VECIMS 2007 – IEEE International Conference on
Virtual Environments, Human-Computer Interfaces, and Measurement Systems
Ostuni – Italy , 25-27 June 2007

A New Graphical Interface For Web Search Engine

Alberto Amato[1], Vincenzo Di Lecce[2], Vincenzo Piuri[3]
[1]II Engineering Faculty of Polytechnic of Bari - Italy
[2] D.I.A.S.S., Polytechnic of Bari, Bari, Italy
[3]Dept Information Technologies, University of Milan, Italy
[e-mail: a.amato@poliba.it, dilecce@poliba.it (corresponding author), piuri@dti.unimi.it]

Abstract
Aim of this work is to present a new human computer interface for web search engines. Despite the noteworthy improvements introduced in the web search engines, their human interface still remain surprisingly based on a textual sorted list. The position of a site in this list expresses its distance from the user's query. The authors propose a graphical user interface for search engines based on the geomorphologic metaphor. This interface makes user aware of the semantic distribution of the web sites retrieved by the search engine. The proposed interface is implemented as a browser plug-in and it is able to work with all the modern search engines.

Keywords *– human computer interface, search engine, semantic distribution*

I. INTRODUCTION

Nowadays, both the amount of information available on the web and the number of internet users are increasing rapidly. These users have markedly different backgrounds, interests, and usage purposes. But they all need some automatic tools allowing them to find their required information. The development of these tools is object of many research works in the field of web mining, due to the unstructured nature of the web pages. Web mining can be broadly defined as the discovery and analysis of useful information from the World Wide Web [1].

The birth of search engines has simplified the users access to web information. They show a set of links related to web pages to user according to his defined query composed of one or more words. The weight of the search engines in the "internet era" becomes clearer analyzing the data from [2]. According to this site, only U.S. Internet users conducted 6.9 billion searches in February 2007. This number is 19% greater than the same value measured in February 2006. From this point of view, it seems that search engines are the gateway for information access on the web.

Nevertheless, the modern search engines show many lacks concerning data consistency, completeness of analysis, precision, accuracy and user interface. There are several reasons for this kind of problems. For example, search engines build their index accessing to a small percentage of the available web pages (the so called surf web). Indeed, the crawlers (the software utilized to build the indexes) do not have access to dynamic web pages or to web pages requiring authorization or a priori registration (the so called hidden web). From the point of view of data consistency, an intrinsic characteristic of a web page is its "volatility". Indeed, in a web site the page content is often modified or deleted daily (an example of such web site can be a commercial web site). It is well known that the time interval between two successive visits of a crawler to the same web site is longer than a day.

Another problem of modern search engines is the poverty of their user interface. Basically, each search engine has a user interface composed of:

- an input text box where users can write a keywords based query,
- an output page showing a sorted list of links to web sites. Under each link there are some sentences extracted from the related web page that can be utilized by user to understand the semantic context of that web pages. In this work we will use the term "link" referring to the whole set of information about a web site given by the search engine. The criterion utilized to sort the list is the relevance of the pages regarding to the key words in the query.

The main problems of the output interface are two:

1. it is textual: perhaps this feature was useful some years ago due to bandwidth limitation, but today it does not have the same sense anymore.
2. each site position in the list does not show any relation with the other sites in the same list, but it presents only a relation with key words in the query.

User interface limits of the modern web search engines are in countertendency in comparison to the improvement of the human-computer interface that are characterizing many other fields of computer applications.

Aim of this work is to present a graphical user interface based on a geomorphologic metaphor. This interface allows users to be aware about the overall distribution of the searched information in the retrieved web sites. This GUI is not dependent on a specific search engine. Indeed the proposed interface works as a browser plug-in. When a user performs a query to a search engine, this will answers with an html page containing the sorted list of results. The proposed interface analyzes this page and shows its graphical representation to user. The user can use this graphical interface to browse the query results.

This work is organized as follows: section II shows a brief overview about the evolution of human-computer interface. Section III presents the adopted method to measure the distances among the links. Section IV shows the technique utilized to obtain a 3D graphical representation of these distances while a description of the proposed interface is proposed in section V. Finally conclusions are reported in section VI.

[1] This work is partially supported by the "New. Ton." Project, co-financed by the ERDF through the INTERREG IIIB ARCHIMED Programme A.1.180 (http://www.newton.interreg.net)

II. HUMAN-MACHINE EVOLUTION OVERVIEW

The rapid and continuous development of the information technology has enormously increased the number of computer applications. Indeed, the computer is becoming more and more present in every day life for many people. In order to allow this computer intrusion in every day life, it is mandatory to study and realize a considerable number of human machine interfaces that satisfies a number of requirements.

An important example of this kind of studies is the evolution of the operating system interface for the PC. Indeed, the first PCs used an operating system with a textual interface, but their evolution uses a graphic user interface.

It is evident that the textual interface has a cardinality that is greater than that of graphic interface. On the other hand for using a textual interface it is necessary for the user to know a great number of commands, while, typically, a graphic user interface is based on the philosophy of the mouse point & click. Despite the loss of cardinality, the graphic interface has won the user's favour and it has given a big contribution to the diffusion of the PC.

Today, the research of friendly user interfaces is a fertile field, because it has a large number of applications. Just for example, the user interface is a driving technology for the successful application of various robotic technologies. Currently, a large part of the robot community does not know what to do for the user's inability to take advantage of robot technologies [3].

Today there are several design standards for human machine interfaces. For example, the Department of Defence MIL-STD-1472E establishes general human engineering criteria for the design and the development of military systems, equipment and facilities. A similar initiative has been taken by NASA to define the interface for space-based systems.

Another application where the development of a friendly user interface has reached marked improvements is the field of the programming languages. Indeed, the introduction of the visual programming paradigm (together to the object oriented programming) has led to the creation of the Rapid Application Interface Development (RAID) environments. This environment allows the programmer to develop the front-end of its applications quickly and to concentrate himself in the developing of the back-end.

A significant field of research has developed technologies that rely on computer-generated displays. While some of these works deal with the display or rendering aspect (i.e. the virtual reality), another significant portion focuses on the use of sophisticated displays to present simulation and high-level control information to the user (i.e. the remote view in remotely operated mobile and underwater vehicles).

Another interesting framework for the study of user interfaces is the World Wide Web. Indeed, in Internet, every site gives to the user a great number of information (i.e. available data for download, image, services like e-commerce, chat-line, sms and so on). The point of question is how to present this information to all the possible users in an effective way (this is the classical communication scheme of one to many). Today this problem is treated very seriously, indeed the development of a web site is a job of group and in every group there is an expert of communication.

Last point is the remarkable improvement resulted by techniques for data analysis and representation leading to the insight of many phenomena. The authors propose to improve the user' perception of a query to a web search engine using a human-machine interface based on one of such techniques and a geomorphologic metaphor.

Metaphor based GUIs have introduced a big improvement in human-machine interface. For example, metaphors are one of the central elements of human-machine interaction since the first 1984 when Apple introduced the metaphor of desktop. Since than the searching for effective metaphors is a topic highly debated to which several solutions are proposed. The spread use of this metaphor after twenty years from its introduction testifies its usefulness. It allows user to act in a familiar context: an office with recycle bins, desktops, folders and files. But, analyzing in deep this interface, it is possible to notice many incoherence. For example:

- It is quite difficult to find a wastepaper on a desktop in real world
- There is no logical connection between a folder and the window that the system shows when user double-click on it
- Typically, in real world the windows are transparent but user cannot see through the application's window.

While the proposed incoherencies are as obvious to disappear for an experienced user, it may be not the same for a beginner.

The authors propose that metaphors are used to create a semiotic ground that supports the user's action. This deals with a perceptual coherence more than a lexical one.

III. THE ADOPTED I.R. SYSTEM

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfy an information need from within large collections (usually stored on computers) [4]. Automated IR systems are often used to reduce information overload. Many universities and public libraries use IR systems to provide access to books, journals, and other documents. IR systems are often related to object and query. An object is an entity which keeps or stores information in a database. User queries are matched to objects stored in the database. A document is, therefore, a data object. Often the documents themselves are not kept or stored directly in the IR system, but are instead represented in the system by document surrogates automatically generated by the same IR system by means of a document analysis. Nowadays there are two approach to document analysis: statistical and semantic. The statistical approach was initially proposed by Lhun. In one of its paper Lhun wrote: "It is here proposed that the frequency of word

occurrence in an article furnishes a useful measurement of word significance. It is further proposed that the relative position within a sentence of words having given values of significance furnish a useful measurement for determining the significance of sentences. The significance factor of a sentence will therefore be based on a combination of these two measurements" [5]. It is interesting to notice that this approach is still used in many modern IR systems.

A Semantic Information Retrieval system uses the notion of semantic similarity (based on lexical and semantic relations) between concepts to determine the relevancy of a certain document. One way of incorporating semantic knowledge into a representation is to map a document's terms to concepts from a certain ontology (e.g. WordNet [6]).

The most spread and known applications of IR are the web search engines. Each search engine answers to a human query with a html pages containing a sorted list of links to web sites or documents. Each link is enriched with some sentences extracted from the related web page. The search engine represents each retrieved web page using a set of sentences that are semantically relevant to user query. From this point of view, computing the distance between two sets of sentences, is possible to obtain a good approximation of the semantic distance between the two relative web pages. This is a common task in the information retrieval (I.R.) process. The most traditional approaches used by I.R. systems to evaluate the distance between two texts (typically the user's query and a document) were based on some functions of the overlap between their terms.

The main problems related to this approach to I.R. are described in [7] and can be summarized in the following points:

- Polysemy: the same word can be used to indicate different things (i.e. the word ring has a means for a bride and another for a boxer)
- Different words or phrases can be used to indicate the same thing (i.e. "psychology" and "science of mind")
- The intrinsic ambiguity of the natural language.

In order to reduce the effects of these problems on the similarity measure performance, many researchers use an external knowledge base [8]. In this work WordNet was utilized. WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations [6].

This work uses WordNet to endow each word, involved in a word to word distance measure, with all its synsets. In this way the distance between two links becomes a function of the overlap of their global synsets.

Let S_i and S_j be the sets of all the synsets regarding the words present in the links i and j. Let $d(S_i,S_j)$ be the distance between S_i and S_j. We define

$$d(S_i, S_j) = \frac{\dim(S_i \cap S_j)}{\max(\dim(S_i), \dim(S_j)} \quad (1)$$

where dim(.) is a function that gives the number of elements of its argument.

IV. MULTIDIMENSIONAL SCALING

In this work, the search engine www.google.ca was utilized to test the proposed interface. Typically this search engine answers to user queries with a set of pages each one containing ten links. The distances among these ten links must be computed in order to generate the proposed graphical interface. This process generates a distance matrix $\in \Re^{10x10}$. By convention, such matrix is categorized as either similarities or dissimilarities, which are opposite poles of the same continuum. A matrix is a similarity matrix if larger numbers indicate more similarity between items, rather than less. A matrix is a dissimilarity matrix if larger numbers indicate less similarity. The distinction is somewhat misleading, however, because similarity is not the only relationship among items that can be measured and analyzed using MDS. In the proposed work the distance matrix is a similarity matrix.

In order to graphically represent the space defined by the similarity matrix, it is necessary to reduce its dimension.

Multidimensional scaling is one of the possible approach to the general problem of "embedding" general metric and non-metric spaces into low-dimensional spaces. Aim of this technique is to study the properties of the original high dimensional space analyzing the low dimensional space. A typical application of this approach is an original space composed of a set of abstract object such as graphs, strings, etc. related through proximity data. In this situation, a low-dimensional (two or three dimensions) representation of inter-object distances can give a graphical space representation.

A widely accepted classification of embedding techniques defines two categories: linear and nonlinear. A typical example of linear embedding technique is the principal component analysis (PCA). It reduces space dimensionality by projecting high-dimensional data onto a low-dimensional subspace. The optimal m-dimensional subspace is selected by rotating coordinate axes to make them coincide with the eigenvectors of the sample covariance matrix, and keeping the m axes along which the sample has the largest variance.

An example of nonlinear embedding techniques are multidimensional scaling (MDS) techniques. The aim of MDS is to construct a low-dimensional map (usually two or three dimensional) in which the distance between any two objects reflects their degree of dissimilarity. This is an iterative method. It starts creating a random mapping of the original points in the desired low dimensional space using a given norm. A stress function that compares proximity values with distances between points in the host space (usually a sum of squared errors function) is used to measure the quality of the embedding, and a gradient descent procedure is applied to improve the embedding until a local minimum of the stress function is reached [9].

Both PCA and MDS attempt to preserve all point to point distances as well as possible; but the latter removes the restriction to linear projections, and arbitrary embeddings are considered. There are many variants of this general approach, in [10] is presented a good review of them.

V. PROPOSED INTERFACE

Figure 1 - The starting condition of the proposed interface

The proposed graphic interface is based on the geomorphologic metaphor. The query results are merged in a three dimensional landscape.

A point on the land surface is identified using three coordinates: longitude, latitude and altitude. Typically the distance between two resorts on a map is computed with a good approximation measuring the distance of their representations on a bi-dimensional map.

In the proposed metaphor each link is represented by a mountain. The user query is represented by the plain. The altitude of each mountain is proportional to the distance of its related link from the user query and the relative link according to the google ranking. The distance, measured in terms of longitude and latitude between two mountains is proportional to the distance between their related links computed using (1).

Moving the mouse pointer on a mountain the user sees the phrases that google associated to that link. Clicking over a mountain the browser is redirected to the related web site.

Figure 1 shows a beta release of the proposed interface. In future works the aspect of this GUI can be enriched using more sophisticated 3D rendering techniques such as texture mapping.

VI. EXPERIMENTS AND CONCLUSIONS

This work presents the beta release of a new graphical interface for web search engine. The interface is based on the well known topographic metaphor. The popularity of this metaphor can make easier understand this interface. Preliminary studies carried out using a human panel (students of Polytechnic of Bari) seem to confirm this hypothesis.

This interface can work as a browser plug-in and it is compatible with every modern web search engine. In order to carry out the experiments for this work, the interface is implemented in java.

The results shown in this work are obtained using the search engine www.google.ca.

In order to evaluate the effectiveness of the proposed interface, a human panel composed of ten students is using it. Each student performs ten queries and for each obtained

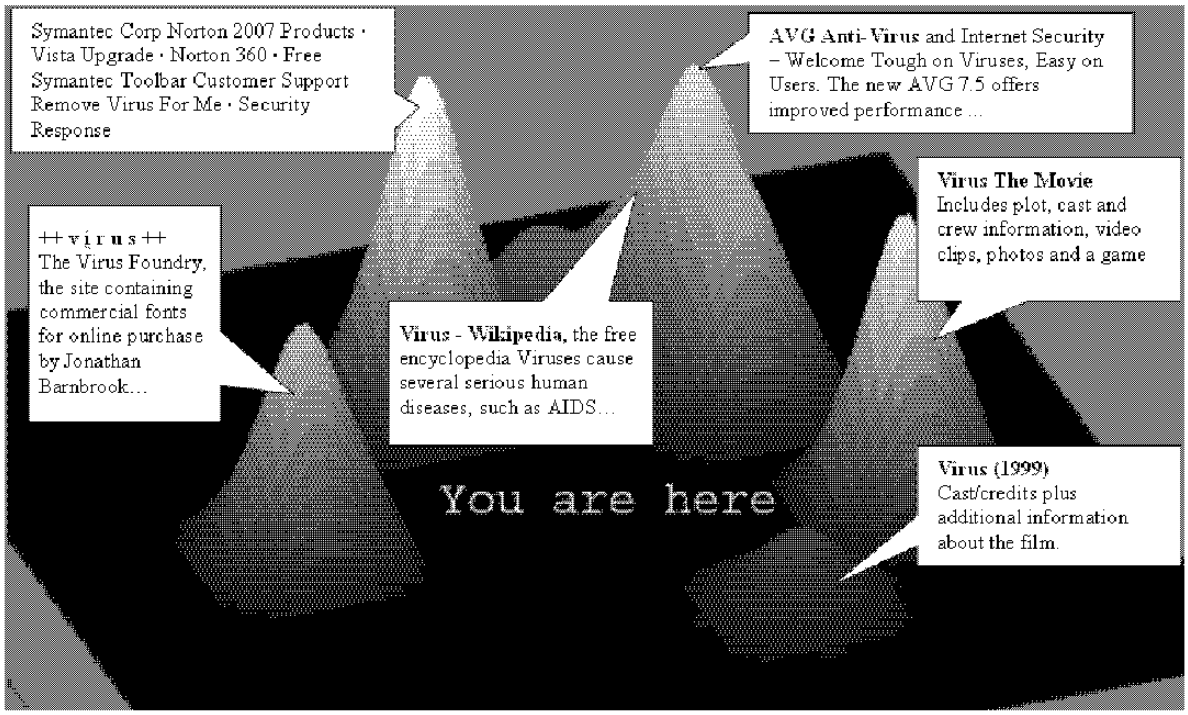

Figure 2 - An example of the proposed interface

result evaluates some parameters such as: spatial distribution of the mountains, semantic mapping between group of links and group of mountains.

Figure 1 and figure 2 show the interface applied to the results obtained using as search key the word "virus". The search engine www.google.ca answers to this query with 10 links that are summarized in the following list:

1. Virus - human diseases - Wikipedia, the free encyclopedia
2. Computer virus - Wikipedia, the free encyclopedia
3. Symantec Corp. - Norton 2007 Products
4. Threat Explorer - Symantec Corp.
5. Virus The Movie
6. AVG Anti-Virus and Internet Security - Welcome
7. What is virus? - A Word Definition From the Webopedia Computer ...
8. Virus (1999)Cast/credits plus additional information about the film.
9. ++ v i r u s ++. The Virus Foundry, the site containing commercial fonts for online purchase by...
10. McAfee - Antivirus Software and Intrusion Prevention Solutions

From a human being's semantic point of view, these 10 web sites can be grouped into four macro areas: human virus (1 link), computer virus (6 links), a movie (2 links) and a site related to character fonts. The results, shown in figure 1 and 2 are automatically generated by the proposed system. They seem to approximate well this semantic partition of the space.

Figure 1 shows the aspect of the proposed GUI when the mouse pointer is out of its window (starting condition). When user moves the mouse on a mountain, a callout, containing the phrases that google associated to that link, appears. Figure 2 shows the semantic partition of the query space computed by the proposed system. The two mountains in the right-bottom corner are related to a movie. The group of mountains in the upper side of the figure concerns human virus and PC virus while the isolated mountain in the left side of the interface is related to the site about character fonts.

Figure 2 shows that the proposed system fails separating the links related to human viruses and computer viruses. But this is a problem related to the I.R. system utilized to generate the proposed interface. This I.R. system is not the focus of the proposed work. From this point of view, also the sorted list proposed by google does not produce a perfect semantic classification of the sites. The focus of this work is the human computer interface that allows the user to see an overview of the semantic distribution of the retrieved web site at a glance.

A possible drawback of a GUI like the proposed one is the number of displayed mountains. Indeed, trying to display many mountains, it is quite difficult to avoid overlaps. But this problem does not affect the proposed interface because it is applied to the results of web search engines. Nowadays, the web search engines organize the search results in html pages with not more than 10 links. The proposed interface was tested using till 25 links giving good results in term of usability.

REFERENCES

[1] R. Cooley, B. Mobasher and J, Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web".

[2] www.searchenginewatch.com

[3] Michael J. McDonald, "*Active Research Topics in Human Machine Interfaces*", SANDIA REPORT (2000).

[4] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.

[5] LUHN, H.P., 'The automatic creation of literature abstracts', IBM Journal of Research and Development, 2, 159-165 (1958)

[6] WordNet: An Electronic Lexical Database. ISBN-10: 0-262-06197-X

[7] A.F. Smeaton, "Progress in the Application of Natural Language Processing to Information Retrieval Tasks", The Computer Journal, 35(3). 268-278

[8] R. Richardson et Al., "Using WordNet for Conceptual Distance Measurement", Proc. BCS.IRSG Colloquium, Glasgow, Scotland.

[9] W. Sammon. A nonlinear mapping for data structure analysis. IEEE Trans. Comput., 18:401–409, 1969.

[10] T. F. Cox and M. A. A. Cox, "Multidimensional Scaling", Chapman and Hall CRC, New York, second edition, 2001.