, to see if any information items have been indexed with 'close by' terms in the relevant dimensions. It turns out that there are four photographs of shops in an adjacent street. Thus the system can traverse relationships in the spatial dimension to assist a query.

Semantic Closeness

Such navigation tools require a distance measurement between terms in each of the index dimensions, which we refer to as 'semantic closeness', a measure of the strength of the relationship between two terms. For SHIC, it is calculated using a simple spreading activation algorithm, a traversal function over the underlying semantic net. Semantic closeness is a function of the traversals that must be undertaken to move from the position of the first term in the index space to the second term. The value ranges from 0 to 1. Each traversal diminishes the semantic closeness of the objects by a cost factor until 0 is reached. In addition, the cost of a traversal in a hierarchical classification is inversely proportional to the level of specialisation of terms involved. As object classes become more specialised the semantic differences between the sub-classes of the object or its instances should become less pronounced. (In the Linnaean classification system:

class/order/genus/species/sub-species

objects at the subspecies level can be considered closer than at the class level.)

A distance measurement offers the opportunity to address imprecise objects within information requests. By applying semantic closeness to generalise terms, notions such as 'around' Taff Street can be implemented.

Related Work

Other hypermedia base architectures have been influenced by a semantic modelling perspective (eg, Garzotto et al 1993, Schnase et al 1993). The generalisation navigation tool above has some similarities with the 'link-oriented' method for defining a collection of nodes that can be navigated by the user, described by Garzotto et al (1994). Both methods employ the semantic notion of typed links to capture important relationships in the informationbase, together with automatic methods of processing those relationships. Both methods assume the user has previously navigated to a particular place in the hypermedia application. For the link-oriented method, this starting place is itself a collection, to which a set of links is applied deriving a further collection of nodes connected to the first by the links specified. For example, from a collection of paintings it derives a collection of artists, and from there might derive a collection of the schools to which the artists belonged. With our best-fit generalisation, the starting place is a particular point in the N-dimensional classification schema. The semantic closeness function is used to produce a set of nearby nodes using the predefined links in the schema. Unlike the 'link-oriented' method, it does not currently allow a selection of particular links to traverse (which allows specific queries over the schema). However it does incorporate a media density measurement to yield terms where there are media items to view.

Navigation via Similarity

Navigation via similarity has been used in several different kinds of applications in information retrieval and hypermedia (Tudhope et al 1995). It places less emphasis on the user of the system being familiar with the nature of the classification systems involved. The user only has to ask "Find me media items similar to this one". The system can address the issues of what represents the media item and what therefore is to be considered similar. "With a distance measure, the inquirer needs less understanding of the underlying taxonomy" (Parunak 1993) . Of course, the similarity measure can only be with regard to the classification systems employed in the system. Parunak describes a measure of the similarity of two sets based around the relative number of elements in common. For any two sets A and B, the measure is

$$\mathrm{SetSim}(A,B) = \sqrt{\frac{|A \cap B|^2}{|A| * |B|}}$$

where $|A|$ is the number of elements in set A. This is a form of a standard similarity function (the cosine coefficient) in information retrieval, here applied to sets. It has the advantage that it can deal with multiple classifications of a media item (multiple SHIC terms for a photograph, for example). The measure produces a similarity value in the range 0 (disjoint sets) to 1 (identical sets).

This approach is sufficient in an application domain where there exists no relationships between index terms. In such situations the presence or absence of a term from a set denotes the presence or absence of that feature of the object. For a match, the set similarity function requires exact equivalence of terms. The work described here extends such similarity coefficients with the notion of semantic closeness, a distance measurement over the index space. As an illustrative example involving two index spaces, consider the following example of two photographs and the index terms assigned in the temporal and hierarchical SHIC classification:

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

Both photographs (Figures1-2) are of groups of people concerned 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 siblings 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.

In extending similarity coefficients to include non-binary distance measurements, we have to consider the possibility of multiple terms and uneven numbers of terms, as in our example. Different approaches are possible (see also Rada 1989). Our current method is to compare each member of each set of terms with all members of the other set and sum the maximum values recorded (see below). The result is normalised by the cardinality of the two sets.

	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

2.75 / (3 + 2) = 0.55

This method takes into account that different index terms may be contributed by different facets of a classification, and attempts where possible to compare like with like.

The temporal coefficient is a simple difference, normalised by the range, though we have experimented with proximity measures between temporal periods (Tudhope et al 1995). 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 high in other dimensions. More work is needed on implementation issues.

Figures 1-2 illustrate how this method can be used as a means of navigation, using our current prototype. Figure 1 shows that the user has navigated (using 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 for conceptual and temporal dimensions. Figure 2 shows the result - the Untouchables had the same SHIC index terms as the Thistles and are thus less interesting than the Spaniards. We would argue that an automatic similarity tool has some advantages where multiple index dimensions are involved, since interactive browsing or refining query parameters may become complicated for a non-specialist user. An example involving the spatial dimension is discussed in Tudhope et al (1995).

Numerical Taxonomy and Navigation

For hypermedia navigation over multiple dimensions, the field of numerical taxonomy is of particular interest. Numerical taxonomy (Sneath and Sokal 1973) is concerned with measures of taxonomic distance with a view to automatically deriving a classification of a set of objects (operational taxonomic units) according to given criteria (phenetic similarity). We however are less concerned with deriving a new classification (although that is an interesting possibility) than with integrating 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 high level navigation tools for an informationbase, which has previously been given classification schema for reasons practical, legal, or conventional.

The navigation method by media similarity described above has the drawback that each classification dimension (geographical, temporal and subject-based in the social history prototype) is treated separately and the intersection of the results returned. However the similarity functions employed are essentially the same as the similarity coefficients in numerical taxonomy. Work in numerical taxonomy on general similarity coefficients gives us the opportunity to define a unified similarity coefficient.

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

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

Sijk is the similarity between the two items on the kth dimension of indexing. Wijk 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 grave suspicion and Gower included it only as an option. However for our purposes, the weighting of dimensions allows users to assign 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 modified the coefficient to incorporate structured index spaces with semantic closeness as described above, and also included a comparison between temporal periods. Thus the same coefficient is applicable to quantitative spatial or temporal metrics, binary indexing of nodes with keywords (applies/not apply), and qualitative multistate or hierarchical classifications of information.

Features that we are looking to achieve include the ability to handle:

- multiple index dimensions
- with mixed data types
- user supplied weights for importance of dimensions
- non-exact matches
- missing terms or invalid comparisons
- multiple classifications in the same dimension
- hierarchical or other structured index spaces,
- and ability to reason over relationships between terms

In some circumstances neither browsing separately in each of several dimensions, nor precise queries will be the answer to a user's needs. A similarity measure along the lines above may be helpful.

Multiple Classifications

We tend to think of classification systems as given or deriving naturally from the world. However, work in the sociology of knowledge and science studies (Barnes 1994) shows the conventional nature of all classifications and sees them as socially and culturally located. Taxonomies often involve a hierarchical ordering of schema terms and carry with them implicit views of what is important. Hooper-Greenhill (1992) poses the questions:

> "If new taxonomies mean new ways of ordering and documenting collection, then do the existing ways in which collections are organised mean that taxonomies are socially constructed rather than being 'true' or 'rational'? Do the existing systems of classifications enable some ways of knowing but prevent others?...Taxonomies within the museum have not been considered in relation to the rational possibilities that they might enable or prevent. Classification within the museum has taken place within an ethos of obviousness. The selection and ordering processes of museums are rarely understood as historically and geographically specific, except at a very rudimentary level."

Feminist studies of science (eg, 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. This need for multiple perspectives has been echoed by workers in the museum profession (Besser 1991, McCorry 1993). One 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 presentation. Museum professionals have pointed out that the depth of detail and rather academic terms employed render it problematic for public use. One possibility is to include both the full SHIC schema for collection management and research-oriented use and also a simpler subset for general public use. The system could automatically translate between them. Other possibilities are to allow different, overlapping subject-indexes on the same data.

Future directions

A semantic modelling approach to hypermedia architecture offers the potential to develop higher level navigation tools that combine aspects of browsing and query modes of interaction and goes some way to meeting needs of users who do not have precisely defined purposes in accessing information, or who make requests in terms other than how the data is actually indexed. Such navigation tools appear less suited to the visitor wishing to get only a general impression of exhibits, spending only a few minutes at each, and who would be better served by simple navigation methods, such as basic browsing or the guided tour. They may be of benefit to the member of the public (or academic researcher) pursuing an interest or independent research, perhaps not very familiar with the classification systems involved, the kind of person likely to seek further information from the curator. They may be suitable to the educational/discovery annexes to main collections provided by some museums to allow interested members of the public to research issues related to the collection in more depth.

Representing the semantic relationships underlying conceptual terms in a separate index space offers the opportunity for a navigation tool (or query processor) to automatically traverse those relationships in order to give the user a wider field of view when interacting with the system. Many manual techniques employed with a thesaurus, such as broadening, narrowing searches, exploring related terms can form part of automatic tools. In the next stage of our research we wish to explore the potential of more elaborate models of relationships in the different index dimensions, drawing on ideas from thesaurus construction and Geographic Information Systems.

Bibliography

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

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

Bates M., Wilde D., Siegfried S. 1993. 'An analysis of the search terminology used by humanities scholars: The Getty online searching project, Report Number 1', The Library Quarterly 63(1): 1-39.

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.

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

Garzotto F., Mainetti, L., Paolini, P., 1993. 'HYPERMILANO: Hypermedia Navigation in a City Information Point'. Proc. International Conference on Museums and Interactive Multimedia (ICHIM'93). Cambridge. pp 95-104.

Garzotto F., Mainetti, L., Paolini, P., 1994. 'Adding Multiple Collections to the Dexter Model'. Proc ACM European Conference on Hypermedia Technology. Edinburgh. pp 70-81.

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

Hooper-Greenhill E. 1992. Museums and the Shaping of Knowledge. Routledge:London.

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.

McCorry H. 1993. 'Thesauri and Terminology Control'. MDA Information 1(1), 12-14.

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.

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.

Salton G. and Allen J. 1993. 'Selective Text Utilization and Text Traversal'. Proc. ACM Conference on Hypertext, Seattle. pp 131-144.

Schnase J., Leggett, J., Hicks D., Szabo R. '1993. Semantic Data Modelling of Hypermedia Associations', ACM Trans on Information Systems, 11(1), 27-50.

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.

Taylor C., Tudhope D., Beynon-Davies P. 1993. 'Hypermedia and the MDA Standards: Crossing the Divide', MDA Information 1(1), 2-8.

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.

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