Interacting with Computers 10 (1998) 107–128

Generalised similarity analysis and pathfinder network scaling

Chaomei Chen[1]

Department of Computer Studies, Glasgow Caledonian University, Glasgow G4 0BA, UK

Abstract

This paper introduces a generic approach to the development of hypermedia information systems. This approach emphasises the role of intrinsic inter-document relationships in structuring and visualising a large hypermedia information space. In this paper, we illustrate the use of this approach based on three types of similarity measurements: hypertext linkage, content similarity and usage patterns. Salient patterns in these relationships are extracted and visualised in a simple and intuitive associated network. The spatial layout of a visualisation is optimised such that closely related documents are placed near to each other and only those intrinsic connections among them are shown to users as automatically generated virtual links. This approach supports self-organised information space transformation based on usage patterns and other feedback such that the visual structure of the information space is incrementally tailored to users' search and browsing styles. © 1998 Elsevier Science B.V.

Keywords: Hypertext; Virtual link structure; Visualisation; Information retrieval

The wide-spread use of the World-Wide Web (WWW) has highlighted the demand for cost-effective accessing of relevant information in large distributed hypermedia environments in areas such as digital libraries, electronic publishing, distance learning and subject-specific knowledge repositories. Previous studies suggest that combining search and navigation in a flexible information retrieval environment may enable users to access information more effectively and intuitively. For example, in the late 1980s, researchers incorporated the notion of hypertext into traditional information retrieval systems (e.g., Refs. [1,2]). More recently, research in information retrieval increasingly emphasises the role of hyperlinks in enhancing the quality of information retrieved (e.g., Ref. [3]).

The highly dynamic and distributed nature of the WWW highlights some fundamental issues to be addressed in the development of large, distributed hypermedia systems [31].

[1] Present address: Department of Information Systems and Computing, Brunel University, Uxbridge UB8 3PH, UK.
E-mail: chaomei.chen@brunel.ac.uk.

PII S0953-5438(98)00015-0

For example, hypermedia components in a traditional hypermedia system are statically connected. A greater flexibility and maintainability can be achieved by using dynamic node–link binding in which the structure of a hypermedia network is determined by one or several distinct sets of link configurations. In fact, there is a rapidly growing interest in open hypermedia services [4,5,30]. A fundamental issue is whether large link configurations can be efficiently generated to represent intrinsic characteristics of the underlying document space.

In this paper, we introduce a generic approach to the development of hypermedia information systems. This approach, called Generalised Similarity Analysis (GSA), emphasises the role of intrinsic inter-document relationships in structuring and visualising a large hypermedia information space [29]. In this paper, we illustrate the use of this approach based on three types of similarity measurements: hypertext linkage, content similarity and usage patterns. The spatial layout of a visualisation is optimised such that closely related documents are placed near to each other and only those intrinsic connections among them are shown to users as automatically generated virtual links. This approach supports self-organised information space transformation based on usage patterns and other feedback such that the visual structure of the information space is incrementally tailored to users' search and browsing styles. Human–computer interaction issues concerning the use of GSA and related techniques such as fisheye views and virtual reality are also addressed.

1. Large hypermedia systems

The use of hypermedia systems has been notoriously associated with two common problems: disorientation (or lost in hyperspace) and cognitive overhead [6]. These problems tend to get increasingly severe with large hypermedia systems. Empirical studies in hypertext have consistently shown that the use of graphical maps and structural overviews can significantly help users to understand how the information is organised and interconnected so that they can use the system more effectively [7]. In this section, we identify a number of theoretical and practical issues concerning structuring and visualising large hypermedia systems. In particular, we emphasise the generic feature of the Generalised Similarity Analysis approach and why GSA differs from existing techniques.

1.1. Information visualisation

There is a growing interest in information visualisation on the WWW. Web-related information visualisation systems fall into two general categories: static and dynamic. Static systems are based on pre-collected data to build visualisations. They often use Internet software agents, commonly known as spiders or wanderers, to acquire information automatically from the WWW. For example, the Navigational View Builder [8] visualises the structure of HTML documents on the WWW by attributes such as author, file-size and keyword. The Navigational View Builder uses a range of information visualisation techniques such as Cone Trees [9] and Perspective Wall [10], but it does not focus on inter-document relationships such as hypertext linkage patterns.

Dynamic systems, on the other hand, dynamically build a graph of WWW documents recently accessed and show users how these documents are related to one another. Examples of dynamic systems include MosaicG [11] and WebNet [12]. Dynamic systems mainly rely on client-side information and focus on representing the history of users' browsing over a short period of time. The techniques described in this paper are originally developed for static systems, but they can be tailored to meet the needs of a dynamic systems.

1.2. Focus and context

Visualisation of a large hypermedia system must address a proper balance between local details and contextual information for users to use the system effectively [13]. It is usually not practical to accommodate information at all the levels into a limited computer display. One may find extremely difficult to understand and interact with large complex graphs.

Furnas' fisheye views model is based on a 'degree of interest' (DOI) function, which assigns a value to each node in accordance with the degree to which a user would be interested in seeing that node [13–15]. Assume that the user is currently at node x_{f}, known as the focal point, the DOI function is defined as

$$\mathrm{DOI}_{\mathrm{fisheye}}(x, x_{\mathrm{f}}) = \mathrm{API}(x) - D(x, x_{\mathrm{f}}),$$

where x is any node in the network, API(x) is the global *a priori* importance of x and $D(x,x_0)$ is the distance between x and x_0. A fisheye view is normally associated with a threshold so that only nodes with sufficient DOI are shown to the user.

API provides a flexible mechanism to define fisheye views based on different preferences. For example, one can define an API function for a Web site and the value of API at a document x, API(x), is the number of times the document x has been visited by external users. Consequently, popular nodes will be highlighted in resultant views in terms of colour and size.

Fisheye views have been traditionally applied to hierarchical structures, in which distances are well defined, with few exceptions on network structures (e.g., Ref. [14]). In this paper, we introduce some generic computational models and techniques that can be used to derive geodesic distances between two nodes in a network. In addition, we particularly emphasise the connection between salient semantic relationships and spatial visualisation representations.

1.3. Information retrieval models

Traditional information retrieval provides a number of concepts and models which can be useful for structuring and visualising large hypermedia, including vector-space retrieval models [16,17], dynamic document space transformation [17] and clustering-based information retrieval models (see Ref. [18]).

A natural starting point for integrating search and navigation-based information retrieval is the classic vector-space model [17]. First, it is an influential retrieval model in traditional information retrieval systems. Second, it provides a natural visualisation model for the underlying document space. In fact, many visual information retrieval systems are based on this model.[2]

[2] See http://www-cui.darmstadt.gmd.de:80/visit/Activities/Viri/visual.html for a classification of visual information retrieval systems.

In this paper, the vector-processing model is an integral part of the generic GSA framework, which also accommodates similarity measures based on hypertext linkage and user behaviour patterns. Savoy [3] utilised information on hypertext linkage to enhance the effectiveness of the vector-space retrieval model, but the extended retrieval model does not take into account navigation patterns associated with the actual use of such hypertext structure. Pirolli *et al.*'s study [19] and the HyPursuit system [20] take into account hypertext linkage and content similarity on the WWW, but they do not fit the use of the vector-space model into a unified framework. Further comparisons between these studies and our framework will be given when we introduce our unifying framework in the next section.

GSA differs from existing approaches in a number of ways. For example, MultiDimensional Scaling (MDS) and scatter plots have been frequently used in information visualisation systems, but the underlying semantic structure is implicit and inter-document connectivity is not readily visible to users. A key component of GSA is the application of Pathfinder network scaling algorithms [23] such that the underlying semantic structure is represented as an associative network of the most salient inter-document relationships. Therefore, GSA provides users with valuable information on salient structural patterns that are not readily available in other systems.

2. Generalised similarity analysis

Generalised Similarity Analysis (GSA) is a unifying framework for extracting structural patterns from a hypermedia information space. A number of intrinsic interrelationships in hypertext, such as hypertext linkage, content similarity and browsing patterns, are consistently incorporated into the generic framework. Pathfinder network scaling is a key component of the framework. In this paper, we illustrate the application of GSA on three types of inter-document relationships.

2.1. Architecture

The architecture of the GSA framework consists of a number of computational models (see Fig. 1). Each of these computational models generates a virtual link structure based on a distinct characteristic. A virtual link structure can also be generated by integrating someor all the component models. One can incorporate additional inter-document relationships into the framework, for example, based on citation and co-citation counts between documents.

The similarity between two documents can be measured psychologically or statistically. In hypermedia systems, some fundamental relationships are hypertext linkage, content similarity and browsing patterns. These relationships are used to estimate document similarities in this paper to illustrate the generic approach.

2.2. Hypertext linkage

A hypertext with N documents, or nodes, corresponds to an $N \times N$ matrix, called the *distance matrix*. The value of the element d_{ij} in the matrix is the distance between node i and j. Botafogo *et al.* [22] introduced two structural metrics, the Relative Out Centrality

(ROC) and Relative In Centrality (RIC) metrics, to identify various structural characteristics of a node. For example, a node with a high ROC can be used as a starting point to reach out for other nodes, whereas a node with a high RIC is easy to get accessed. The structure of the hypertext can be transformed to one or more hierarchies with a high-ROC node as the root. Botafogo *et al.* [22] suggested that large hierarchies may be displayed with fisheye views, which balance local details and global context [13].

HyPursuit is a hierarchical network search engine based on semantic information embedded in hyperlink structures and document contents [20]. HyPursuit considers not only links between two documents, but also how their ancestor and descendant documents are related. For example, if two documents have a common ancestor, they are regarded more similar to each other. In HyPursuit, document similarity by linkage is defined as a linear combination of three components: direct linkage, ancestor and descendant inheritance.

Pirolli *et al.* [19] characterise documents in a Web locality, a closed subset of WWW documents, by feature vectors based on attributes such as the number of incoming and outgoing hyperlinks of a document, how frequently the document was visited and content similarities between the document and its children. They used these feature vectors to categorise the nature of a page and predict the interests of visitors to that page.

In this study, document proximity is defined based on similarities between documents. The document similarity by hypertext linkage in this study is defined as follows:

$$\mathrm{sim}_{ij}^{\mathrm{link}} = \frac{\mathrm{link}_{ij}}{\sum_{k=1}^{N} \mathrm{link}_{ik}},$$

where link_{ij} is the number of hyperlinks from document D_i to D_j in a collection of N documents from the WWW, for example, from a particular server or on a specific topic. Higher-order interrelationships with ancestors and descendants are not considered because they can be resolved by Pathfinder network scaling algorithms. This definition allows asymmetrical as well as symmetrical relationships between documents. The Pathfinder network scaling algorithms can handle both symmetric and asymmetric data. Without losing generality, we assume that these measures are symmetric unless we state otherwise.

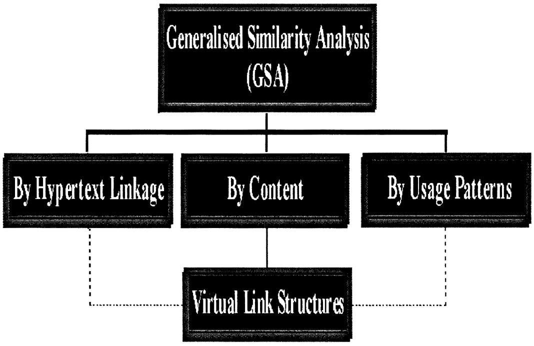

Fig. 1. The architecture of generalised similarity analysis.

According to this definition, a similarity of 0 between two documents implies $\text{link}_{ij} = 0$, which means that one document is not linked to the other at all. On the other hand, a similarity of 1 implies $\text{link}_{ik} = 0$ for all the $k \neq j$, which means that the two documents are connected by hyperlinks to each other but not to any other documents.

2.3. Content similarity

The vector-space model [16], originally developed for information retrieval, is a powerful framework for analysing and structuring documents. In this model, each document is represented by a vector of terms. Terms are weighted to indicate how important they are in representing the document. The distance between two documents can be determined according to corresponding vector coefficients. A large collection of documents can be split into a number of smaller clusters such that documents within a cluster are more similar than documents in different clusters. By creating links between documents that are sufficiently similar, Salton *et al.* [16] automatically generated semantically-based hypertext networks using the vector-space model.

In this study, we use the well-known *tf* × *idf* model, term frequency times inverse document frequency, to build term vectors. Each document is represented by a vector of T terms with corresponding term weights. The weight of term T_k to document D_i, is determined by

$$w_{ik} = \frac{tf_{ik} \times \log\left(N/n_k\right)}{\sqrt{\sum_{j=1}^{T} \left(tf_{ij}\right)^2 \times \log\left(N/n_j\right)^2}},$$

where tf_{ik} is the occurrences of term T_k in D_i, N is the number of documents in the collection (such as the size of a WWW site), and n_k represents the number of documents containing term T_k. The document similarity is computed as follows based on corresponding vectors $D_i = (w_{i1}, w_{i2}, \ldots, w_{iT})$and $D_j = (w_{j1}, w_{j2},\ldots,w_{jT})$:

$$\text{sim}_{ij}^{\text{content}} = \sum_{k=1}^{T} w_{ik} \times w_{jk}.$$

HyPursuit [20] also used a modified version of the vector-space model. However, the weight function in HyPursuit does not include collection frequency n_k, whereas we use the complete vector-space model on a specific collection of documents retrieved from the WWW. The vector-space model used by Pirolli *et al.* [19] was restricted to the existing hyperlink structure in that content similarities were only considered between hyperlinked documents. In our study, content similarities are considered across the entire collection of documents in order to find out under-represented patterns.

2.4. State-transition patterns

There is a growing interest in incorporating usage patterns into the design of large distributed hypermedia systems and notably on the WWW. Access logs maintained by many WWW servers provide a valuable source of empirical information on how users actually access the information on a server and what documents appear to attract the attention of users. Sequential patterns of browsing indicate, to some extent, document relatedness perceived by users. For example, the number of users who followed a hyperlink connecting two documents in the past were used by Pirolli *et al.* [19] to estimate the degree of relatedness between the two documents.

The dynamics of a browsing process can be captured by state transition probabilities. Transition probabilities can be used to indicate document similarity with respect to browsing. Using transition probabilities has some advantages. For example, the construction of the state transition model is consistent with linkage- and content-based similarity models. In this study, one-step transition probability p_{ij} from document D_i to D_j is estimated as follows:

$$p_{ij} = \frac{f(D_i, D_j)}{\sum_{k=1}^{N} f(D_i, D_k)},$$

where $f(D_i,D_j)$ is the observed occurrences of a transition from D_i to D_j and$\sum_k f\ (D_i, D_k)$ is the total number of transitions starting from D_i. Transition probability p_{ij} is used to derive the similarity between document D_i and D_j in the view of users:

$$\mathrm{sim}_{ij}^{\mathrm{usage}} = p_{ij}.$$

2.5. Meta-similarities

The similarity between two sets of similarity measures, i.e. a meta-similarity, can be measured in several ways. For example, the meta-similarity between linkage-based similarities and content term-based similarities can be measured by the squared sum of differences in corresponding measures as follows:

$$\mathrm{Distance(link,\ content)} = \sum_{i,j=1}^{N} \left(\mathrm{sim}_{ij}^{\mathrm{link}} - \mathrm{sim}_{ij}^{\mathrm{content}}\right)^2.$$

In this paper, we compute both Pearson's and cosine correlation coefficients among the three sets of inter-document similarities associated with a website. One can combine distinct sets of similarity measures such that the resultant virtual link structure reflects the strong influence of some specific characteristics (see Fig. 2).

For example, one may tailor an existing hyperlink structure so that the underlying semantic structure based on content term-similarities are taken into account using the following generic formula:

$$\mathrm{sim}_{ij}^{\mathrm{combined}}(\omega_{ij}) = \frac{\omega_{ij}\cdot\mathrm{hyperlinks}_{ij}}{\sum_{k=1}^{N} \omega_{ik}\cdot\mathrm{hyperlinks}_{ik}},$$

$$\mathrm{sim}_{ij}^{\mathrm{link+content}} = \mathrm{sim}_{ij}^{\mathrm{combined}}(\omega_{ij} = \mathrm{sim}_{ij}^{\mathrm{content}}),$$

where the resultant similarity $sim_{ij}^{\mathrm{link+content}}$ weights existing hypertext linkage by corresponding content similarities. In this paper, we focus on the overall architecture of the

GSA framework and interactions between component similarity models will be further investigated in our future work.

2.6. Pathfinder networks

The Pathfinder network scaling algorithm is a structural and procedural modelling technique which extracts underlying patterns in proximity data and represents them spatially in a class of networks called Pathfinder Networks (PFNETs) [23,24]. The essential concept underlying Pathfinder networks is pairwise similarity. Similarities can be obtained based on a subjective estimation or a numerical computation. Pathfinder provides a more accurate representation of local relationships than techniques such as multidimensional scaling (MDS).

The topology of a PFNET is determined by two parameters q and r and the corresponding network is denoted as PFNET(r,q). The q-parameter constrains the scope of minimum-cost paths to be considered. The r-parameter defines the Minkowski metric used for computing the distance of a path. The weight of a path with k links is determined by weights w_1,w_2,…,w_k of each individual link as follows:

$$W(P) = \sqrt[r]{\sum_{i=1}^{k} w_i^r},$$

The q-parameter specifies that triangle inequalities must be satisfied for paths with $k \leq q$ links:

$$w_{n_i n_{i+1}} = \sqrt[r]{\sum_{i=1}^{k-1} w^r_{n_i n_{i+1}}}, \quad \forall k \leq q.$$

When a PFNET satisfies the following 3 conditions, the distance of a path is the same as the weight of the path:

1. the distance from a document to itself is zero;
2. the proximity matrix for the documents is symmetric; thus the distance is independent of direction;

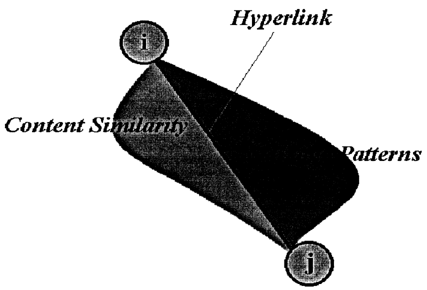

Fig. 2. Structuring mechanisms and link adjustment.

3. the triangle inequality is satisfied for all paths up to q links. If q is set to the total number of nodes less one, then the triangle inequality is universally satisfied over the entire network.

The number of links in a network can be reduced by increasing the value of parameter r or q. The geodesic distance between two nodes in a network is the length of the minimum-cost path connecting the nodes. A minimum-cost network (MCN), PFNET($r = \infty$, $q = N - 1$), has the least number of links.

The graph layout of a Pathfinder network is determined based on the spring model described in Ref. [25]. There is a growing interest in spring models in information visualisation [26] because the idea is simple and intuitive. In a spring model, nodes are connected by weighted links, or proximity measures as in this study. These nodes are forced into place by spring energy transformed from the weights. As the overall spring energy in the system is minimized, the graph gradually takes shape. Resolving spring models usually requires the computational complexity of $O(N^2)$. As the number of nodes in the graph grows, more efficient solutions are necessary. One possible solution is to use the divide-and-conquer strategy, in which a large information space can be rapidly and recursively split into smaller clusters with simple classification algorithms until the size of clusters becomes insignificant for those computationally expensive algorithms.

The following example illustrates the process of Pathfinder network scaling on a collection of concepts about living things. The proximity between two concepts was rated on a one-hundred point scale from 0 to 99. The higher the score, the larger the semantic distance between the two concepts. For example, the *dog–deer* proximity was 35, *animal–dog* was 13 and *animal–deer* was 11. Table 1is part of the proximity matrix.

Fig. 3 shows part of the Pathfinder network, rendered in Virtual Reality Modelling Language (VRML), in which *animal* was connected to *dog* and *deer*, but *dog* and *deer* were not directly connected. A *dog–deer* link (with a weight of 35) would violate the triangular inequality among *animal*, *dog* and *deer*, because the cost of the *dog–deer* link (35) is greater than the sum of *dog–animal* and *deer–animal* (13 + 11 = 24). Thus, the Pathfinder network preserves the most salient relationships.

The major advantage of Pathfinder networks is that salient relationships among

Table 1

A proximity matrix of a set of living things, animals and plants

	Living	Animal	Blood	Bird	Feathers	Robin	Chicken	Hairs	Dog	Deer
Living										
Animal	13									
Blood	29	26								
Bird	18	25	47							
Feathers	51	44	54	8						
Robin	23	34	43	12	17					
Chicken	17	23	43	18	15	27				
Hairs	45	33	55	65	36	67	67			
Dog	15	**13**	39	35	73	44	41	20		
Deer	15	**11**	37	41	73	42	41	44	**35**	
Bats	22	28	35	22	70	32	45	47	43	46
Antler	41	31	48	73	56	65	68	49	66	11

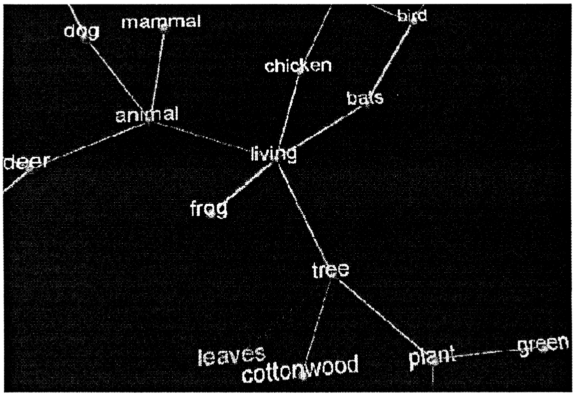

Fig. 3. The Pathfinder network of concepts on living things based on minimum-cost paths.

documents are highlighted. This type of information filtering improves the clarity and quality of the information produced by information visualisation systems based on spring models. Users are able to see how documents are related to each other.

3. Extracting structures

In this section, GSA is applied to departmental WWW sites and conference proceedings on the WWW. We analyse inter-site connectivity of computer science departmental WWW sites in 13 universities because computer science departments in general have established infrastructure and they are more experienced in developing WWW documents.

3.1. Data collection

HTML documents were automatically retrieved from 13 WWW sites by HARVEST's Gatherers. HARVEST[3] is an integrated set of tools to gather, extract, organise, index and search relevant information on the Internet [27]. HARVEST provides 2 useful subsystems, Gatherers and Brokers, to collect and index subject-specific information on the WWW. We used HARVEST Release 1.4 with an HP-UX operating system. HARVEST's Gatherers downloaded HTML documents. The boundary of a WWW site was determined by pattern matching rules which instruct a HARVEST Gatherer to retrieve documents from valid URLs, Unique Resource Locators, on specific WWW servers. Full-text papers were automatically retrieved from the online proceedings of CHI'96 on the WWW[4] and 46 valid papers were subsequently used.

[3] WWW address: http://harvest.cs.colorado.edu

[4] WWW address: http://www.acm.org/sigchi/chi96/proceedings/

3.2. Data analysis

We used HARVEST to extract attribute–value information from HTML documents retrieved from several WWW sites. HARVEST supports a type-specific extraction algorithm, i.e., a summarizer, to digest the data. We modified the HTML summarizer to focus on extracting keywords from predominant structural elements of a document. The full-text version of the HTML summarizer was also used. For example, higher weights were given to words that appeared in HTML markups such as < `head` >, < `title` > and < `a href` >. A stopword list was compiled based on all the terms occurred in particular document collections.

We analysed the connectivity among the 13 departmental sites in terms of incoming and outgoing hyperlinks for each site. Based on the connectivity map, we chose a few small-to-medium sites to apply the GSA analysis, namely, hypertext linkage, content similarity and usage patterns. Minimum-cost networks ($r = \infty$, $q = N - 1$) were normally used in this study because they represent salient relationships in proximity data. Final Pathfinder networks were generated on PC by PCKNOT from Interlink, Inc., New Mexico. Similarity matrices were submitted to PCKNOT.

4. Results

This section presents Pathfinder networks derived from the GSA study. Each numbered box in a graphical representation of a PFNET corresponds to a document. Some Pathfinder networks are rendered in VRML to illustrate local details and overall structures. Standard 0.005 spring-energy threshold was used throughout the study to generate Pathfinder networks.

4.1. Inter-site connectivity

Inter-site connectivity was computed in terms of the number of links between one site and another. The connectivity was represented by a 13×13 asymmetric proximity matrix. Pathfinder is particularly suitable to deal with asymmetric proximity data. Fig. 4 is the connectivity map of the 13 sites in Scotland.

Sites connected with shorter links have more hyperlinks between them, for example, `Napier` and `Edinburgh`. Such connectivity maps can be used to identify the distribution of expertise in specific areas.

4.2. Hypertext linkage

Fig. 5 shows the structure of a WWW site (SITE_A) according to hypertext linkage. Pathfinder extracted 189 salient relationships from 1,503 initial similarity measures. The spring energy in this PFNET is less than 0.005 (four isolated nodes are not shown).

Some nodes are more special than others. For example, Node 22, 80 and 25 in Fig. 5 led to document clusters on an HTML tutorial, collaborating researchers and object-oriented programming at the site, respectively.

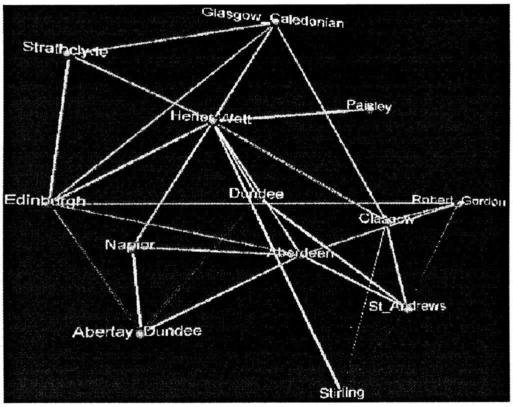

Fig. 4. A connectivity map of 13 departmental WWW sites, shown as PFNET($r = \infty$, $q = N - 1 = 12$).

4.3. Content similarity

Two Pathfinder networks were generated for papers in the CHI'96 proceedings based on the vector-space model. The structure in Fig. 6, PFNET($r = 2$, $q = 1$), is based on all the connecting paths derived from the vector-based content similarity model, where $q = 1$ implies the inclusion of a path is independent from any other paths. In Fig. 7, the similarity between two papers was considered in the context of all the connecting paths between the two documents. Only the minimum-cost path was preserved in the final network to represent the salient relationship. The resultant graph is a natural candidate for an overview map of the information space.

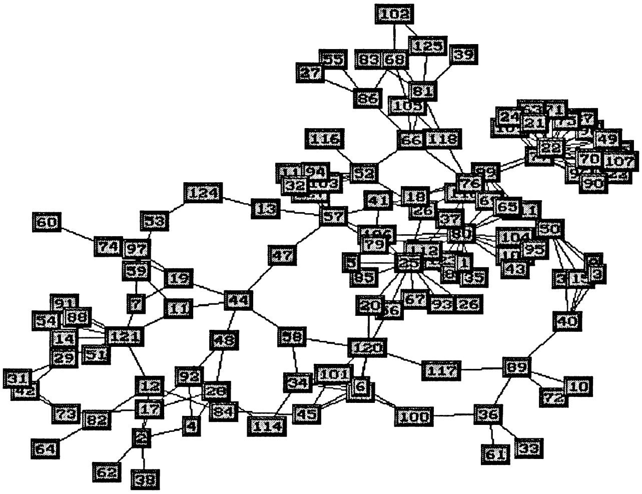

Fig. 5. The structure of SITE$_A$ (partial) by hyperlinks, shown as a PFNET($r = \infty$, $q = N - 1 = 126$) with 189 links.

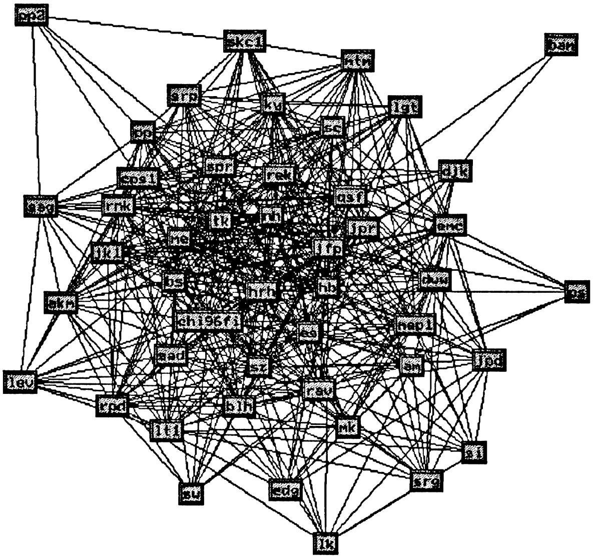

Fig. 6. CHI'96 papers structured according to all the connecting paths (Stress < 0.005, Links = 516).

A graph representation takes shape as the overall spring energy reduces below a threshold given in advance. Fig. 8 shows the node placement process for CHI'96 papers at 6 discrete points. The value of spring energy at each point is given at the right-hand corner. For example, at an early stage, the weighted graph had spring energy of 0.999 and the energy was reduced to 0.900, 0.500, 0.200, 0.100 and eventually the threshold 0.005.

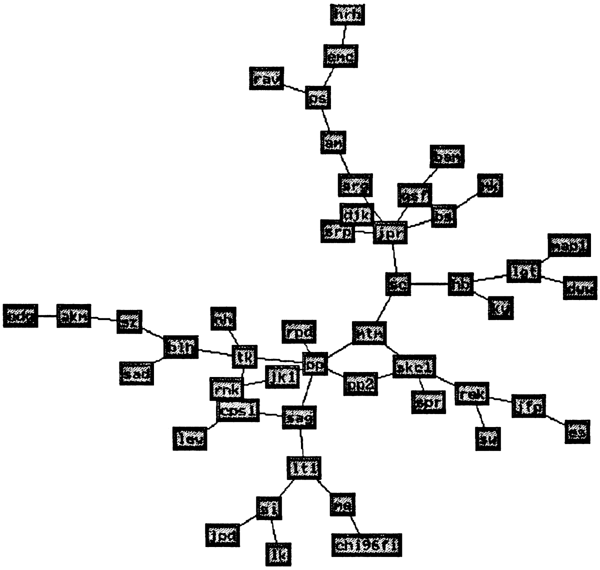

Fig. 7. CHI'96 papers structured according to minimum-cost paths (Stress < 0.005, Links = 47).

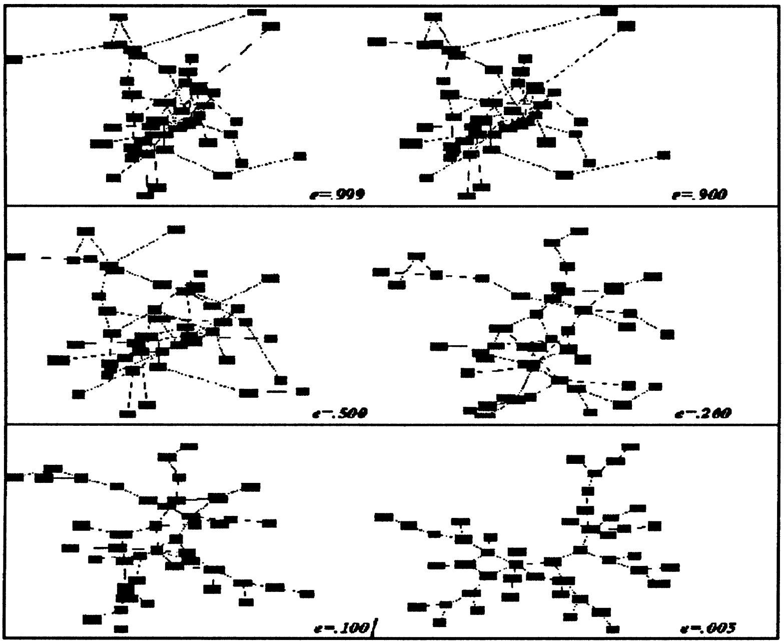

Fig. 8. The graph layout is being optimised as the spatial configuration represents the underlying similarity model more accurately.

Fig. 9 shows a PFNET for another departmental WWW site $SITE_B$ with 172 HTML documents. The network has 172 nodes and 242 links. The screen display becomes crowded even if we used numerical IDs in the network for corresponding documents. See Section 5 for further discussions on practical issues concerning the display of a large network.

We have analysed the structure and content of the WWW site A to explore generalised similarity analysis (GSA) for structuring and visualising a Web site. To what extent the resultant clusters are consistent with users' views of interests as they actually visit the WWW site? We analysed state-transition patterns based on access logs maintained on a CERN HTTP server at $SITE_A$.

Fig. 10 shows 3 Pathfinder networks corresponding to 3 bi-monthly access log data between September 1996 and January 1997 associated with external users' access to the author's homepage. A number of predominant cycles emerged from the graph. In fact, there seems to be some correspondence between a cycle and a set of documents of a particular type. For example, the largest cycle corresponds to top-level documents regarding general information about the homepage (Node 7), the page counters and plans. The cycle (17,19,20,21,6,15) corresponds to some research papers. The cycle (21,22,23,33,6) corresponds to documents used in teaching. It also seems larger cycles correspond to deeper browsing sequences, whereas small cycles tend to relate to more specific topics and shorter browsing sequences. Node 0 is an artificial node to indicate the end of a browsing sequence.

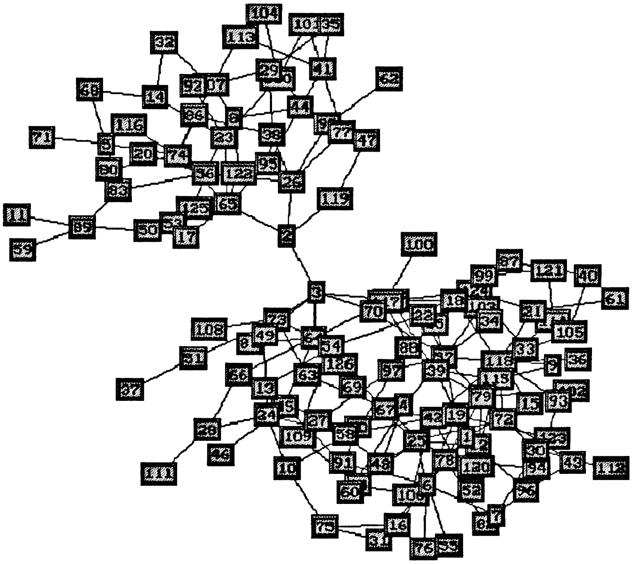

Fig. 9. The structure of SITE$_B$ by content similarity, shown as a PFNET($r = \infty$, $q = N - 1 = 171$) with 242 links.

A total of 22,209 access requests were made between 30 July and 31 September, 1996 by 1,125 user times. The behaviour of top 30 most active users was used as a basis of establishing representative behavioural patterns in terms of first-order state transitions. These 30 users count 10.7% of all the users ever visited the site during this period of time. The number of pages visited by the top 30 users range from 13 to 115. Fig. 11 shows a PFNET derived from similarities based on first-order state-transition probabilities. Cluster A is enlarged as Cluster A*.

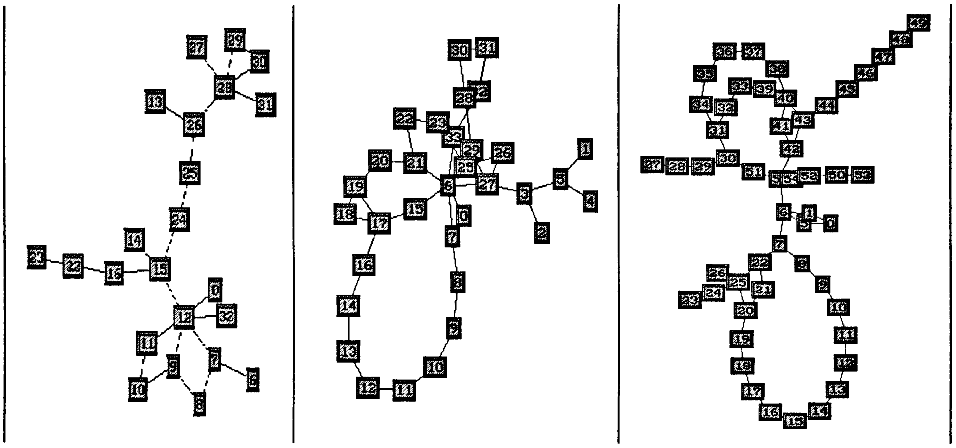

Fig. 10. The structure of a subset of SITE$_A$, containing WWW pages maintained by the author, by usage patterns (Stress < 0.005) (3 bi-monthly snapshots between September 1996 and January 1997).

The spike at the lower left half and the ring in Cluster A^* essentially associate with an M.Sc. student's project on Web-based interface design. The spike at the upper right half corresponds to some research papers on hypertext.

4.4. Meta-similarities

A meta-similarity is an overall estimate of the strength that two similarity variables are related. To illustrate this concept, we computed Pearson's and cosine correlation coefficients among three sets of similarities associated with the website SITE_A according to hyperlinks, content terms and usage patterns. A total of 127 valid documents from the SITE_A were included in our study. The linkage–content meta-similarity has the highest score on both Pearson's and cosine correlation coefficients ($r = 0.3201$ and $r_c = 0.4682$, $N = 127$) (see Table 2). The linkage–usage meta-similarity has the lowest score on both Pearson's and cosine correlation coefficients ($r = 0.0184$ and $r_c = 0.0644$, $N = 127$).

We analysed the changes in usage patterns associated with a collection of documents maintained by the author on the WWW over six consecutive months between August 1996 and January 1997. By comparing usage pattern-based similarity measures between adjacent months, we found that the meta-similarity increased from 0.1967 to 0.4586 over the six months. It appears to be a trend that the meta-similarity is increasing with time (see Fig. 12). A possible explanation is that usage patterns become increasingly similar as the underlying structure settles down at least for frequently visited documents. Experimental studies and a thorough examination of specific documents and associated usage patterns may lead to more insights into the pattern.

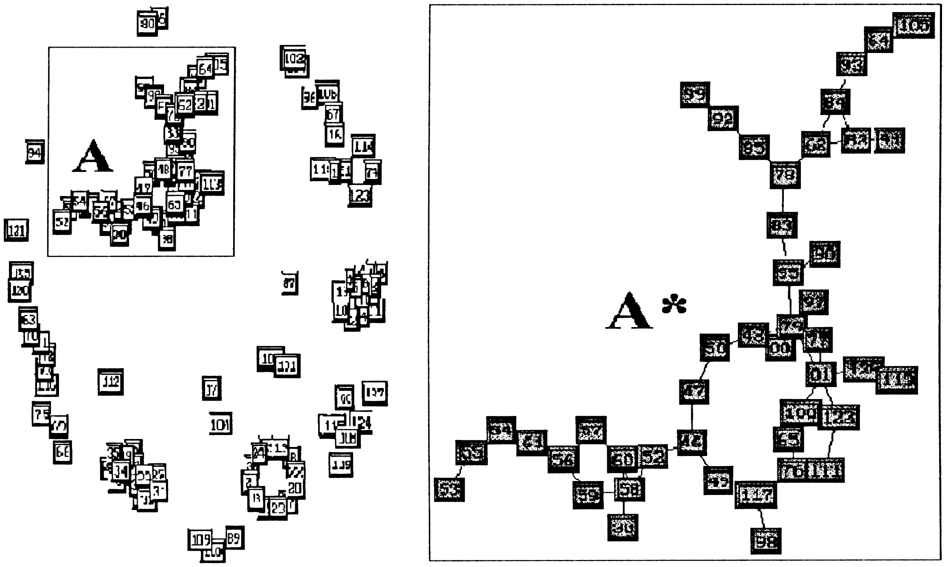

Fig. 11. The structure of SITE_A by state-transition patterns, shown as a PFNET($r = \infty$, $q = N - 1$) with 67 links.

Table 2
Peaeson's and cosine correlations coefficients among similarities based on linkage, content and usage patterns associated with SITE$_A$

SITE$_A$ (N = 127)	Linkage	Content	Usage
Mean	0.0735	0.1671	0.0020
S.D.	0.1413	0.3121	0.0357

SITE$_A$ (N = 127)	Pearson	Sig.	Cosine
Linkage-Content	0.3201	0.000	0.4682
Linkage-Usage	0.0184	0.017	0.0423
Content-Usage	0.0429	0.000	0.0644

5. Discussion

This section discusses strengths and limitations of the GSA framework for structuring and visualising large hypermedia information spaces with respect to existing techniques.

5.1. Strengths and limitations of GSA

GSA extracts and visualises a number of types of salient relationships in a hypermedia information space, such as hypertext linkage, content similarity and usage patterns illustrated in this paper. The aim of this work is to provide a generic approach to the development of a wide range of hypermedia systems, including large, distributed

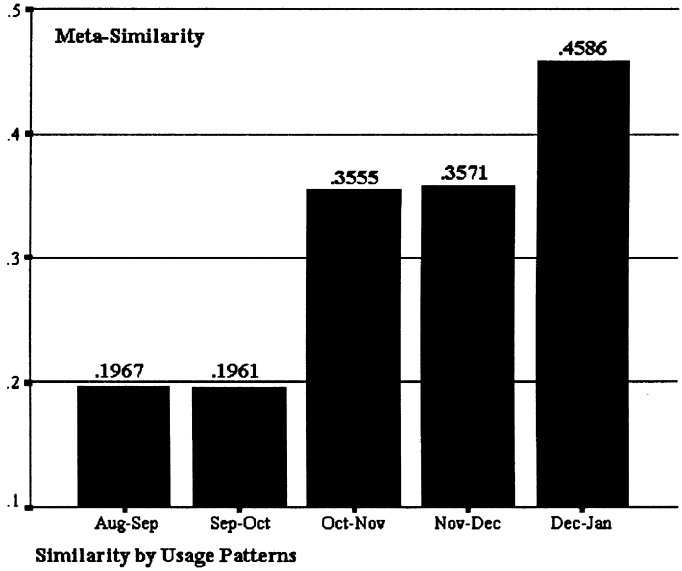

Fig. 12. Meta-similarities of adjacent monthly usage pattern-based similarities over six months.

general-purpose hypermedia systems and specific application systems for information retrieval, digital libraries and electronic publishing.

In GSA, the underlying semantic structure is explicitly represented. Graphical user interfaces based on such representations not only provide valuable navigational cues for users to use an information space more effectively, but also allow users to perform a wider range of direct manipulation tasks on virtual link structures in order to understand the information space from a number of perspectives.

GSA is a unifying framework in that each component model associated with a particular type of similarity measures in GSA can be used independently as well as in association with other component models. In contrast, related work such as Ref. [19] used feature vectors that represent the collective effects of heterogeneous attributes. Users may not have the flexibility to explore a particular dimension of the information space.

Many existing information visualisation techniques mainly focus on isolated characteristics of documents and their use in an information space, such as file-size, last modification time and single-page visit counts. GSA emphasises the fundamental role of inter-document relationships in structuring and visualising hypermedia systems. On the other hand, only the most salient relationships are shown in associative network representations and subsequent virtual reality-based visualisations. In the future, these two approaches, within- and between-documents, should be incorporated.

A dynamic approach to the visualisation of browsing history is often based on the rich information obtained locally from the client-side of the WWW (see Refs. [11,12]). A static approach, on the other hand, tends to focus on the overall structure, which often require information only remotely available from the server-side of the WWW. The two approaches are different in terms of scope, lifetime and granularity of resultant structural visualisation. GSA is originally designed as a static approach. It involves a number of computationally expensive algorithms and time-consuming data collection tasks, especially for analysis based on distributed sources of data, but it can be tailored to facilitate some tasks supported by a dynamic approach.

A common problem encountered by many information processing techniques with the WWW is to what extent techniques can scale up to deal with increasingly large-sized collections of documents. A promising strategy is to divide and conquer: rapidly and recursively split a large collection of documents into a number of smaller clusters of more focused documents until the resulting computational complexity can be dealt with by existing special-purpose algorithms. We are exploring appropriate classification and clustering schemes for experimentation with GSA on large document collections.

5.2. Graphical user interface and virtual reality design

A number of user interface design issues must be addressed in order to incorporate resultant structural models and visualisations into practical systems. A static Pathfinder network becomes increasingly cluttered as the number of documents in the underlying information space increases. Two possible solutions are particularly attractive: fisheye view models and virtual reality-based user interfaces.

In essence, fisheye views are information filters based on the importance of an object in the information space and the pre-defined distance between the present standing point of

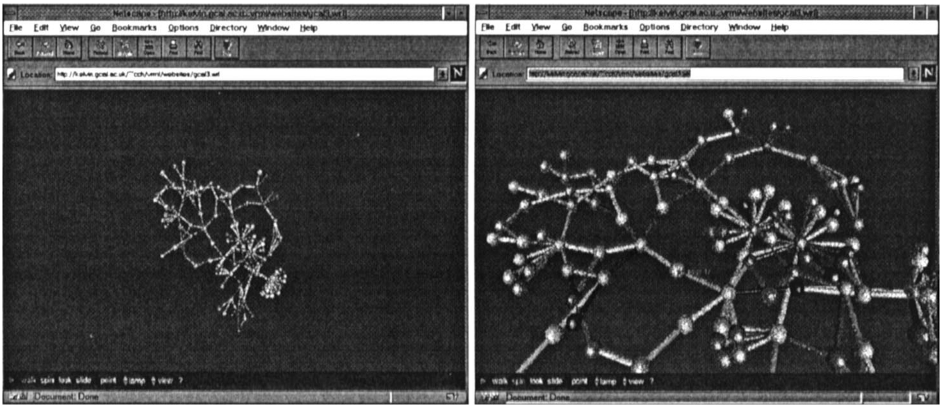

Fig. 13. A seamless movement in a virtual reality-based user interface in VRML (SITE_A).

the user and the object. For example, one may consider a fisheye view model defined as follows:

$$\mathrm{DOI}_{\mathrm{fisheye}}(x, x_f) = \mathrm{PageCount}(x) - D(x, x_f),$$

where x_f is the focal point and x is a node in the minimum-cost Pathfinder network. The *a priori* importance (API) is defined as the corresponding page visit count PageCount(x). $D(x,x_f)$ is the minimum-cost path connecting the two points x and x_f. With an appropriate threshold to the DOI function, a fisheye view can simplify graphical user interfaces of the structural visualisation.

Using virtual reality-based user interfaces for the structural visualisation has several advantages. Users can freely and seamlessly move back and forth to adjust a wide range of views from close-up views of local details to a birds-eye view of the overall structure to suit their needs. We automatically generated virtual reality-based user interfaces in VRML. Users can interact with the underlying information space more intuitively through Web browsers. Fig. 13 shows the structural visualisation of the website SITE_A rendered in VRML. The VRML interface enables users to perform direct manipulation tasks on the virtual structure as a whole as well as on individual objects. Point-and-click on a sphere will take the user to the corresponding document on the WWW.

5.3. Evaluation

GSA introduces some new ways of organising, visualising and using a hypermedia system. The evaluation of this generic approach requires re-examination of existing theories and methodologies of usability evaluation in human–computer interaction as well as traditional effectiveness evaluation criteria in information retrieval. We are particularly interested in the effects of individual differences that have been found significant in traditional hypertext systems and that are closely related to visual information processing, including spatial ability and cognitive styles [7].

While recall and precision remain the standard and the most popular evaluation measures in information retrieval, many other factors may be more important than recall and precision in practice [17,18,28]. The overall usability of an interactive information system, such as easy to learn, easy to use, system reliability and response time, is essential for users' satisfaction and acceptance.

Structuring and visualisation in GSA takes into account usage patterns. In a dynamic document space transformation process, a document space is incrementally optimised in response to various feedback such as usage patterns and the quality of search results. Will the three types of inter-document similarities converge as the structure of the underlying information space is optimised to match users' preferences and semantic characteristics? Will users be able to navigate more effectively with the help of the structural visualisation? Will the visualised structure provide an ideal meeting point for query- and navigation-based information retrieval? We will be investigating these questions and other human factor issues in our subsequent studies.

6. Conclusions

The GSA framework described in this paper has several distinct features for structuring and visualising hypermedia spaces. A number of conclusions are drawn based on these features. (1) Virtual link structures automatically generated by GSA can be used for reinforcing existing hyperlinks, identifying possible missing links and suggesting new hyperlinks. GSA provides a basis for combining query- and navigation-based information retrieval. GSA visualises the vector-space model for navigation. The provision of the vector-space readily supports query-based information retrieval. (2) Proximity-based virtual link structures and Pathfinder network scaling provide a natural basis for structuring and visualising hypermedia systems. (3) Spring-energy models provide a natural means of determining the spatial layout of an associative network. More efficient algorithms will be explored in the future to enable GSA to handle larger datasets. (4) Virtual reality-based user interfaces allow seamless transformation of structural visualisations across a range of levels of detail. More direct manipulation tasks may be enabled with the resultant structural visualisation.

References

[1] W.W. Croft, H. Turtle, A retrieval model for incorporating hypertext links, in: Proc. of the ACM Hypertext'89, Pittsburgh, PA, ACM Press, 1989, pp. 213–224.

[2] D.B. Crouch, C.J. Crouch, G. Andreas, The use of cluster hierarchies in hypertext information retrieval, in: Proc. of the ACM Hypertext'89, Pittsburg, PA, ACM Press, 1989, pp. 225–237.

[3] J. Savoy, An extended vector-processing scheme for searching information in hypertext systems, Information Processing and Management 32 (2) (1996) 155–170.

[4] L. Carr, G. Hill, D. De Roure, W. Hall, Open information services, Computer Networks and ISDN Systems 28 (1996) 1027–1036.

[5] K. Gronbaek, N.O. Bouvin, L. Sloth, Designing Dexter-based hypermedia services for the World-Wide Web, in: Proc. of the ACM Hypertext'97, Southampton, England, ACM Press, 1997, pp. 146–156.

[6] J. Conklin, Hypertext: An introduction and survey, IEEE Computer (1987) 17–41.

[7] C. Chen, R. Rada, Interacting with hypertext: A meta-analysis of experimental studies, Human–Computer Interaction 11 (2) (1996) 125–156.

[8] S. Mukherjea, J. Foley, Visualizing the World-Wide Web with the Navigational View Builder, in: Proc. of the World-Wide Web Conf. (WWW95), 1995. WWW address: http://www.igd.fhg.de/www/www95/papers/44/mukh /mukh.html.

[9] G.G. Robertson, J.D. Mackinlay, S.K. Card, Cone trees: Animated 3D visualisations of hierarchical information, in: Proc. of CHI'91, New Orleans, LA, ACM Press, 1991, pp. 189–194.

[10] J.D. Mackinlay, G.G. Robertson, S.K. Card, The perspective wall: Detail and context smoothly integrated, in: Proc. of CHI'91, New Orleans, LA, ACM Press, 1991, pp. 173–179.

[11] E.Z. Ayers, J.T. Stasko, Using graphic history in browsing the World Wide Web, in: Proc. of the 4th International World-Wide Web Conference, Boston, 1994. WWW address: http://www.w3.org/pub/Conferences/ WWW4/Papers2/270/.

[12] A. Cockburn, S. Jones, Which way now? Analysing and easing inadequacies in WWW navigation, International Journal on Human–Computer Studies 45 (1996) 105–129.

[13] G. Furnas, Generalized fisheye views, in: Proc. of CHI'86, ACM Press, 1986, pp. 16–23.

[14] K. Fairchild, S. Poltrok, G. Furnas, Semnet: Three-dimensional graphic representations of large knowledge bases, in: R. Guindon (Ed.), Cognitive Science and its Applications for Human–Computer Interaction, Lawrence Erlbaum, 1988.

[15] Y.K. Leung, M.D. Apperley, A review and taxonomy of distortion-oriented presentation techniques, ACM Transactions on Computer–Human Interaction 1 (2) (1994) 126–160.

[16] G. Salton, J. Allan, C. Buckley, Automatic structuring and retrieval of large text files, Communications of the ACM 17 (2) (1994) 97–108.

[17] G. Salton, M.J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill Inc., New York, 1983.

[18] W.M.Shaw Jr., , R. Burgin, P. Howell, Performance standards and evaluations in IR test collections: Cluster-based retrieval models, Information Processing and Management 33 (1) (1997) 1–14.

[19] P. Pirolli, J. Pitkow, R. Rao, Silk from a sow's ear: Extracting usable structures from the Web, in: Proc. of CHI'96, 1996. WWW address: http://www.acm.org/sigchi/chi96/proceedings/papers/Pirolli_2/pp2.html.

[20] R. Weiss, B. Velez, M. Sheldon, C. Nemprempre, P. Szilagyi, A. Duda, D. Gifford, HyPursuit: A hierarchical network search engine that exploits content–link hypertext clustering, in: Proc. of the ACM Hypertext'96, Washington, DC, 1996. WWW address: http://www.psrg.lcs.mit.edu/ftpdir/papers/.

[21] Nancy et al. (1996). Author please provide details.

[22] R. Botafogo, E. Rivlin, B. Shneiderman, Structural analysis of hypertexts: Identifying hierarchies and useful metrics, ACM Transactions on Office Information Systems 10 (2) (1992) 142–180.

[23] N.J. Cooke, K.J. Neville, A.L. Rowe, Procedural network representations of sequential data, Human–Computer Interaction 11 (1) (1996) 29–68.

[24] J.E. McDonald, K.R. Paap, D.R. McDonald, Hypertext perspectives: Using Pathfinder to build hypertext systems, in: R.W. Schvaneveldt (Ed.), Pathfinder Associative Networks: Studies in Knowledge Organization, Ablex Publishing Corporation, Norwood, NJ, 1990, pp. 197–212.

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

[26] M. Chalmers, A linear iteration time layout algorithm for visualising high dimensional data, in: Proc. of IEEE Visualization Conference, San Francisco, 1996. WWW address: http://www.ubs.com/ubilab/Publications/Cha96a.html.

[27] C.M. Bowman, P.B. Danzig, U. Manber, F. Schwartz, Scalable Internet resource discovery: Research problems and approaches, Communications of the ACM 37 (8) (1994) 98–107.

[28] W.M.Shaw Jr., , R. Burgin, P. Howell, Performance standards and evaluations in IR test collections: Vector-space and other retrieval models, Information Processing and Management 33 (1) (1997) 15–36.

[29] C. Chen, Structuring and visualising the WWW by generalised similarity analysis, in: Proc. of the ACM Hypertext'97, Southampton, England, 1997, ACM Press, pp. 177–186.

[30] H. Davis, J. Hey, Automatic extraction of hypermedia bundles from the digital library, in: Proc. of Digital Libraries’95, 1995. WWW address: http://csdl.tamu.edu/DL95/davis.html.

[31] J.E. Pitkow, C.M. Kehoe, Emerging trends in the WWW user population, Communications of the ACM 39 (6) (1996) 106–108.