PERGAMON

Information Processing and Management 35 (1999) 401–420

INFORMATION PROCESSING & MANAGEMENT

Visualising semantic spaces and author co-citation networks in digital libraries ☆

Chaomei Chen*

Department of Information Systems and Computing, Brunel University, Uxbridge UB8 3PH, UK

Abstract

This paper describes the development and application of visualisation techniques for users to access and explore information in a digital library effectively and intuitively. Salient semantic structures and citation patterns are extracted from several collections of documents, including the ACM SIGCHI Conference Proceedings (1995–1997) and ACM Hypertext Conference Proceedings (1987–1998), using Latent Semantic Indexing and Pathfinder Network Scaling. The unique spatial metaphor leads to a natural combination of search and navigation within the same semantic space in a 3-dimensional virtual world. Author co-citation patterns are visualised through a number of author co-citation maps in attempts to reveal the structure of the hypertext, including an overall co-citation map of 367 authors and three periodical maps. These maps highlight predominant research areas in the field. This approach provides a means for transcending the boundaries of collections of documents and visualising more profound patterns in terms of semantic structures and co-citation networks.

1. Introduction

Digital libraries have drawn attention to a wide variety of perspectives ranging from technical to social dimensions (Fox, Akscyn, Furuta & Leggett, 1995; Fox & Marchionini, 1998). Advances in digital libraries have highlighted several challenging issues that must be addressed. For example, how should the usability of an evolving digital library be maintained?

* *E-mail address:* Chaomei.Chen@brunel.ac.uk (C. Chen)

☆ All the figures used in this article are available in colour on the WWW at: http://purl.lib.vt.edu/dlib/pubs/IPM1999Chen

0306-4573/99/$ - see front matter
PII: S0306-4573(98)00068-5

Are there more intuitive and effective ways of retrieving and exploring information in a digital library? What is the potential of virtual reality techniques in supporting more engaging and intuitive interactions between users and digital libraries?

In traditional libraries, books on similar topics are likely to be placed near each other. A similar spatial metaphor would be desirable in digital libraries. However, finding an intuitive and meaningful way of organising information in a digital library is not a trivial task. Advances in information visualisation and automatic construction of hypertext provide opportunities to improve the accessibility of information in digital libraries. Rao et al. (1995) summarise a collection of information visualisation tools developed at Xerox PARC and suggest that these tools should be readily accessible to users of digital libraries in order to foster rich interaction between users and information in digital libraries.

Spatial metaphors are by far the most popular design principle for information visualisation, for example, SemNet (Fairchild, Poltrock & Furnas, 1988), BEAD (Chalmers, 1992), LyberWorld (Hemmje, Kunkel & Willett, 1994), Starfield (Ahlberg & Shneiderman, 1994), VR–VIBE (Benford, Snowdon, Greenhalgh, Ingram, Knox & Brown, 1995), and SPIRE (Hetzler, Harris, Havre & Whitney, 1998a). In parallel, artificial neural network techniques have been used to generate self-organised feature maps to facilitate information retrieval (Lin, 1997; Lin, Soergel & Marchionini, 1991).

In this paper, we describe the design and use of several visualisation tools to help users to access and explore information effectively and intuitively in a digital library. Salient semantic structures and co-citation patterns are extracted from collections of ACM publications in the fields of Human-Computer Interaction (HCI) and Hypertext. These structures can be visualised such that users are able to search as well as navigate within the same virtual world. Co-citation patterns may reveal the structure of a field. Author co-citation maps can be used to highlight threads of topics and predominant research areas in hypertext.

The rest of this paper is organised as follows. First, we present a review of the literature concerning information visualisation and digital libraries. Then, we introduce fundamental techniques such as Latent Semantic Indexing (LSI) and Pathfinder Network Scaling. Several examples of structural visualisation are included to illustrate the strengths of these techniques in helping users to understand information in a digital library. We also analyse the structure of the hypertext as a research field based on a number of author co-citation maps. Finally, we discuss the findings and implications of this approach on the development of digital libraries in general.

2. Related work

Envision is one of the pioneering multimedia digital libraries of computer science literature to meet the needs of computer science researchers, teachers, and students at all levels of expertise (Fox, Hix, Nowell, Brueni, Wake & Rao, 1993; Heath, Hix, Nowell, Wake, Averboch, Labow, Guyer, Brueni, France, Dalal & Fox, 1995). Envision was designed with full text searching and full content retrieval capabilities. In particular, search results from the Envision database can be visualised as a matrix of icons (Nowell, France, Hix, Heath & Fox, 1996). These matrices of icons enable users to assess the results of a search graphically based

on a variety of document attributes. For example, a year-by-relevance layout can reveal peaks and valleys of document-query relevance at a glance. The design of the Envision user interface used various colours and shapes to convey important characteristics of documents. For example, the colour of an icon indicates the degree of relevance: the most relevant documents would be in orange; documents marked by users as useful would be in red; and documents marked by users as not useful would be in white. A star icon indicates that the document is in the top 35% of relevance range; a diamond is in the next 35% of the range; and a triangle denotes a document in the bottom 30% of the range. The notion of peaks and valleys of relevance has also been used in more recent visualisation systems, especially ones with various spatial metaphors.

Spatial metaphors are among the most popular design options in information visualisation. As mentioned earlier, many influential information visualisation systems are based on a spatial metaphor. In this paper, we will concentrate on closely related works in this rapidly growing field.

LyberWorld (Hemmje, Kunkel & Willett, 1994) is a well-known example of applying metaphors of spatial navigation in abstract information spaces. The design of LyberWorld focuses on a network representation of an information space. Such networks consist of two types of nodes: document nodes and term nodes, and three types of links: document-term links, term-term links, and document-document links. Navigation in LyberWorld relies on the concepts of content and context spaces. The content space is the entire search space, whereas the context space is the sub-space where the user has visited. Instead of using computationally expensive graph layout algorithms to map the content network into a 3-dimensional space, LyberWorld derives hierarchies from the network representation of the content space and maps these hierarchies onto a visualisation tool, called NavigationCones. NavigationCones illustrate the navigation history of a user in the content space along content-oriented search paths, which connect different documents in the content space.

More recently, Pacific Northwest National Laboratory in the USA has developed a suite of information visualisation tools, known as Spatial Paradigm for Information Retrieval and Exploration (SPIRE) (Hetzler, Harris, Havre & Whitney, 1998a; Hetzler, Whitney, Martucci & Thomas, 1998b). The design is based on an intuitive spatial metaphor. SPIRE aims to enable users with little special domain knowledge to use these visualisation tools in order to explore and discover topical themes and other underlying relationships among documents. In particular, SPIRE supports two types of visualisation: Galaxies and Themescape. Galaxies provides an overview of the entire set of documents; documents are represented as a galaxy of star clusters in the night sky. Like many other visualisation systems using spatial metaphors, similar documents are represented by stars near each other in the galaxy, whereas documents about different topics are separated by large distances. Themescape visualises the presence of specific themes across different documents as mountains in a relief map of natural terrain. The height of a peak indicates the relative strength of a given topic in the document set. Similar themes are grouped into neighbourhoods, whereas unrelated themes are separated by large distances. One may use these visualisation tools to identify unanticipated relationships and examine changes in topics over time. One may also examine these thematic spaces over time so as to understand vast interrelated dynamic changes simply not possible to detect using traditional approaches. Orendorf and Kacmar (1996) also describe a spatial approach to

organising digital libraries. However, their work took advantage of geographical layout in their organisation. Structuring abstract digital documents in general presents a challenging issue and our work aims to tackle this issue.

Classic vector space models and probabilistic retrieval models are by far the most popular options in visualising abstract information spaces for retrieval and exploration. An interesting alternative, lexical chaining, is described by Green (1998) in an attempt to deal with two major linguistic factors that may undermine the effectiveness of traditional information retrieval models, namely, synonymy and polysemy. Both synonymy and polysemy can cause problems because of the ambiguity of the meaning of a particular term. A lexical chain consists of a sequence of semantically related words in the text. For example, if the words 'apple' and 'fruit' appear in a document, then both words should be included in a chain because we know that apple and fruit are related concepts. Lexical chains in text can be recovered using any lexical resource, such as Roget's Thesaurus (Chapman, 1992) and the WordNet database (Beckwith, Fellbaum, Gross & Miller, 1991), that specifies how words are connected according to their meanings. The similarity between two documents, therefore, can be measured by the similarity between lexical chains associated with these documents. It appears to be a promising alternative to the existing information visualisation paradigms.

A unifying framework, Generalised Similarity Analysis (GSA), was developed in earlier work for structuring and visualising complex information spaces (Chen, 1997, 1998b). GSA includes a set of modelling and visualisation tools to uncover a variety of structures inherited in a collection of documents, for example, content-based similarity, cross reference-based similarity, and usage pattern-based similarity. A key element in our approach is the use of Pathfinder network scaling techniques (Schvaneveldt, Durso & Dearholt, 1989). The role of Pathfinder network scaling is to extract the most salient links and eliminate redundant or counter-intuitive links. Pathfinder has some desirable features over techniques such as multidimensional scaling (MDS).

In earlier work (Chen, 1997), inter-document similarities were computed based on the classic vector space model with $tf \times idf$ weighting (Salton, Allan & Buckley, 1994). However, the vector space model is subject to an assumption that terms used in document vectors are independent. It has been realised that this assumption may over-simplify the interrelationship between the use of a term and its context. In this paper, we incorporate Latent Semantic Indexing (LSI) (Deerwester, Dumais, Landauer, Furnas & Harshman, 1990) into our framework in order to obtain a more accurate picture for underlying semantic structures.

3. Visualising semantic spaces

Two structuring and modelling techniques are crucial to our work: Latent Semantic Indexing (LSI) and Pathfinder Network Scaling. In this section, we first describe the data sets used in our subsequent analysis and visualisation. Then we explain the basic concepts associated with LSI and Pathfinder techniques and their roles in our approach.

3.1. Document collections

Our work focuses on two major sources: the ACM SIGCHI conference series and the ACM

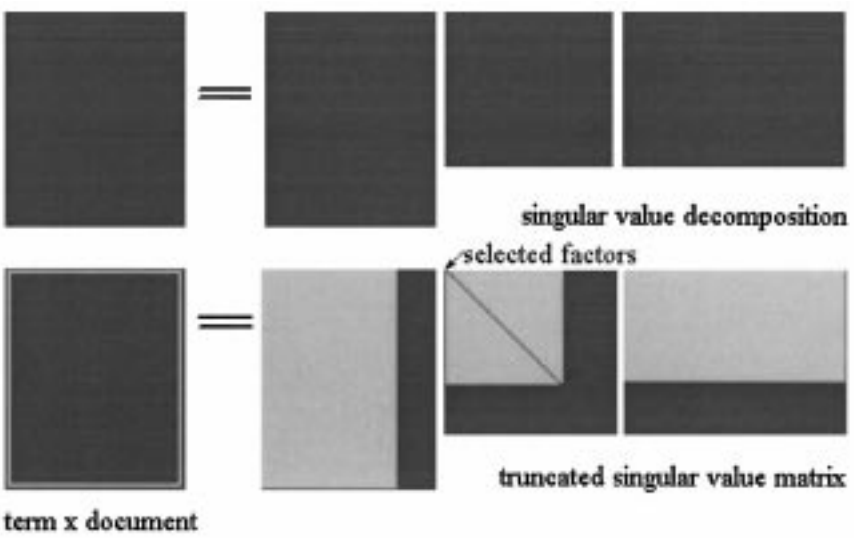

Fig. 1. Singular Value Decomposition (SVD) and a truncated SVD matrix.

Hypertext conference series. The ACM SIGCHI collection contains 169 papers published in three recent conference proceedings, namely CHI '95[1], CHI '96[2] and CHI '97[3]. The ACM Hypertext collection includes all the papers published in the ACM Hypertext Conference Proceedings (1987–1998). We also analyse the CACM collection with LSI, which is well understood by the information retrieval community, but in this paper our focus is on the first two collections of documents.

3.2. Latent semantic indexing

The meaning of a word can best be understood within its context. Latent Semantic Indexing (LSI) is designed to overcome the so-called vocabulary mismatch problem in information retrieval (Deerwester, Dumais, Landauer, Furnas & Harshman, 1990; Dumais, 1995). LSI is based on the assumption that the underlying semantic structure in a collection of documents is often obscured by words chosen in a retrieval process, and that the latent semantic structure can be uncovered with statistical techniques.

In LSI, a semantic space is constructed based on a large matrix of term-document association observations. LSI uses a mathematical technique called Singular Value Decomposition (SVD). One can approximate the original term × document matrix with a truncated SVD matrix. A compelling claim is that, using LSI, one can retrieve documents

[1] http://www.acm.org/sigchi/chi95/
[2] http://www.acm.org/sigchi/chi96/
[3] http://www.acm.org/sigchi/chi97/

relevant to a query but without any words in common with the query (Deerwester, Dumais, Landauer, Furnas & Harshman, 1990; Dumais, 1995). A proper truncation can capture the essential structure in the original data so that the recall and precision of information retrieval can be improved (Fig. 1).

In general, LSI can reduce the dimensionality of a data set considerably. In Fig. 2(a) there is a 2D-scatter plot of the ACM SIGCHI collection, containing 169 documents. This data set appears to be relatively well captured by the first two dimensions. In contrast, Fig. 2(b) shows a scatter plot of the CACM collection, containing more than 3200 documents. A large number of documents are plotted near to the origin, suggesting that their positions in the semantic space cannot be adequately represented within a 2D space.

3.3. Pathfinder network scaling

Pathfinder network scaling is a structural modelling technique originally developed by (Schvaneveldt, Durso & Dearholt, 1989) for the analysis of proximity data in psychology. It simplifies a complex network representation of proximity data to a much more concise and meaningful network – only the most important links in the network are preserved. The resultant networks are called Pathfinder networks (PFNETs).

Pathfinder relies on the so-called triangle inequality to eliminate redundant or counter-intuitive links. This process is called Pathfinder network scaling. The weight of a path is defined based on the Minkowski metric. Given a link in the network connecting nodes n_x and n_y, if an alternative path in the network also connects the two nodes but the weight is greater, then the interrelationship between the two nodes is better captured by the alternative path. This particular link therefore becomes redundant or even counter-intuitive and should be pruned from the network.

The topology of a PFNET is determined by two parameters r and q and the resultant Pathfinder network is denoted as PFNET(r, q). The weight of a path is defined based on the Minkowski metric with the r-parameter. The q-parameter specifies that the triangle inequality must be maintained

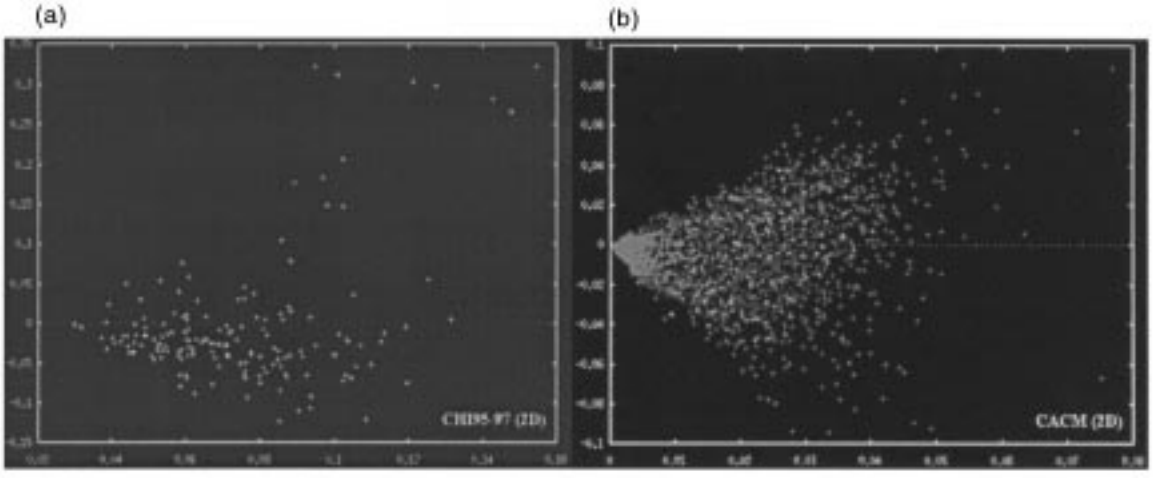

Fig. 2. (a) Scatter plot (CHI) and (b) Scatter plot (the CACM collection).

against all the alternative paths with up to q links connecting nodes n_1 and n_k:

$$w_{n_1 n_k} \leq \left(\sum_{i=1}^{k-1} w^r_{n_i n_{i+1}}\right)^{\frac{1}{r}} \forall\, k = 2,3,\ldots,q$$

For a network with N nodes, the maximum value of the q-parameter is $N - 1$. PFNET($r = \infty$, $q = N - 1$) consists of the least number of links. In such a network, each path is a minimum-cost path. If there are more paths than one connecting the same pair of nodes, these paths must have the same weight. We normally impose the tightest triangle inequality ($q = N - 1$) in order to achieve a concise Pathfinder network for visualisation purposes. In these cases, the triangle inequality must be maintained throughout the entire network.

A Pathfinder network can be generated based on an existing minimal spanning tree (MST) of the original network by including additional links, provided new links do not violate the triangle inequality. In fact, the minimum-cost Pathfinder network (MCN) is the set union of all the possible MSTs. Therefore, the structure of an MCN is unique for each original proximity network. Our software allows us to choose an MST instead of a PFNET to represent a large network.

Fig. 3 illustrates how the triangle inequality filter works and how its outcome should be interpreted. Suppose there are three papers: A, B, and C. Paper A describes LSI. Paper B is about information visualisation. Paper C applies LSI to an information visualisation design. The relationship between papers A and B is established by the content of Paper C. Therefore the path along links y and z reflects the nature of this relationship more profoundly than link x does. Link x becomes redundant and should be removed.

Graphical representations of Pathfinder networks are generated using force-directed graph-drawing algorithms (Kamada & Kawai, 1989; Fruchterman & Reingold, 1991). These algorithms are increasingly popular in information visualisation because they tend to lay out similar nodes near each other, and put dissimilar ones farther away from each other. Similar algorithms are used by BEAD (Chalmers, 1992) and SPIRE (Hetzler, Harris, Havre & Whitney, 1998a).

The value of Pathfinder in our work is its ability to reduce the number of links in a meaningful way. Pathfinder network scaling usually results in a concise representation of

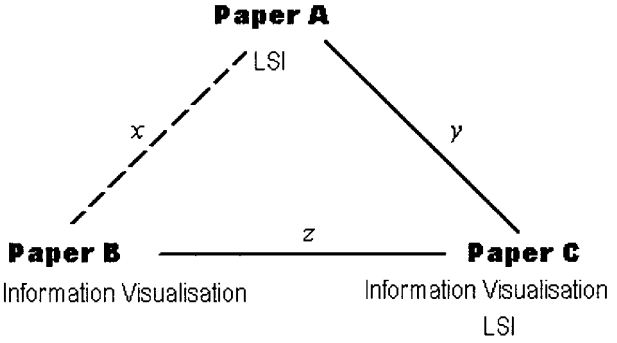

Fig. 3. Triangle inequality: links y and z are more fundamental than x.

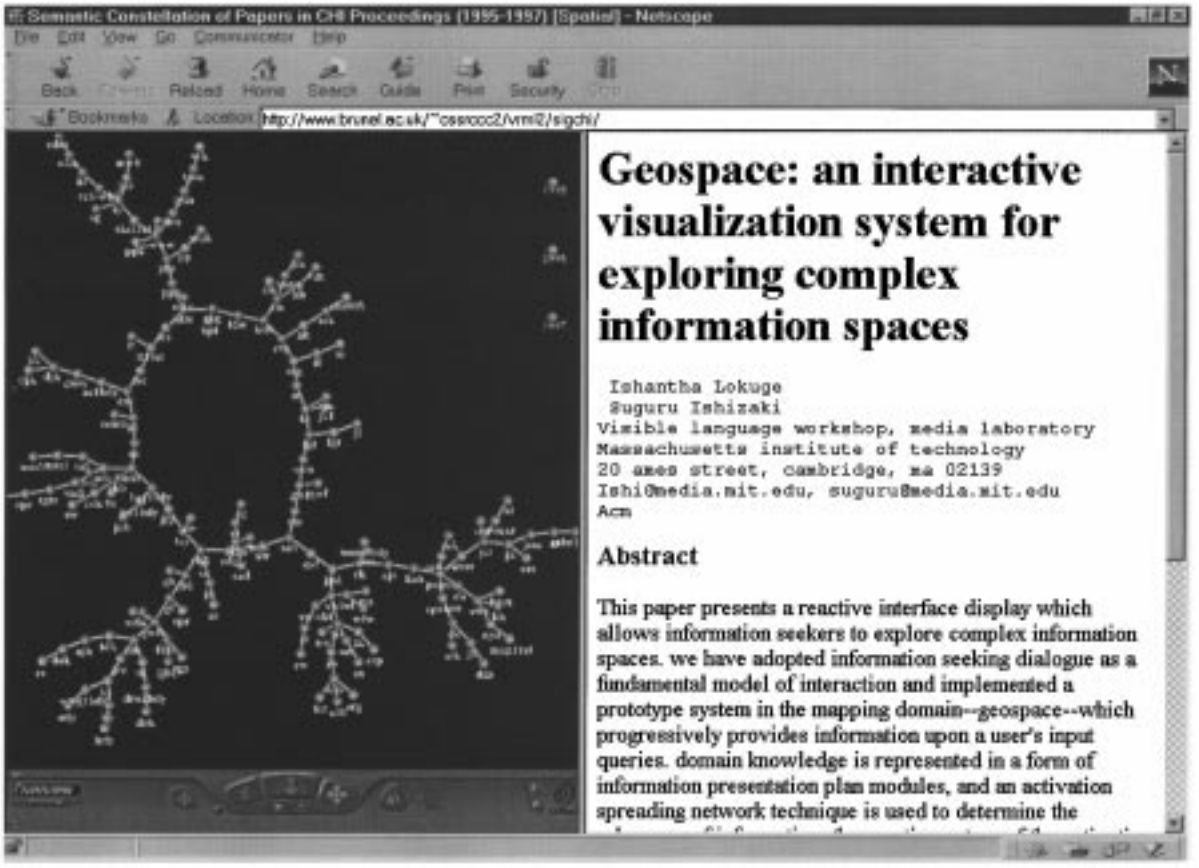

Fig. 4. The user interface can easily fit into users' browsers.

clarified proximity patterns, which is a desirable feature for visualising a complex structure. Pathfinder networks provide not only a fuller representation of the salient semantic structures than minimal spanning trees, but also a more accurate representation of local structures than multidimensional scaling techniques.

3.4. Exploring a semantic space in virtual worlds

A document-document similarity matrix is generated by LSI using log entropy term weighting (Deerwester, Dumais, Landauer, Furnas & Harshman, 1990). Indexing is based on the title, authors' affiliations, abstract, and list of keywords provided with each document. This similarity matrix is submitted to Pathfinder network scaling and is subsequently rendered as a virtual-reality model in a 3-dimensional space.

Table 1 summarises how various objects in a digital library are rendered. For example, a document is visualised as a sphere in the virtual world; the year of publication or the source of literature is indicated by the colour of the sphere. Links in the underlying Pathfinder networks are depicted as cylinders. If users click on a document sphere, the content of the document will

Table 2
Spearman correlation coefficients between similarity ranking in LSI and spanning distance in the Pathfinder network (one-tailed significance). DID: Document ID

Query	Spearman	Sig.	Query	Spearman	Sig.	Query	Spearman	Sig.
visualization www	$r=0.816$	$p=0.002$	spatial map	$r=0.467$	$p=0.87$	digital library	$r=0.043$	$p=0.453$
DID	Similarity	Distance	DID	Similarity	Distance	DID	Similarity	Distance
20	0.529833	0	45	0.724254	0	18	0.479292	0
168	0.411085	1	70	0.369563	8	34	0.432195	19
161	0.405141	2	164	0.361105	1	12	0.392344	20
154	0.363778	2	119	0.260759	2	143	0.365243	11
147	0.357083	9	79	0.252059	8	4	0.343401	15
44	0.349539	2	38	0.248563	4	46	0.282718	12
32	0.311178	4	71	0.149930	8	148	0.263064	15
122	0.275670	7	101	0.140158	6	54	0.240167	18
140	0.260205	3	160	0.138844	17	84	0.192334	17
76	0.245697	15	22	0.124602	6	31	0.191008	12

be downloaded and displayed on their WWW browsers, regardless of whether these documents are stored locally or remotely.

Users are able to explore the virtual structure in a variety of ways, for example, by zooming in and out in the virtual world. In this spatial metaphor, the results of a search are natually integrated into the scene. As shown in Figs. 5 and 7, the results from LSI are superimposed over the global semantic map to mark the locations of documents relevant to the search. The magnitude of query-document relevance is indicated by the height of a spike raising from the sphere of the document.

Fig. 4 is a screen dump that shows a semantic structure extracted from the ACM SIGCHI collection (1995–1997)[4]. Each sphere in the virtual world is labelled with a unique ID number. The colour of a sphere indicates the year of publication: light blue is for 1995, light green for 1996, and light red for 1997. When the mouse cursor moves over a sphere in the structure, the title of the corresponding document will be displayed to the user. If the user clicks on the sphere, the content of the document will be displayed in a frame next to the frame containing the virtual world.

The overall structure consists of a central ring and a number of associated branches. Examining the contents of documents in the semantic structure reveals that each branch contains documents similar to each other. The top branch, for example, contains documents on information visualisation, the WWW, and related issues. Documents on the centre ring appear to be more generic than leaf-documents of a branch. Since only the shortest paths between any two documents are included in the semantic structure, the central ring is essential in

[4] http://www.brunel.ac.uk/~cssrccc2/vrml/acm/spatial/spatial2.html

establishing connections between documents across different branches. This pattern also is echoed in author co-citation maps.

The use of force-directed graph drawing algorithms usually provides an optimal spatial layout. In our study, the convergence of a spatial configuration to the underlying semantic structure is achieved if the stress is less than 0.005. Spearman's correlation coefficients between document-query similarity ranking in LSI and the spanning distance in the Pathfinder network are illustrated with examples based on three search queries (see Table 2).

A large, statistically significant correlation was found with the search query on visualization and *WWW* ($r = 0.816$, $P = 0.002$); a medium-sized correlation was found with the search on *spatial map* ($r = 0.467$, $P = 0.087$); but no correlation was found with the search on *digital library*. It appears that visualisation and the WWW are well-represented topics in CHI conferences, whereas the topic of digital libraries is not. Nevertheless, a thorough investigation is required to fully understand the profound relationship between a semantic space and its spatial network representation.

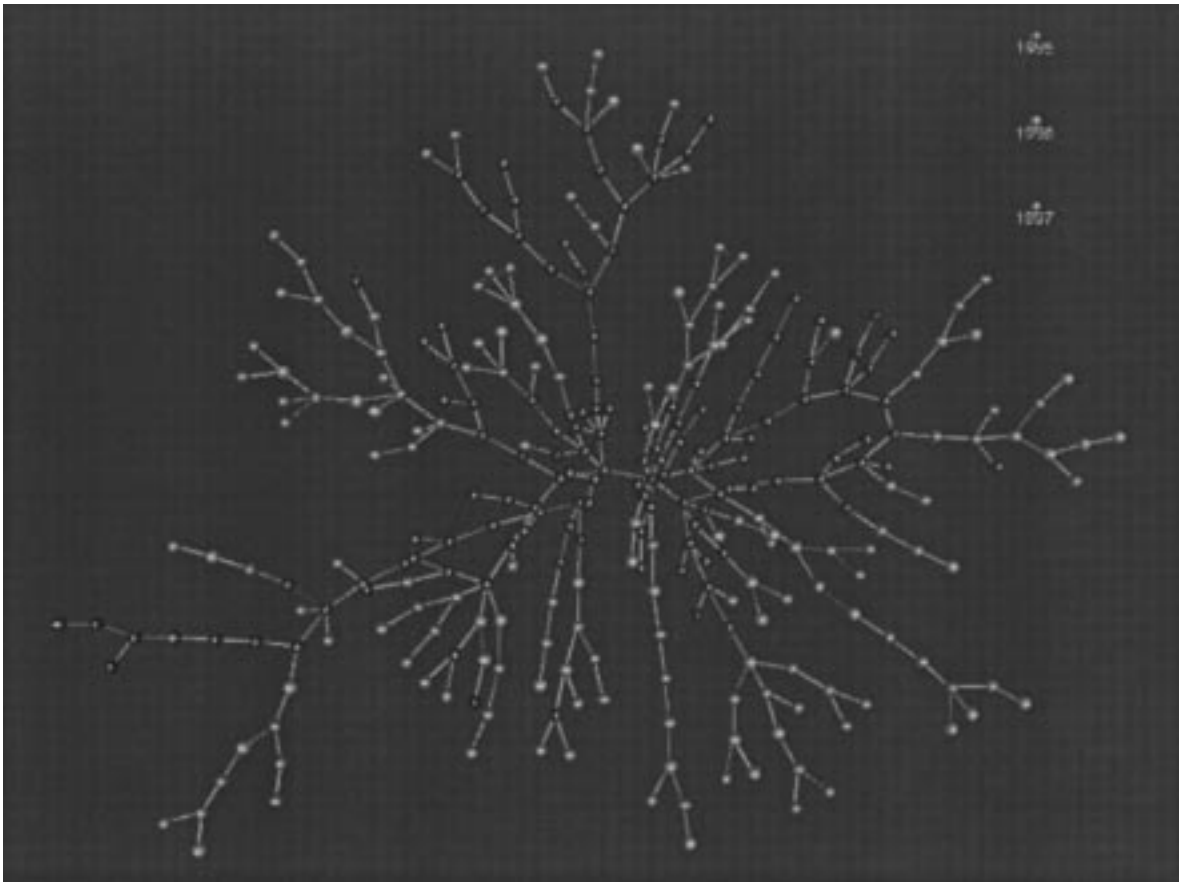

Fig. 6. A total of 304 papers from three sources are merged into the same coherent semantic space in VRML, including 169 CHI papers (light blue, light green, and purple), 127 ACM HTC papers (red) and panels and the author's own 8 papers (blue).

In Fig. 5, locations of the top 10 documents in response to the search for *digital library* and *spatial map* are marked in the semantic space by spikes. The height of each spike indicates the strength of the corresponding query-document similarity. For example, the best result for search *spatial map* (72%) is the highest spike at the far side of the scene; the document is entitled "Multimodal Interfaces for Dynamic Interactive Maps".

A heterogeneous virtual structure is shown in Fig. 6, which is derived from three different papers: CHI papers (light blue: 1995; light green: 1996; and light red: 1997); the ACM Hypertext Compendium (Akscyn, 1991) (red); and the author's previous publications (dark blue). Coloured maps highlight trends in various research areas and their interrelationships. Such virtual structures would allow users to access documents across different online resources, while the original resources remain intact. One may build and maintain a personalised digital library with such a self-organised virtual structure.

4. Visualising predominant research areas in hypertext

Citation analysis provides an effective way to characterise the impact of researchers' publications in a field (Garfield, 1994). In this section, we visualise author co-citation patterns in an attempt to reveal topical trends and predominant research areas in the field of hypertext. Once again, Pathfinder network scaling plays an important role in visualising co-citation patterns.

4.1. Author co-citation analysis

In this study, we are interested in predominant research areas in the field of hypertext and how author co-citation maps can provide a means of visualising and organising the literature in a more intuitive way. Our focus is on themes and research areas that may run across individual documents in the literature.

The Atlas of Science was constructed on the basis of journal co-citation patterns in biochemistry and molecular biology within a one-year interval (ISI, 1981). Distinct clusters of articles (102), called research front specialties, were identified as a snapshot of the field. More recently, ISI developed Sci–Map software for users to navigate the citation network (Small, 1994, 1997). Unfortunately, the work at ISI largely remains unknown to the hypertext and digital library communities, for whom a major concern is to make the literature of a subject domain widely accessible.

Author co-citation analysis (ACA) focuses on interrelationships among authors in the literature instead of individual publications. White and McCain (1998) present an extensive author co-citation analysis of information science, including a factor analysis, based on publications in 12 key journals in information science over a 23-year period (1972–1995). The top 120 authors, ranked by citation counts, were included in their analysis. Several maps were generated for the top 100 authors using multidimensional scaling (MDS). Two major research areas were identified within information science: experimental retrieval and citation analysis. Remarkably, these two areas have little overlap in terms of their memberships.

Author co-citation is a rigorous grouping principle. Co-citation patterns are reinforced by the views of researchers expressed in their publications (White & McCain, 1998). Visualising such patterns should lead to more insights into how a field like hypertext is structured.

A number of author co-citation analyses have used MDS to visualise the co-citation patterns. However, this option is often limited by the capacity of MDS routines implemented in statistical packages such as SPSS. In fact, White and McCain (1998) had to limit their visualisation to the top 100 authors because of such constraints. In this study, we are able to map 367 authors without imposing a hard cut-off threshold.

4.2. Author co-citation maps

The structure of the literature of hypertext, more precisely, a snapshot of the literature, is extracted from the entire ACM Hypertext collection (1987–1998) using techniques such as LSI and Pathfinder network scaling (Chen, 1998a, 1998b; Chen & Czerwinski, 1998). This structure is visualised in VRML 2.0 according to author co-citation patterns computed from the bibliography of each paper in the conference series (1989–1998). To be included in the co-citation analysis, authors must have five or more citations in the ACM Hypertext collection. Three hundred and sixty-seven authors were selected based on this criterion. In order to identify emergent trends over the last decade, the conference series was divided into three sub-periods. Each sub-period consists of three consecutive conferences (Table 3). The number of authors included in each sub-period is 196, 195 and 195, respectively.

Major research areas in hypertext are identified in a factor analysis. SPSS for Unix Release 6.1 was used because of the computation-intensive nature of the analysis. The raw co-citation counts were transformed into Pearson's correlation coefficients. Pearson's r was used as a measure of similarity between author pairs, because, according to (White & McCain, 1998), it registers the likeness in shape of their co-citation count profiles over all other authors in the set.

A number of co-citation maps were automatically generated with various hypertext reference links embedded. The name of an author in the co-citation map is linked to a file in which bibliographical details are available. These links are useful for us to examine bibliographical details and determine the nature of a research area. Co-citation patterns in the three sub-periods were visualised with the same tools.

Table 3
Three sub-periods of the series

Sub-Periods	I (1989–1991)	II (1992–1994)	III (1996–1998)
Number of authors	196	195	195
Conferences	Hypertext '89	ECHT '92	Hypertext '96
	ECHT '90	Hypertext '93	Hypertext '97
	Hypertext '91	ECHT '94	Hypertext '98

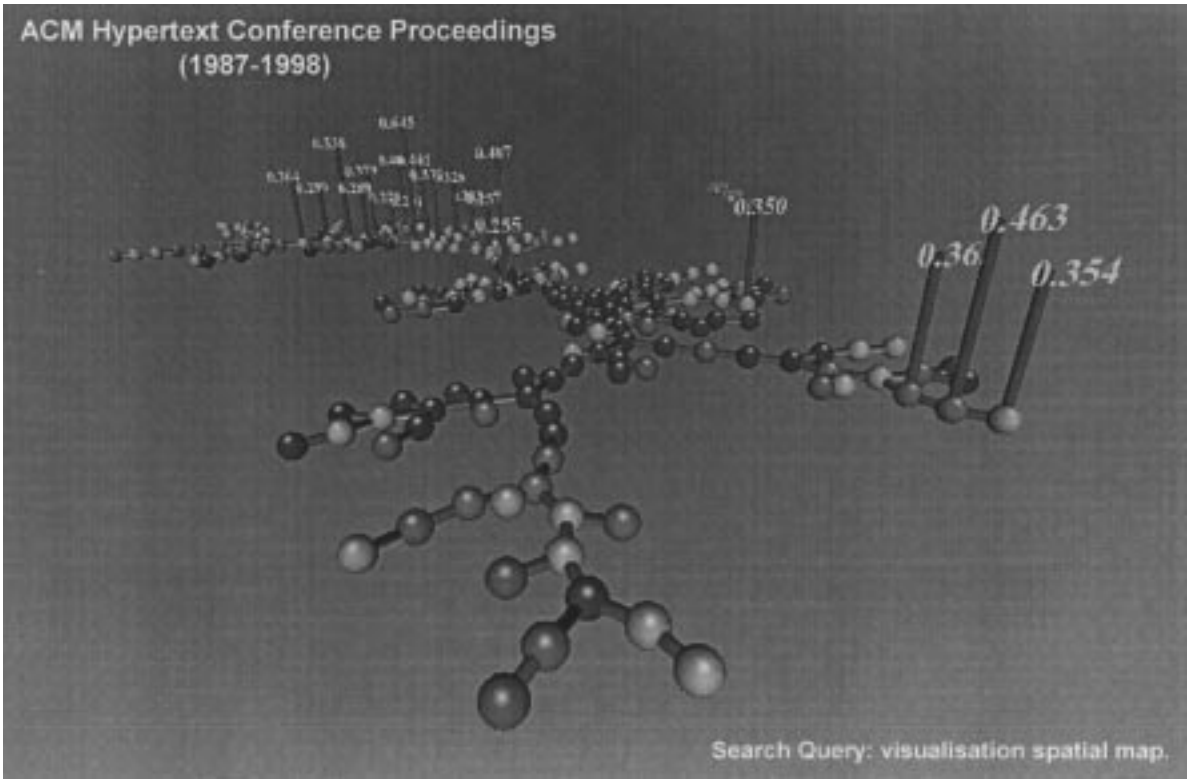

Fig. 7. A content-similarity map of the ACM Hypertext conference proceedings (1987–1998).

4.3. Results

4.3.1. The hypertext literature

The ACM Hypertext collection, containing 269 full papers published between 1987 and 1998, is visualised in VRML 2.0 (see Fig. 7). Each sphere in the scene represents a paper in the collection. The colour of a sphere indicates the 'age' of the paper: older papers are darker and younger ones are lighter. A group of closely connected light-coloured spheres may indicate an emergent research subject, whereas a concentration of darker spheres may correspond to topics found in the early years of hypertext research.

In Fig. 7, the results of a search using LSI are also displayed in the scene. The search query contains three terms: visualisation, *spatial*, and *map*. The higher a red spike on a sphere, the more similar the document is to the query. In Fig. 7, these spikes form two clusters: one is relatively new (at the far end) with many recently published documents (lighter coloured ones) and the other (at the near end) with papers published some years ago.

The centre of the semantic structure is occupied by a large number of dark spheres—these are papers published in the early years of ACM Hypertext conference series. This pattern suggests that some papers play a more important role than others do in connecting papers across different areas. This pattern becomes more apparent in author co-citation maps.

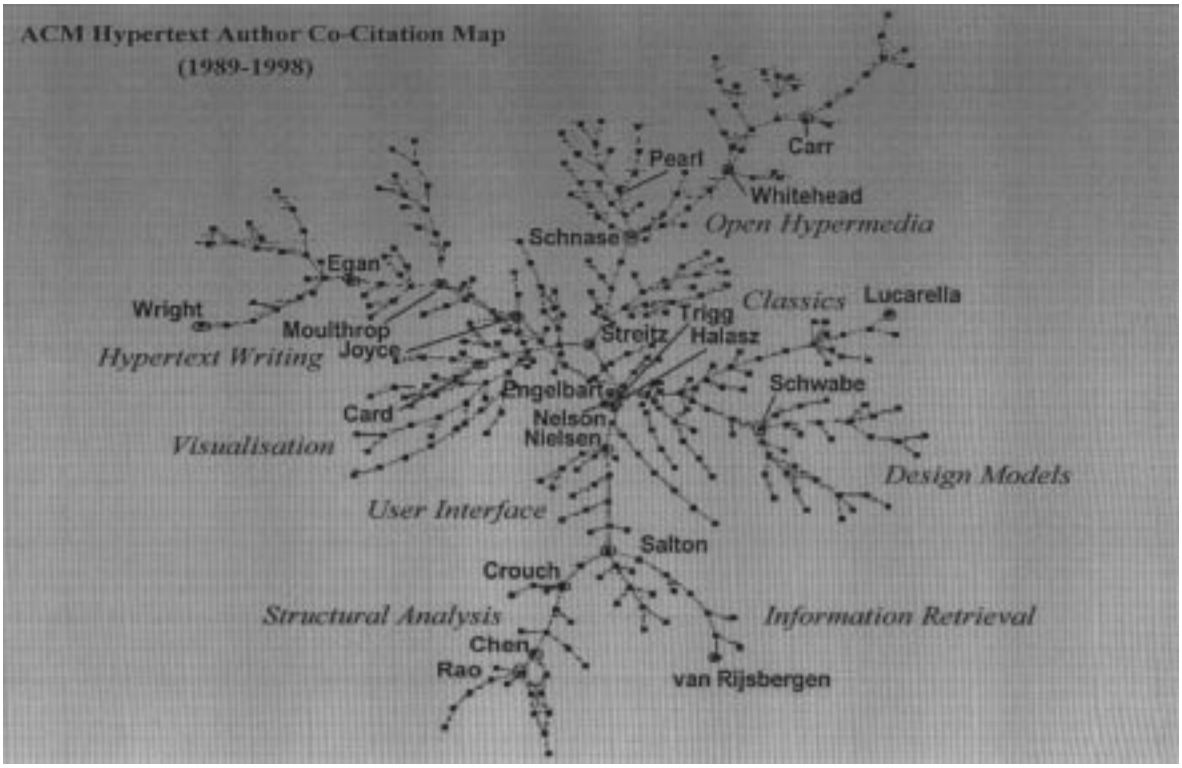

Fig. 8. The overall author co-citation map and predominant research areas based on the ACM Hypertext collection (1989–1998).

4.3.2. Factor analysis

Thirty-nine factors were extracted by factor analysis, explaining 87.8% of the variance. The top four factors alone explain 52.1% of the variance. Each extracted factor corresponds to a research area in hypertext. These research areas are categorised based on the profiles of the top 20 authors, ranked by their factor loadings (Table 4). The research areas corresponding to the first four factors are identified as (1) classic hypertext, (2) information retrieval, (3) graphical user interfaces and information visualisation, and (4) links and linking mechanisms.

4.3.3. Overall author co-citation map

The overall author co-citation map is annotated in Fig. 8. The position of an author is significant if it connects many authors in different areas of the map. An author at a point where several branches join is likely to be a highly predominant researcher.

Indeed, names such as *Engelbart*, *Nelson*, *Halasz*, *Trigg* and *Streitz* appear at such positions. They play an essential role in connecting several groups of authors together. For example, if we remove *Streitz* from the map, much of the insights into the local structure of the author co-citation network would be lost. This explanation also reinforces the semantics of the triangle inequality in Pathfinder network scaling. The following major research areas are identified on the map based on the reputations of authors at branching nodes:

1. Classics
2. Design Models
3. Hypertext Writing
4. Information Retrieval
5. Open Hypermedia
6. Information Visualisation
7. Structural Analysis
8. User Interface

For example, in Fig. 8 *Salton* is at a branching node which connects three branches to the main body of the map. Salton's work in information retrieval and automated hypertext generation is well known in both information science and hypertext. In contrast, leaf nodes indicate authors who might have unique positions in the field. For example, *van Rijsbergen*, another famous name in information retrieval, is at the leaf node of a branch underneath *Salton*. Therefore this branch is associated with an area in hypertext that has a strong link with information retrieval. This map is incorporated into an interactive user interface on the WWW. Users are able to explore these maps in virtual reality and access the content of a document directly.

4.3.4. Periodical author co-citation maps

Three periodical author co-citation maps are generated in attempts to capture the evolution of the hypertext field since 1989 (see Fig. 9). They are Map A (1989–1991), Map B (1992–1994), and Map C (1996–1998). In Map A, predominant authors are those associated with classic hypertext systems, such as NoteCards, Intermedia, KMS, and Microcosm. Map B corresponds to the second period (1992–1994), and highlights the role of SEPIA, a successful system developed at GMD. Six members of GMD occupied the centre of the map. In the overall author co-citation map, Open Hypermedia was identified as a predominant research area. In the second period, *Pearl* became the branching node of the Microcosm group. *Leggett* and the *Pearl*'s branch joined the mainstream of the map. This movement indicated emergent research in open hypermedia. This example shows that author co-citation maps can be used to highlight the impact of particular aspects of one's work.

In the most recent period (1996–1998), we expected to identify the impact of the World-Wide Web, given the rapid advances of the Web and the growth of the WWW as a research field. However, our initial analysis found little evidence of the Web impact, which could also be interpreted as the lack of overlaps between the WWW and Hypertext communities. Further work is needed in this area. For example, visualising the WWW literature along with the Hypertext literature may lead to more insights into the cross-domain interrelationships.

5. Conclusion

We have described a novel approach to the development of intuitive and engaging digital

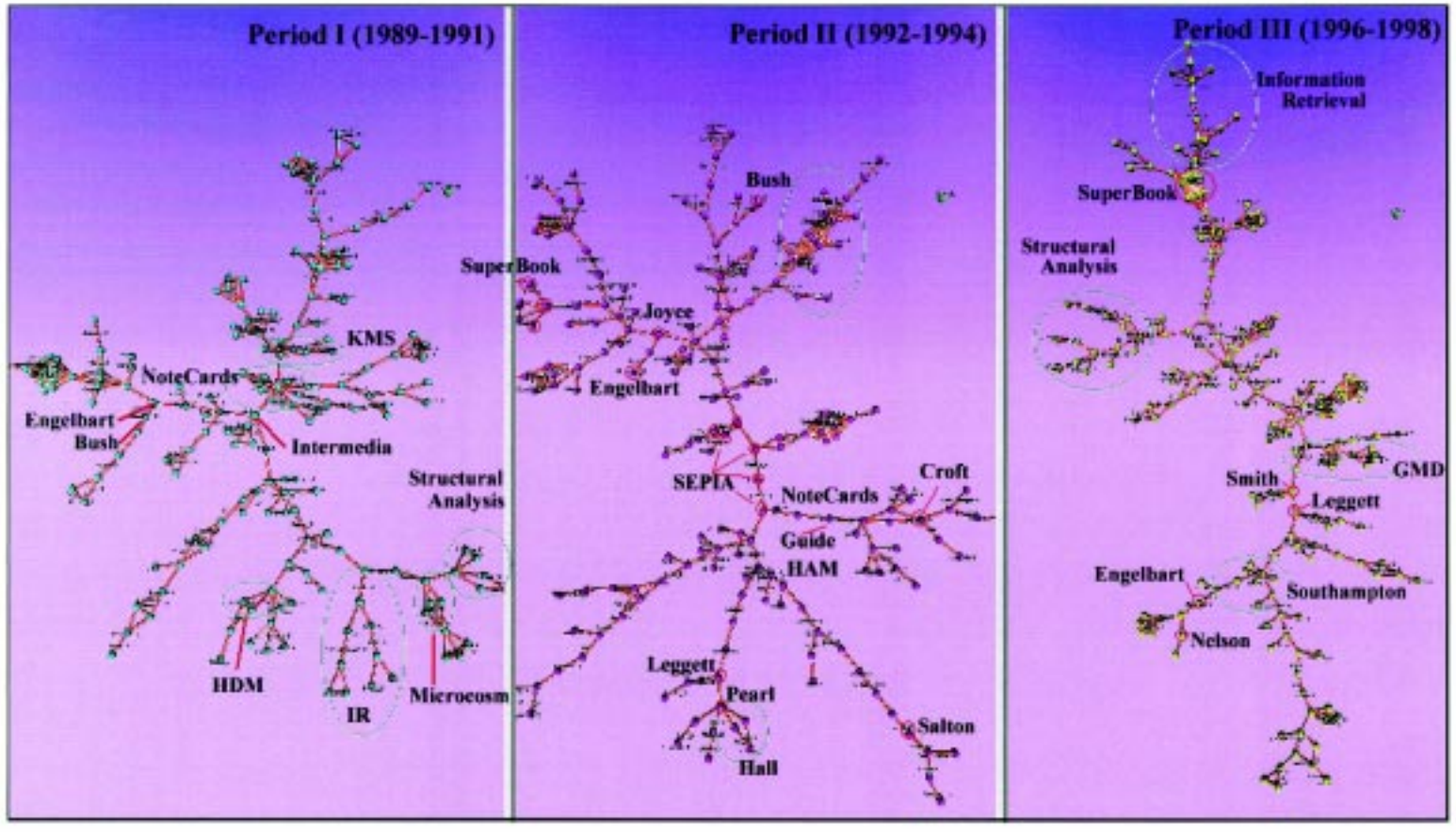

Fig. 9. Three snapshots of the evolution of the hypertext field: Map A (1989–1991), Map B (1992–1994) and Map C (1996–1998).

libraries, and we have explained the value of techniques such as LSI, Pathfinder networking scaling and virtual reality. This approach provides a means of transcending the boundaries of collections of documents and visualising more profound patterns in terms of semantic structures and co-citation networks. The spatial metaphor also lends itself to a natural combination of search and navigation within the same semantic space.

Author co-citation networks provide a snapshot of a scientific field as reflected through publications in the literature. More importantly, co-citation patterns offer a valuable alternative to the existing visualisation paradigms, which largely rely on the analysis of term distributions in a document collection. Author co-citation maps generated in our study are informative and revealing.

More work is needed in areas such as evaluating the usability of such visualisation paradigms in realistic digital libraries and investigating the role of individual differences in exploring a complex semantic space (Vicente & Williges, 1988; Dillon & Watson, 1996; Benyon & Höök, 1997; Chen & Czerwinski, 1997). We have been pursuing a number of projects addressing individual differences and social dimensions of virtual environments (Chen, Cole & Thomas, 1998; Chen & Czerwinski, 1998). We will continue the development and refinement of this approach in order to provide users with more effective, intuitive, and engaging digital libraries in particular and electronic working environments in general.

Acknowledgements

This work was supported in part by the British Engineering and Physical Sciences Research Council (EPSRC) under the Multimedia and Networking Applications Programme (Research Grant No. GR/L61088). Bell Communication Research kindly provided the software for Latent Semantic Indexing. Special thanks go to Les Carr at the University of Southampton for generating author co-citation data, and Janet Cole at Brunel University for her assistance in preparing the ACM Hypertext collection. The author would like to thank Edward Fox, Gary Marchionini, and reviewers for their helpful comments and suggestions.

References

Ahlberg, C., & Shneiderman, B. (1994). Visual information-seeking—Tight coupling of dynamic query filters with starfield displays. Proceedings of CHI '94, pp. 313–317, Boston, MA.

Akscyn, R. (Ed.). (1991). *The ACM Hypertext Compendium*. New York: ACM Press.

Beckwith, R., Fellbaum, C., Gross, D., & Miller, G.A. (1991). WordNet: a lexical database organized on psycholinguistic principles. In U. Zernik (Ed.), Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon, pp. 211–231. Lawrence Erlbaum Associaties.

Benford, S., Snowdon, D., Greenhalgh, C., Ingram, R., Knox, I., & Brown, C. (1995). VR–VIBE: a virtual environment for co-operative information retrieval. *Computer Graphics Forum*, *14* (3), C.349–C.360.

Benyon, D., & Höök, K. (1997). Navigation in information spaces: Supporting the individual. Proceedings of Human-Computer Interaction: INTERACT'97, pp. 39–46.

Chalmers, M. (1992). BEAD: Explorations in information visualisation. Proceedings of SIGIR '92, pp. 330–337, Copenhagen, Denmark.

Chapman, R.L. (Ed.). (1992). Roget's International Thesaurus (5th edition): Harper Collins.

Chen, C. (1997). Structuring and visualising the WWW with Generalised Similarity Analysis. Proceedings of the 8th ACM Conference on Hypertext (Hypertext '97), pp. 177–186, Southampton, UK.

Chen, C. (1998a). Bridging the gap: the use of Pathfinder networks in visual navigation. *Journal of Visual Lanuages and Computing*, *9* (3), 267–286.

Chen, C. (1998b). Generalised similarity analysis and pathfinder network scaling. *Interacting with Computers*, *10* (2), 107–128.

Chen, C., Cole, J., & Thomas, L. (1998). Fostering social interaction in a shared semantic space for collaborative learning. Proceedings of World Conference of the WWW, Internet and Intranet (WebNet '98) Orlando, Florida, USA.

Chen, C., & Czerwinski, M. (1997). Spatial ability and visual navigation: an empirical study. *New Review of Hypermedia and Multimedia*, *3*, 67–89.

Chen, C., & Czerwinski, M. (1998). From latent semantics to spatial hypertext: an integrated approach. Proceedings of the 9th ACM Conference on Hypertext and Hypermedia (Hypertext '98), pp. 77–86, Pittsburgh, PA.

Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, *41* (6), 391–407.

Dillon, A., & Watson, C. (1996). User analysis in HCI: the historical lessons from individual differences research. *International Journal of Human-Computer Studies*, *45* (6), 619–637.

Dumais, S.T. (1995). Using LSI for information filtering: TREC-3 experiments. Proceedings of the 3rd Text REtrieval Conference (TREC3).

Fairchild, K., Poltrock, S., & Furnas, G. (1988). SemNet: Three-dimensional graphic representations of large knowledge bases. In: R. Guidon (Ed.), Cognitive Science and its Applications for Human-Computer Interaction, pp. 201–233. Lawrence Erlbaum Associates.

Fox, E., Hix, D., Nowell, L., Brueni, D., Wake, W., Heath, L., & Rao, D. (1993). Users, user interfaces, and objects: Envision, a digital library. *Journal of the American Society for Information Science*, *44* (5), 480–491.

Fox, E. A., Akscyn, R. M., Furuta, R. K., & Leggett, J. J. (1995). Digital libraries. *Communications of the ACM*, *38* (4), 22–28.

Fox, E. A., & Marchionini, G. (1998). Toward a worldwide digital library: introduction. *Communications of the ACM*, *41* (4), 28–32.

Fruchterman, T. M. J., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software—Practice and Experience*, *21*, 1129–1164.

Garfield, E. (1994). Scientography: mapping the tracks of science. *Current Contents: Social and Behavioural Sciences*, *7* (45), 5–10.

Green, S.J. (1998). Automated link generation: Can we do better than term repetition? Proceedings of the 7th International World-Wide Web Conference Brisbane, Australia.

Heath, L., Hix, D., Nowell, L., Wake, W., Averboch, G., Labow, E., Guyer, S., Brueni, D., France, R., Dalal, K., & Fox, E. (1995). Envision: a user-centered database of computer science literature. *Communications of the ACM*, *38* (4), 52–53.

Hemmje, M., Kunkel, C., & Willett, A. (1994). LyberWorld — A visualization user interface supporting fulltext retrieval. Proceedings of the 17th Annual International ACM–SIGIR Conference on Research and Development in Information Retrieval, pp. 249–259, Dublin, Ireland.

Hetzler, B., Harris, W.M., Havre, S., & Whitney, P. (1998a). Visualizing the full spectrum of document relationships. Proceedings of the 5th International ISKO Conference: Structures and Relations in the Knowledge Organization Lille.

Hetzler, B., Whitney, P., Martucci, L., & Thomas, J. (1998b). Multi-faceted insight through interoperable visual information analysis paradigms. Proceedings of IEEE Information Visualization '98.

ISI (1981). ISI atlas of science: Biochemistry and molecular biology, 1978/80. Philadelphia, PA: Institute for Scientific Information.

Kamada, T., & Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters*, *31* (1), 7–15.

Lin, X. (1997). Map displays for information retrieval. *Journal of the American Society for Information Science*, *48* (1), 40–54.

Lin, X., Soergel, D., & Marchionini, G. (1991). A self-organizing semantic map for information retrieval. Proceedings of SIGIR '91, pp. 262–269, Chicago, IL.

Nowell, L.T., France, R.K., Hix, D., Heath, L.S., & Fox, E.A. (1996) Visualzing search results: Some alternatives to query-document similarity. Proceedings of the 19th Annual ACM SIGIR Conference, 67–75, Zurich, Switzerland.

Orendorf, J. and Kacmar, C. (1996). A spatial approach to organizing and locating digital libraries and their content. Proceedings of the 1st ACM international conference on Digital libraries (DL '96). pp. 83–89, Bethesda, MD.

Rao, R., Pedersen, J. O., Hearst, M. A., Mackinlay, J. D., Card, S. K., Masinter, L., Halvorsen, P.-K., & Robertson, G. G. (1995). Rich interaction in the digital library. *Communications of the ACM*, *38* (4), 29–39.

Salton, G., Allan, J., & Buckley, C. (1994). Automatic structuring and retrieval of large text files. *Communications of the ACM*, *37* (2), 97–108.

Schvaneveldt, R.W., Durso, F.T., & Dearholt, D.W. (1989). Network structures in proximity data. In: G. Bower (Ed.), The Psychology of Learning and Motivation, Vol. 24, pp. 249–284. Academic Press.

Small, H. (1994). A SCI–MAP case study: Building a map of AIDS research. *Scientometrics*, *30* (1), 229–241.

Small, H. (1997). Update on science mapping: Creating large document spaces. *Scientometrics*, *38* (2), 275–293.

Vicente, K. J., & Williges, R. C. (1988). Accommodating individual differences in searching a hierarchical file system. *International Journal of Man-Machine Studies*, *29*, 647–668y.

White, H. D., & McCain, K. W. (1998). Visualizing a discipline: an author co-citation analysis of information science, 1972–1995. *Journal of the American Society for Information Science*, *49* (4), 327–356.