# Visualization for Information Retrieval based on Fast Search Technology

**Mamoon H. Mamoon[1], Hazem M. El-Bakry[2], Amany A. Salama[3]**
Faculty of Computer Science & Information system, Mansoura University, Egypt
Corresponding author, e-mail: helbakry5@yahoo.com, amanysalama90@yahoo.com

***Abstract***

*The core of search engine is information retrieval technique. Using information retrieval system backs more retrieval results, some of them more relevant than other, and some is not relevant. While using search engine to retrieve information has grown very substantially, there remain problems with the information retrieval systems. The interface of the systems does not help them to perceive the precision of these results. It is therefore not surprising that graphical visualizations have been employed in search engines to assist users. The main objective of Internet users is to find the required information with high efficiency and effectiveness. In this paper we present brief sides of information visualization's role in enhancing web information retrieval system as in some of its techniques such as tree view, title view, map view, bubble view and cloud view and its tools such as highlighting and Colored Query Result.*

## 1. Introduction

The typical generic scenario for searching, retrieving, and displaying information is the following. A user has an information need about a certain topic. With a user interface he/she formulates a query to the system [1]. The query starts an action in the system (search engine, information retrieval (IR) system, digital library, or other) [2]. The system will retrieve (or not) objects and will display them with appropriate messages and layouts in the same graphical user interface (GUI) where the user entered the query (3). Finally, the user decides if the documents are relevant or not. He/she can either exit the system because the information was found or refine the query and start again [2]. Information retrieval (IR) is the task of representing, storing, organizing, and offering access to information items [1]. The problem for search engines is not only to find topic relevant results, but results consistent with the user's information need. How to retrieve desired information from the Internet with high efficiency and good effectiveness is become the main concern of internet user-based [3].

Search engines interfaces are intuitive and in some cases restricted by the nature of the WWW. There is a limited use of color, no pull-down menus, and limited user interaction. The typical input interface is a simple box where the user fills the terms to search plus button to submit the query. The visualization process of the answers can be text only or more rich and complex with the use of a graphical metaphor. In the text only approach, the user gets a list of the top 10 or 20 best documents that potentially contains the information. The list usually contains the title, its URL, size, date, and an abstract of no more than 4 lines of the document. The user opens each document until the desired information is finally located. This is not a problem when the target document is located in the first 20 answers. It becomes a problem when the output of a query is a list of hundreds or thousands of documents. A graphical metaphor presents a rich interface in which the user can browse, filter, process, and reformulate the query [2]. User behavior, performance and attitude were recorded as well as usability problems. The system had few usability problems and users liked the visualizations, but recall performance was poor. The reasons for poor/good performance were investigated by examining user behavior and search strategies. Better searchers used the visualizations more electively and spent longer on the task, whereas poorer performances were attributable to poor motivation, difficulty in assessing article relevance and poor use of system visualizations [15].

Hence, visualization is an effective tool to partially solve data overload problems in WWW retrieval when answers contain hundreds of documents. The visualization of quantitative information consist of principles to help achieve the main goal: communicate complex ideas with clarity, precision, and efficiency [2].

*Received January 14, 2013; Revised March 7, 2013; Accepted March 18, 2013*

This survey paper contains problems that faced web information retrieval system whether because of the web nature or user activity or searching process itself. Then, how the search engine works and models of information retrieval. Next, the meaning of visualization, information visualization as one of its application and how it enhances web information retrieval system. Finally, real systems used information visualization tool in reducing and solving some of web information retrieval system's problems.

## 2. Problem Definition

The World Wide Web is a huge, widely distributed, global source for information services, hyper-link information, access and usage information and web site content and organization [4]. There is a huge quantity of text, audio, video, and other documents available on the Internet, on about any subject. Users need to be able to find relevant information to satisfy their particular information needs. There are two ways of searching for information: to use a search engines or to browse directories organized by categories (such as Yahoo Directories). There is still a large part of the Internet that is not accessible (for example private databases and intranets) [1]. By all measures, the web is enormous and growing at a staggering rate, which has made it increasingly intricate and crucial for both people and programs to have quick and accurate access to web information and services [4]. It is not surprising that about 85% of internet users surveyed claim to be using search engines and search services to find specific information of interest [5, 6]. The same surveys show, however, those users are not satisfied with the performance of the current generation of search engines; the slow speed of retrieval, communication delays, and poor quality of retrieved results (e.g., noise and broken links) are commonly cited problems [5]. Search engines have played a key role in the World Wide Web's infrastructure as its scale and impact have escalated. Although search engines are important tools for knowledge discovery on the web, they are far from perfect. The poor quality of retrieved results, handling a huge quantity of information, addressing subjective and time-varying search needs, finding fresh information and dealing with poor quality queries are commonly cited glitches [4].

There are many problems with different reasons which it can be by the web nature, users, search engine and hardware.

### 2.1. Problem when Interacting with the Web (web nature):

a. The "abundance" problem:

With the phenomenal growth of the web, there is an ever increasing volume of data and information published in numerous web pages. According to world wide websize.com, the indexed web contains at least 27.56 billion pages (Sunday, 24 august, 2008) [4], 27.87 billion pages (Sunday, 22 June, 2008) [6] and about 8 billion web pages were indexed by Google in 2005 [1].

b. Web search results usually have low precision and recall:

For finding relevant information, the search services is generally a keyword-based, query-triggered process which results in problems of low precision (difficulty to find relevant information) and low recall (inability to index all information available on the web).

c. Lack of personalization of information and limited customization to individual users:

Most knowledge on the web is presented as natural-language text with occasional pictures and graphics. This is convenient for human users to read and view but difficult for computers to understand. It also limits the state of art search engines, science they cannot infer contextual meaning. For example the occurrence of word 'bat' refers to a bird or to a cricket bat. These factors uphold the inevitable creation of intelligent server and client-side systems that can effectively mine for knowledge both across the internet and in particular web localities [4].

d. Heterogeneity:

- Information/data of almost all types exist on the web, e.g., structured tables, texts, multimedia data, etc.
- Much of the web information is semi-structured due to the nested structure of HTML code.
- Much of the web information is linked
- The web is noisy: a web page typically contains a mixture of many kinds of information, e.g., main contents, advertisement, navigational panels, copyright notices [4, 6].
- Much of the web information is redundant [6].

e. Dynamics:

The freedom for anyone to publish information on the web at anytime and anywhere implies that information on the web is constantly changing. It is a dynamic information environment whereas traditional systems are typically based on static document collection [4, 6]. This dynamic nature guarantees that at least some portions of any manuscript on the subject will de out-of-date before it reaches the intended audience, particularly URLs which are referenced [5].

f. Duplication:

Several studies indicate that nearly 30% of the web's content is duplicated, mainly due to mirroring [4, 6].

g. A comprehensive coverage of all of the important topics is impossible, because so many new ideas are constantly being proposed and either quickly accepted into the internet mainstream or rejected [5].

Different between IR and IR on web = challenges = problem definition that facing searchers and developers because of web nature.

### 2.2. Problems about Users (Information Searching Activity)

The typical Information Retrieval (IR) systems now available are characterized by a representation of a request for information (query) and the system usually responses with a set of results which most closely matches the request. Whatever representation of a request the seeker has to formulate, he has often to face with problems related to the clear specification of his information needs [7]. If the search is performed in a distributed and heterogeneous environment as web, the search becomes harder: the seeker anxiety grows up according to the heterogeneity and the amount of information available in World Wide Web. It generates the following problems [8]:

a. All users are not created equal:

Different users may use different terms to describe similar information needs; the concept of "what is relevant" to a user has only become more and more unclear as the web has matured and more diverse data have become available. Because of this, it is of key interest to search services to discover sets of identifying features that an information retrieval system can use to associate a specific user query with a broader information need [4].

b. The ambiguity of the natural language (English or other languages) that makes it difficult to have perfect matches between documents and user queries [1].

c. User search behavior:

The users have different expectations and goals such as informative, transactional and navigational. Often they compose short, ill-defined queries and impatiently look for the results mainly in the top 10 results [6].

d. Problem of vocabulary: "Which term to use?" The difference in terms of knowledge and perception between the information providers and the seeker has been modeled in terms of informative space and cognitive space. The former is defined as a set of object and relations among them held by the system whereas the latter is defined as a set of concepts and relations among them held by individual. Information providers organize their resources according to their knowledge and to the vocabulary that concurs in building the “informative spaces”. If seekers have a different knowledge background, or a different purpose, then his cognitive space has a poor overlapping to the information space. This make reasonable to assume they will use different terms to identify the same concept. So they have to discover which the proper terms to express a query in the information space.

e. Query formulation/refinement: “how to modify the query to find more relevant information?".

f. Seeker anxiety: The gap between what the seeker understands and what he thinks he should understand generates anxiety. This happens whenever information does not fulfill his needs.

g. Seeker and provider relationship: seeker and provider have different skill levels and different domain of knowledge. Moreover there is usually no direct interaction among them.

h. Seeker knowledge: the seeker has often only a perception of his information needs. He has a limited knowledge of what he is looking for.

i. Database selection: "which search engine to select?" The problem is well known in the WWW because the actual search engines are able to cover a limited portion of the web resources. The seeker has to decide which search engine to make use of.

j. Information overload: "how to explore many retrieved documents?" user still has to face with huge amount of candidates, which are all pertinent to what he is looking for. He needs to be supported in the analysis of heterogeneous information sources to be able to choose the most suitable ones for his purpose [8].

k. Query coordination: The seeker may need to be supported in the choice for queries. Human behavioral studies during the search activity have shown that the user is lazy and usually tends to create short queries and rarely adopts Boolean expression in his query criteria. Whenever the seeker needs information, which might seriously affect the results of his work, he is forced to a deeper search.

### 2.3. Problems of Searching

The unprecedented growth of available data coupled with the vast number of available online activities has introduced a new wrinkle to the problem of search: it is now important to attempt to determine not only what the user is looking for, but also the task they are trying to accomplish and the method by which would prefer to accomplish it [4].

#### 2.3.1. Problems faced search engine in IR process on the web

1. There are many publicly available search engines, but users are not necessarily satisfied with:
   a. The different formats for inputting queries.
   b. Speeds of retrieval.
   c. Presentation formats of the retrieval results.
   d. Poor quality of retrieved information [5, 6].

In particular, speed (i.e., search engine and retrieval time plus communication delays) has consistently been cited as " the most commonly experienced problem with the web" in the bi-annual WWW surveys conducted at the Graphics, visualization, and Usability Center of Georgia Institute of Technology 63% to 66% of web users in the past three surveys, over a period of year and a half were dissatisfied with the speed of retrieval and communication delay, and the problem appears to be growing worse. Even though 48% of the respondents in the April 1998 survey upgraded modems in the past year, 53% of the respondents left a website while searching for product information because of "slow access". "Broken links" registered as the second most frequent problem in the same survey. Other studies also cite the number one and number two reasons for dissatisfaction as "slow access" and "the inability to find relevant information" respectively [5].

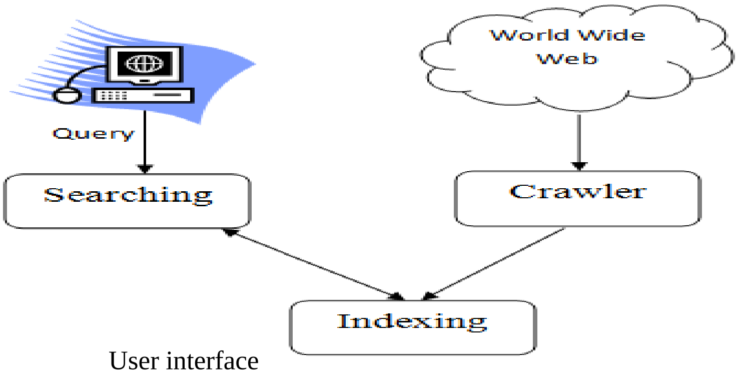
2. Limited query interface based on keyword-oriented search:
   It is hard to extract useful knowledge out of information available because the search service used to find out specific information on the web is retrieved-oriented, whereas to extract potentially useful knowledge out of it, is a data-mining oriented, data-triggered process [4].
3. Indexing web pages to facilitate retrieval is a much more complex problem than with classical databases because of: a) The enormous number of existing web pages and their rapid increase. b) Frequent updating. c) Removal of spurious information (e.g., newsgroup discussions, FAQ postings) [5]. d) Handling a huge quantity of information, addressing subjective and time-varying search needs. e) Finding fresh information. f) Dealing with poor quality queries [6].

So we can summarize challenges that face motivating researchers in web IR in improved system that retrieve the most relevant information available on the web to better satisfy a user's information need, or in the other words, combination of challenges that stem from traditional information retrieval and challenges characterized by the nature of the World Wide Web.

## 3. Web Information Retrieval

### 3.1. How Web Search Engines Work

A search engine operates in the following order: Web crawling, Indexing, and Searching, as declare in Figure 1. Web search engines work by storing information about many web pages, which they retrieve from the HTML itself. These pages are retrieved by a Web crawler (also known as a spider — an automated Web browser which follows every link on the site). Exclusions can be made by the use of robots.txt. The contents of each page are then analyzed to determine how it should be indexed (for example, words can be extracted from the titles, page content, headings, or special fields called Meta tags). Data about web pages are stored in an index database for use in later queries. A query can be a single word. The index helps find information as quickly as possible. Some search engines, such as Google, store all or part of the source page (referred to as a cache) as well as information about the web pages, whereas others, such as AltaVista, store every word of every page they find. This cached page always holds the actual search text since it is the one that was actually indexed, so it can be very useful when the content of the current page has been updated and the search terms are no longer in it. This problem might be considered a mild form of linkrot, and Google's handling of it increases usability by satisfying user expectations that the search terms will be on the returned webpage. This satisfies the principle of least astonishment, since the user normally expects that the search terms will be on the returned pages. Increased search relevance makes these cached pages very useful, even beyond the fact that they may contain data that may no longer be available elsewhere. When a user enters a query into a search engine (typically by using keywords),where the channel connection between his and the system is user interface the engine examines its index and provides a listing of best-matching web pages according to its criteria, usually with a short summary containing the document's title and sometimes parts of the text. The index is built from the information stored with the data and the method by which the information is indexed. Most search engines support the use of the Boolean operators AND, OR and NOT to further specify the search query. Boolean operators are for literal searches that allow the user to refine and extend the terms of the search. The engine looks for the words or phrases exactly as entered. Some search engines provide an advanced feature called proximity search, which allows users to define the distance between keywords. There is also concept-based searching where the research involves using statistical analysis on pages containing the words or phrases you search for. As well, natural language queries allow the user to type a question in the same form one would ask it to a human. A site like this would be ask.com.



User interface

Figure 1. How Search Engine Works

The usefulness of a search engine depends on the relevance of the result set it gives back. While there may be millions of web pages that include a particular word or phrase, some pages may be more relevant, popular, or authoritative than others. Most search engines employ methods to rank the results to provide the "best" results first. How a search engine decides which pages are the best matches, and what order the results should be shown in, varies widely from one engine to another. The methods also change over time as Internet usage changes and new techniques evolve. There are two main types of search engine that have evolved: one is a system of predefined and hierarchically ordered Figure 1.

Keywords that humans have programmed extensively. The other is a system that generates an "inverted index" by analyzing texts it locates. This first form relies much more heavily on the computer itself to do the bulk of the work [10].

User interface:

Information seeking has become increasingly interactive as tools and services on the WWW have evolved. Thus, there is more to searching than typing in a query and waiting for the search engine to display a set of possible web pages. The only way to achieve substantial advances in search and browse capabilities is to combine research and development in human-computer interaction with research and development in information retrieval to create highly interactive systems that engage the user in defining their needs iteratively and going beyond retrieval to understanding the corpus and the retrieved information [9]. The current user interface and its tool and evaluation in detail in user interface section, and its more activity and its developing in information visualization section.

### 3.2. Web Information Retrieval Models

Retrieval models form the theoretical basis for computing the answer to a query. A Retrieval Model is a formal representation of the process of matching a query and a document. The model of Web IR can be defined as a set of premises and an algorithm for ranking documents with regard to a user query. More formally, a Web IR model is a quadruple [D, Q, F, R (qi,dj)] where D is a set of logical views of documents, Q is a set of user queries, F is a framework for modeling documents and queries, and R(qi,dj) is a ranking function which associates a numeric ranking to the query qi and the document dj. The model is characterized by four parameters:

1. Representations for documents and queries, which define the model.
2. Matching strategies for assessing the relevance of documents to a user query, which involves learning parameters from query.
3. Methods for ranking query output.
4. Mechanisms for acquiring user-relevance feedback.

Retrieval models can describe ***the Computational process***, for example, how the documents are ranked and note that how documents or indexes are stored is implementation. The Retrieval models can also attempt to describe ***the User process***, for example, the information need and interaction level. The Retrieval variables are usually depicted by queries, documents, terms, relevance judgments, users & information needs. They can have an explicit or implicit definition of relevance.

**First Dimension: Computational Process: The Mathematical Basis**

According to the first dimension, the models can be classed into three types: set theoretic, algebraic and probabilistic models. In the following sections, we describe instances of each type.

**1. Set theoretic models**

Documents are represented by sets that contain terms. Similarities are derived using set-theoretic operations. Implementations of these models include the Standard Boolean Model, the Extended Boolean Model and the Fuzzy Model. The strict Boolean and fuzzy-set models are preferable to other models in terms of computational requirements, which are low in terms of both the disk space required for storing document representations and the algorithmic complexity of indexing and computing query-document similarities.

**2. Algebraic models**

Documents are represented as vectors, matrices or tuples. These are transformed using algebraic operations to a one-dimensional similarity measure. Implementations include the Vector Space Model and the Generalized Vector Space Model. The strength of this model lies in its simplicity. Relevance feedback can be easily incorporated into it. However, the rich expressiveness of query specification inherent in the Boolean model is sacrificed.

**3. Probabilistic Models**

Document's relevance is interpreted as a probability. Documents and queries similarities are computed as probabilities for a given query. The probabilistic model takes these term dependencies and relationships into account and, in fact, specifies major parameters such as the weights of the query terms and the form of the query document similarity. Due to its simplicity and efficient computation, the Vector Model is the most widely used model in IR. The model requires term-occurrence probabilities in the relevant and irrelevant parts of the document collection, which are difficult to estimate. However, this model serves an important function for characterizing retrieval processes and provides a theoretical justification for practices previously used on an empirical basis (for example, the introduction of certain term-weighting systems).

**Second Dimension: User Process**: **The Relevance Basis**

Another dimension of defining different categories of Web IR models can be based on their applications as follows:

**1. Classical models**

- Query languages, indexing (Boolean)
- Introducing ranking and weighting (Vector Space).

**2. Topical relevance models**

- IR as Bayesian classification, relevance information, tf.idf weights (BM25)
- Probabilistic models of documents, queries, topics (Language Modeling).

**3. User relevance models**

- Combinations of evidence, features, query language (inference network, Inquery).

**4. Linear feature-based models**

- Learning weights, arbitrary features, optimizing effectiveness measures (Ranking SVM, Linear Discriminant, MRF)
- “Learning to Rank”, learning ranking rather than classification, preferences [6].

### 3.3. User Interface

The current user interface in most search engine as Google, yahoo!, … etc called List View. The List View (Figure 2) is the classic search results view. It contains a list of files that, based on Boolean logic, match the users query. Each file name is shown along with the number of “hits” from the query. A hit is defined as one occurrence of one query term in the file contents. The files are listed in descending order of the total number of hits.

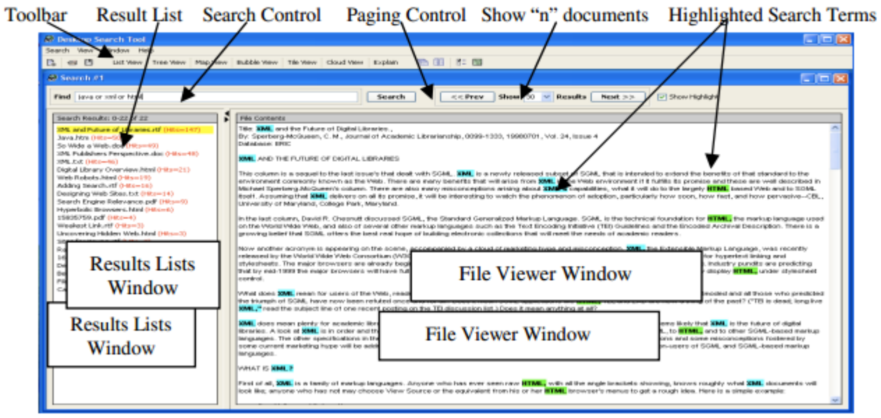


Figure 2. List View

This view also contains a File Viewer window that will display the textual contents of a selected file in order to allow quick review of the file contents. Matching query terms are highlighted in different colors to aid the user identify where these terms occur in the document. Any non-textual content, such as images, etc. in the actual document are not displayed. Similarly, some of the source document formatting will be lost as only line and paragraph breaks are preserved in the extraction process.

Double clicking on a selected file displays the original document in a separate window. Such a view, although elementary, is simple, intuitive, provides clarity and a quick preview of the documents. Unsurprisingly, most evaluation participants gave strong support to this view as both easy to use (89% who Agree or Strongly Agree) and useful (86%) in reviewing the results. The evaluators liked the highlighting of the search terms in the file viewer and clear indication of the number of hits per result file, and suggested improvements related to more flexible sorting of the results and more document and result information to be made easily available. These results confirm that the basic de-sign and operation of the desktop search engine is effective and useful [11].

The following describe efforts to improve search interfaces by incorporating visual information into display using techniques from the field of information visualization.

## 4. Visualization

Visualization is any technique for creating images, diagrams, or animations to communicate a message. Visualization through visual imagery has been an effective way to communicate both abstract and concrete. Applications of visualization are scientific visualization, educational visualization, information visualization, knowledge visualization, product visualization, systems visualization, visual communication, and visual analytics [14].

Most web search engines are text-based. They display results from input queries as long lists of pointers, sometimes with and sometimes without summaries of retrieved pages. Future commercial systems are likely to take advantage of small, powerful computers and will probably have a variety of mechanisms for querying non-textual data (e.g., hand drawn sketches, textures and colors, speech) and better user interfaces to enable users to visually manipulate retrieved information [5]. From that the role of information visualization appears as declare in the following.

### 4.1. Information Visualization

Information visualization is all about making data visible or more precisely, the patterns that are hidden in the data. This is a method of presenting data or information in non-traditional and interactive graphical forms. By using 2-Dor 3-Dcolor graphics, text and animation, these visualizations can show the structure of information, allow one to navigate through it, and modify it with graphical interactions [13].

Chaomei chen writes "information visualization aims to maximize our perceptional and cognitive abilities to make sense of visual-spatial representations". Information visualization strives to make the information more accessible and less structured to improve usability. In the Web, Information Visualization provides visualization approaches to manage big amount of information in a summarized way and graphical interaction techniques to manipulate the search results [7]. The human perceptual system is highly attuned to images, and visual representations can communicate some kinds of information more rapidly and effectively than text. The goal of information visualization (INFOVIS) is to translate abstract information into a visual form that provides new insight about that information [12]. And is not pictures, but insight, It’s not about looking at pictures; it’s about interacting with them to "amplify cognition".

Information visualization joins the human’s capacity of visual thinking and the computer’s capacity of analytical computing, thereby building a bidirectional visual and interactive interface between human user and the information resources. Very few information visualization applications do away with text altogether. The goal is to find the representation appropriate for a particular task. In many situations text remains the best form of representation. But we all know from experience that many complex ideas are best represented visually. Justas movies did not eliminate the novel; information visualization will not eliminate the need for text.

Information visualization will only succeed if it solves the scalability problem. This view assumes that the really big problems are the only interesting ones, and the only hard ones. It also assumes that if the data set has billions of elements, it is important to display all of those elements at once. In many situations the real challenge is to narrow the billions down to a more reasonable and manageable subset. This is where data mining begins to play an important role. Size and scalability are important issues, but it is a mistake to think that information visualization only applies to extreme problems.

Information visualization is about speed. It is sometimes said that information visualization aims to help us move from slow reading to faster visual perception, and that it can help us deal with information overload by allowing us to process more information faster. This is only true up to a point.

Information visualization is about insight, not pictures. Insight means understanding and creating knowledge and learning. Those processes often require reflection, combination, and rearrangement. The speed element of information visualization aims to reduce the cognitive load of certain tasks so that larger, more complex tasks become possible. Particular tasks may be made more efficient, but information visualization can also open up a range of new tasks that were previously impossible or simply not feasible because they were too burdensome [13].

Guidelines for designing information visualizations are available from writers such as Few (Few, 2006, Few, 2009) and Tufte (Tufte, 1983, Tufte, 1990b). Some of these guidelines overlap with guidelines from graphic design, including the need to present information clearly, precisely, and without extraneous or distracting clutter. Other guidelines relate to the special purposes of visualization. Good visualizations use graphics to organize information, highlight important information, allow for visual comparisons, and reveal patterns, trends, and outliers in the data. Visualization guidelines are also derived from principles of human perception, and urge the designer to be aware of the perceptual properties which can affect the design [12].

### 4.2. Technique for Interactive Visualization

The challenge is to improve efficiency and effectiveness of search and result selection. In this respect, the metaphor that “a picture paints a thousand words” neatly encapsulates the concept that well presented graphical views can convey large amounts of complex information in a simple and easy to understand manner. It is therefore not surprising that graphical visualizations have been employed in search engines to assist users. While each of the individual visualization might not be new by itself, we believe that the seamless integration of these views and value-added functionality in them are novel to assist in the results review, selection and query refinement.

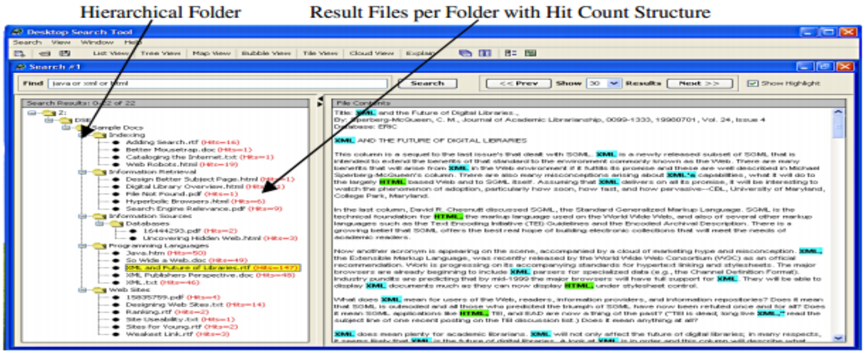
The design of the user interface was based on the following research premises:

- Visualizations can assist users to search for documents.
- Different visualizations can be used to support different elements of the searching process (results review and query reformulation).
- Different graphical techniques can be used to assist users to visualize different kinds of information.
- Visualizations work best when they are kept simple.

The search engine GUI has a plug-in view architecture that allows different views to be created independent of the searching mechanism [11]. Several interactive techniques are important to information visualization [12]. Five views were constructed for use and evaluation: Tree View, Map View, Bubble View, Tile View and Cloud View.

1) Tree View

The Tree View is similar to the List View (in Figure 3) except the result files are organized based on their underlying folder structure. For each file in the results list, all of its parent folders are added to the folder hierarchy (avoiding duplicates). The Result files are then added into the tree at the appropriate folder for their physical location.


Hierarchical Folder
Result Files per Folder with Hit Count Structure
Desktop Search Tool
Search View Window Help
List View
Tree View
Map View
Bubble View
Tile View
Cloud View
Explain
Search #1
Find
java or xml or html
Search
<< Prev
Show
30
Results
Next >>
Show Highlight
Search Results: 0-22 of 22
Z:
DSE
Sample Docs
Indexing
Adding Search.rtf (Hits=16)
Better Mousetrap.doc (Hits=1)
Cataloging the Internet.txt (Hits=1)
Web Robots.html (Hits=19)
Information Retrieval
Design Better Subject Page.html (Hits=1)
Digital Library Overview.html (Hits=1)
File Not Found.pdf (Hits=1)
Hyperbolic Browsers.html (Hits=6)
Search Engine Relevance.pdf (Hits=9)
Information Sources
Databases
16444293.pdf (Hits=2)
Uncovering Hidden Web.html (Hits=3)
Programming Languages
Java.htm (Hits=50)
So Wide a Web.doc (Hits=49)
XML and Future of Libraries.rtf (Hits=147)
XML Publishers Perspective.doc (Hits=48)
XML.txt (Hits=46)
Web Sites
15835759.pdf (Hits=4)
Designing Web Sites.txt (Hits=14)
Ranking.rtf (Hits=2)
Site Useability.txt (Hits=1)
Sites for Young.rtf (Hits=2)
Weakest Link.rtf (Hits=3)
File Contents

Title: XML and the Future of Digital Libraries.,
By: Sperberg-McQueen, C. M., Journal of Academic Librarianship, 0099-1333, 19980701, Vol. 24, Issue 4
Database: ERIC

XML AND THE FUTURE OF DIGITAL LIBRARIES

This column is a sequel to the last issue's that dealt with SGML. XML is a newly released subset of SGML that is intended to extend the benefits of that standard to the environment commonly known as the Web. There are many benefits that will arise from XML in the Web environment if it fulfills its promise and these are well described in Michael Sperberg-McQueen's column. There are also many misconceptions arising about XML's capabilities, what it will do to the largely HTML based Web and to SGML itself. Assuming that XML delivers on all its promise, it will be interesting to watch the phenomenon of adoption, particularly how soon, how fast, and how pervasive--CDL, University of Maryland, College Park, Maryland.

In the last column, David R. Chesnutt discussed SGML, the Standard Generalized Markup Language. SGML is the technical foundation for HTML, the markup language used on the World Wide Web, and also of several other markup languages such as the Text Encoding Initiative (TEI) Guidelines and the Encoded Archival Description. There is a growing belief that SGML offers the best real hope of building electronic collections that will meet the needs of academic readers.

Now another acronym is appearing on the scene, accompanied by a cloud of marketing hype and misconception. XML, the Extensible Markup Language, was recently released by the World Wide Web Consortium (W3C) as an official recommendation. Work is progressing on its accompanying standards for hypertext linking and stylesheets. The major browsers are already beginning to include XML parsers for specialized data (e.g., the Channel Definition Format). Industry pundits are predicting that by mid-1999 the major browsers will have full support for XML. They will be able to display XML documents much as they can now display HTML, under stylesheet control.

What does XML mean for users of the Web, readers, information providers, and information repositories? Does it mean that SGML is outmoded and all those who predicted the triumph of SGML have now been refuted once and for all? Does it mean SGML applications like HTML, TEI, and EAD are now a thing of the past? ("TEI is dead, long live XML," read the subject line of one recent posting on the TEI discussion list.) Does it mean anything at all?

XML does mean plenty for academic librarians. XML will not only affect the future of digital libraries; in many respects, it seems likely that XML is the future of digital libraries. A look at XML is in order and this column will describe what

Figure 3. Tree View

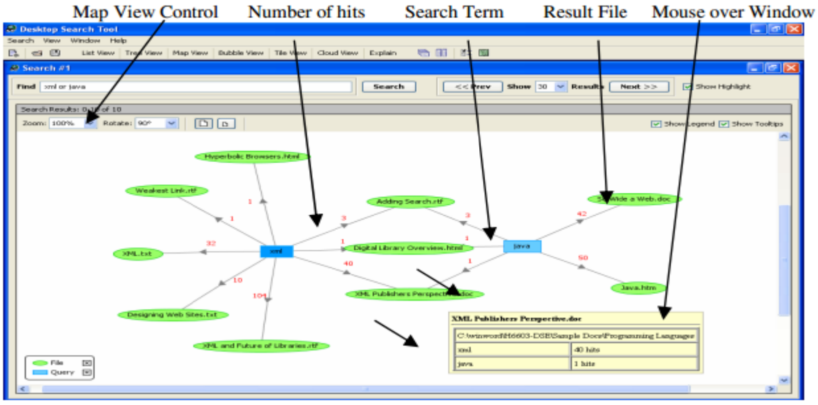
This view is very similar to the Microsoft Windows Explorer view. However, only files that match the query string are displayed and only the parent folders of these files are included in the tree. The purpose of this view is to use the physical file structure as part of the results display. If users have taken the time to organize their documents into meaningful folders and hierarchies then this information may be useful when reviewing results. This view is particularly suited for thesaurus or taxonomy based folder organizations where documents are stored in the respective nodes of this organization scheme. As such, related documents would already have been assessed and organized into folder hierarchies that will help users to quickly zoom into documents of interest. With the familiarity of Windows Explorer, participants strongly indicated that this view was easy to use (93%) and useful (91%) in reviewing the results. They found the view clear and obvious. 87% of them acknowledged that if they had organize their documents logically in folders, then this view would be especially useful for them.

This confirms the design premise that the user's folder structure is a useful aid to present search results, as well as a means to logically organize information in thesaurus/taxonomy-like structures that can support browsing as well as searching.

2) Map View

The Map View (Figure 4) provides an overview of the relationship between the query terms and the result files. Each query term is depicted as a blue rectangle and each result file as a green ellipse. Lines link related query terms and results files. These are annotated (in red) with the number of occurrences of the query term in the result file.

The view can be zoomed and rotated and individual shapes can be moved around on screen to obtain views that are more legible and avoid cluttering. If the mouse is moved over a query term it will display a popup window that lists all the result files that contain the query term along with their respective number of hits (not shown in Figure 4). Similarly, if the mouse is moved over a result file then a popup window will display all the query terms found in this file with their respective hit counts (Figure 4).



Find: xml or java

Show 30 Results

Search Results: 0-10 of 10

Zoom: 100%  Rotate: 90°

XML Publishers Perspective.doc

| C:\winword\H6603-DSE\Sample Docs\Programming Language | |
|---|---|
| xml | 40 hits |
| java | 1 hits |

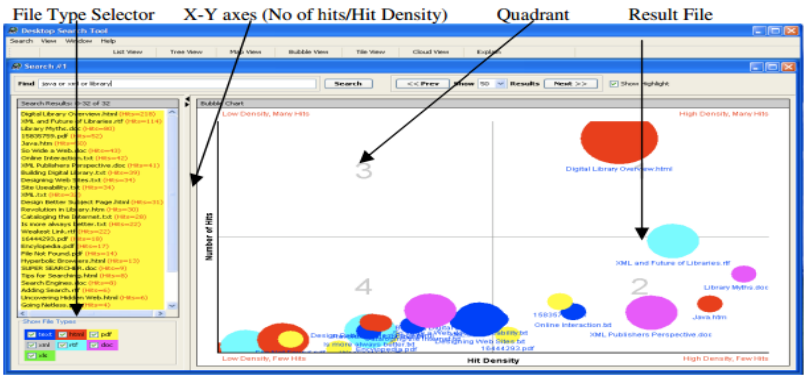
Figure 4. Map View

This view shows how individual query terms affect the results and which files contain one or more query terms. This bird's eye view can be used to detect problems in the query specification if the required results are not as expected. It will clearly show the relative influence of each query term in producing the result files and therefore help the user in deciding whether the query needs to be reformulated and how to do so. The evaluation found that slightly over half of the evaluators (51%) agreed or strongly agreed that the Map View was useful in reviewing their query results and reformulating their query. The distribution of responses for ease of use and usefulness are very similar. Qualitative comment analysis indicated that the most useful aspect noted by the evaluators (35%) was the ability to see an overview of the relationship between the query terms and the results files. This was the design premise for the Map View – to provide a clear overview of the query and its effect on matched results. However, the view can become very crowded for complex Boolean queries with a large number of items displayed resulting in overlapping of the graphic objects. A significant number of evaluators (36%) indicated that this caused confusion.

3) Bubble View

Boolean logic systems make it difficult to judge the relevance of a result file. The total number of hits alone is not necessarily a good guide to relevance especially when document length is taken into consideration. Therefore, it neither is desirable to normalize this measure to take into account document size. In this work, a hit density is calculated as the number of hits per 1,000 searchable terms (non-stop words) in the document.

The intention of the Bubble View (Figure 5) is to help the user better assess the relevance of different documents. The axes of the graph are the number of hits and the calculated hit density. These measures are used to distribute the documents along each axis as they provide good document discrimination in order to achieve a better visualization. The diameter of the bubble is determined by the number of query terms present in the result file and its color represents its file type.



Find: java or xml or library

Show: 50

☑ Show Highlight

Search Results: 0-32 of 32

Digital Library Overview.html (Hits=218)
XML and Future of Libraries.rtf (Hits=114)
Library Myths.doc (Hits=60)
15835759.pdf (Hits=52)
Java.htm (Hits=50)
So Wide a Web.doc (Hits=43)
Online Interaction.txt (Hits=42)
XML Publishers Perspective.doc (Hits=41)
Building Digital Library.txt (Hits=39)
Designing Web Sites.txt (Hits=34)
Site Useability.txt (Hits=34)
XML.txt (Hits=32)
Design Better Subject Page.html (Hits=31)
Revolution in Library.htm (Hits=30)
Cataloging the Internet.txt (Hits=28)
Is more always better.txt (Hits=22)
Weakest Link.rtf (Hits=22)
16444293.pdf (Hits=18)
Encyclopedia.pdf (Hits=17)
File Not Found.pdf (Hits=14)
Hyperbolic Browsers.html (Hits=13)
SUPER SEARCHER.doc (Hits=9)
Tips for Searching.html (Hits=8)
Search Engines.doc (Hits=8)
Adding Search.rtf (Hits=6)
Uncovering Hidden Web.html (Hits=6)
Going Netless.[illegible] (Hits=4)

Figure 5. Bubble View

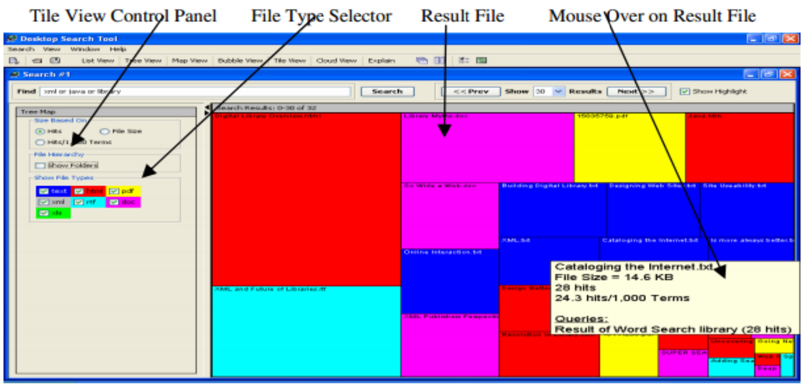
Quadrant 1 is expected to contain the most relevant documents as both the number and density of hits is greatest. Correspondingly, quadrant 4 will be expected to contain the least relevant documents, as both the hit count and density are smallest. The display suggests that documents should be explored in priority according to the Quadrant numbers. The view therefore attempts to provide an overview of document relevance for a given query and aids the review of documents most likely to be relevant to the query.

The evaluation results show that 46% of the evaluators found this view useful in reviewing their query results. The majority found the position (65%) and size (59%) of the bubbles gave them useful information, which supports the concept of this view as a means to convey several dimensions about the relevance of the results documents. The comments analysis showed that useful features were the ability to get a quick and easy overview of the relevancy of the results and the ability to see the hit density.

The major confusion factors related to the display of a large number of result documents where the titles overlap and become unreadable and the display was found to be very cluttered and messy. Suggestions for potential improvement relate mainly to improving the layout to increase clarity and for help on how to interpret the view.

4) Tile View

The Tile View (figure n.6) presents each result file as a colored tile using a Tree map. A Tree map is “a space-constrained visualization of hierarchical structures”. The size of each tile is determined by a measure such as Total Number of Hits, File Size, and Hit Density (Hits per 1,000 searchable terms). Using the control panel, the user can change the measure used to determine the size of a tile. As before, the color of a tile is determined by its file type and the display can be restricted to certain file types.



Find: xml or java or library

Show: 30

☑ Show Highlight

◉ Hits ○ File Size
○ Hits/1,000 Terms

□ Show Folders

Search Results: 0-30 of 32

Cataloging the Internet.txt
File Size = 14.6 KB
28 hits
24.3 hits/1,000 Terms

Queries:
Result of Word Search library (28 hits)

Figure 6. Tile View

The Tile View can optionally include the folder hierarchy of the results files (not shown in figure). In this variant, all the result files in a specific folder are grouped together in a "super tile". Each folder is enclosed within a tile representing its parent so that the entire folder structure of the results files can be displayed. This is an alternative display of the tree view but with value-added information in the tiles. The purpose of the Tile View is to allow users to review the results visually and judge their relevance based on different criteria with larger tiles denoting the most relevant documents.

The results of the evaluation of the Tile View show over half the evaluators (59%) agreed or strongly agreed that the Tile View was useful in reviewing their results. Over two thirds found the tiles to be obvious and easy to understand (68%) and the ability to use different criteria to control their sizing was found to be useful (69%). This supports the design objective for this view to easily support the use of different criteria for judging the relevance of the results documents. The ability to group files by folders also received strong support with 75% of evaluators agreeing or strongly agreeing that this was useful. The comments analysis indicated that useful features were the ability to change tile size based on different criteria, the ability to group files by folder and the use of color to distinguish file types.

5) Cloud View

The Cloud View (Figure 7) is adapted from the Tag Clouds popular on social networking sites such as Flicker. A Tag Cloud is a weighted list which contains the most popular tags used on that site and the relative popularity of each tag is indicated by changing its font size. It is thus easy to see the most popular tags. The Cloud View creates a Word Cloud based on the (indexable) content of the result files. The file contents are examined and stop words and non-indexable terms are removed. The words are then stemmed and a simple term count of the documents contents. The top 300 terms are then displayed in a Word Cloud as they represent the most common indexable terms.

Only files selected in the Results List (in the left hand window) have their contents included in the Word Cloud. If the selection of files is changed, the Word Cloud is dynamically refreshed with information based on the new selection of files. When the user clicks on a word in the Word Cloud a popup menu appears offering the choice to expand (OR), restrict (AND) or exclude (NOT) the word from the cur-rent query or to create a new search (NEW) using the selected word (Figure 7).

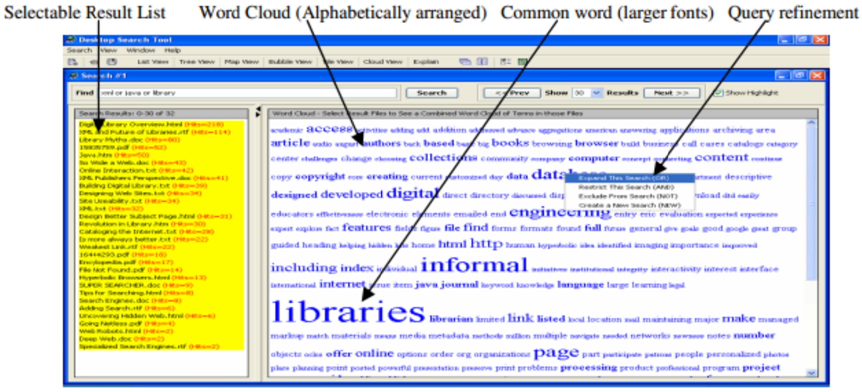


Figure 7. Cloude View

The purpose of the Cloud View is to provide users with the most common words (300 in this instance) found in the selected files in the result lists thereby providing an idea of the contents of the result files (i.e. basically a concordance) and information on potential words that can be used in the query refinement process.

The evaluation results for the Cloud View showed that nearly two thirds of the evaluators found the Cloud View useful in reformulating their query (63%) and easy to use (61%). However, the distribution profile for Question 37 (usefulness of Cloud View) is different with a bi-polar distribution, with a peak for Disagree and Agree. This implies that the evaluators were split into two groups, a conclusion strongly supported by a review of the comments. Those evaluators who scored the usefulness of the Cloud View very low (Strongly disagree or

Disagree) reported a lot of confusion as to the contents of the Cloud. In other words, they did not find the view useful because they did not understand what it does. Those who did rated it highly. This implies that some users had not seen this type of visualization before nor understood its potential [11].

### 4.3. Web Visualization/Visualization Tool/Visualization in WWW

Web visualization tools have been used to help users maintain a "big picture" of the retrieval results from search engines, web sites, a subset of the web, or even the entire web. The most well known example of using the tree-metaphor for web browsing is the hyperbolic tree developed by Xerox PARC. These visualization systems, machine learning techniques are often used to determine how web pages should be placed in the 2-D or 3-D space [4]. There is a study show how the existing tools to browse the WWW adopt visualization to satisfy seeker needs. It has been limited to some of the most well known tools such as Kartoo, Grokker, Web Theme [14], Aduna AutoFocus. To achieve this purpose the following research activities have been performed:

1. Identification of the main functionalities provided by these tools.
2. Analysis both of the correlation among these functionalities and of the problems in the information search [7].

Now, How Typical Visualization Tool Works?

1. Visualization tool takes set of key words from user and gives to search engine.
2. Search engine gives results to visualization tool as query per document.
3. In each Query, frequent words, no of occurrences of each frequent word, URL is there.
4. Creates concepts by taking some combinations of frequent words.
5. Do text clustering by using concepts.
6. Displays whole documents by using some visualization technique [13].

The results of these activities are summarized in Table 1 and Table 2. Table 1 is the result of the first activity. It illustrates the association between the tools (columns) and some of their functionalities (rows). It have identified some heterogeneous functionalities: graphical visualization functionalities (Hierarchical Visualization, Clustering Visualization, Map Based Visualization), graphical interaction functionalities (Visualization Manipulation, Graphical Selection) and those functionalities that are a combination of them (Highlighting, Colored Query Result, Filter Result Representation, co-occurring term interaction/visualization). In the following, a description for each of them is provided:

- Hierarchical Visualization: the visualization represents its content according to different levels of granularity. This allows browsing the information at different levels of detail (as Grokker).
- Clustering Visualization: the content is visualized (grouped) according to some similarity criteria. The groups can be obtained either by applying a clustering algorithm (galaxy view) or according to properties specified by the user (cluster map).
- Map Based Visualization: it imitates the geographical map appearance; the content is organized according to thematic terms or co-occurrence criteria, which are represented as peaks in the map (i.e. Kartoo represents the isograms and the name of the mountains respectively as concentric isolines and thematic terms on the top of them).
- Visualization Manipulation: the interaction between user and the graphical representation allows to re-organize the elements displayed, to move them and to add new ones (i.e. Grokker and Kartoo allow to add a new web site to the search and to insert it in the displayed graph according to user needs).
- Graphical Selection: the selection of a single (Grokker, Aduna AutoFocus, Kartoo) or many elements at a time allows the user to select different information source such as URI, PDF or DOC document in Grokker, Aduna Autofocus, Kartoo or data as in Web Theme.
- Highlighting: whenever an element of the visualization is selected, all the sources related to such element are highlighted too. Aduna AutoFocus and Kartoo allow highlighting the related co-occurring terms, whereas Grokker permits the highlighting both of the related co-occurring terms and of the related elements in the visualization.
- Colored Query Result: Web Theme allows to query the visualized data set and to set a particular color to each result set. This facilitates the comparison among different queries (results).

- Filter Results Representation: some filters can be applied to the contents shown in the visualization. For instance, Grokker allows filtering on the rank, on the domain and on the source, whereas Kartoo allows filtering on the co-occurring terms.
- Co-Occurring Terms Visualization: As users tend to formulate their queries using common words, a statistical thesaurus expands these queries with other highly frequent terms that should help the user in discriminating relevant documents.

Table 1. Functionalities Provided by Some Existing Tolls to Browse the WWW

| | | Grokker | Aduna AutoFocus | Kartoo | Web Theme |
|---|---|---|---|---|---|
| Graphical visualization | Hierarchical Visualization | √ | | | |
| | Clustering Visualization | √ | √ | | √ |
| | Map Based Visualization | | | √ | √ |
| raphical Interactio | Visualization Manipulation | √ | √ | √ | |
| | Graphical Selection | √ | √ | √ | √ |
| Interaction and Visualization | Highlighting | √ | | | |
| | Colored Query Result | | | | √ |
| | Filter Results Representation | √ | | √ | |
| | Co-Occurring Terms Interaction/Visualization | √ | √ | √ | |

Table 2. Tools Functionalities and How they Satisfy Seeker Needs

| | | database | vocabulary | Query formulation/refinement | Information overload | Query coordination |
|---|---|---|---|---|---|---|
| Graphical visualization | Hierarchical Visualization | | | √ | √ | |
| | Clustering Visualization | | √ | √ | √ | |
| | Map Based Visualization | | √ | √ | √ | |
| Graphical Interaction | Visualization Manipulation | | | √ | √ | |
| | Graphical Selection | | | √ | | |
| Interaction and Visualization | Highlighting | | | | √ | |
| | Colored Query Result | √ | | | | √ |
| | Filter Results Representation | | | √ | √ | |
| | Co-Occurring Terms Interaction/Visualization | | √ | √ | | |

Table 2 is the result of the activities to identify the contribution of the functionalities to solve problems related to seeker needs (information overload, query formulation, vocabulary, and database selection). It is possible to argue that:

- Graphical visualization functionalities: provide different results. They give a structured organization of information offering the user an overview of the available information relieving the information overload problem. They support the query formulation/ refinement: a correct and rapid understanding of search results is the prerequisite to have a successfully query refinement. Graphical visualization functionalities provide useful hints to solve the vocabulary problem by map based and clustering visualizations. They show co-occurring terms as cluster representative or in map representation permitting to learn which terms belong to the information space and how terms are related to each other.

- Graphical interaction facilitates the information overload and query formulation /refinement problems: visualization manipulation supports in the analysis of results by modifying the layout, whereas graphical selection provides a visual and intuitive way to select results user is interested to.
- Finally, the functionalities based on the integration between interaction and visualization techniques support in the entire problem mentioned about. In particular, functionalities as Colored Query Result allow comparing the results of different queries supporting in the queries coordination problem. Whenever the information about which search engines have found a result is maintained, such functionality can be exploited to compare the results coming from different search engines supporting the solution of database problem. [7]

## 5. Conclusion and Future Work

Despite the success of web as a preferred or defacto source of information, the retrieval of information from the web is still an unsolved problem with many different applications probably undiscovered. Specifically, the operative challenges motivating researchers in web IR include problems relating either to data quality or user satisfaction. The problems facing successful web information retrieval are a combination of challenges that stem from traditional information retrieval and challenges characterized by the nature of the World Wide Web.

The ultimate challenge of web IR research is to provide improved systems that retrieve the most relevant information available on the web to better satisfy a user's information need.

In researcher's journey to overcome most of the previous problems, they accept data mining, annotation, semantic web and visualizing the retrieval results as a helpful techniques utilize in facing web information retrieval process' problems. Some of these problems can't be solved but do the effort to adapt with them. These are abundance, dynamic, and heterogeneity because they are a web information' characteristic. The most challenges when interacting with the web are: (1) Attempt to determine not only what the user is looking for, but also the task they are trying to accomplish and method by which they would prefer to accomplish. (2) Creating new knowledge out of the information available on the web. These described as challenges that are difficult needs a lot of skill and effort to do.

Of course, there is always the new development, and it will be exciting to see what that future brings to user's search, like *nature language queries*; Users could express their queries in natural language, not just as keywords. This requires deeper syntactic and semantic analysis of the queries and the documents. Allowing the user to orally describe the information need into a microphone is a more natural way to interact with a search engine. *Intelligent and adaptive web services*; problems which can be tackled by these agents include: finding and filtering information, customizing information, and automating completion of simple tasks or perform some other service without (the user's) immediate presence and on some regular schedule, and adaptive web site automatically improves their organization and presentation based on user access data. Also Multimedia Queries, Knowledge Retrieval, Using and building Arabic language in IR system.

## References

[1] Diana Inkpen. *Information Retrieval on the Internet*. 2008.

[2] Omar Alonso, Ricardo Baeza-Yates. *Model for visualizing large answer in WWW retrieval.*

[3] Hazam M El-Bakry, Nikos Mastorakis. *Fast Information Retrieval from Web Pages.* Proc. of the 7th WSEAS Int. Conf. On Computational Intelligence, Man-Machine Systems and Cybernetics. 2008; 229.

[4] Mps Bhatia, Akshi Kumar Khalid. Information retrieval and machine learning: supporting technologies for web mining research and practice. *Webbology*. 2008; 5(2).

[5] Mei Kobyashi, Koichi Takeda. Information retrieval on the web. draft paper LINK. 2000.

[6] Mps Bhatia, akshi Kumar Khalid. A primer on the web information retrieval paradigm. *Journal of theoretical and applied information technology*. 657-662.

[7] Riccardo Albertoni, Alessio Bertone, Monica Martino. semantic web and information visualization. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.61.6774&rep=rep1&type=pdf , 26/2/2013

[8] A Spoerri. *How Visual Query Tools Can Support Users Searching the Internet.* Proceedings of Eighth International Conference on Information Visualization IEEE. 2004: 329-334.

[9] Gary Marchionini. User interface for information retrieval on the WWW. INFORUM2005: 11th conference on professional information resources, prague. 2005.

[10] http://en.wikipedia.org/wiki/Search_engine, last visit: 5/3/2013

[11] Schubert Foo, Douglas Hendry. Desktop Search Engine Visualization and Evaluation", D.H.-L. Goh et al. (Eds.): ICADL. *Springer-Verlag Berlin Heidelberg*. 2007; LNCS 4822: 372–382.

[12] Marti Hearst. Search user interfaces. published by Cambridge university press. 2009.

[13] Kotaiah Choudary Ravipati. Visualization of Web Search Results in 3D. seminar report

[14] http://en.wikipedia.org/wiki/Visualization_%28computer_graphics%29, last visit: 8/3/2013

[15] AG Sutclife, M Ennis, J Hu. Evaluating the effectiveness of visual user interfaces for information retrieval. *Int. J. Human-Computer Studies*. 2000; 53: 741-763.