

# Structuring and Visualising the WWW by Generalised Similarity Analysis

*Chaomei Chen*

Department of Computer Studies  
Glasgow Caledonian University  
Glasgow, G4 0BA, UK  
Tel: +44 141 3313288  
E-mail: cch@gcal.ac.uk

## ABSTRACT

This paper describes a generic approach to structuring and visualising a hypertext-based information space on the WWW. This approach, called Generalised Similarity Analysis (GSA), provides a unifying framework for extracting structural patterns from a range of proximity data concerning three fundamental relationships in hypertext, namely, hypertext linkage, content similarity and browsing patterns. GSA emphasizes the integral role of users' interests in dynamically structuring the underlying information space. Pathfinder networks are used as a natural vehicle for structuring and visualising the rich structure of an information space by highlighting salient relationships in proximity data. In this paper, we use the GSA framework in the study of hypertext documents automatically retrieved over the Internet, including a number of departmental WWW sites and conference proceedings on the WWW. We show that GSA has several distinct features for structuring and visualising hypertext information spaces. GSA provides some generic tools for developing adaptive user interfaces to hypertext systems. Link structures derived by GSA can be used together with dynamic linking mechanisms to produce a number of hypertext versions of a common information space.

**KEYWORDS:** WWW, Pathfinder networks, structural analysis, information visualisation, sequential behavioural patterns

## 1 INTRODUCTION

Navigating and authoring large, distributed information spaces on the World-Wide Web (WWW) raises a number of practical issues. Some of these issues are common among hypertext systems and some are specific to information access across the Internet. Empirical evidence shows that some well-known problems with hypertext systems, such as disorientation and cognitive overhead, also exist in the use

of the WWW [1, 9, 17]. These problems may become even more unbearable when it takes several seconds or even longer for users to download information from the Internet in rush hours. Georgia Institute of Technology's WWW User Surveys [17] shows that 69.1% of users regarded the delay in downloading Web pages as a major problem and 34.5% of users identified the difficulty of finding an existing page. In particular, 14.3% of the users reported the difficulty of visualising where they have been and where they can go and 6.5% identified the classic hypertext problem — lost in hyperspace. The memory overload remains a problem when navigating the WWW. Users on the Web often drift into unintentional “web-surfing”[9].

Previous research in hypertext navigation has suggested that the use of graphical or spatial overviews can help users to find their way in hyperspace [7]. There is a growing interest in information visualisation on the WWW. For example, the Navigational View Builder [16] visualises the structure of HTML documents on the WWW by attributes such as author, file-size and keyword. It parses HTML documents to obtain required information. However, some required meta-information may not always be available. The Navigational View Builder uses a range of information visualisation techniques such as ConeTrees and Perspective Wall. On the other hand, it does not focus on fundamental relationships typically found in hypertext systems, such as hypertext linkage patterns. Ideally, spatial relationships in visualisation should be determined by some psychological judgements of proximity, such as similarity, dissimilarity and relatedness.

Systems that operate in a batch mode are static in that they each time produce discrete representations of an evolving information space. Static systems normally draw on the rich information available across the WWW and prior extraction and analysis of overall structures. These systems often use Internet software agents, commonly known as spiders or wanderers, to acquire information automatically from the WWW.

Dynamic systems, such as MOSAICG [1] and WEBNET [9], dynamically build a graph of WWW documents recently

Permission to make digital/hard copies of all or part of this material for personal or classroom use is granted without fee provided that the copies are not made or distributed for profit or commercial advantage, the copyright notice, the title of the publication and its date appear, and notice is given that copyright is by permission of the ACM, Inc. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires specific permission and/or fee.

Hypertext 97, Southampton UK  
© 1997 ACM 0-89791-866-5...\$3.50

accessed and show users how these documents are related to one another. Dynamic systems mainly rely on client-side information and focus on representing the history of users' browsing over a short period of time. The dynamic approach may be useful for locating previously visited documents, while the static approach may be more appropriate for users' understanding of the overall structure of a large information space.

Currently, although dynamically generated documents do exist on the WWW, such as search results returned by search engines, most hypertext documents on the WWW contain hyperlinks that are physically embedded in documents. This static node-link binding mechanism makes it difficult for structuring and re-structuring an evolving hypermedia network [4]. The significance of being able to impose a set of hyperlinks dynamically on a collection of documents has been recognised. Dynamic node-link binding has greater potential for increasing the flexibility and maintainability of a hypermedia network [4].

In this paper, we present a unifying framework for structuring and visualising documents according to three types of similarity measures in hypertext systems. We adapt techniques from hypertext structural analysis, information retrieval and state-transition analysis of usage patterns in order to extract various similarity patterns, and represent these patterns as a special type of associative networks known as Pathfinder networks. This approach is applied to a number of departmental WWW sites in British universities and conference proceedings on the WWW. We also discuss how the three similarity models can be integrated and iteratively evaluated.

## 2 ACCESS LARGE HYPERMEDIA SYSTEMS

In this section, we describe some related work and information visualisation models designed to reveal structural insights of a large hypertext information space on the WWW. Pirolli, Pitkow and Rao's study [18] and HyPursuit [20] are two notable examples of taking into account hypertext linkage, content similarity and usage information on the WWW. Further comparisons between these studies and our framework will be given when we introduce our unifying framework in the next section.

Research in hypertext suggests that users may benefit from a number of different presentations of a particular topic. Link structures of different natures may effectively convey salient interrelationships associated with an underlying information space. This study is to find a way of generating and integrating various link structures such that the semantics associated with each type of link structure can be preserved and presented to users. User interfaces based on such link structures will allow users to access information from various perspectives as in a hypertext system that supports typed links.

A large hypermedia system can lead to a complex overview of its underlying structures. Users found it extremely difficult to understand and interact with such 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 [14, 12]. Assume that the user is currently at node  $x_0$ , known as the focal point, the DOI function is defined as

$$DOI_{\text{fisheye}}(x, x_0) = API(x) - D(x, x_0)$$

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 can be generated with a threshold so that only nodes with sufficient DOI are displayed in the view.

API provides a flexible mechanism to define fisheye views based on different proximity measures. By choosing a different API function, one can produce a fisheye view which emphasizes a particular type of structural patterns[12]. For example, the number of times that a node has been visited can be used to define a user-centred fisheye view, in which popular nodes will be highlighted for easy access.

In order to produce a graphical user interface which naturally represents underlying relationships in a hypertext system, one must investigate salient patterns associated with these relationships. In this paper, we focus on extracting underlying relationships in a hypertext information space and representing resultant patterns for structuring and visualising the information space. Existing techniques such as fisheye views can be subsequently incorporated into such systems with improved spatial configuration mechanisms.

## 3 GENERALISED SIMILARITY ANALYSIS

Generalised Similarity Analysis (GSA) is a unifying framework for extracting structural patterns from a range of proximity data regarding a hypertext-based information space. A number of fundamental interrelationships in hypertext, such as hypertext linkage, content similarity and browsing patterns, can be naturally integrated within this framework. Pathfinder network scaling techniques are used for extracting and representing the most essential relationships. The GSA framework is introduced in the following subsections, including three types of document similarity measures used in this study and Pathfinder networks.

### 3.1 Document Similarities

Proximity relationships, such as similarity and relatedness, can be measured psychologically or statistically. In hypertext, some fundamental relationships are hypertext linkage, content similarity and browsing patterns. These relationships are used to estimate document similarities in this study.

*By 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. [2] 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. suggested that large hierarchies may be displayed with fisheye views, which balance local details and global context [12, 14].

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, two documents that share common ancestors will be regarded more similar than documents otherwise equally connected but without such common ancestors. In HyPursuit, document similarity by linkage is defined as a linear combination of three components: direct linkage, ancestor and descendant inheritance.

Pirolli, Pitkow and Rao [18] developed a model which characterises documents on the WWW by various attributes associated with these documents, such as the number of incoming and outgoing hyperlinks of a document, how frequently the document was downloaded from the hosting WWW server and content similarities between the document and its children. By controlling the weight associated to each attribute, they suggested that such Web document feature vectors can be used to categorise the nature of a page and to predict the interests of visitors to that page.

In this study, document proximity is measured as similarities between documents. For example, similarity 1 means that the corresponding documents are identical, whereas 0 means they are completely different. The document similarity by hypertext linkage in this study is defined as follows:

$$sim_{ij}^{link} = \frac{link_{ij}}{\sum_{k=1}^N link_{ik}}$$

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. This definition also takes into account the overall connectivity of the document  $D_i$ , which can be related to the ROC metric defined in [2]. For simplicity, ancestors and descendants are not considered.

*By content similarity.* The vector-space model [19], 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. Thus the distance between two documents can be determined by comparing 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. Salton et al. produced a system for a fully automatic generation of semantically based hypertext networks using the vector-space model by creating links between documents that are sufficiently similar [19].

In this study, we use the well-known  $tf \times idf$  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

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

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  $D_i = (w_{i1}, w_{i2}, \dots, w_{iT})$  and  $D_j = (w_{j1}, w_{j2}, \dots, w_{jT})$ :

$$sim_{ij}^{content} = \sum_{k=1}^T w_{ik} \cdot w_{jk}$$

We used up to 500 most frequently occurring significant terms in each collection to create the corresponding vector space. Stopwords were removed in advance.

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. In [18], the vector-space model is restricted to the existing hypertext structure in that content similarities were considered between documents that were connected by existing hyperlinks, while in our study content similarities are considered across the entire collection of documents in order to find out under-represented patterns.

*By 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 in [18] to indicate the degree of relatedness between the two documents.

We have applied a state transition approach to extracting behavioural patterns of users with a hypertext system [6]. The dynamics of a browsing process can be captured by state transition probabilities. Transition probabilities can be used to indicate document similarity in the nature of 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 derived the similarity between document  $D_i$  and  $D_j$  in the view of users:

$$sim_{ij}^{usage} = \frac{p_{ij}}{\sum_{k=1}^N p_{ik}}$$

### 3.2 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) [10, 15]. The essential concept underlying Pathfinder networks is pairwise similarity, for example, between documents in this study. 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)[10]. Pathfinder has been applied to a number of human-computer interaction problems [10].

Patterns in proximity data are represented by links in PFNETs. 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  $w_1, w_2, \dots, w_k$  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:

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

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.
3. The triangle inequality is satisfied for all paths with 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 distance between nodes in a network is the length of the minimum-length path connecting the nodes; such a path is known as the geodesic connecting the nodes. A minimum-cost network (MCN), PFNET( $r=\infty, q=n-1$ ), has the least number of links.

The graph layout algorithm that generates the graphical representation of a Pathfinder network is based on the spring model described in [13]. There is a growing interest in spring models in information visualisation (see [5]) because the idea is simple and intuitive. In a spring model, nodes are connected by weighted links, for example, proximity measures 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. The version of the Pathfinder implementation used in this study can handle a network with 200~300 nodes, while some websites we sampled have 4,000~5,000 valid documents. We are investigating alternative algorithms based on techniques such as stochastic algorithms [5] and simulated annealing.

In a recent study [8], neural-network techniques were used to categorise WWW documents into related groups based on most frequently occurring terms in these documents. Relationships among these groups were represented by a self-organised feature map, in which each area corresponds to a group of documents. The major problem with this approach is that the nature of the map is difficult for users to interpret. On the other hand, users prefer the idea of a map that would allow them to browse the Web visually.

The major advantage of Pathfinder networks is that salient relationships among documents are extracted by patterns associated with minimum-cost paths. 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.

#### 4 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.

##### 4.1 Data Collection

HTML documents were automatically retrieved from 13 WWW sites by HARVEST's Gatherers. HARVEST<sup>1</sup> is an integrated set of tools to gather, extract, organise, index and search relevant information on the Internet [3]. HARVEST provides 2 useful subsystems, Gatherers and Brokers, to collect and index subject-specific information on the WWW. We installed HARVEST Release 1.4 on an HP workstation with HP-UX 9000 operating system directly connected to the Internet. HARVEST's Gatherers were used to download and digest HTML documents using a modified extraction algorithm, known as summarizers in HARVEST, which determines what types of information should be collected. The boundary of a WWW site was determined by some pattern matching rules which instruct a HARVEST Gatherer to retrieve documents from valid URLs, Unique Resource Locators, on specific WWW servers. CHI'96 papers were automatically retrieved from the WWW<sup>2</sup>.

##### 4.2 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 salient structural elements of a document. The fulltext 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>, <anchors> and <lists>. A stopword list was compiled based on the complete list of terms that occurred in the collection of documents.

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 conduct the Generalised Similarity 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.

#### 5 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. Standard 0.005 spring-energy threshold was used throughout the study to generate Pathfinder networks.

##### 5.1 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  $13 \times 13$  asymmetric proximity matrix. Pathfinder is particularly suitable to deal with asymmetric proximity data. Figure 1 is the connectivity map of the 13 sites in Scotland.

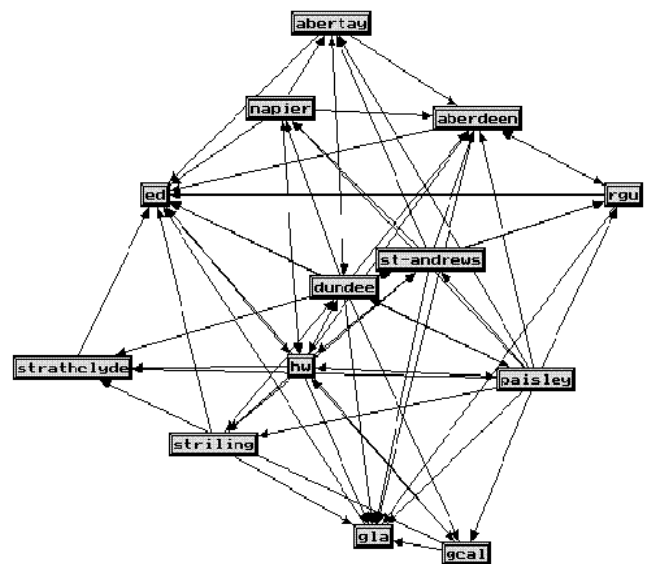


Figure 1. A connectivity map of 13 departmental WWW sites, shown as PFNET( $r=\infty$ ,  $q=N-1=12$ ).

Sites connected with shorter links have more hyperlinks between them. Since the matrix is asymmetric, the connectivity map is a directed graph. For example, the short, directed link from the site at Glasgow Caledonian University (gcal) to Glasgow University (gla) indicates there are many hyperlinks from site gcal to site gla. Such connectivity maps can be used to identify the distribution of expertise in specific areas.

The average number of HTML documents at a sampled departmental WWW site was 1,672 (max. =4,177; min. = 8) with 1,552 associated links. 92% of these links were HTTP links (1,414); 5% were FTP (61) and 3% were Gopher (46). The average number of FORMs used at a departmental site was 12, which may reflect the extent to which users can

<sup>1</sup> <http://harvest.cs.colorado.edu>

<sup>2</sup> <http://www.acm.org/sigchi/chi96/proceedings/>

interact with a WWW document. 51% of the HTTP links terminated at a *uk* server and 29% pointed to a *us* server. Within *uk* domains, the majority of links from these WWW sites led to academic domains rather than commercial domains (*ac.uk/co.uk*= 10). In contrast, the ratio of links to *edu* to that to *com* domains was 10:12.

### 5.2 Structure by Hypertext Linkage

Figure 2 shows the structure of a WWW site (*SITE<sub>A</sub>*) 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).

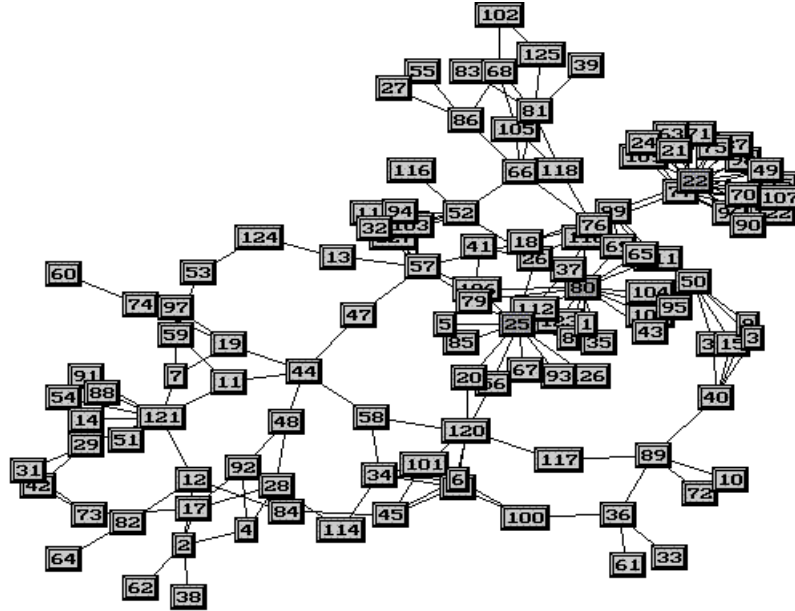


Figure 2. The structure of *SITE<sub>A</sub>* (partial) by hyperlinks, shown as a PFNET( $r=\infty$ ,  $q=N-1=126$ ) with 189 links.

Some nodes are more special than others. For example, Node 22, 80 and 25 in Figure 2 led to document clusters on an HTML tutorial, collaborating researchers and object-oriented programming at the site, respectively.

### 5.3 Structure by Content Similarity

Two Pathfinder structures were generated for papers in the CHI'96 proceedings. Each paper is represented by a vector of 108 significant terms selected from the list of most frequently occurred terms. The structure in Figure 3, 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. The relationship between two papers was considered across all the paths connecting the two documents and only the minimum-cost path was used to represent the salient relationship with an improved clarity. The resultant graph is a natural candidate for an overview map of the information space.

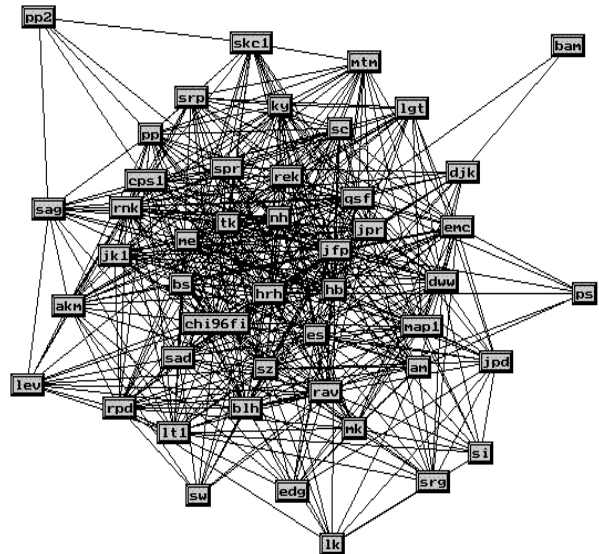


Figure 3. CHI'96 papers structured according to all the connecting paths (Stress < 0.005, Links=516).



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.

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. Figure 8 shows a PFNET derived from similarities based on first-order state-transition probabilities. Cluster A is enlarged as Cluster A\*.

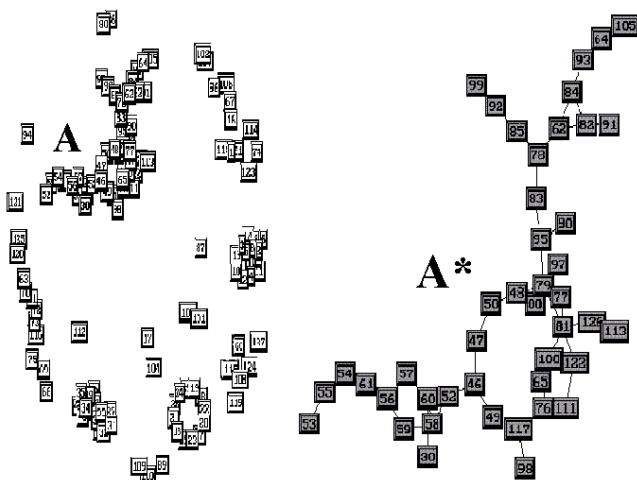


Figure 8. The structure of SITE<sub>A</sub> by state-transition patterns, shown as a PFNET( $r=\infty$ ,  $q=N-1$ ) with 67 links.

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.

## 6 DISCUSSION

This section discusses strengths and limitations of the GSA framework, compared with existing techniques for structuring and visualising information on the WWW as well as issues concerning the use of GSA in practice and its evaluation in a larger context.

### 6.1 Strengths and Limitations of GSA

GSA extracts salient relationships from a hypertext-based information space. The three types of relationships, hypertext linkage, content similarity and usage patterns, are of fundamental importance to structuring and using hypertext systems. A wide range of hypertext systems, such as many large and evolving information spaces on the WWW, face similar problems regarding the usability and

maintainability of hypertext networks. GSA is particularly designed to deal with these problems.

GSA has some distinct features. (1) GSA emphasizes that users can substantially benefit from explicit, graphical representations of salient relationships in hypertext systems, and these graphical representations should be incorporated into user interfaces so as to reduce cognitive burdens on users in browsing. (2) Each component model in GSA can be used independently for extracting structures of a particular type so that users may contrast patterns in distinct characteristics. In contrast, related work such as [18] combines various features into a monolithic feature vector. Consequently, the resulting inter-document relationship is a combined effect of a range of factors. Users may not be able to assess how documents are related along a specific dimension. (3) GSA focuses on relationships that are particularly essential for hypertext systems and these relationships are preserved in resulting network representations. Many existing information visualisation techniques are based on storage information such as file-size and last modification time, and often use hierarchical structures as the basis of visualisation. Differences between the two approaches should be evaluated by further empirical studies.

In general, GSA is a static approach in that it operates in a batch mode, as opposed to a dynamic approach such as [1] and [9]. A dynamic approach is often based on the rich information obtained locally from the client-side of the WWW, whereas a static approach focuses on global structures, which are often derived from information available from the server-side of the WWW. The two approaches are different in terms of scope, lifetime and granularity of resultant structural visualisation. What is an appropriate combination of the two approaches should be further studied in the future.

One of the common problems encountered at several stages in this study is to scale up the structural mechanisms to handle a small-to-medium sized collections on the WWW. An appealing approach is to start with conventional structural analysis so that a large collection of documents can be split into a number of smaller clusters that can be handled with reduced computational complexity.

### 6.2 Usability Issues

A number of usability issues must be considered in order to incorporate resultant structural models into practical systems. For example, a Pathfinder network becomes increasingly cluttered as the number of documents in the underlying information space increases. There are several possible ways to deal with this issue. One is to use existing display techniques such as fisheye views, which provide adequate access to specific local information as well as contextual structure. In particular, the API function, *a priori*, for a fisheye view can be defined as follows:

$$DOI_{fisheye}(x, x_f, z) = D(x, z) - D(x, x_f)$$

where  $x_f$  is the focal point and  $x$  is any node in the network. The  $z$  node is a reference point of a fisheye view. In each connected component of a Pathfinder network, the minimum-cost path between any two nodes is now always available so that this metric space offers a convenient and intuitive way to construct a fisheye view. By choosing different reference points, one can define a number of fisheye views to meet different needs. A node with a high relative out centrality (ROC) at a WWW site can be used as the reference point for a fisheye view. In this case, the WWW site is visualised with reference to the degree of accessibility. Alternatively, the most frequently visited HTML document at the site can be used as the reference point to visualise the structure of the WWW site.

Another obvious way is to utilise virtual-reality based user interfaces, in which users can fly through the virtual space from one node to another. Similar documents are naturally placed near to each other in the space. Users can gain a birds-eye view of the global structure by moving up to a higher view point in the sky and have a close look by moving down to a view point closer to the target document. An interesting landscape metaphor is presented in [5].

### 6.3 Integration and Evaluation

GSA provides a framework in which several structural patterns based on distinct characteristics can be compared. Integration and evaluation of these models should provide further insights into how the GSA framework can be improved and how its usability can be empirically assessed. There are several obvious options that the three types of link structures can be integrated, such as the mean, the maximum or the minimum of three similarity measures associated with the same pair of documents, namely, an integrated proximity measure  $U_{ij}$  can be defined as one of the follows:

$$U_{ij} = \text{mean}_{linkage, content, usage} (sim_{ij}^{linkage}, sim_{ij}^{content}, sim_{ij}^{usage}),$$

$$U_{ij} = \text{max}_{linkage, content, usage} (sim_{ij}^{linkage}, sim_{ij}^{content}, sim_{ij}^{usage}),$$

$$U_{ij} = \text{min}_{linkage, content, usage} (sim_{ij}^{linkage}, sim_{ij}^{content}, sim_{ij}^{usage}).$$

The choice of a particular integration should be based on an understanding of how these similarity models relate to each other. Regression techniques can be used to investigate if these similarity models are related in a particular form. For example, to find out to what extent usage patterns can be explained by underlying information structures, one can fit and examine a regression model as follows:

$$sim_{ij}^{usage} = \beta_{linkage} sim_{ij}^{linkage} + \beta_{content} sim_{ij}^{content} + \epsilon$$

Resulting correlation coefficients may provide further insights into how these similarities should be integrated.

## 7 CONCLUSION

The GSA framework described in this paper has several distinct features for structuring and visualising hypertext-based information spaces. A number of conclusions are drawn based on these features. (1) Link structures generated by GSA can be used for reinforcing existing hyperlinks, identifying possible missing links and suggesting new hyperlinks in hypertext systems. In the future, the GSA approach can be combined with dynamic linking mechanisms such as the one described in [4] so that a number of hypertext versions can be automatically generated based on an underlying information space. (2) Proximity-based link structures lead to a natural spatial representation which can be used as a graphical overview of a hypertext system together with techniques such as fisheye views and virtual reality. Initial results show that these graphical overviews are intuitive. A future topic is to assess the extent to which a graphical representation conforms with the perception of users. (3) Higher-order triangle inequalities provide an efficient knowledge elicitation rule for extracting salient relationships in proximity data. Our experience suggests that triangle inequalities should be imposed globally to reduce redundant or misleading information. (4) A class of spring-energy models provide a natural means of determining the layout of spatial representations of an associative network. More efficient graph layout algorithms will be explored in the future to enable GSA to handle larger datasets.

## 8 ACKNOWLEDGEMENT

The author would like to thank Dr Andy Cockburn and anonymous reviewers for their constructive comments and suggestions on an early version of the paper.

## REFERENCES

1. Ayers, E. Z., and Stasko, J. T. Using graphic history in browsing the World Wide Web. In *Proceedings of the 4th International World-Wide Web Conference* (Boston, December, 1994). <http://www.w3.org/pub/Conferences/WWW4/Papers2/270/>
2. Botafogo, R., Rivlin, E., and Shneiderman, B. Structural analysis of hypertexts: Identifying hierarchies and useful metrics. *ACM Transactions on Office Information Systems*, 10, 2 (1992), 142-180.
3. Bowman, C. M., Danzig, P. B., Manber, U., and Schwartz, F. Scalable Internet resource discovery: Research problems and approaches. *Commun. ACM* 37, 8 (1994), 98-107.
4. Carr, L., Hall, W., and De Roure, D. Microcosm extensions to the World-Wide Web. <http://vim.ecs.soton.ac.uk/www.html>
5. Chalmers, M. A linear iteration time layout algorithm for visualising high dimensional data. In *Proceedings of*

- IEEE Visualization Conference* (San Francisco, Oct. 1996). <http://www.ubs.com/ubilab/Publications/Cha96a.html>
6. Chen, C. Behavioural patterns of writing with collaborative hypertext: A state-transition approach. In *People and Computers XI*, M. Sasse, R. Cunningham, and R. Winder, Eds. Springer-Verlag, London, 1996, pp. 265-279.
  7. Chen, C. and Rada, R. Interacting with hypertext: A meta-analysis of experimental studies. *Human-Computer Interaction*, 11, 2 (1996), 125-156.
  8. Chen, H., Schuffels, C., and Orwig, R. Internet categorization and search: A self-organizing approach. *J. Visual Communication and Image Representation*, 7, 1 (Mar. 1996), 88-102.
  9. Cockburn, A., and Jones, S. Which way now? Analysing and easing inadequacies in WWW navigation. *Int. J. Human-Computer Studies* 45(1996), 105-129.
  10. Cooke, N. J., Neville, K. J., and Rowe, A. L. Procedural network representations of sequential data. *Human-Computer Interaction*, 11, 1(1996), 29-68.
  11. Davis, H., and Hey, J. Automatic extraction of hypermedia bundles from the digital library. In *Proceedings of Digital Libraries'95*. <http://csdl.tamu.edu/DL95/davis.html>
  12. Fairchild, K., Poltrok, S., and Furnas, G. Semnet: Three-dimensional graphic representations of large knowledge bases. In *Cognitive Science and its Applications for Human-Computer Interaction*, R. Guindon, Ed. Lawrence Erlbaum, 1988.
  13. Kamada, T., and Kawai, S. An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31, 1(1989), 7-15.
  14. Leung, Y. K., and Apperley, M. D. A review and taxonomy of distortion-oriented presentation techniques. *ACM Transactions on Computer-Human Interaction*, 1, 2 (June 1994), 126-160.
  15. McDonald, J. E., Paap, K. R., and McDonald, D. R. Hypertext perspectives: Using Pathfinder to build hypertext systems. In *Pathfinder Associative Networks: Studies in Knowledge Organization*, R. W. Schvaneveldt, Ed. Ablex Publishing Corporation, Norwood, NJ., 1990, pp.197-212.
  16. Mukherjea, S., and Foley, J. Visualizing the World-Wide Web with the Navigational View Builder. In *Proceedings of the World-Wide Web Conference (WWW95)*. <http://www.igd.fhg.de/www/www95/papers/44/mukh/mukh.html>
  17. Pitkow, J. E., and Kehoe, C. M. Emerging trends in the WWW user population. *Commun. ACM*, 39, 6(1996), 106-108.
  18. Pirolli, P., Pitkow, J., and Rao, R. Silk from a sow's ear: Extracting usable structures from the Web. In *Proceedings of CHI'96* (1996). [http://www.acm.org/sigchi/chi96/proceedings/papers/Pirolli\\_2/pp2.html](http://www.acm.org/sigchi/chi96/proceedings/papers/Pirolli_2/pp2.html)
  19. Salton, G., Allan, J., and Buckley, C. Automatic structuring and retrieval of large text files. *Commun. ACM*, 17, 2 (1994), 97-108.
  20. Weiss, R., Velez, B., Sheldon, M., Nemprenpre, C., Szilagyi, P., Duda, A., and Gifford, D. HyPursuit: A hierarchical network search engine that exploits content-link hypertext clustering. In *Proceedings of Hypertext'96* (Washington, DC. March, 1996). <http://www.psrg.lcs.mit.edu/ftpdir/papers/>