A Novel Visualization Model for Web Search Results

Tien N. Nguyen, *Member, IEEE* and Jin Zhang

Abstract—This paper presents an interactive visualization system, named *WebSearchViz*, for visualizing the Web search results and facilitating users' navigation and exploration. The metaphor in our model is the solar system with its planets and asteroids revolving around the sun. Location, color, movement, and spatial distance of objects in the visual space are used to represent the semantic relationships between a query and relevant Web pages. Especially, the movement of objects and their speeds add a new dimension to the visual space, illustrating the degree of relevance among a query and Web search results in the context of users' subjects of interest. By interacting with the visual space, users are able to observe the semantic relevance between a query and a resulting Web page with respect to their subjects of interest, context information, or concern. Users' subjects of interest can be dynamically changed, redefined, added, or deleted from the visual space.

Index Terms—Visualization model, Web search results, movement, speed.

1 Introduction

1.1 Motivation

Information visualization [5] offers the unique means that enables users to handle abstract information by taking advantage of their visual perception capabilities. By presenting information visually, it is possible for human beings to use more of their perceptual abilities in understanding and processing information. This ability of the human mind to rapidly perceive visual information makes information visualization a powerful and necessary tool for information discovery. Information visualization is a highly efficient way for human beings to directly perceive data and discover knowledge and insights. The application of the visualization techniques to Web search really broadens the horizon and the impacts of information exploration on the World Wide Web (WWW) environment.

The majority of existing Web search engines present relevant Web pages in a list with titles and short paragraph descriptions extracted from the texts of the pages. In practice, browsing the ranked list can become rather tedious and unproductive [22]. With a linear way of presenting Web search results, it is difficult for users to group results that are relevant to particular topics or their current subjects of interest. Here, Web search result visualization comes into the picture. Web search result visualization is not merely a simple way of information presentation, displaying results for a query. It also provides an interactive environment for users to explore, discover, and analyze information. Visualization environments offer an intuitive context with which users can make a decision on whether pages are relevant or irrelevant. It allows them to effectively grasp the needed information.

Although most of these visualization environments are successful in using state of the art graphic techniques for demonstrating the connection among the query and resulting Web pages, they pay little attention to developing an effective mechanism to guide users in finding the needed data among many Web search results. In addition, users' keywords for Web search might not always ensure that the results exactly match to what they are looking for. A word can have multiple meanings depending on different contexts. The order of keywords used in a query also affects the final outcomes of a search. For example, the queries "information system" and "system information" result in different outcomes. More importantly, context information on the current topic of interest is crucial and can be used to guide users to their desired information and help them in the decision about which links they should further explore. For example, with a query like "information visualization", the output could contain several groups of results corresponding to different aspects of information visualization. Assume that a user at one moment is interested in visualization techniques for software evolution, he/she could not find the desired Web page because it is not listed in the first few dozen results. He/she has to modify the query to include additional keywords (e.g. "software" and "evolution"), hoping to get the desired page. If he/she still could not locate the desired page, additional keywords must be introduced as the context information in order to narrow the set of resulting pages. That is, with the existing Web search framework, users' topics of interest or contextual information can only be introduced as a different query. Furthermore, the user must use the same sequence of keywords in order to re-produce the same results later on. However, the subjects of interest change over time. Thus, the user can easily lose his/her orientation during a Web navigation and Web search process.

Web search result visualization tools are expected to provide users semantic views to help them understand the semantic relations between a query and Web search result pages. However, existing Web search visualization tools do not show the degree of relevance between a resulting page or a group of resulting pages and the query in the context of users' subjects of interest. To accommodate the changes of preferences, the visualization of those views must be controllable and customizable based on users' interests or information needs. In other words, it is necessary to have a visualization model for Web search results that is capable of adapting to users' subjects of interest and contextual information.

1.2 Our Approach

It is believed that the use of a metaphor in the construction of a visual space would facilitate users to understand the visualization environment, to comprehend the underlying theory, to shorten learning process, to increase users' interest for the system, and to make a full use of perception capacity of human being in navigation.

Our model attempts to use the solar system to explain its visual space. In the solar system every planet or asteroid has its own orbit and a constant moving speed. All planets and asteroids rotate around the sun in the universe. They are attracted to each other by gravity. It is gravity that determines orbits and moving speed of a planet or asteroid. Our model works pretty much like the solar system. In our visual space, the defined central point – a query – is regarded as the sun, while scattered subjects of interest and Web search results are regarded as planets and asteroids. When a subject icon moves around, Web pages relevant to the query revolve around the central point (i.e. the sun) as well. The rotation speed of each Web page icon may be different and is determined by the *semantic strength* (i.e. the degree of relevance) between the moving subject and the Web page. Scattered Web pages are gravitated by the central point. The gravity in the visual

- *Tien N. Nguyen is with Electrical and Computer Engineering Department at Iowa State University, E-mail: tien@iastate.edu.*
- *Jin Zhang is with University of Wisconsin at Milwaukee, E-mail: jzhang@uwm.edu.*

Manuscript received 31 March 2006; accepted 1 August 2006; posted online 6 November 2006.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

1077-2626/06/$20.00 © 2006 IEEE Published by the IEEE Computer Society

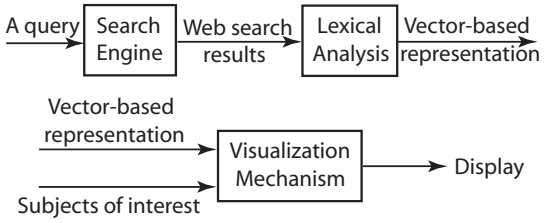

Fig. 1. Architecture

space is also defined as the semantic strength between the query and a resulting Web page. The closer to each other, the more relevant they are and vice versa. The subjects are dynamic and can be modified, added, and deleted to accommodate the change of users' preferences.

Our visualization model provides users an interactive, browsing, navigational, retrieval, and visual environment where the semantic relations among a query, Web pages, and subjects are presented. Locations, colors, and spatial distances among graphical objects in our visual space are used to represent the semantic relationships. The speed of object movement is also employed to illustrate the degree of relevance between Web search result pages and the selected topic of interest. Notice that most existing visualization systems have utilized animation and movement only for decoration purposes, rarely as the means to convey the semantic information among objects [25]. The uniqueness of our visualization model is to use *movement*, *speed*, and *distance* to visualize the degree of relevance among a query and Web search result pages with respect to users' subjects of interest and contextual information.

Figure 1 shows the general architecture of our visualization process for Web search results. First of all, a query is sent to a search engine such as Google, Yahoo, Alta Vista, etc. The returned search results are usually expressed in term of a list of hyperlinks referring to Web pages. Our visualization tool then follows those links and retrieves all result pages. Together with the original query, those pages will be lexically analyzed and their representations are created in accordance with the vector-based model [21]. The inputs for our visualization mechanism include those representation data for Web search results and the selected subjects from users. Finally, the tool will display the interactive visual space for the Web search results.

The next section discusses related work. Section 3 describes our lexical analysis process. Section 4 presents our representation for queries, Web search result pages, and topics of interest. It also explains how we compute the semantic relevance between a query and a Web page, and between a Web page and a subject of interest. The visualization model is described in Section 5. Section 6 describes *WebSearchViz*, an Web search visualization environment that is built based on our model. Our experimental studies are presented in Section 7. Conclusions appear in the last section.

2 Related Work

There has been much research focusing on the visualization of WWW search results. Lighthouse [22] is an on-line interface for a Web-based information retrieval system. It accepts queries from a user, collects the retrieved documents from the search engine, organizes and presents them to users. The system integrates two known presentations of the retrieved results: the ranked list and clustering visualization. It accepts the users' input and adjusts the document visualization accordingly. Documents in Lighthouse are clustered if they are semantically related to each other. However, users' subjects of interest and contextual information cannot be integrated. In WebSearchViz, subjects of interest and contextual information are controlled by users and the visual space will be changed in accordance with changes to subjects and contextual information. RankSpiral [30] enables users to rapidly scan large numbers of documents and their titles in a single screen. It uses a spiral mapping that can handle information density. Focus+Context interactions enable users to examine document clusters or groups in more details. Categorization and clustering techniques for search results in Vivisimo [32] are based on *just-in-time* tagging methods, while WebSearchViz uses a vector-based similarity measure to compute the degree of relevance between pages (see Section 4).

Sparkler [14] combines a bull's eye layout with star plots, where a document is plotted on each star spoke based on its rankings by the different queries or search engines. Torres [31] *et al* use a similar spiral layout to Lighthouse to display relevant images to the query. WebBook/WebForage [6] allow the search results to be organized and manipulated in various ways in a 3D space. The visual space in WebBook contains groups of pages in the form of books, which can allow users to express an elementary form of sense-making by grouping and ordering. WebQuery [7] visualizes the results of a query along with all pages that link to or are linked to by any page in the original result set. It explores the hyperlink connectivity among Web pages in visualizing Web search results. Angelaccio and Buttarazzi [2] introduce an XML-based visualization system that provides a semantic visualization by means of dynamic fields that are associated to each node. CardVis [27] visually displays Web search results using the card metaphor. Focus+Context visualization technique is nicely applied for the search space of WWW queries. However, it is not equipped with a Web browser as in WebSearchViz. Therefore, it lacks the close interoperability between the visualization and browsing environments.

NIRVE [9] contains a "document spiral" tool, which places the highest ranked document in the center. Subsequent document icons are placed and spaced along the spiral proportional to their relative document scores. Kartoo [19] creates a 2D map of the highest ranked documents and also displays the key terms. Grokker [13] uses nested circles or rectangles to visualize a hierarchical grouping of the search results. MetaSpider [8] uses a self-organizing 2D map approach to organize the documents. Mukherjea *et al* [28] provide the visualization for both text and image search results via 3D bird-eye views. INSYDER [25] is a visual information seeking system for the Web. The aim of INSYDER is to find business information on the Web. Its author evaluates different visualization techniques including HTML-List, ResultTable, ScatterPlot, BarGraph and SegmentViews. Roberts *et al* [29] implement two designs based on the bracketing concept that successfully visualize Web search results. They utilize detail-in-context and multiple view techniques to display search result data.

Furthermore, several systems for visualizing Web sites and Web-based information have been developed. Examples include Navigational View Builder [26], Harmony Internet Browser [1], Narcissus [16], and WebCutter [24]. WebPath [11] provides visualization for users' browsing history and visited Web pages. VisageWeb [17] is an information-centric user interface to the WWW that was built within the Visage data visualization environment. VR-VIBE [4] and VisIT [20] use three-dimensional visual spaces to offer much richer information than the traditional linear presentation structure. Other visualization models for information retrieval systems such as SenseMaker [3] can also be applied to the Web search results. SenseMaker is an interface for information exploration across heterogeneous sources in a digital library. Cat-a-Cone [15] is a 3D interface that integrates searching and browsing of very large category hierarchies with their associated text collections. WebSearchViz environment shares some fundamental components with our previous work, *WebStar* [35]. WebStar aims to visualize the navigational *structure* among *hyperlinked* Web pages in order to help users in keeping track of their navigational paths. Hyperlink structures among pages are visualized via lines connecting page icons, which form multi-angle, star-shaped structure in a display sphere. During a navigation process, users change their topics of interest. This change is reflected in the visual space via moving and shifting of the current focus page and topics of interest.

3 Lexical Analysis

An important component of our tool is its lexical analyzer. The lexical analyzer parses the contents of Web pages returned from a search engine. The goal of this process is to build the vector-based representation for queries and Web search result pages. In general, only single words rather than phrases are extracted and kept as access entries. The extracted keywords must go through a keyword filter process, which

filters out insignificant words such as common or grammatical words. The first phase of this lexical analysis process is to eliminate insignificant words. The studies have shown that the most common 250 to 300 words in English may account for 50% or more of any given text [21]. Those common words are *the*, *of*, *and*, *a*, *with*, *via*, and so on. Such stop words are considered insignificant if they are separated from contexts as individual words. Therefore, these words are not considered as keywords in the system. The second phase of this process is the cut-off phase. This phase is to remove words whose occurrences fall below a certain threshold in the entire set of resulting Web pages. Words appearing only once across the whole collection can be removed since they are very likely to be misspelling or mistyped words (for example, typographic errors [21]). The first phase is done on a document basis while the second is carried out on the entire document set.

To increase the precision of the semantic representation of those extracted keywords, a process called *keyword normalization* is applied. Each word extracted from the pages is normalized to its regular form. For example, plural form, past form, past participle form, and continuous form of an extracted term are replaced by their regular form. A term can correspond to multiple lexical varieties, for instance, the normalized form of "computing" and "computed" is "compute". In fact, the two un-normalized terms represent similar meanings but different forms. Without normalization processing, these un-normalized terms will be extracted and used as different entry terms in the keyword list. In other words, terms with similar concepts but different forms are scattered in different positions in the database. It will increase overhead of the system because the same meaning terms show up multiple times in the keyword list. In addition, it may lead to computational inaccuracy when the similarity between Web pages is calculated. Therefore, this keyword normalization process would improve the precision of our document relevance measures. This task is performed via the online lexical reference database WordNet [10].

To take advantage of the structural information in Web pages in determining the importance of a keyword, we also implement an adjustment mechanism. The main purpose of HTML is to carry style or layout information for hypertext. However, HTML tags also add values to the words, and can be used to semantically distinguish the words they describe to some extent. For example, sentences described by HTML tag H1 are more important than the ones by tag BODY, and they are more generic than the ones by H2 or H3 [18]. It is believed that the words appearing in HTML headers, titles, in bold, italic, underline faces are more important because people are prone to use conclusive or summary terms in title or header, and use bold, italic, underline faces to emphasize terms. In addition, in a Web page, one can use Dublin Core Elements to intentionally convey semantic information [33]. Among the defined core elements of Dublin Core Metadata, title, author, subject (keyword), and description are identified for the term weight computation. Considering these factors, we come up with a simple adjustment mechanism to improve the term weighting process. The raw frequency of a term occurring within header, title, bold, underline, strike, italic tags is automatically adjusted by a modifier. For example, a word appearing on the title would be counted as if it occurs three times. The term distribution information is used in the process of building the representation for pages, queries, and subjects.

4 Vector-based Model and Term Weights

This section describes our representation for Web pages, queries, and subjects using a data structure called *Document Term Matrix* [21].

4.1 Document Term Matrix

The rows of the matrix correspond to Web pages and the columns correspond to keywords that were extracted by the lexical analyzer. Words appearing in the query are also added into the keyword list. Web pages in this context are also referred to as *documents*. The computation process for the matrix's cells is based on our term significance measure algorithm [34]. That algorithm is modified to integrate both width and depth characteristics of a term distribution in the collection of Web search result pages. The width characteristic of term distribution is the term distribution within the whole document collection (e.g. the number of documents containing the term). The depth characteristic of a term distribution refers to its distribution within the documents containing the term (including the number of terms in those documents). In general, our term weighting measure takes into account the following dimensions: 1) term frequency in a document collection; 2) frequency retrieval characteristics of term significance; 3) document collection characteristics; and 4) term distribution, including both its depth and width characteristics at the document collection level. The following formula defines the term significance in a document:

$$W_{ik} = \begin{cases} c^{-(f_{ik}-f_{ia})^2} * log[\frac{N*D_k}{d_k*L_k}] & : \quad d_k \neq 0 \\ 0 & : \quad d_k = 0 \vee L_k = 0 \end{cases} \tag{1}$$

where W_{ik} is term significance of the term k in document i, f_{ik} is the raw frequency of term k in document i, f_{ia} is the middle value of the frequency range in document i, N is the number of documents in the document collection, D_k is the number of all terms in documents containing term k, d_k is the number of documents containing term k, L_k is the number of term k in the document collection, and c (> 0) is a constant used to adjust the impact of term frequencies on the weight. If d_k is not equal to zero, L_k is not equal to zero either; if d_k or L_k is equal to zero, the corresponding W_{ik} is defined as zero.

Luhn [23] suggested that the terms located in the middle of a frequency range had a relatively stronger distinguishing ability than those located in the two ends of the frequency range. The first part of Equation 1 describes this phenomenon ($c^{-(f_{ik}-f_{ia})^2}$). This mathematical term manifests the effect of term frequency and frequency characteristics on term significance. We use a constant c to the power of $-(f_{ik}-f_{ia})^2$ rather than $1/(f_{ik}-f_{ia})^2$ to soften the effects of the variable changes on term significance. Another benefit of using a constant c is that by changing the value of constant c, we can control the degree to which the term frequency and frequency characteristics impact on term significance.

In the second part of Equation 1 ($log[\frac{N*D_k}{d_k*L_k}]$), the ratio of N to d_k and the ratio of D_k to L_k fairly reflect the influences of the term width and depth distribution characteristics, respectively. The applications of parameters N and D_k in the second part are the considerations of both the document collection characteristics and the characteristics of documents containing term k, respectively. The reason that we employ a logarithm of the ratio rather than a ratio is that this strategy can moderate the influence of variable changes on term significance. This second part implies that the larger the number of documents containing a term, the less the impact of the term width distribution on the term significance, and vice versa. It also suggests that the more frequently a term occurs within documents containing it, the less the impact of the term depth distribution on term significance term, and vice versa; and the larger the number of all terms in documents containing the term, the stronger the impact of the term on its importance, and vice versa. The aforementioned analyses also show that the smaller the number of documents containing a term and its term occurrence in a document collection, the better that term as a discriminator. Thus, with the addition of term depth distribution, this term significance measure is more precise than existing measure algorithms.

The term significance model is based on term distribution characteristics. A term distribution can be divided into two levels: one is at the individual document level (i.e. the first part of (1)) and another is at the whole document collection level (i.e. the second part of (1)). The frequency retrieval characteristic of a term within a document is one of the most important factors determining its significance at the individual document level while both its depth and width characteristics indicate directly its distribution characteristics at the entire document collection level. In other words, the two parts consist of a complete term distribution in a document collection. They are integrated and influence each other in terms of information retrieval.

4.2 Semantic Strength

To compute the semantic strength (i.e. the degree of relevance) between the query and a resulting Web page, or between two Web pages,

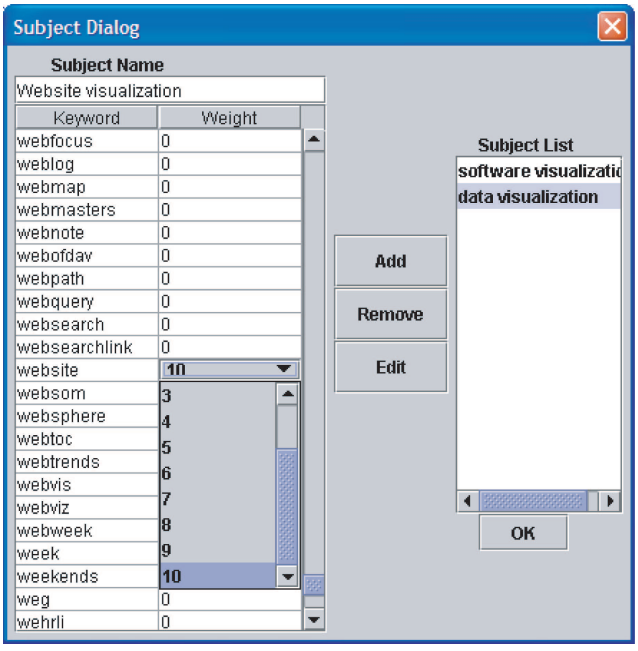

Fig. 2. Subjects of Interest

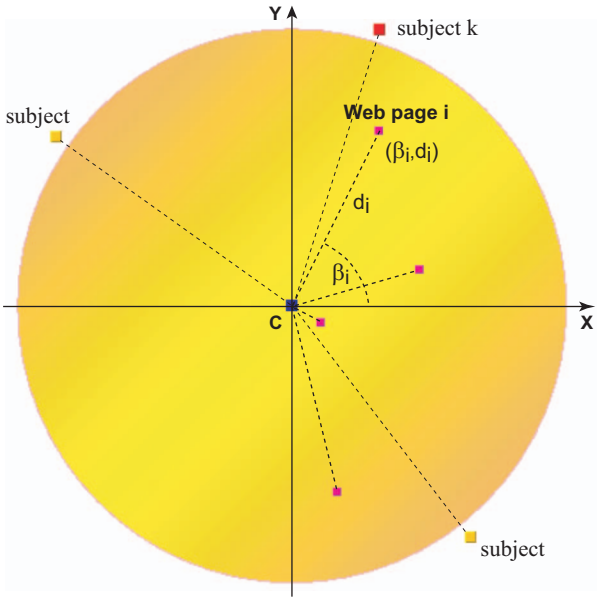

Fig. 3. Visualization Space

we use the iso-content-based angle similarity measure [36]. The measure takes the strengths of both the distance and direction of compared objects into consideration. Note that a query is represented by a vector whose coefficients correspond to keywords in the keyword list of the document term matrix. On the other hand, a document (i.e. a Web page) is represented by a row in the document term matrix, which is also based on the same keyword list. The semantic strength between the query and a Web page is defined as:

$$S_i = a^{|d|} \times \cos(\delta_i) \tag{2}$$

where d is the Euclidean distance between the vector representing the query and the vector representing the Web search result page. δ_i is the angle formed by the central point and the Web page against the origin in the document term matrix. Constant a is ranging from 0.8 to 0.99 [36]. From Equation 2, we can see that the semantic strength value S_i is between 0 and 1. The semantic relevance between two pages or between a topic of interest and a page will also be computed via this formula. A topic of interest (i.e. a subject) is represented as a vector. Details of that representation will be discussed next.

4.3 Subjects of Interest

Users' information needs change diversely during a search. A query and Web pages may generally involve multiple topics while users might be interested in only parts of them. Providing a mechanism to adjust users' interest or information needs is a must. To address this, we add an important dimension to our visualization model: the *subjects* that represent users' topics of interest. Subjects are positioned together with the pages in the visual space, enabling users to observe and analyze search results with respect to their interests. A dynamic subject can accommodate changes to users' information needs from different perspectives.

A subject can be defined in term of a vector in a similar manner as a Web page in the vector-based model. To define a subject, users can assign a weight to each individual term based on the keyword list that is extracted from the Web search result pages and the query. That keyword list is also used to build the document term matrix as described. The weights reflect the users' information need. The greater the value that is assigned to a term, the more important it is to users. Each term can be weighed in a scale from 0 to 10. During the interaction with the visual space, users can dynamically redefine subjects, add new subjects, or delete existing ones. Figure 2 shows an example of the definition of the subject "Website visualization". The weights that are assigned by users will be normalized into [0,1]. Multiple subjects can be defined. Users have the control whether a defined subject will be displayed in the visual space or not.

5 Visualization Model

5.1 Description

The visual space is defined as a display circle in a two-dimensional space (see Figure 3). The radius of this circle can be controlled by users via the zoom in/out function. Change of the radius leads to a proportional change of Web page distribution in the visual space. The query is always situated in the center. Subjects are initially anchored on the circumference. By default, subjects are evenly distributed. However, users can freely relocate them on any positions on the circumference. Users can control the number of displayed subjects by adding or discarding them from the visual space. In addition, users can click on a subject icon to edit or revise the definition of the subject when their preferred topics or contextual information change.

The position of a Web page icon in the visual space is affected by two parameters. The first parameter is the *distance* (d_i) that is defined as the semantic strength between the query and the Web search result page. That is, a relationship between the query and the Web search result page in the visual space is represented by a radiate line from the center to its icon. The radiate lines can be set to either visible or invisible in the visual space. The length of a radiate line is truly dependent on the semantic relevance between the query and the Web search result page. And the shorter the radiate line, the stronger it is relevant to the query, and vice versa. This strength reflects the gravitational pull of the query on the Web page (see Figure 3). The length of the corresponding radiate line from a Web page icon i to the central point is decided by the following formula:

$$d_i = R \times (1 - S_i) \tag{3}$$

where R is the radius of the visual space circle, and S_i is the semantic strength between the query and a Web search result page. Since S_i is between 0 and 1, d_i is between 0 and R. The other parameter that affects the position of a Web page icon i in the visual space is the *angle* β_i of that icon against the horizontal X-axis. It is affected by both physical locations of subject icons on the circumference and the semantic relevance between the Web page and the subjects. That is, the impact of all subjects on a Web page is considered in terms of the

angle β_i of a Web page icon against the X-axis in the visual space (see Figure 3). That angle is computed as follows:

$$\beta_i = \begin{cases} \frac{\sum_{k=1}^{m}(\alpha_k \times S_{ki})}{\sum_{k=1}^{m} S_{ki}} & : \sum_{k=1}^{m} S_{ki} \neq 0 \\ 0 & : \sum_{k=1}^{m} S_{ki} = 0 \end{cases} \quad (4)$$

where β_i is the angle of the page i against the X-axis, α_k is the angle of subject k against the X-axis, m is the number of subjects on the circle, S_{ki} is the semantic strength between the page i and the subject k, which is computed via Equation 2.

If Web page i is irrelevant to any of the current subjects, then the angle β_i is defined as zero to avoid meaningless β_i. Thus, Web pages that are not relevant to any of the current subjects will stay put and be projected onto the X-axis. This would help users quickly recognize those irrelevant pages in the visual space. Based on the aforementioned equations (2,3,4), both the projection distance and projection angle of a Web search result page to the central point (i.e. the query) in the visual space are clearly defined, therefore, the Web page icon can be uniquely projected onto the visual space.

5.2 Graphical Object Rotations

In this visualization system, users can activate a subject icon on the display circle and enable it to rotate evenly around the display circle. Since subjects affect the projection angle of a projected Web page icon in the visual space, the rotation of a subject icon would have an impact on the visual distribution of Web page icons. The change of distribution can be used to judge whether the Web pages are relevant to the moving subject. Now let us examine the impact of a moving subject on each individual Web page within the visual space. Assume that the moving speed of a specified subject T_r is θ degrees per second. Its angle in the visual space is α_r. The dynamic angle of a Web page icon W_i can be expressed as follows:

$$\beta_i = \frac{\sum_{k=1,k\neq r}^{m}(\alpha_k \times S_{ki}) + (\theta \times t + \theta_0) \times S_{ri}}{\sum_{k=1}^{m} S_{ki}} \quad (5)$$

where t is a time variable and θ_0 is an initial angle of the subject T_r against the X-axis in the visual space. From Equation 5, we have

$$\frac{d\beta_i}{dt} = \frac{\theta \times S_{ri}}{\sum_{k=1}^{m} S_{ki}} \quad (6)$$

This equation indicates that the rotation speed of a Web search result page caused by a moving subject, is linear. In other words, all affected Web pages will move evenly around the center point of the visual space. Different Web pages may correspond to different moving speeds. The moving speed of a Web page is affected primarily by the similarity (or relevance) between the moving subject and that Web page (S_{ri}). Since the relevance between the query and the Web page is a constant, the orbit of the moving Web page is a circle and its center point is the origin of the visual space.

It is clear that when there is no relevance between a moving subject and a Web page (i.e. the semantic strength between them is zero), the Web page stays in the same position as the subject icon moves because

$$\frac{d\beta_i}{dt} = \frac{\theta \times S_{ri}}{\sum_{k=1}^{m} S_{ki}} = 0.$$

If a Web page is only relevant to the moving subject and it is *not* relevant to *other* displayed subjects, then it has the same moving speed as that of the moving subject because

$$\frac{d\beta_i}{dt} = \frac{\theta \times S_{ri}}{\sum_{k=1}^{m} S_{ki}} = \frac{\theta \times S_{ri}}{0 + S_{ri}} = \theta.$$

In this case, that Web page icon is always located on the line formed by the center point and a moving subject icon because

$$\begin{aligned} \beta_i &= \frac{\sum_{k=1,k\neq r}^{m}(\alpha_k \times S_{ki}) + (\theta \times t + \theta_0) \times S_{ri}}{\sum_{k=1}^{m} S_{ki}} \\ &= \frac{0 + (\theta \times t + \theta_0) \times S_{ri}}{0 + S_{ri}} \\ &= \theta \times t + \theta_0 \\ &= \alpha_r. \end{aligned}$$

Equation 6 suggests that the more relevant a Web page to the moving subject, the faster it moves; and vice versa. Also, the relevance between a Web page and other subjects affects its moving speed because the sum of all similarities between the Web page and all subjects is the divisor in that equation. The divisor $\sum_{k=1}^{m} S_{ki}$ in conjunction with the similarity between the Web page and the moving subject (S_{ri}) as well as the subject moving speed θ ultimately determines the moving speed of the Web page icon in the visual space.

5.3 Projection Ambiguity and Visual Occlusion

Projection ambiguity refers to the phenomena that many objects (Web pages, in this case) in a high-dimensional space are projected onto a limited, low-dimensional space. Projection ambiguity and/or the presence of a large number of objects in the visual space can cause visual occlusion. We consider the following cases. First of all, if two Web pages W_i and W_j have different similarity measures with respect to the query, their corresponding objects will have different distances to the center of the visual space. Thus, if those objects revolve, they will do so in different orbits. Secondly, if the two pages have the same measure, they will have the same projection distance value. In this case, let us consider their projection angles. If these two Web pages are not relevant to any displayed subject, they are projected onto the same point on the X-axis since their projection angles have the values of zero. Since those objects will not move at all, the system automatically clusters them together and displays a special representative icon. Those pages will be listed on demand.

If these two Web pages are relevant to some subject(s) of interest, let us consider their corresponding projection angles $\beta_i = \frac{\sum_{k=1}^{m}(\alpha_k \times S_{ki})}{\sum_{k=1}^{m} S_{ki}}$ and $\beta_j = \frac{\sum_{k=1}^{m}(\alpha_k \times S_{kj})}{\sum_{k=1}^{m} S_{kj}}$. If they are equal regardless of whether a subject icon is moving and what positions of subject icons are, we must have $\frac{S_{ki}}{\sum_{k=1}^{m} S_{ki}} = \frac{S_{kj}}{\sum_{k=1}^{m} S_{kj}}$ for all k from 1 to m. Note that these two pages have the same projection distance. Replacing the definitions of similarity measure and projection distance into that set of equations, we can derive that the vectors representing for W_i and W_j must be extremely similar in the document term matrix space. Therefore, pages are rarely projected onto the same points when objects revolve. That is, if W_i and W_j are not identical, their overlapping icons can be disambiguated via moving or changing the positions of subject icons.

However, with many pages, graphical objects might still partially overlapped. To address this, we provide the following features. First of all, the automatic clustering/grouping functionality allows users to group pages whose similar measures are close to each other. Then, with the zoom in/out feature, they can zoom out for a broader overview or zoom in to reveal relationships among page icons in a particular visual cluster/area for a greater detail. Also, WebSearchViz provides the *focus page shifting* functionality. If users are interested in examining a particular page, they can make it become the new *center* of the visual space. The visual space is accordingly changed. The surrounding objects in the space now represent relevant Web pages to the center page.

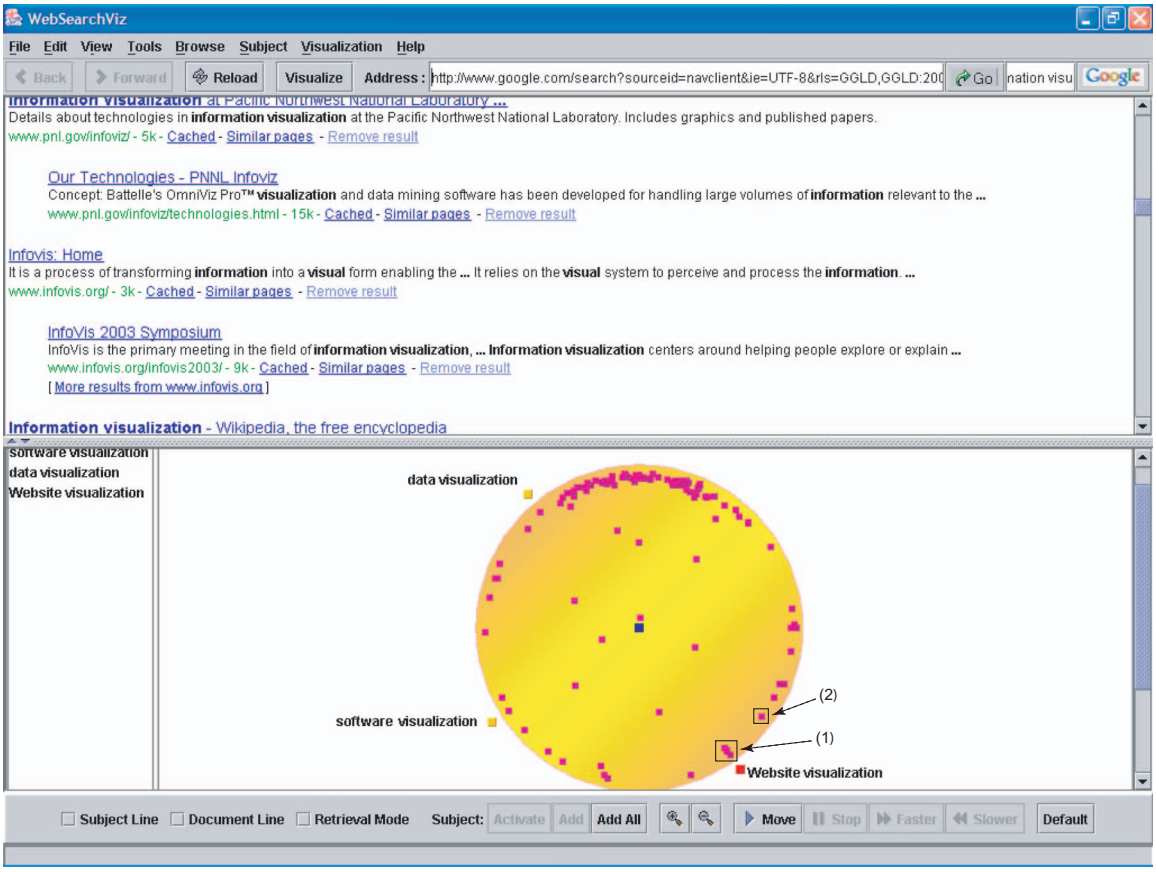

Fig. 4. WebSearchViz Visualization Environment

Users can observe the semantic relationships between the center page and relevant pages with respect to different topics of interest. Furthermore, when a page icon gets a mouse focus, a tool tip is displayed and the icon itself becomes larger and is highlighted. Users can also color a page icon to keep track of it when it revolves or when the visual space is updated due to the changes to the subjects or to the center.

6 WebSearchViz Visualization Environment

We have applied this visualization model to build an interactive, Web-based visualization environment for Web search results, named *WebSearchViz*. It is enhanced with a Web browser (the top part of Figure 4). This browser is re-used from our previous work, *WebStar*, a visualization tool for hyperlink structure [35]. This section focuses only on WebSearchViz's visualization functionality.

A user enters a query into the system. The system sends the query to the Google search engine [12] and fetches the Web search result pages into a local disk. For example, in Figure 4, the query "information visualization" was entered. The collected pages are lexically analyzed. Then, the user is asked to provide the subjects of interest as in Figure 2. Assume that he/she defines three subjects: "software visualization", "data visualization", and "Website visualization". The user assigns weights for keywords that correspond to each subject. Note that those keywords are also collected from Web search result pages. For convenience, the subjects can be saved into the system for later use. He/she can choose any subjects to be displayed in the visual space. The others will not be shown, however, still remain in the system unless the user explicitly deletes them. The user can also limit the number of Web pages to be displayed since Google could potentially return a huge number of hits for each query. During the interaction with the visual space, the user is able to modify, add, delete, or redefine subjects at will. The visual space will be updated accordingly. These operations on subjects are simple since they do not require a repeat of the lexical

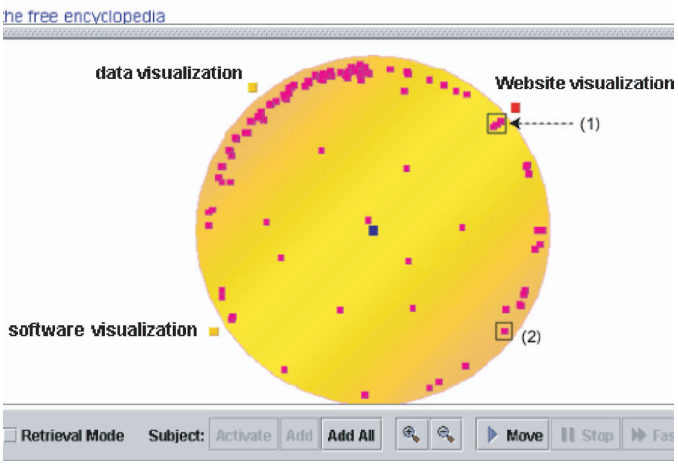

Fig. 5. Semantic Relevance and Subject Rotation

analysis process on Web pages or the re-calculation of the document term matrix. The reason is that the document and keyword sets are unchanged and only the vectors representing the modified subjects need to be updated. Thus, these operations involve only the re-computation of similarity measures between the modified subjects and Web pages, and the re-rendering of the visual space.

The subject rotation operation is the most distinguished and novel feature. The subject rotation adds one dynamic dimension to the visual space: both speed and movement of an object in the visual space can represent meaningful semantic relationships. Assume that now the user would like to know which resulting Web pages are most relevant

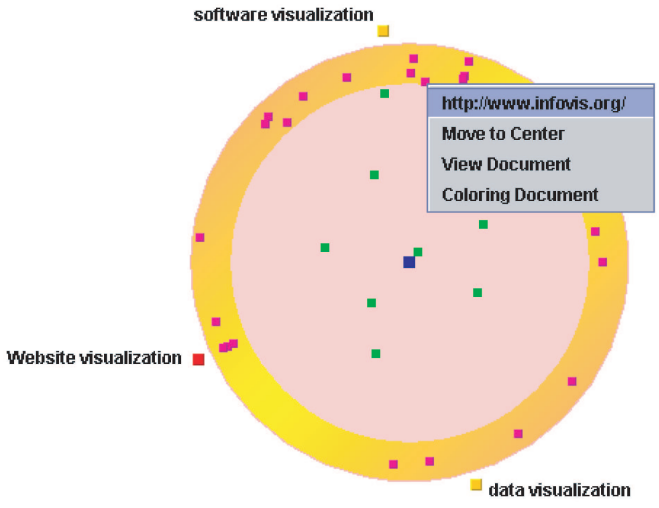

Fig. 6. Filtering and Grouping

to the query with respect to a particular subject, he/she can select that subject and activate the movement operation. The subject icon will start moving around the circle. After a subject is rotated, it would impact Web page icons in the visual space. The faster a Web page icon moves in the visual space, the more relevant it is to the subject; and vice versa. If a Web page icon stays still (but not on the horizontal X-axis), it suggests that the rotating subject has no impact on it but it is still relevant to other displayed subjects. If a Web page is irrelevant to all subjects in the visual space, it would stay on the horizontal X-axis and does not move at all.

The moving picture gives the user better sense of Web search result pages from a different perspective and it makes it easier and more intuitive for the user to decide which Web page or group of pages he/she should explore next. WebSearchViz provides the user with both automatic rotation and manual subject rotation modes. The automatic mode does not need any human intervention. In the automatic mode, the system also allows the user to have the control over the speed of a subject icon by making it move faster or slower. On the other hand, the primary advantage of the manual mode (i.e. a subject icon is dragged around via mouse) is that users can view the impact of a moving subject on a local area in the visual space by changing both the moving speed and the direction of a subject icon.

Let us consider the active subject "Website visualization" in the Figure 4 and Figure 5. The snapshots from WebSearchViz are captured at different time during the counter-clockwise movement of the subject icon "Website visualization". The user notices that the group of Web pages denoted by (1) moves in accordance with the subject icon. Thus, they are the pages that best reflect the topic "Website visualization". On the other hand, the page icon denoted by (2) is almost staying at the same place. Thus, that page is not quite relevant to that topic.

In this example, there is no page icon that stays still on the horizontal X-axis since all the returned pages from Google are somewhat related to the query ("information visualization") and all of displayed subjects are concerned with "visualization".

Notice that when a subject is put onto the visual space and it rotates around the circle, it is not easy for users to distinguish moving Web pages affected by the rotating subject. In addition, after the contents of selected subjects change, the position of a Web page icon in the visual space may also change. For tracking page objects in the visual space, WebSearchViz provides the coloring feature, allowing users to "mark" them with any preferred colors. Users are able to select a page and display them in the browser. Also, any page can be made the center of the visual space via the focus page shifting feature.

The semantic strength between each individual Web page and the query might be different. WebSearchViz also allows users to set a threshold on the semantic strength between a Web page and the query. More relevant Web pages will be kept and highlighted in the visual space (see Figure 6). Users select a threshold by creating a filter circle. Web pages whose semantic strengths are greater than the threshold are scattered within that circle. Then, users can choose to display only those pages in the visual space and discard other pages. WebSearchViz also provides the grouping functionality. From the visual space, users are able to select a group of Web pages, and then ask the system to visualize only those pages. The visual space will be accordingly updated. More details on WebSearchViz environment can be found at `http://home.eng.iastate.edu/~tien/WebSearchViz`.

7 Experimental Studies

To evaluate the time efficiency of our environment, we have conducted an experimental study. We used the Google search engine and the set of query words from the 1.7 million-query log excerpt from Excite. Although the number of hits on average per query is very large (over one million), our experimental study processed only the first 10,000 Web pages from Google. Note that WebSearchViz environment allows users to control the number of pages that are displayed in the visual space. The experiment was carried out on a 3.0GHz Intel Pentium 4 with 1GB RAM and 300GB hard disk. We measured the time to parse, to index, and to compute representation vectors. For that set of Web pages, it takes 0.7 sec to parse and to index, and 0.5 sec to generate the document term matrix. In a 10,000-page matrix, there are about 54,000 keywords. The rendering time for the visualization is very small. The first time rendering takes only 1.2 sec. When users interact with the system, the changes to activated subjects cause the redrawing of the visual space. However, it is almost instantly since we redraw only the graphical objects that would be affected by changes.

The preliminary results from our usability study is also very promising. We conducted the study on about 20 undergraduate students. The participants have backgrounds in both natural sciences and social sciences, and they are regularly using Internet search engines. The participants were required to use WebSearchViz to do Web searching. Questionnaires were prepared to gather their feedbacks. They were also asked to evaluate the accuracy of results from our visualization tool. Even though all results returned from Google are relevant to the query, the pages that users really look for with respect to their subjects of interest might not be revealed until using our visualization tool. Our experiment aims to evaluate how well our tool helps users in finding those desired Web pages. The recall and precision results are very satisfactory. The recall ratio (i.e. the ratio of the number of relevant results found by the tool to the total number of actual results in the collection) is 89%. Precision ratio (i.e. the ratio of the number of relevant Web pages found by the tool to the total number of pages found by the tool) is 92%.

The usability results also show that the users are more successful with WebSearchViz (visualization + ranked list) than they would be by following only the ranked list. Most of participants like to manually move the subject of interest in the visual space than using the automatic rotation feature. They also like to use the grouping and filtering features and then to explore further by the focus page shifting feature. Another positive feedback from participants is the short learning curve characteristic. They felt that the Web search result visualization tool in WebSearchViz being used as an enhancement to a normal Web browser is a good idea. They did not feel that they have to learn to work with a totally new Web search environment. They do not have to invoke an external Web browser as in other visualization tools for Web search.

8 Conclusions

Searching on the World Wide Web has become an important part in daily life for many of us. Applying visualization for Web search results would tremendously help users in finding the desired information. However, the subjects of interest and contextual information, which change over time, are not integrated into existing visualization tools for Web search results. This paper presents a visualization model/environment that addresses that issue.

In our model, physical location, spatial distance, color, and movement of graphical objects are used to represent the degree of relevance between a query and relevant Web pages in the context of users' subjects of interest. Its distinguished feature is the use of *movement* and *speed*, which adds a novel dimension into the visual space in order to illustrate the semantic connections among a query and Web search results, with respect to users' topics of interest. The metaphor in our visualization model is the solar system with its planets and asteroids revolving around the sun in the universe. A query is located at the center and regarded as the sun, while scattered subjects of interest and Web search results are regarded as planets and asteroids. When a subject object moves around, Web pages relevant to the query with respect to the subject also revolve around the center. The rotation speed of each Web page is determined by the degree of relevance between the moving subject and the Web page. Subjects of interest can be dynamically defined, changed, added, or deleted. By interacting with the visualization space, users are able to observe the semantic relevance between a query and a result Web page with respect to a particular subject of interest, context information, or concern.

Our experimental studies show the efficiency, effectiveness, and usefulness of the WebSearchViz environment. Other potential applications of our visualization paradigm include the domains of bibliographic citation analysis and interactive hypertext visualization. Our future work is to investigate a corresponding 3D visualization model and a way of integrating the movements of multiple subjects at the same time.

ACKNOWLEDGEMENTS

We would like to thank the anonymous reviewers for their valuable and insight comments.

REFERENCES

[1] K. Andrews. Visualising cyberspace: information visualisation in the Harmony Internet browser. In *Proceedings of the 1st IEEE Symposium on Information Visualization (InfoVis'95)*, pages 97–104. IEEE Computer Society, 1995.

[2] M. Angelaccio and B. Buttarazzi. Multiple Web Search Visualization using Dynamic Fields. In *Proceedings of 8th International Conference on Information Visualization (IV'04)*, pages 920–924. IEEE Computer Society, 2004.

[3] M. Q. W. Baldonado and T. Winograd. Sensemaker: an information-exploration interface supporting the contextual evolution of a user's interests. In *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI'97)*, pages 11–18. ACM Press, 1997.

[4] S. Benford, D. Snowden, C. Brown, G. Reynard, and R. Ingram. Visualizing and Populating the Web: Collaborative Virtual Environments for Browsing, Searching and Inhabiting Webspace. *Computer Networks and ISDN Systems*, 29(1):1751–1761, 1997.

[5] S. K. Card, J. Mackinlay, and B. Shneiderman. *Readings in Information Visualization: Using Vision to Think*. Interactive Technologies Series. Morgan Kaufmann Publishers, 1999.

[6] S. K. Card, G. G. Robertson, and W. York. The WebBook and the Web Forager: an information workspace for the World-Wide Web. In *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI'96)*, pages 111–117. ACM Press, 1996.

[7] J. Carriere and R. Kazman. WebQuery: Searching and Visualizing the Web through Connectivity. In *Proceedings of the 6th International World Wide Web Conference (WWW'97)*, pages 701–711. ACM Press, 1997.

[8] H. Chen, H. Fan, M. Chau, and D. Zeng. MetaSpider: Meta-Searching and Categorization on the Web. *Journal of the American Society for Information Science and Technology*, 52(13):1134–1147, 2001.

[9] J. Cugini, C. Piatko, and S. Laskowski. Interactive 3D Visualization for Document Retrieval. In *Proceedings of the Workshop on New Paradigms in Information Visualization and Manipulation, ACM Conference on Information and Knowledge Management (CIKM '96)*. ACM Press, 1996.

[10] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[11] E. Frecon and G. Smith. WebPath – A Three-Dimensional Web History. In *Proceedings of the 4th IEEE Symposium on Information Visualization (InfoVis'98)*, pages 3–10. IEEE Computer Society, 1998.

[12] Google. http://www.google.com/.

[13] Grokker. http://www.groxis.com/.

[14] S. Havre, E. Hetzler, K. Perrine, E. Jurrus, and N. Miller. Interactive Visualization of Multiple Query Results. In *Proceedings of the 7th IEEE Symposium on Information Visualization (InfoVis'01)*, pages 105–112. IEEE Computer Society, 2001.

[15] M. A. Hearst and C. Karadi. Cat-a-Cone: an interactive interface for specifying searches and viewing retrieval results using a large category hierarchy. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '97)*, pages 246–255. ACM Press, 1997.

[16] R. Hendley, N. Drew, A. Wood, and R. Beale. Narcissus: visualising information. In *Proceedings of the 1st IEEE Symposium on Information Visualization (InfoVis'95)*, pages 90–96. IEEE Computer Society, 1995.

[17] M. Higgins, P. Lucas, and J. Senn. VisageWeb: Visualizing WWW Data in Visage. In *Proceedings of the 5th IEEE Symposium on Information Visualization (InfoVis'99)*, pages 100–107. IEEE Computer Society, 1999.

[18] HTML 4.01 Specification. http://www.w3.org/TR/html4/.

[19] Kartoo. http://www.kartoo.com/.

[20] D. Kauwell. Visualization of Internet search results and archived information using VisIT. In *Proceedings of the World Conference on Educational Multimedia, Hypermedia and Telecommunications (EDMEDIA'99)*, pages 1521–1522. Association for Advancement of Computing in Education (AACE), 1999.

[21] R. Korfhage. *Information Storage and Retrieval*. Wiley Computer Pub, 1997.

[22] A. Leuski and J. Allan. Lighthouse: Showing the Way to Relevant Information. In *Proceedings of the 6th IEEE Symposium on Information Visualization (InfoVis'00)*, pages 125–129. IEEE Computer Society, 2000.

[23] H. Luhn. The automatic creation of literature abstract. *IBM Journal of Research and Development*, 2(4), 1958.

[24] Y. Maarek and I. Shaul. WebCutter: A System for Dynamic and Tailorable Site Mapping. In *Proceedings of the 6th International World Wide Web Conference (WWW'97)*, pages 713–722. ACM Press, 1997.

[25] T. Mann and H. Reiterer. Evaluation of Different Visualizations of Web Search Results. In *Proceedings of the 11th International Workshop on Database and Expert Systems Applications (DEXA'00)*, pages 586–590. IEEE Computer Society, 2000.

[26] S. Mukherjea and J. D. Foley. Visualizing the World Wide Web with the Navigational View Builder. *Computer Networks and ISDN Systems. Special issue on the Third International World Wide Web Conference*, 27(6):1075–1087, 1995.

[27] S. Mukherjea and Y. Hara. Visualizing World-Wide Web Search Engine Results. In *Proceedings of 3th International Conference on Information Visualisation (IV'99)*, pages 400–405. IEEE Computer Society, 1999.

[28] S. Mukherjea, K. Hirata, and Y. Hara. Visualizing the Results of Multimedia Web Search Engines. In *Proceedings of the 2nd IEEE Symposium on Information Visualization (InfoVis'96)*, pages 64–65. IEEE Computer Society, 1996.

[29] J. C. Roberts and E. Suvanaphen. Visual Bracketing for Web search Result Visualization. In *Proceedings of 7th International Conference on Information Visualisation (IV'03)*, pages 264–269. IEEE Computer Society, 2003.

[30] A. Spoerri. RankSpiral: Toward Enhancing Search Results Visualizations. In *Proceedings of the 10th IEEE Symposium on Information Visualization (InfoVis'04)*, pages 18–19. IEEE Computer Society, 2004.

[31] R. Torres, C. Silva, C. Medeiros, and H. Rocha. Visual structures for image browsing. In *Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM'03)*, pages 49–55. ACM Press, 2003.

[32] Vivisimo - Search Done Right. http://vivisimo.com.

[33] S. Weibel. The Dublin Core metadata initiative: mission, current activities, and future directions. *D-Lib Magazine*, 6(12), 2000.

[34] J. Zhang and T. N. Nguyen. A new term significance weighting approach. *Journal of Intelligent Information Systems*, 24(1):61–85, 2005.

[35] J. Zhang and T. N. Nguyen. WebStar: A Visualization Model for Hyperlink Structures. *Journal of Information Processing and Management*, 41(1):1003–1018, 2005.

[36] J. Zhang and E. Rasmussen. Developing a New Similarity Measure from Two Different Perspectives. *Journal of Information Processing and Management*, 37(2):279–294, 2001.