A Web Search Methodology for Different User Typologies

Marco Alfano , Biagio Lenzitti

Abstract: *Search engines and directories are the main tools used to find desired information into the ocean of digital contents that is the Web. However, they are not presently able to understand the user specific needs and starting knowledge because their inability to simulate the processes of human mind. Natural Language Processing, Folksonomy, Semantic Web and Serendipitous Surfing are some of the recent research fields towards understanding and satisfying the different user needs.*

This work aims to add one step more to this evolution path by presenting a new web search methodology that allows to create new knowledge paths based on user specific requirements. Thus, we consider different web-searcher typologies such as "basic searchers", "deep searchers" and "wide searchers" with different search expectations and starting knowledge. Some preliminary experiments have been performed to validate this methodology and build a new search engine around it.

Key words: *Information Search and Retrieval, Clustering, Search Process, Web Engine.*

INTRODUCTION

The *World Wide Web* is a huge set of interlinked hypertext documents accessed via the Internet. With a web browser, one can view web pages that may contain text, images, videos, and other multimedia and navigate between them using hyperlinks [14].

It is not always easy to find the information of interest in this chaotic, and non homogeneous world that is continuously evolving and where anybody can freely introduce contents without any censorship or quality control. Search engines and directories are born to help users to navigate into this ocean of information and reach the desired "harbour" [10]. However, present search tools are not able to understand users specific needs and their starting knowledge because they cannot simulate the processes of human mind. As a consequence, search engines are evolving and becoming "smarter" so to be able to understand the real user needs. *Natural Language Processing*, *Folksonomy*, *Semantic Web* and *Serendipitous Surfing* are some of the recent research fields used to evolve search techniques in this direction.

This work aims to add one step more to this evolution path by presenting a new search methodology that allows users new exploration possibilities. The main objective is to create new knowledge paths on the web based on user specific searching requirements [11]. Thus, we consider different searcher typologies such as a "basic searcher" who knows little about a topic and will look for more information, a "deep searcher" who will look for specific details on a topic that he/she already knows and a "wide searcher" who will look for expanding his/her knowledge domain with topics that are loosely related to the starting topic.

The paper is organized as follows. The second chapter presents a short survey on web search tools together with their evolution trends. The third chapter describes the basic principles of the search methodology oriented to the different searcher typologies described above. The final chapter presents the preliminary experimental results obtained by applying our methodology with some conclusions and future work.

SEARCH TOOLS AND EVOLUTION TRENDS

Subject directories and search engines (with their derivates) have long been the most used search tools for finding information on the web [2], [9]. Their main characteristics are shortly described below.

Subject directories (Yahoo!, DMOZ, About.com)

These are hierarchical databases with references to web sites. They are created and maintained by human editors who review and select sites for inclusion on the basis of previously determined selection criteria or ontologies. They allow users to perform

searches starting from macro-categories and specializing them by going down along a tree structure. Directories require a significant human effort and tend to be smaller than search engine databases, typically indexing only the home page or top level pages of a site. Nowadays, they are much less used than search engines.

Search Engines (Google, Yahoo! Search, Exalead)

They use huge databases of web page files that have been assembled automatically. They compile their databases by employing "spiders" or "robots" ("bots") to crawl through web space from link to link, identifying and analysing pages. Once the spiders get to a web site, they typically index most of the words on the publicly available pages at the site.

In ranking web pages, search engines follow a set of rules that may vary from one engine to another. Their goal, of course, is to return the most relevant pages at the top of their lists. The first search engines looked for the location and frequency of keywords and phrases in the document body and, sometimes, in the tags. The newer ones, as Google, assess popularity by the number of links that are pointing to sites; the more links, the greater the popularity, i.e., value of the page [3].

Metasearch engines (Clusty, Dogpile)

These engines do not crawl the web compiling their own searchable databases. Instead, they search the databases of several individual search engines simultaneously (often this search includes smaller, less well known search engines and specialized sites). They present the results of their searches either in a single list (from which duplicate entries have been removed) or multiple lists as received from each engine (duplicate entries may appear).

Deep web engines (DeepPeep, DeepDyve)

They act on a large portion of the Web that usual search engine spiders cannot, or may not, index. This part of the web has been called the "Invisible Web" or the "Deep Web" and includes, among other things, pass-protected sites, documents behind firewalls, archived material, the contents of certain databases and information that is not static but assembled dynamically in response to specific queries [1].

The main limitations of the tools presented above are the pertinence of the found information with the user search (mainly for search engines) and the potential information overload due to the amount of found results. This is due to the inability of the search tools to understand the user specific needs and starting knowledge because they cannot simulate the processes of human mind. Moreover, the Web has evolved towards the concept of collective participation to site contents and Web 2.0 presents many environments oriented to information and knowledge sharing [12]. The Web thus becomes the ideal environment for a comparison between search paths automatically created by engines and paths created by users with common interests. This spontaneously creates new logic links among semantic areas that at first sight looked unrelated. Web 3.0 will go further by not simply archiving data but elaborating their meaning thus easing the semantic search [13]. Moreover, data contextualization will be optimized by the so called "Data Web", i.e., the transformation of the Web in structured databases with data published using formats such as XML and RDF.

The understanding of user needs together with the evolution of the Web is bringing, as a natural consequence, to the development of new research fields that influence the innovation of the search strategies:

- *Natural Language Processing* (P.I.Q.U.A.N.T. [6], AskJeeves) where the query overcomes the rigid limitations of traditional search-engine queries by developing a particular search technique capable to answer user questions [8].

- *Folksonomy* (Del.icio.us, Technorati, Gataga) that labels web resources through the natural creation of meaningful relations among concepts in the different semantic areas. The bloggers community is the main "place" in the Web that allows the rise of these connections by sharing not only links but also tags and keywords freely associated to the signalled information. The tags of a site allow the creation of a content map through which classifying the site contents [7].
- *Semantic Web* (Hakia, Swoogle) based on the qualitative analysis of a site contents to "understand" the site subject through a formal logic and correctly place it into the SERPs (search engine report pages) [3].
- *Serendipity* (BananaSlug), is often connected to scientific discoveries (e.g., penicillin and X rays), i.e., the discovery of something unsearched and unexpected while searching for something else. The web navigation of a user is often "serendipitous", i.e., with a specific aim but ready to catch surprises, novelties and the unexpected [5].
- Other techniques, such as *clustering* (Clusty, iBoogie, Exalead, Vivisimo, Northern Light), *maps* (Kartoo, WebBrain, MapNet) and *search customization* (Google) help users to focus on their needs.

A SEARCH METHODOLOGY FOR DIFFERENT USER TYPOLOGIES

As discussed above, the recent research fields that influence search engines mainly try to somehow "read" the mind of users so to understand what they are really looking for. Users, however, have different needs when performing a search (especially for knowing or learning). Although the specialized search engines seen above can satisfy specific user needs, they are not easily accessible to "average" users who must first understand their specific requirements, then find the proper search engine and, finally, learn its specific syntax.

To overcome this limitation, we have developed a new search methodology that tries to satisfy the needs of different user typologies without imposing specific requirements on the user. In what follows we consider users more as "learning searchers" rather than "focused searchers". A "learning searcher" does not have a specific request but explores and navigates on the web to increase his/her knowledge (e.g., a user who wants to learn more about Wolfgang Amadeus Mozart). A "focused searcher", instead, knows exactly what is looking for and uses a general-purpose search engine to find it (e.g., an airline web site for booking a flight).

For simplicity we consider three main categories of learning searchers:

- the *basic searcher* has a small or no knowledge of the searched topic and uses a direct way to reach the objective of his/her analysis, i.e., looks for information strictly correlated to the searched keywords;
- the *deep searcher* has a good knowledge on the searched topic and desires to deepen his/her knowledge on the topic, i.e., looks for information that provides the details of the searched keywords;
- the *wide searcher* is not so interested to focus on a topic details but rather prefers to expand his/her knowledge domain by looking for topics that are loosely related to the searched keywords.

It should be noted that a user along the navigation path can change his/her needs becoming alternately a basic, a deep or a wide searcher.

We assume that a user starts his/her search on a topic by choosing, as usual, an initial keyword(s). We then consider a collection of *n* documents (web pages) that contain the searched keyword(s). We take a target, put the keyword(s) at the centre and distribute all the other *m* words present in the documents as arrows in the target (Fig. 1). The arrow a_{ij} indicates that the word i is present in document j and the distance d_{ij} represents, for document j, the averaged spatial distance between the word i and the keyword (it is an

average because both the word and the keyword could appear more than once in the document). In Fig. 1 we have drawn a radius for each document and put all the words belonging to that document at the different distances along the radius.

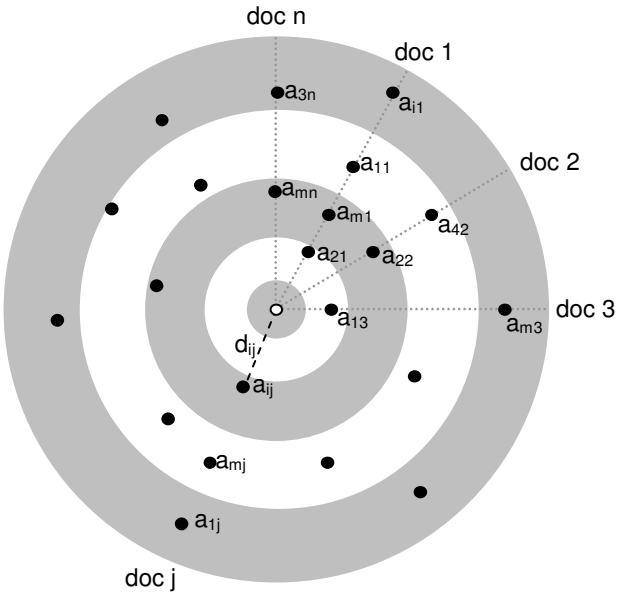

Fig. 1. Word distribution in the target.

For each word, we look at the distribution of the arrows in the target and, depending on the distance and number of the arrows, we can consider three main correlation categories (Fig. 2):

- target a represents a close and recurring word and indicates "strong correlation";
- target b represents a close and sporadic word and indicates "deepening";
- target c represents a far and recurring word and indicates "widening".

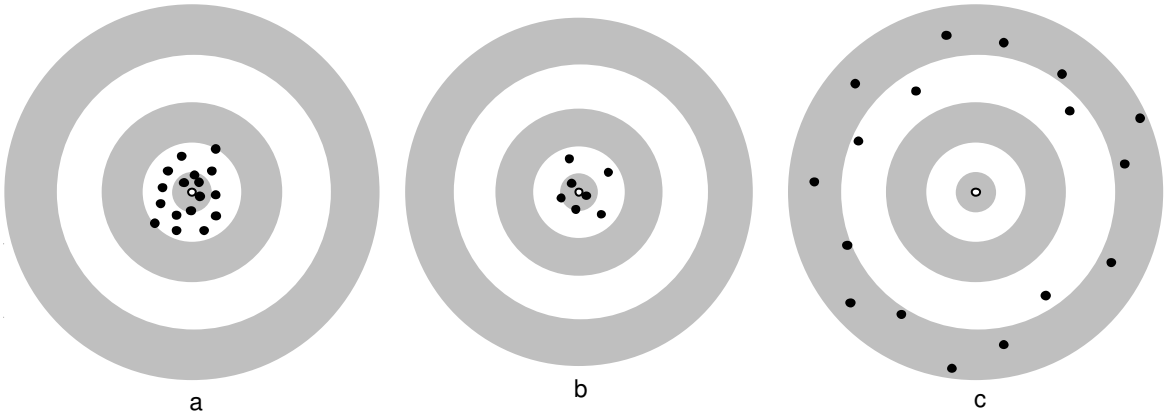

Fig. 2. Correlation categories: strong correlation (a), deepening (b), widening (c).

Each word has then a specific correlation with the initial keyword(s) and will allow a specific type of navigation. Thus, the "strongly correlated" words are terms that are close and recurring and they are likely to be conceptually close to the initial keyword(s). They will be used by a basic searcher for an understanding of the related knowledge domain. The "deepening" words are terms that are close but sporadic so they are likely to be correlated to the initial keyword(s) but will describe specific details because of their scarceness. They will be used by a deep searcher to go deep inside the knowledge domain. The "widening" words are terms far but recurring so they are likely to have a loose correlation to the initial

keyword(s) but important nevertheless to retrieve meaningful information positioned at the border of the of the semantic domain. They will be used by a wide searcher to expand the knowledge domain.

We plan to use the described methodology for the development of a new search engine. It will provide, for each category, both the found words and the related web pages. To this end, for each area we build a k-ary weighted tree with a single level where the root is the keyword(s) and the leaves are the correlated words. The weight of each edge is a linear combination of the distance between the word and the initial keyword and the word occurrences (Fig. 3).

The tree represents a "reference" page and the engine will rank the web pages of the collection creating a similar tree for each page and comparing it with the "reference" tree. The web page that has the minimum distance from the "reference" page will be first in the SERP and so on for the other pages.

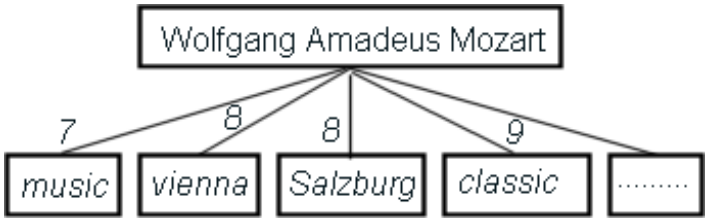

Fig. 3. K-ary weighted tree for the "strong correlation" category.

This search engine will allow the user to choose among different paths depending on his/her interests and objectives and focus on the analysis of the contents that belong to the selected search typology. The found pages and words will bring the user to compare his/her knowledge with this new information and to reason on how to continue the search towards a refining of the investigation (e.g., using the initial keyword and one of the suggested words for a new search) or towards new and unexpected discoveries based on the serendipity principle.

CONCLUSIONS AND FUTURE WORK

We have run preliminary experiments with a few keywords of different disciplines to verify the accuracy and usefulness of our methodology. To this end, we have downloaded the fifty web pages returned by Google for some specific keywords. For each web page, we have extracted the text and eliminated tags and common words. We have then computed the average of the distances between each remaining word of the page and the keyword. We have repeated this procedure for each web page so to create a target with the arrows such as the one shown in Fig. 1.

Table 1 shows the results obtained for the following keywords: *Wolfgang Amadeus Mozart*, *Computer Networks* and *Napoleon*. Note that for each category (strong correlation, deepening and widening) we have taken the first five words with the highest indexes.

We are presently refining our methodology running more experiments to start, as said above, implementing it in a search engine. We are analysing the fourth word category (far and sporadic words) to understand whether and how it can be useful to the user learning path. Moreover, we are planning to use semantic analysis to eliminate similar words such as composer and composers or network and networking as shown in Table 1. Finally, we are thinking on how to eliminate words (beside the common words and the similar ones) that may have no relevance to the user navigation path. This is a very critical issue because we want to be careful not to delete words that can lead users to serendipity discoveries but, on the contrary, we would like to facilitate them.

Table 1. Word categories found for different keywords.

keyword	strong correlation	deepening	widening
Wolfgang Amadeus Mozart	*1. music* *2. Salzburg* *3. Vienna* *4. piano* *5. works*	*1. composer* *2. musical* *3. opera* *4. symphony* *5. composers*	*1. time* *2. composed* *3. age* *4. young* *5. name*
Computer Networks	*1. network* *2. internet* *3. networking* *4. wireless* *5. information*	*1. data* *2. systems* *3. research* *4. communication* 5. *protocols*	*1. computers* *2. local* *3. area* *4. peer* 5. *server*
Napoleon	1. *French* 2. *France* *3. Bonaparte* *4. history* 5. *war*	1. *battle* *2. revolution* *3. military* *4. army* *5. napoleonic*	1. art *2. hotel* *3. usa* *4. years* *5. review*

REFERENCES

[1] Barbosa L. and Freire J. Locating Hidden-Web Entry Points. Proc. of the 16th International World Wide Web Conference. Banff, 2007.

[2] Barker J. and Kupersmith J. Finding Information on the Internet: A Tutorial. http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/FindInfo.html. 2008.

[3] Barroso L. A., Dean J. and Hölzle U. Web search for a planet: The Google cluster architecture. IEEE Micro n. 23, vol. 2, pp. 22–28. 2003.

[4] Berners-Lee T., Hendler J. and Lassila O. The Semantic Web. Scientific American, may 2001.

[5] Campos J. and Dias de Figueiredo A. Searching the Unsearchable: Inducing Serendipitous Insights. Proceedings of the Fourth International Conference on Case-Based Reasoning. 31 July 2001, Vancouver, Canada.

[6] Chu-Carroll J., et al. IBM's PIQUANT II. Proceedings of TREC2005.

[7] Hotho A. Information Retrieval in Folksonomies: Search and Ranking. Lecture Notes in Computer Science, vol. 4011. Springer Berlin, 2006.

[8] Jurafsky D. and Martin J.H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall, 2000.

[9] Koch P. and Koch S. A short and easy search engine tutorial. http://www.pandia.com/goalgetter/index.html. 2006

[10] Manning C.D., Raghavan P. & Schütze H. An Introduction to Information Retrieval. Cambridge University Press, 2009.

[11] Navarro-Prieto, R., Scaife, M., and Rogers, Y. Cognitive Strategies in Web Searching. Proceedings of the 5th Conference on Human Factors & the Web, 1999.

[12] O'Reilly T. What is Web 2.0. http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html. 2005.

[13] Spalding S. How To Define Web 3.0. http://howtosplitanatom.com/news/howto-define-web-30-2/. 2007

[14] Wikipedia. World Wide Web. http://en.wikipedia.org/wiki/World_wide_web.

ABOUT THE AUTHORS

Marco Alfano, PhD, Anghelos Centre on Communication Studies, Palermo Italy, Phone: +39 091 341791, E-mail: marco.alfano@anghelos.org.

Assist. Prof. Biagio Lenzitti, Dipartimento di Matematica ed Applicazioni, University of Palermo, Phone: +39 091 6040427, E-mail: lenzitti@math.unipa.it.