Visualizing Web Site Comparisons

Bing Liu, Kaidi Zhao and Lan Yi
School of Computing
National University of Singapore/Singapore-MIT Alliance
3 Science Drive 2
Singapore 117543
{liub, zhaokaid, yilan}@comp.nus.edu.sg

ABSTRACT
The Web is increasingly becoming an important channel for conducting businesses, disseminating information, and communicating with people on a global scale. More and more companies, organizations, and individuals are publishing their information on the Web. With all this information publicly available, naturally companies and individuals want to find useful information from these Web pages. As an example, companies always want to know what their competitors are doing and what products and services they are offering. Knowing such information, the companies can learn from their competitors and/or design countermeasures to improve their own competitiveness. The ability to effectively find such business intelligence information is increasingly becoming crucial to the survival and growth of any company. Despite its importance, little work has been done in this area. In this paper, we propose a novel visualization technique to help the user find useful information from his/her competitors' Web site easily and quickly. It involves visualizing (with the help of a clustering system) the comparison of the user's Web site and the competitor's Web site to find similarities and differences between the sites. The visualization is such that with a single glance, the user is able to see the key similarities and differences of the two sites. He/she can then quickly focus on those interesting clusters and pages to browse the details. Experiment results and practical applications show that the technique is effective.

Categories and Subject Descriptors
H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia – *architectures, navigation, theory, user issues.*

General Terms
Human Factors, Design, Algorithms.

Keywords
Visualization, Web site comparison, business intelligence, user-interface, browsing.

1. INTRODUCTION
With the advance of the Web and popularity of e-commerce, companies, organizations and individuals are increasingly using the Web to conduct businesses, and disseminate information on a global scale. A large amount of important and even sensitive information is now published on the Web by these companies, organizations and individuals. The amount of information is still growing at a phenomenal rate. With all this information publicly available, it offers a great opportunity for companies, organizations and individuals to get to know and to learn from each other, and to find useful/interesting information from each other's Web pages. This is not only important for individuals, but also important for businesses. In a business environment, knowing one's competitors is of crucial importance to the survival and growth of any business. Business intelligence information used to be very hard to find before the Web was available and before e-commerce became popular. Now, much of the information is accessible from the Web. However, this does not mean it is easy to obtain useful information from the competitors' Web sites. In fact, it is still a difficult and time-consuming task.

In our interactions with industrial executives, we found that they often spent a large amount of time browsing through their competitors' Web sites in order to find what their competitors are doing so that they can learn from their competitors and to design effective measures to improve their competitiveness. Manual browsing is still the dominating technique. However, manual browsing is a hard and very time-consuming task because of the following main reasons:

1. A commercial site often has a large number of pages (hundreds, thousands or more), which makes it very difficult for manual browsing without any automated assistance. Even for a relatively small Web site, the number of Web pages could overwhelm the human user, not to mention that the number of pages in a typical site is still growing at an alarming rate.
2. People often use anchor texts of hyperlinks to help them focus and decide whether to go to lower level pages to obtain further details. However, anchor texts can be quite ambiguous and misleading. This situation becomes worse if one browses through the Web site of a competitor from a different country due to culture differences.
3. Different companies may organize the same information very differently. One company may use one page and another may use a few pages. Some even put different aspects of the same information in different categories. This makes it hard to obtain the complete information about a particular item.

All these factors and more make manual browsing a difficult task. In many situations, even if a page exists in a Web site, the user may not be able to find it because of the complex link structure of the site. Although some sites provide site search to enable the user to find related information using keywords, the precision of such searches is often very low. In addition, searching does not help

WWW 2002, May 7-11, 2002, Honolulu, Hawaii, USA.
ACM 1-58113-449-5/02/0005.

the user find unanticipated or unexpected information, i.e., those pieces of information that the user does not know of and thus is unable to issue any search query.

Broadly speaking, companies often want to find the following types of pages from their competitors:

Similar pages: These pages in the competitor site are quite similar to some pages at the user site. Such pages allow the user to know how the competitor presents the same information and/or conduct the same business, from which the user can learn from his/her competitor and exploit any weakness of the competitor. If such pages can be highlighted to the user, he/she can quickly focus his/her attention on them to perform detailed analysis, e.g., browsing the actual pages.

Different pages: These pages exist at the competitor site, but not at the user site (or vice versa). Such pages are often very interesting, as companies always want to know what their competitors have or are doing that they do not have or are not doing.

Despite the practical importance of this task, little research work has been done in this area. Although there are many existing techniques that help one find useful information from the Web as we will see in the related work section, they are not able to deal with this problem effectively. In this paper, we propose a novel technique to assist the user in analyzing his/her competitors' sites through comparing the user site and competitors' sites to highlight those similar and different pages and/or information.

The basic idea of the proposed technique is as follows: Given a user site U, a competitor site C, the proposed technique performs the following steps:

1. Crawl the pages from the competitor site (assume the pages of the user site are available).
2. Combine the pages in U and the pages in C, i.e., $A = U \cup C$.
3. Cluster all the pages in A using a hierarchical clustering algorithm (see Section 3.1). This results in a cluster tree at different levels of details. At the bottom of the tree, we have all the individual pages in A. As we move up the tree, we have clusters and each cluster (or node) covers more and more pages, i.e., the clusters become larger and larger (we also have fewer and fewer clusters).
4. Visualize the clustering results (see Section 3.2). Since the clustering results are represented as a tree, this step visualizes the cluster tree. The pages from different sites are represented with different colors. Thus, one part of a cluster can be in one color and the other part in another color. This enables the user to clearly see similar and different pages from U and C. In addition, each cluster in the tree also contains a rich set of summary information about the cluster.

Using a clustering technique to group Web pages is not a new idea. Clustering has been used to provide a summary of a site or a set of Web pages [e.g., 1, 10, 34]. Each cluster center is often used to represent the pages in the cluster. Since the number of clusters is usually much smaller than the original number of pages, human users can obtain an overview of the site or the pages, and selectively drill down to see those interesting clusters. Our proposed technique is different. There are two novel ideas:

1. We use clustering to compare Web sites to help the user find similar and different pages, which are what the user is often interested in [21].
2. We combine the pages from both U and C sites into one single set A, and cluster them together. Coupled with color effects (see above), similar and different pages and/or clusters from U and C can be visualized very clearly. Specifically, after clustering, we can see three types of clusters (note that individual pages can be seen as the smallest clusters) from the visualization:

   Pure C clusters: These clusters contain only C pages, which tell us that the competitor has some useful information that the user site does not have. These pages are often very useful as we discussed above because they often represent unexpected or unknown information (e.g., products and/or services) from the competitor.

   Pure U clusters: These clusters contain only U pages, which show that the user site has some useful information that the competitor does not have. These pages are also useful as the user company can take advantage of this fact in marketing to differentiate them from their competitor.

   Mixed clusters: Each of these clusters contains pages from both C and U. These clusters show that both sites have some similar pages. Such clusters are also useful. They allow the user to carry out focused study to find the finer commonalities and differences of the two sites.

Combining the pages from both U and C sites into a single set A, and cluster them together is proven to be a crucial idea. If we do not do this but cluster the pages from each site separately, it will be very hard to find the above interesting pages because there is no common ground for comparison. Our visualization system also makes it easy to see all the interesting clusters with a single glance due to the coloring effects, i.e., different colors are used to represent C pages and U pages and also their associated clusters. Enhanced with additional features and summary information, the visualization system gives user the ability to visually explore those interesting pages/clusters without browsing through a large number of uninteresting ones.

It is important to note that when we say that the user analyzes the competitor site, we do not mean that the user has to analyze all the pages of the site. Usually, for a particular user, he/she is only interested in certain segments (those related to his/her current tasks) of the competitor's site. Our system can simply treat these segments as the pages in the competitor site to compare with some segments of the user site. Note also, in the above discussion, we only mentioned the comparison of one U site and one C site. In fact, the proposed technique can be used to visualize and compare any number of U sites and C sites at the same time. However, our experience shows that one to one comparison each time is often more effective.

In order to find truly interesting pages for the user, we also allow the user to input his/her existing domain knowledge, which is currently in the form of domain specific *stoplist* (*stopwords*). That is, the user can input some stopwords so that these words will not participate in the clustering process since they may be too common in a particular domain. For example, in a travel domain, words such as "airline", "travel", etc., are of no significance. Using them only blurs the cluster boundaries.

So far, a number of experiments have been conducted with our

system, which is called VSComp (*Visual Site Comparison*). It has also been used in two practical domains. The results show that the system is very effective.

2. RELATED WORK

The related work encompasses a number of areas, namely, clustering, visualization, and information extraction and discovery. We discuss them in turn below.

Web page clustering has been studied by many researchers. For example, [34] studies clustering of a large number of Web pages; [28] studies parallel clustering; [14] studies clustering of web users and [1] studies document clustering with user interactions. Clustering is often used to summarize a large number of pages to facilitate user focusing [34, 1, 10]. We have discussed the differences of our work with normal Web page clustering in the introduction section. We will not repeat them here.

In the Web context, visualization has also been explored extensively. For example, [30, 7, 34] study visualization of search results; [6, 17, 15] study visualization of the user's web browsing experience and surfing history; [5] studies visualization of a page in a summarized form; [23] studies visualization of Web structures; [9, 13] study visualization and tracking of Web structure changes; and [11, 19] provide visual supports for web querying. Clearly, these works are different from ours as we visualize the comparison of Web sites.

With regard to finding useful information from the Web, our technique is also different from existing approaches. Existing methods focus on what the user wants or specifies explicitly. These techniques include keyword-based search, wrapper-based information extraction, user preferences, Web and XML queries, and resource discovery. In keyword-based search, the user specifies some keywords, and a search engine (e.g., Yahoo!, Excite, Alta Vista, and Google) finds those Web pages that contain the keywords, and ranks them according to various measures [4, 12]. In Web information extraction [e.g., 2, 26], a wrapper or a specific extraction procedure is built automatically or manually for a Web page to extract some specific pieces of information requested by the user. User preference based approaches are commonly used in push type of systems [e.g., 33], where the user specifies what categories of information are interesting to him/her. The system then gives him/her only those types of information in the user-specified preference categories. In Web query based approaches, database query language SQL is extended and modified so that it can be used to query semi-structured information resources, such as XML documents and Web pages [e.g., 22, 19]. Web resource discovery finds Web pages related to the user's requests [e.g., 8, 12]. This approach uses techniques such as link analysis and text classification algorithms to find relevant pages. The pages can also be grouped into authoritative pages, and hubs.

All these existing approaches essentially view the process of finding useful information from the Web as a *query-based process*, although the queries may be of different forms, search query, information extraction query, preference query, Web, XML or semi-structured data query, and resource query. These approaches suffer from a major shortcoming. They do not help the user find unexpected information, which is unknown to the user or contradicts the user's existing beliefs. They can only find anticipated information because user queries can only be derived from the user's existing knowledge space. Yet, a lot of information that does not meet the user's queries may also be of interest to the user. These pieces of information are often novel. Our approach is able to highlight unanticipated pages and clusters, e.g., pages that exist in C but not in U, which are often very useful. Our method is also able to help the user find what he/she wants, as we will see in subsequent sections.

In [21], we reported a Web comparison system. The system uses information retrieval and data mining techniques to compare keywords in U pages and C pages to identify those potentially interesting pages. The current work is different from that in [21] in two main aspects: (1) the work in [21] does not have a visualization component. All the comparison results in [21] are listed on the screen, which are hard to comprehend. The user is not given much freedom to choose what he/she wishes to see. The system presented in this paper is much more flexible. Due to its intuitive visualization system, the user is able to view and to choose whatever interests him/her. The visualization only highlights those potentially interesting pages and clusters. (2) [21] does not use clustering. The returned result is often very large, which overwhelms the user. Clustering helps to summarize similar pages. The user can see big pictures first before focusing on interesting details. This is important when a large number of pages are involved. In fact, it is these shortcomings of the existing system in [21] that motivated this work.

Our work is also related to the interestingness research in data mining [20, 27, 31, 25]. The issue of interestingness is stated as follows: Many data mining algorithms often produce too many patterns or rules, and most of the patterns or rules are of no interest to the user. Due to the large number of rules, it is very difficult, if not impossible, for the human user to analyze them in order to find those truly interesting ones for application use. Automated assistance is needed. Past research has proposed a number of techniques [20, 27, 31, 25] for the purpose. These approaches are, however, not suitable for the Web. The reason is that rules are structured and have clear syntax and semantics, while information on the Web is semi-structured. Different methods are thus needed to help the user find interesting information from Web pages.

3. THE PROPOSED TECHNIQUE

This section presents the proposed technique, which helps the user compare and browse two Web sites. Let $U = \{u_1, u_2, ..., u_w\}$ be the set of pages in the user's Web site, and $C = \{c_1, c_2, ..., c_v\}$ be the set of pages in the competitor's Web site (which are crawled from the competitor site). As mentioned in the introduction section, our technique first combines the pages in U and C, i.e., $A = U \cup C$. It then clusters all the pages in A. After that, it visualizes the clustering results. In the following sub-sections, we discuss each step in turn. The focus is on visualization as clustering is a well-known technique, and we only used an existing clustering algorithm. Although combining the pages in U and C can be trivially done, it is a crucial idea that makes the subsequent clustering and visualization meaningful, i.e., makes it possible to show the similarities and differences of the two sites naturally.

3.1 Hierarchical Clustering

In our work, we use agglomerative hierarchical clustering, which produces a nested sequence of clusters [18, 32] like a tree structure (also called a dendrogram or a cluster tree). Singleton clusters (individual pages in our case) are at the bottom of the tree. One root cluster is at the top, which covers all the pages. The

clustering process starts building the dendrogram from the bottom level, and merging the most similar (or nearest) cluster pairs at each level up until all the pages are merged into a single cluster (i.e., the root cluster). Figure 1 shows a schematic example. At the bottom of the tree, we have 5 clusters (5 Web pages). Cluster 6 is formed by merging clusters 1 and 2 (assume they are most similar to each other). Cluster 8 is formed by merging clusters 6 and 3, etc. As we move up the cluster tree, we have fewer and fewer clusters. Since the whole clustering tree can be stored in the clustering process, the user can choose to view clusters at any level of the tree.

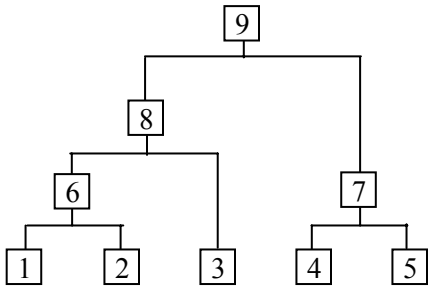

Figure 1. Hierarchical clustering

The reason that we choose this hierarchical clustering method is precisely because the cluster tree allows the user to view clusters at any level of details. It makes user analysis and visualization of the two Web sites very convenient. That is, the user can drill down and roll up to see clusters and pages at any level of granularity. This process is also facilitated by our visualization system, i.e., the users can "click, change, and see" the results at any level on the fly. In each cluster node, valuable pieces of summary information are also be stored in the clustering process, which will be very useful subsequently (as we will see later).

If we used a clustering technique that requires a fixed number of clusters as input, e.g., *k*-means clustering [18, 32], we will not only have a problem in providing the initial number of clusters *k*, but also give the user no flexibility to see the clustering at different levels of granularity. Additionally, Web pages are usually organized hierarchically based on their contents. Using hierarchical clustering fits this naturally, which facilitates the discovery of interesting pages and information.

In this work, we only use textual information in a Web page for clustering. Each Web page is thus treated as a text document. The document representation scheme that we employ is the widely used vector space model [3, 29]. This representation is commonly used in information retrieval. The assumption of the representation is that similarities, differences and the main contents of text documents can be represented by *keywords* (also called *index terms*) that appear in the documents.

In our clustering, the similarity or distance measure, which decides how pages are clustered, is the popular cosine measure used in document clustering. The details about the cosine measure can be found in [3].

Note that before clustering, we also remove words in a *stoplist* [3]. The stoplist contains words that have too low content discrimination power, e.g., "a", "the" and "and". Removal of such words reduces the length of the text documents, and also makes the similarity comparison of documents more accurate. Word stemming [3, 29] is also applied to reduce the remaining words to their stems by suffix stripping.

Domain Specific Stoplist: Apart from the general stoplist that contains those common words that are not representative of document contents, the proposed technique also allows the user to input a domain specific stoplist. This domain specific stoplist contains those words that are very common in the application domain, e.g., "airline" and "travel" in the travel domain, and also the company name of each site. These words tend to blur the cluster boundary. Our experiments show that when a domain specific stoplist is used, the resulting clusters are much more informative and accurate.

3.2 Visualization

The visualization component of the proposed technique displays the cluster tree, which gives an intuitive view of the site comparison. It takes advantage of the superb visual capability of human users to enable them to spot interesting patterns, pages, and information easily. The tree structure naturally allows the user can drill down or roll up the cluster tree. Each node of the tree represents a cluster (at a particular level of granularity) and each link represents a parent-child relationship. Each parent cluster covers all the children clusters below, i.e., Web pages associated with a parent cluster include all the pages associated with its children clusters. Figure 2 shows an example tree (only part of a full tree). The number next to each cluster node represents the cluster ID, which facilitates the user to remember which cluster he/she is analyzing. From the visualization, the user can also easily see which clusters are closer to each other. For example, in Figure 2, we can clearly see that pages 3 and 4 are clustered together before they are clustered with page 5. We can also see that the first 7 pages are closer to one another than the pages from 8 to 13.

By default, we display the whole tree on the screen. When the tree is too large, scroll bars can be used to see all the hidden clusters. Zoom-in and zoom-out mechanisms are also provided for the user to focus on a specific area or to see a global picture. Additionally, a "cut" operation may be performed to remove lower level clusters (for focusing). Figure 3 shows an example cut of Figure 2, where only the top 5 levels of clusters are displayed.

A full screen dump from one of our real-life applications is shown in Figure 4. Each cluster is represented with a small rectangle. As the user moves the mouse to point to a cluster, summary information associated with the cluster is displayed in 3 separate windows (to be discussed below).

The key features of our clustering visualization are as follows:

1. Different colors are used to represent pages in *C* and *U* sites. This enables the visualization system to reveal similarities and differences among the pages in *U* and *C* clearly. By default, we use green for *U* pages and red for *C* pages (the user can change these two colors).

2. Each cluster (represented with a rectangle) is partitioned into two parts, which represent the proportions of pages from *U* and *C* covered by the cluster respectively. The two parts are also colored accordingly using green and red colors. For example, if a cluster covers 5 pages from the *U* site and 10 pages from the *C* site, then the rectangle will have 1/3 of the area colored in green, and 2/3 of the area colored in red.

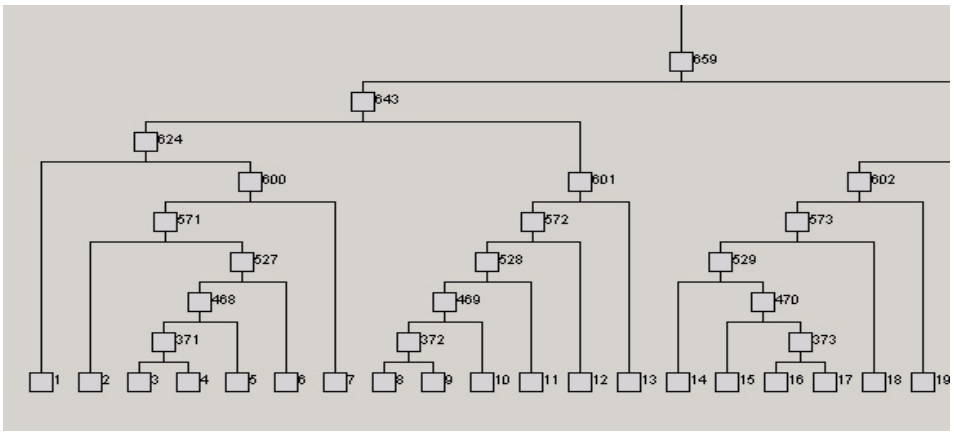

Figure 2. Part of a dendrogram or cluster tree

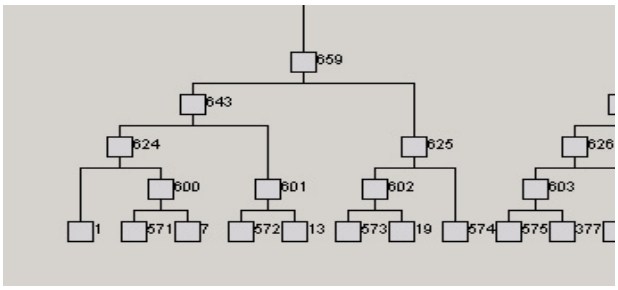

Figure 3. Result of a cut of the top 5 levels

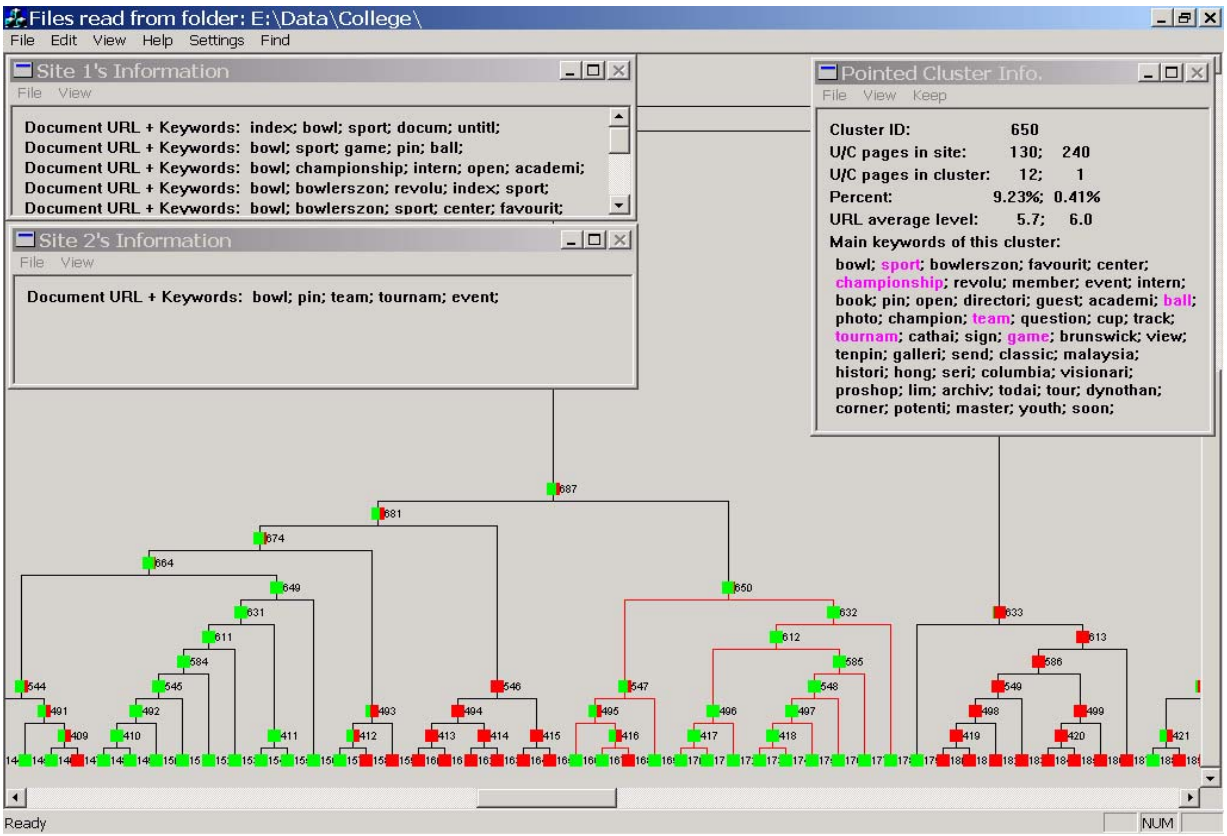

Figure 4. A full screen example

3. Each cluster is also associated with some summary information. When the user moves the mouse to point to a cluster, the summary information of the cluster will be shown in three small windows (Figure 4). The first window lists the URLs of the pages from the *U* site (denoted as Site 1 on the screen), and optional top-ranking keywords of the pages. The second window lists the URLs of the pages from the *C* site (denoted as Site 2 on the screen) and the keywords from each *C* page. The user can click on any URL to open the page for browsing. Note that due to confidentiality reasons, we could not reveal our user site names. The site's URLs are masked with "Document URL" on the screen. The third window (on the right) gives a few other pieces of summary information:

   a). Cluster ID: This is the ID of the cluster that the user is currently focusing on. The user can use this ID number to find the cluster in the tree.

   b). *U*/*C* pages in site: These two numbers give the total number of pages in the *U* site and the total number of pages in the *C* site respectively.

   c). *U*/*C* pages in cluster: These two numbers give the number of *U* pages and the number of *C* pages in the cluster respectively.

   d). Percent: These two numbers give the percent of *U* pages and the percent of *C* in the cluster with respect to the total number of pages in *U* and the total number of pages in *C* respectively. The user can use these two numbers to see how much emphasis each site is putting on the particular cluster (or topic).

   e). URL average levels: These two numbers give the average depth of the *U* pages and the average depth of the *C* pages in the cluster. This allows the user to see how deep some pages are buried in each site.

   f). Main keywords of this cluster: This gives the main keywords of the cluster. In other words, these keywords characterize the cluster. With these keywords, the user will know what this cluster is about and decide whether it is interesting to focus on this cluster and drill down to obtain further details. The system can also compare this cluster with a previous cluster and highlight the common keywords (e.g., keywords in purple in Figure 4).

4. Besides the three small windows, the system allows the user to keep the summary information window and add more. Sometimes, the user may want to study multiple clusters (especially in the same sub-tree) at the same time. He/she can use the "keep" menu option to keep the current cluster's summary information windows on the screen, and navigate to other clusters. Two examples are shown in Section 5.

5. A "find" function is also provided to enable the user to find those clusters that contain a set of user-specified keywords. These clusters are then highlighted on the screen. This function can be used to quickly focus on those interesting clusters. Section 5 gives an example.

Using the proposed clustering and visualization technique, the user can easily identify the following interesting clusters (which have been discussed in the Introduction section, and we list them here again for completeness):

- Pure *C* clusters: Rectangles representing such clusters are colored completely in red. These clusters tell the user that the competitor has some pages that the user site does not have. These pages are often very interesting.
- Pure *U* clusters: Rectangles representing such clusters are colored completely in green. These clusters show that the user site has some pages that the competitor does not have.
- Mixed clusters: Rectangles representing such clusters are partly colored in red and partly colored in green. These clusters show that both sites have some similar pages.

4. SYSTEM ARCHITECTURE

We have implemented a Web sites comparison and visualization system based on the proposed technique. The system is called VSComp, which is coded in Visual C++ under the Window's environment. It consists of 4 main components (a block diagram of the system architecture is shown in Figure 5):

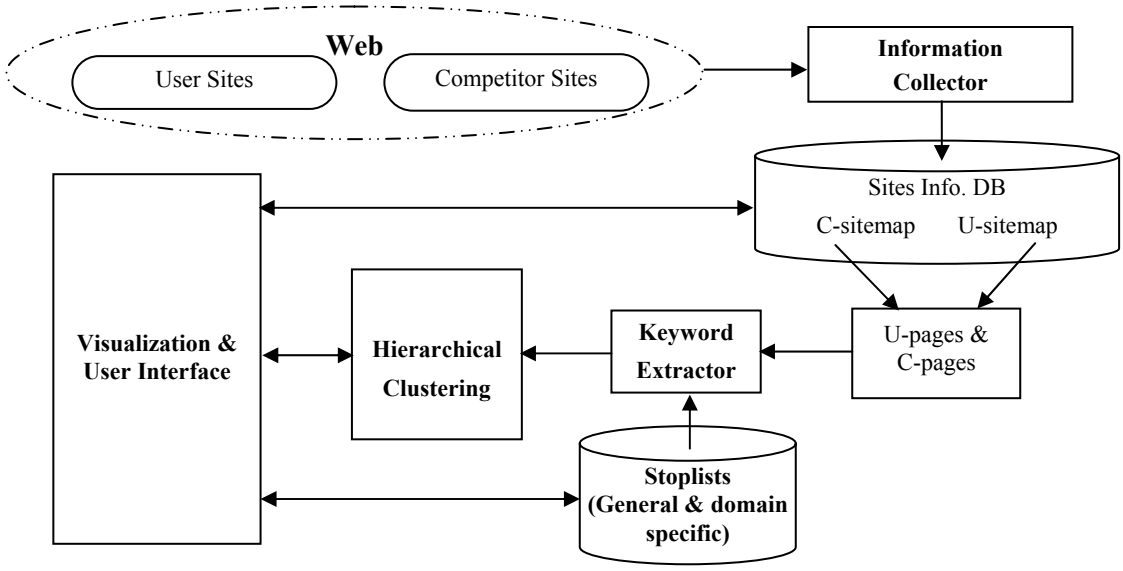

Figure 5. VSComp system architecture

1. A multi-thread information collector: It crawls a Web site to download all its pages (we use a similar crawling method as that in [24]). After crawling, a sitemap [21] is built to allow the user to choose any subset of pages for comparison.
2. A keyword extractor: It extracts keywords from a Web page, and performs the standard operations of eliminating *stopwords*, and word *stemming*. We use the Smart system [29] for this purpose.
3. A hierarchical clustering component: It clusters the user selected *C* pages and *U* pages together for site comparisons.
4. A visualization and user interface component: It visualizes the results and allows the user to interact with the system.

5. EVALUATION

This section evaluates the proposed technique and system. We first use two real-life examples from two different domains to show the working of the system and to illustrate how different functions provided by the system can be exploited for finding interesting *C* pages and information from them. We then compare VSComp with a previous system through our experiences.

5.1 Running Examples

In the two running examples, the first user site is a travel site. Its competitor site is provided by the user. At the time of writing this paper, there are 352 and 241 pages at these two sites respectively. The second user site is a junior college site. The competitor site is another junior college site. At the time of writing this paper, there are 130 and 240 pages in these two sites respectively.

Note that in the visualization, we could not reveal the user site names due to confidentiality reasons. Their sites' URLs are masked with "Document URL" on the screen (as we saw in Figure 4). However, keywords from the Web documents will be shown to facilitate understanding.

The proposed technique and visualization are versatile and can be used to find many types of interesting pages and information. In the following, we only show a few main types to illustrate the use of the system. Note also that some unnecessary parts of the figures are discarded to save space (the user can rearrange objects on the screen to give a better view).

Unexpected *C* pages (or clusters): These pages or clusters exist at the *C* site but not at the *U* site. As discussed previously, such pages are unexpected and often very interesting. In our

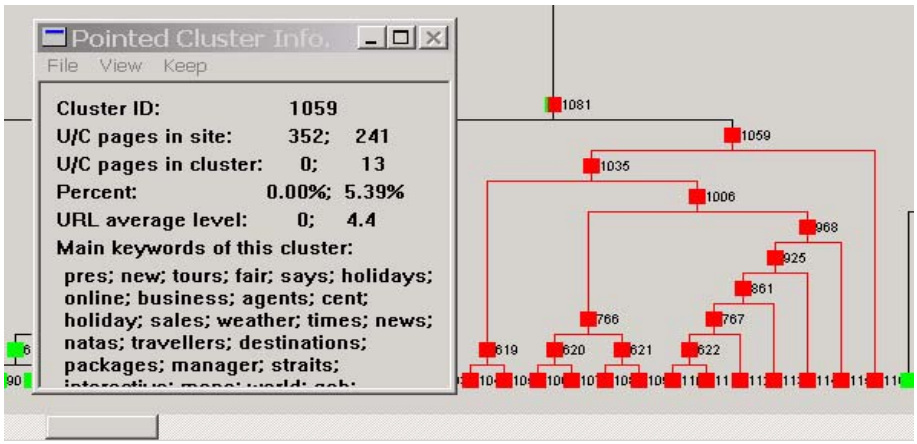

Figure 6. An example of unexpected *C* pages: travel application

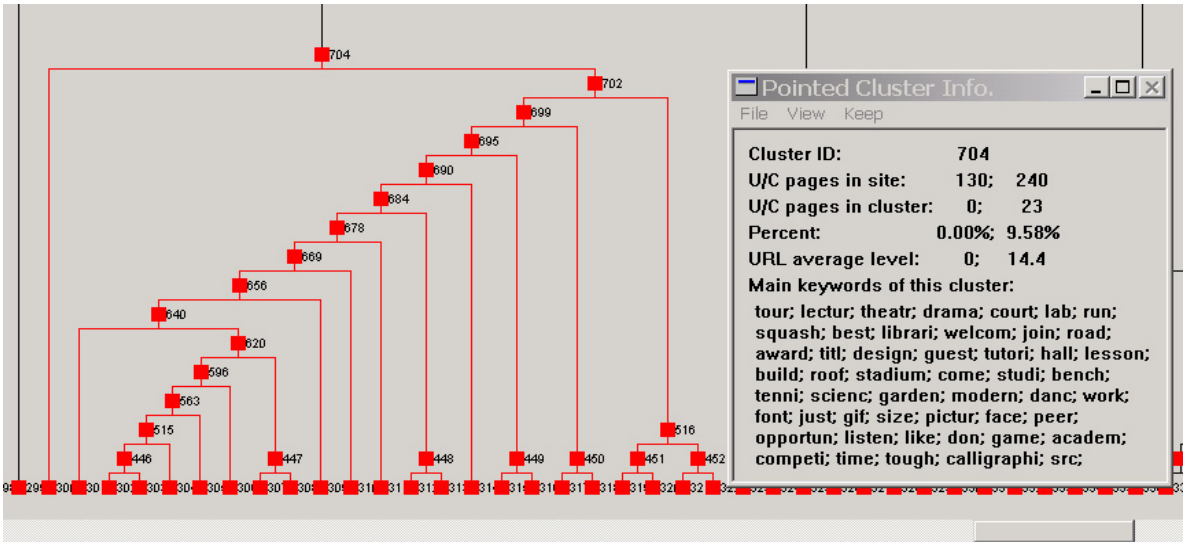

Figure 7. An example of unexpected *C* pages: college application

visualization, these pages or clusters are colored completely in red. In Figure 6, we highlighted cluster 1059, which has 13 *C* pages (5.39% of the total *C* pages) and it is a pure red cluster - all pages come from the competitor's site. As related pages are clustered together, this gave the user a surprise because it means that the competitor site used 5.39% of their Web pages to devote to something that the user site did not have at all. It was found (from the keywords on the screen and also the actual pages) that these pages were about the company's history, news and press information (it can also be seen from the other display windows that provide URLs, which are not given here.). The user site did not have any such press release or company news information. The user felt that they should add such information to their site to improve their publicity and public relation.

Another example comes from the college domain. Figure 7 shows the example. It was found that the competitor site used 9.58% of their total pages on tours of their college, while the user site did not provide any such online tour of their college. The user found that this was quite important as such online tour pages can help prospective students (or their parents) know their college better before applying to the college.

Different emphases: Different companies often have different emphases of their businesses. These emphases reflect their current directions and/or strengths. In our system, this kind of information is easily noticed and discovered. One simply finds those clusters that have very unbalanced numbers of pages from the two sites. Different color proportions in each cluster on the screen reveal such information clearly. Figure 8 shows such an example in our travel domain.

The highlighted cluster (1020) consists of 12 *U* pages, but only 1 *C* page from the competitor. The keywords show that the highlighted cluster is mainly about travel packages to Tokyo, Japan. A detailed study revealed that the *C* site had only one package to Japan. The user company can make use of this information in promotion to stress their Japan tours and to differentiate them from their competitors.

A similar example is also shown in the college application in Figure 4. In cluster 650, we can see that the user college has 12 pages on their bowling team, while the competitor college has only a single page on it.

Expected pages: In most cases, the user knows something about their competitors. He/she may want to confirm his/her knowledge about the competitor. In the travel application, the user believed that their competitor had similar tour packages to Europe as theirs. He then used the "find" function in the system to find all the Web pages on Europe tours. The results were quite satisfactory. In Figure 9, the pages on Europe tours are highlighted with circles. Note that the cluster IDs are not shown in Figure 9 due to figure size limitations. In our system, the user has the option to hide all cluster IDs on the screen.

As the user rolls up in the cluster tree, he comes to the

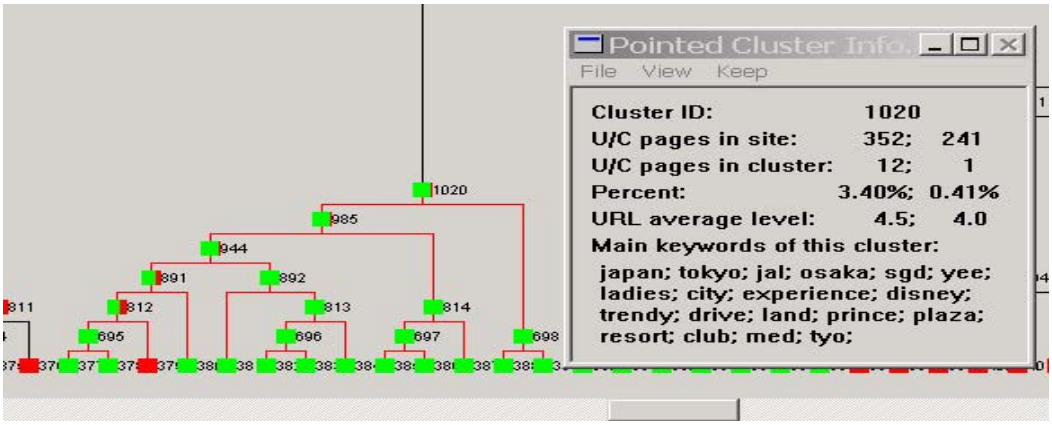

Figure 8. An example of unbalanced business emphasis: travel application

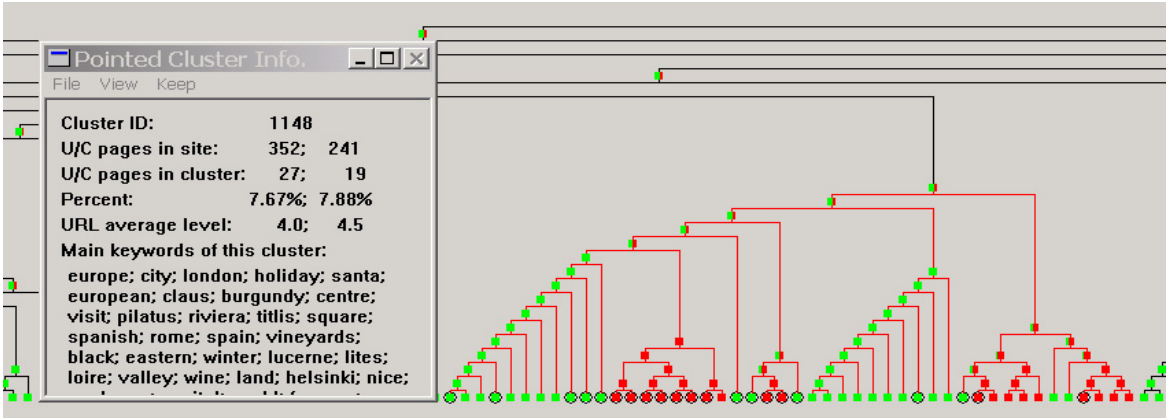

Figure 9. Finding expected pages: travel domain

highlighted cluster (1148), which shows that both the user and competitor sites have similar numbers of pages dedicated to tours to European countries (7.67% and 7.88% respectively).

Finer differences: Here we show that one can drill down from a high level cluster to its lower level clusters to find interesting details. The function "keep" is very useful here as it allows information windows from multiple clusters to be displayed on the screen at the same time. An example from the travel domain is given in Figure 10. The "keep" option is used 4 times on 4 different clusters. The screen thus shows 5 cluster information windows simultaneously. The windows were rearranged manually to make maximum use of the space.

The highlighted high-level cluster is cluster 1107. It is about tour packages to Canada. Both the user site and the competitor site have such packages. From the corresponding display window, we can see that the two sites have 10 *U* pages and 6 *C* pages respectively, which are 2.84% (10/352) and 2.48% (6/241) of their total pages. The user then drilled down to the children of cluster 1107. 4 clusters are shown on the screen (not at the same level), 1021, 894, 702 and 403 (the cluster ID 403 is too small to be seen on the screen). We observe that cluster 1021 is mainly from the user site and is devoted to travel package to the city of Vancouver. Cluster 894 is from the competitor site and is devoted to Vancouver too. Cluster 403 and 702 both come from the user site and are about Montreal and Toronto trips respectively. These clusters reveal that Vancouver tour is common to both the user site and the competitor cite. However, the user company also provides Montreal and Toronto tours. The competitor site does not.

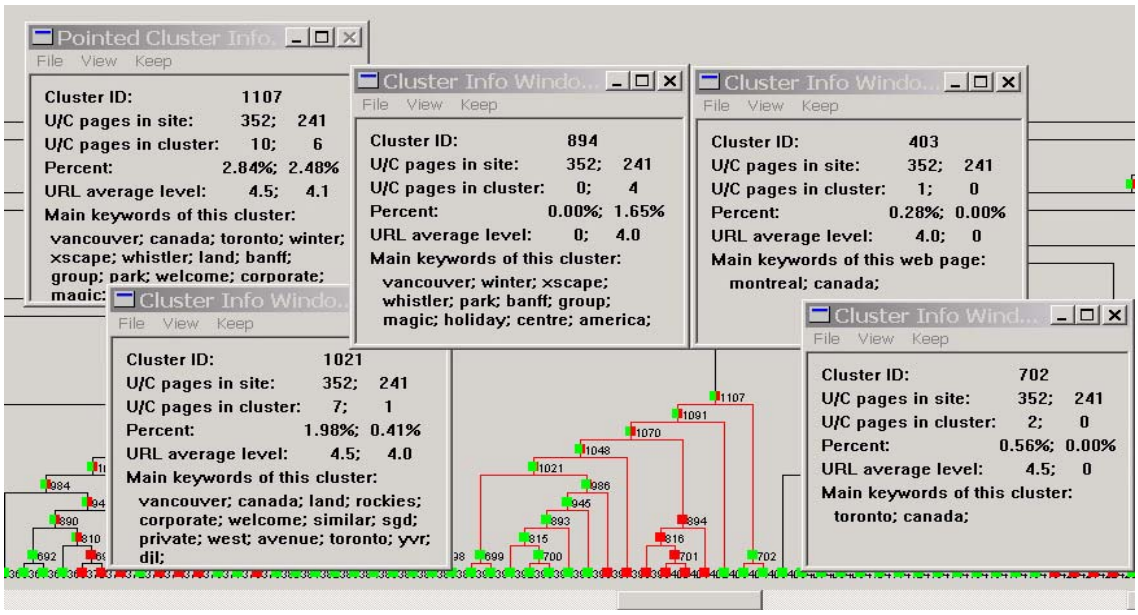

Figure 10. Drilling down the cluster tree: travel application

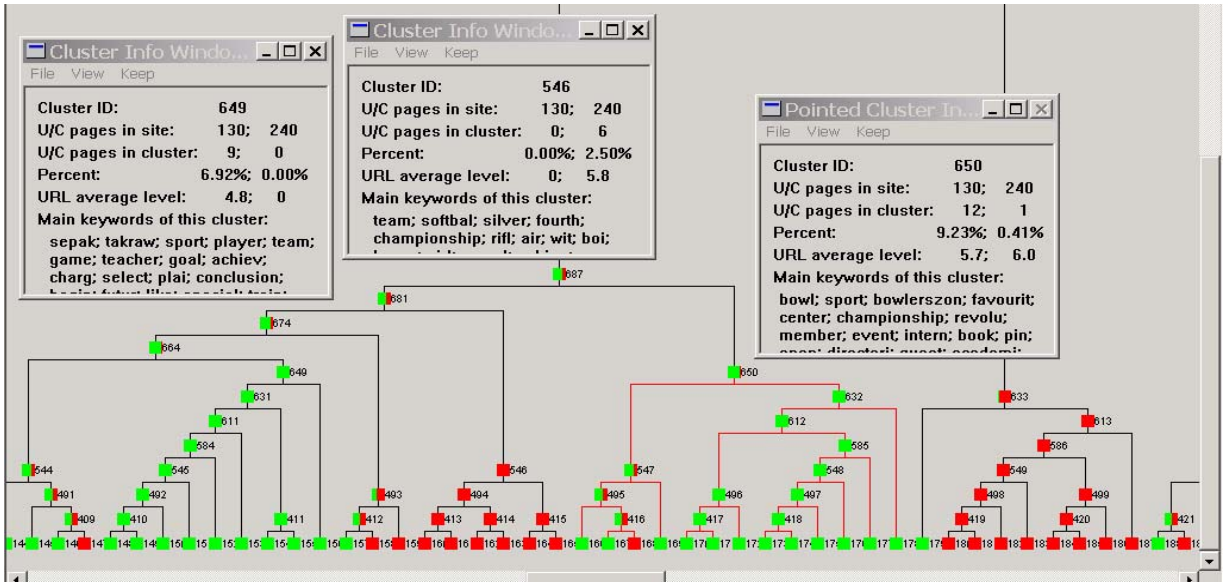

Figure 11. Drilling down the cluster tree: college application

These were later confirmed by browsing the pages. Knowing such information is quite useful for the user company in their promotion and marketing.

We now show an example from the college application (Figure 11). The high level cluster that the user was interested in was the co-curricular activities (not shown on screen due to figure size limitations). When the user drilled down to find details, he found that cluster 544 was about soccer; cluster 649 was about sepak takraw (a type of sports from Malaya); cluster 493 was about squash; cluster 546 was about softball; cluster 650 was about bowling; cluster 633 was about music band. Note that only two summary windows are kept on the screen in Figure 11 to give the reader a better view. From the visualization result, it is clear that the user college emphasizes on sepak takraw and bowling, while it lacks softball and is also weak in music.

In summary, we can say that the proposed technique makes it easy for the user to find many types of useful information from the two sites. This is because our visualization system is able to highlight those potentially interesting aspects of the comparison to the user. The user simply chooses the areas to focus on.

5.2 Application Experiences

Since the proposed technique deals with something subjective, it is difficult to have an objective measure of its performance. Here, we report our application experiences and compare our VSComp system with an existing system (called WebCompare) [21], which, to the best of our knowledge, is the only system that is able to perform Web site comparisons. We have carried out a number of experiments involving our users to check whether the new system is indeed superior. Our users were from two organizations: a travel agency, and an educational institution. Each of them compared their company site with a competitor site. Some of the examples have been shown in Section 5.1.

WebCompare [21] compares two sites by using information retrieval and data mining techniques to analyze the keywords in *U* pages and *C* pages and to rank the *C* pages according to various criteria. In terms of techniques, VSComp is different from WebCompare as VSComp uses clustering and visualization, which are not used in WebCompare. These differences result in important consequences. Our experiences show that the new system is superior to the old in two important aspects:

Flexibility: Ranking pages (done in WebCompare) does not give the user sufficient flexibility to explore those aspects that may be of interest to him/her. Those pages that are ranked high may not be those that interest the user because interestingness of information is quite subjective. Ranking is also a global operation, which sorts all the pages in the *C* site to produce a single list. This list tends to drown those interesting local areas. In our applications of WebCompare, we found that when the top ranking pages were not that interesting, users did not know what to do and became frustrated. In the new system, this does not happen, as the user is able to inspect clusters and pages from anywhere and at any level of granularity. That is, he/she can pick any interesting clusters or pages (i.e., local areas) for further analysis (see Figures 6, 7 and 8). The color effects of the clusters and hierarchical clustering make this very convenient. This cannot be done in WebCompare, which tends to dictate what pages the user should see (through ranking).

Information overload: WebCompare often returns too much information which overwhelms the user due to its use of the data mining method called association mining (which often produces a huge number of patterns [20]). Without a summarization and visualization system, it is very hard for the user to inspect them in order to find something interesting. Thus, the users often give up. The new system is friendlier. Its clustering system is able to summarize a large number of pages into a small number of clusters to give the user a global picture. He can then choose to drill down to some interesting aspects.

We are unable to directly compare the results of the two systems because VSComp does not output any ranking. It only clusters and shows those potentially interesting spots on the screen. It is up to the user to select clusters or pages for further analysis.

The new system was shown to be much more versatile and friendlier due to its flexibility and its visualization system. Without the system, it was very tedious to browse through many pages to fish for something interesting. The users often gave up after browsing some top-level pages. Our system helped them perform a more complete analysis of their competitor sites. In these applications, many pieces of useful information were uncovered, and some of them were shown in Section 5.1.

6. CONCLUSIONS

In this paper, we studied the problem of comparing Web sites in order to find useful/interesting pages from these sites. A novel technique is proposed for the purpose. It combines clustering and visualization to highlight those potentially interesting pages from the two Web sites. The key idea of the proposed approach is that Web pages from the two sites are combined first, and then clustered and displayed together. This naturally reveals those interesting pages, i.e., similar and different pages in the two sites. Enhanced with color effects of the visualization, the user is able to obtain a clear picture of those interesting spots. He/she can then focus his/her attention on those areas to browse for further details. Our applications demonstrated that the technique is effective and useful in practice. In our future work, we plan to study how hyperlinks and metadata in the Web pages can be utilized to build a more intelligent system for Web site comparison and analysis.

7. REFERENCES

[1] Allan, J., Leouski, A. V. and Swan, R. C. "Interactive Cluster Visualization for Information Retrieval". Tech. Rep. IR-116, Uni. of Mass., Amherst, 1997.

[2] Ashish, N. and Knoblock, C. "Wrapper Generation for Semi-structured Internet Sources". Workshop on Management of Semistructured Data, Ventana Canyon Resort, Tucson, Arizona. 1997.

[3] Baeza-Yayes, R. and Ribeiro-Neto, B. Modern Information Retrieval. Addison Wesley. 1999.

[4] Brin, S. and Page, L. "The Anatomy of a Large-Scale Hypertextual Web Search Engine". WWW-7, 1998.

[5] Brown, M. H., Marais, H., Najork, M. A. and Weihl, W. E. "Focus+Context Displays of Web Pages: Implementation Alternatives". WWW-6. 1997.

[6] Cadez, I., Heckerman, D., Meek, C., Smyth, P. and White, S. "Visualization of Navigation Patterns on a Web Site Using Model-Based Clustering". KDD-2000, 2000.

[7] Carey, M., Kriwaczek, F. and Ruger, S. M. "A

Visualization Interface for Document Searching and Browsing". Proc of NPIVM 2000, 2000.

[8] Chakrabarti, S., Berg, M. van den and Dom, B. "Focused crawling: a new approach to topic-specific Web resource discovery". WWW-8, 1999.

[9] Chen, Y. F. and Koutsofios, E. "WebCiao: A Website Visualization and Tracking System." WebNet97, 1997.

[10] Crouch, D. B., Crouch, C. J. and Andreas, G. "The Use of Cluster Hierarchies in Hypertext Information Retrieval". Hypertext'89, 1989.

[11] Davulcu, H., Freire, J., Kifer, M. and Ramakrishnan, I.V. "A Layered Architecture for Querying Dynamic Web Content". SIGMOD'99, 1999.

[12] Dean, J., and Henzinger, M.R. "Finding Related Pages in the World Wide Web". In Proceedings of WWW-8. 1999.

[13] Douglis, F., Ball, T., Chen, Y. F. and Koutsofios, E. "The AT&T Internet Difference Engine: Tracking and Viewing Changes on the Web". World Wide Web Journal, Vol. 1. No.1. Baltzer Science Publishers, Jan. 1998.

[14] Fu, Y., Sandhu, K. and Shih, M Y. "Clustering of Web Users Based on Access Patterns." In Proceedings of the 1999 KDD Workshop on Web Mining. 1999.

[15] Hasan, M., Mendelzon, A. and Vista, D. "Visual Web Surfing with Hy+." CASCON'95, 1995.

[16] Hersovici, M., Jacovi, M., Marrek, Y. S., Pelleg, D., Shtalhaim, M. and Ur, S. "The shark-search algorithm – An application: tailored Web site mapping." WWW-7, 1998.

[17] Hong, J. and Landay, J. "WebQuilt: A Framework for Capturing and Visualizing the Web Experiences." WWW-10, 2001.

[18] Jain, A. K., Murty, M. N. and Flynn, P. J. "Data Clustering: A Review". ACM Computing Surveys, 1999.

[19] Li, W. S. and Shim, J. "Facilitating complex Web queries through visual user interfaces and query relaxation". WWW-7, 1998.

[20] Liu, B., Hsu, W. and Ma, Y. "Pruning and Summarizing the Discovered Associations." KDD-99, 1999.

[21] Liu, B., Ma, Y. and Yu, P. S. "Discovering Unexpected Information from Your Competitor's Web Sites". KDD-01, 2001.

[22] Mendelzon, A., Mihaila, G. and Milo, T. "Querying the World Wide Web." International Journal on Digital Libraries, 1(1):54-67, 1997.

[23] Munzner, T. and Burchard, P. "Visualizing the Structure of the World Wide Web in 3D Hyperbolic Space". Proceedings of VRML'95, 1995.

[24] Najork, M. and Wiener, J. L. "Breadth-First Search Crawling Yields High-Quality Pages". WWW-10, 2001.

[25] Padmanabhan, B. and Tuzhilin, A. "Small is Beautiful: Discovering the Mining Set of Unexpected Patterns". KDD-2000. 2000.

[26] Papakonstantinou, Y., Gupta, A., Garcia-Molina, H. and Ullman, J. "A Query Transition Scheme for Rapid Implementation of Wrappers". Proc. 4th International Conference on Deductive and Object-Oriented Databases, 1995.

[27] Piatesky-Shapiro, G. and Matheus, C. "The Interestingness of Deviations". KDD-94. 1994.

[28] Ruocco, A. and Frieder, O. "Clustering and Classification of Large Document Bases in a Parallel Environment". Journal of the American Society for Information Science, 48(10): 932-943, 1997.

[29] Salton, G. and McGill, M. J. Introduction to Modern Information Retrieval. McGraw-Hill, 1983.

[30] Sebrechts, M. M., et al. "Visualization of Search Results: A Comparative Evaluation of Text, 2D, and 3D Interfaces". SIGIR'99, 1999

[31] Silberschatz, A. and Tuzhilin, A. "What Makes Patterns Interesting in Knowledge Discovery Systems". IEEE Trans. on Know. And Data Eng. 8(6), 1996.

[32] Steinbach, M., Karypis, G. and Kumar, V. "A Comparison of Document Clustering Techniques". In KDD Workshop on Text Mining, 2000.

[33] Underwood, G., Maglio, P. and Barrett, R. "User-Centered Push for Timely Information Delivery". WWW7, 1998.

[34] Zamir, O. and Etzioni, O. "Grouper: a Dynamic Clustering Interface to Web Search Results". WWW-8, 1999.