U-Search: A meta engine for creation of knowledge paths on the web

Marco Alfano, Biagio Lenzitti

Abstract: *The main tools used to find digital contents in the Web are search engines and directories but they are not presently able to understand the user specific needs and starting knowledge. This work presents "U-Search" a new meta engine that allows to create knowledge paths on the Web based on specific user requirements and knowledge levels. To this end, we consider different searcher categories 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 meta engine will suggest words and web pages correlated to each of those searcher categories thus creating new knowledge paths tailored to the real user needs.*

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

INTRODUCTION

Search engines and directories are born to help users to navigate into the ocean of non homogeneous and continuously evolving information that is the *World Wide Web* [2], [3], [8], [9]. The main limitations of those tools are the pertinence of 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 search tools to understand user specific needs and starting knowledge. Moreover, the Web has evolved towards the concept of collective participation to site contents and in Web 2.0 we find many environments oriented to information and knowledge sharing [10]. The Web thus becomes the ideal environment for a comparison between search paths automatically generated by engines and paths manually created by users with common interests. This spontaneously creates new logic links among semantic areas that at first sight looked unrelated.

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 innovate the search strategies. For example, *Natural Language Processing* [7] is used by AskJeeves, *Folksonomy* [6] is used by Del.icio.us, Technorati and Gataga, *Semantic Web* [4] is used by Hakia and Swoogle, *Clustering* is used by Clusty, iBoogie, and Exalead and *Serendipity* [5] is used by BananaSlug. Other novel techniques found in some search engines are *Maps* used by Kartoo, WebBrain and MapNet and *Search customization* used by Google.

This paper, based on some ideas discussed in [1], aims to add one step more to this evolution path by presenting "U-Search" a new meta engine that allows to create knowledge paths on the Web based on specific user requirements.

To this end, we consider different searcher categories 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 meta engine will suggest words and web pages correlated to each of those searcher categories thus creating new knowledge paths tailored to the real user needs.

The paper is organized as follows. The second chapter describes the basic principles of our search methodology. The third chapter presents the architectural details of U-Search and the fourth chapter describes its implementation and some experimental results. The final chapter presents some conclusions and future work.

U-SEARCH METHODOLOGY

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 categories 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 Napoleon). 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:

- a *basic searcher* has a small or no knowledge of the searched topic and wishes to understand more, i.e., looks for information strictly correlated to the searched keyword(s);
- a *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 keyword(s);
- a *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 keyword(s).

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 have created a search model based on the searcher typologies described above. To this end, 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) and the *m* words present at least in one document (not considering the common words). For each word, we evaluate the number of documents containing the word together with the number of occurrences of the word in each document. We then make the following categorization:

- a word that appears in many pages with many occurrences can be used for "understanding";
- a word that appears in many pages with a few occurrences can be used for "deepening" the knowledge;
- a word that appears in a few pages with many occurrences can be used for "widening" the knowledge.

Each word has then a specific correlation with the initial keyword(s) and will allow a specific type of navigation. Thus, the "understanding" words 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 (e.g., "Bonaparte" and "french" for the "Napoleon" keyword). The “deepening” words are terms that are likely to have a loose correlation to the initial keyword(s) (in terms of number of occurrences) but appear in many documents so they are likely to represent specific topics inside the semantic domain (e.g., "military" and "Italy" for the "Napoleon" keyword). The “widening” words are terms that are likely to have a strong correlation (in terms of number of occurrences) to the initial keyword(s) but appear only in a few documents so they are likely to represent specific topics at the border of the of the semantic domain (e.g., "pope" and "Rome" for the "Napoleon" keyword).

U-SEARCH META ENGINE

“U-Search” is a meta engine that implements the search methodology presented above. It allows a user to specify one or more keywords and provides words and related web pages for each category. Fig. 1 shows the basic architecture of “U- Search”.

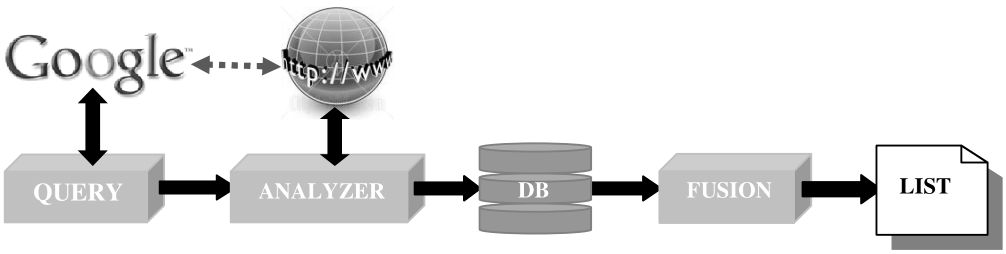


Fig. 1. U-Search Architecture.

The “QUERY” module takes the keyword(s) specified by the user and the number n of documents (web pages) to be analyzed. It then searches this keyword(s) through Google and takes the first *n* results creating a collection $D=\{doc_1, doc_2, \dots doc_n\}$ of *n* documents with the related links.

For all links, the “ANALYZER” module retrieves the related web pages, cleans them (by removing tags and common words) and stores the remaining words $W=\{w_1, w_2, w_3, \dots. w_m\}$, present at least in one document, together with their number of occurrences in the “DB” database.

The "FUSION" module uses the sets *D of n* documents *and W* of *m* words. For each w_i and for each doc_j it creates the matrix $A=\{a_{ij}\}$ of occurrences of w_i in doc_j reported in Table 1. It then calculates ND_i, i.e., the number of documents in which w_i appears and $WO_i = \sum_{j=1}^{n} a_{ij} / ND_i$, i.e., the total number of occurrences of w_i divided by number of documents that contain w_i.

Table 1. Occurrences matrix.

	doc_1	doc_2	...	doc_j	...	doc_n
w_1	a_{11}	a_{12}	...	a_{1j}	...	a_{1n}
w_2	a_{21}	a_{22}	...	a_{2j}	...	a_{2n}
...	...	...	...	...	...	...
w_i	a_{i1}	a_{i2}	...	a_{ij}	...	a_{in}
...	...	...	...	...	...	...
w_m	a_{m1}	a_{m2}	...	a_{mj}	...	a_{mn}

Finally, the “LIST” module takes the lower and upper thresholds for the number of documents (*NDLT* and *NDUT* respectively) and the lower and upper thresholds for the word occurrences (*WOLT* and *WOUT* respectively), as specified by the user in the input page, and places each word in one of the three correlation categories based on the following rules (Fig. 2):

- words w_i with $ND_i \geq \mathit{NDUT}$ and $WO_i \geq \mathit{WOUT}$ go to the "understanding" category;
- words w_i with $ND_i \geq \mathit{NDUT}$ and $WO_i < \mathit{WOLT}$ go to the "deepening" category;
- words w_i with $ND_i < \mathit{NDLT}$ and $WO_i \geq \mathit{WOUT}$ go to the "widening" category;
- all other words are not listed.

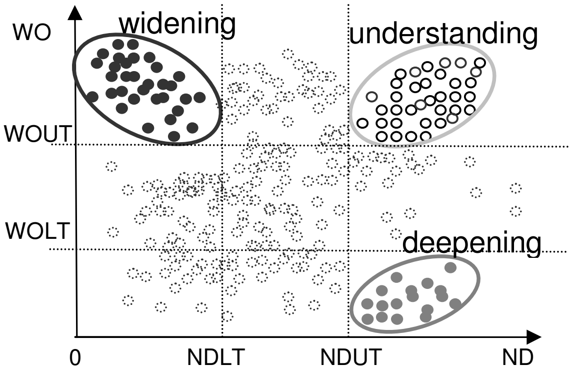
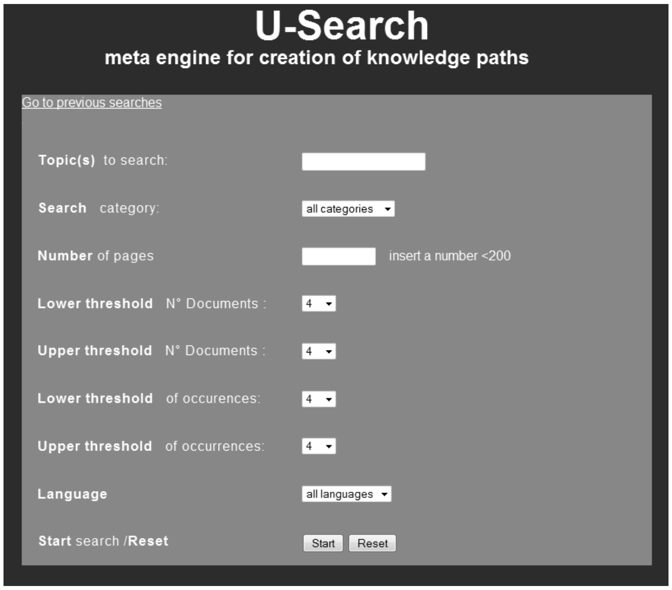


Fig. 2. Word distribution and correlation categories.

At the end of this process, U-search will present the user with a list of words grouped in the different categories and a list of links to web pages for each category, so that the user can choose the topic(s) to continue his/her learning path and explore the related web pages. Note that the web pages associated to a category are the first five pages with the highest number of occurrences of the words in that category.

U-SEARCH IMPLEMENTATION AND EXPERIMENTAL RESULTS

We have implemented the U-Search meta engine using the PHP language and MySQL database. It can be found at the address "http://www.citc.unipa.it/usearch". Fig. 3 shows the input page where the user can specify the keyword(s), search category (understanding, deepening or widening), number of pages to be analyzed, lower and upper thresholds for number of documents and word occurrences, together with the language of the web pages (currently english and italian).



U-Search

meta engine for creation of knowledge paths

Go to previous searches

Topic(s) to search:	
Search category:	all categories ▾
Number of pages	insert a number <200
Lower threshold N° Documents :	4 ▾
Upper threshold N° Documents :	4 ▾
Lower threshold of occurences:	4 ▾
Upper threshold of occurrences:	4 ▾
Language	all languages ▾
Start search /**Reset**	Start Reset

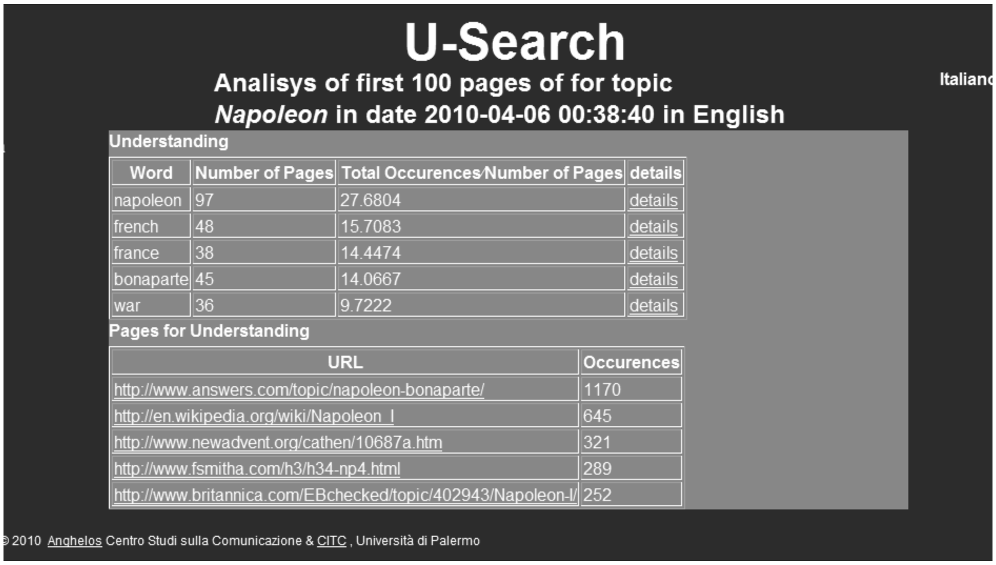
Fig. 3. U-Search input page.

We have run some experiments using keywords from different disciplines and Table 2 shows the results obtained when looking for "Napoleon", "Mozart", "Galaxies" and "Bulgaria" keywords in a set of one hundred pages.

Table 2. Results for "Napoleon", "Mozart", "Galaxies" and "Bulgaria" keywords.

Keyword	*Understanding*	*Deepening*	*Widening*
Napoleon	1. french 2. france 3. bonaparte 4. war 5. army	1. napoleonic 2. military 3. history 4. general 5. Italy	1. pope 2. napol 3. louis 4. paris 5. rome
Mozart	1. music 2. opera 3. piano 4. wolfgang 5. amedeus	1. classical 2. vienna 3. requiem 4. musical 5. composer	1. concertos 2. chamber 3. allegro 4. choral 5. soprano
Galaxies	1. galaxy 2. stars 3. milky 4. way 5. spiral	1. universe 2. light 3. gas 4. elliptical 6. hubble	1. ngc 2. seyfert 3. hole 4. black 5. Harvard
Bulgaria	1. bulgarian 2. sofia 3. country 4. world 5. government	1. european 2. travel 3. economy 4. macedonia 5. varna	1. nuclear 2. empire 3. cup 4. war 5. party

Fig. 4 shows the actual answer page for the "understanding" category of the "Napoleon" keyword. As said above, five web pages are also listed for this category.



U-Search

Analisys of first 100 pages of for topic
Napoleon in date 2010-04-06 00:38:40 in English

Italiano

Understanding

Word	Number of Pages	Total Occurences⁄Number of Pages	details
napoleon	97	27.6804	details
french	48	15.7083	details
france	38	14.4474	details
bonaparte	45	14.0667	details
war	36	9.7222	details

Pages for Understanding

URL	Occurences
http://www.answers.com/topic/napoleon-bonaparte/	1170
http://en.wikipedia.org/wiki/Napoleon_I	645
http://www.newadvent.org/cathen/10687a.htm	321
http://www.fsmitha.com/h3/h34-np4.html	289
http://www.britannica.com/EBchecked/topic/402943/Napoleon-I/	252

© 2010 Anghelos Centro Studi sulla Comunicazione & CITC , Università di Palermo

Fig. 4. U-Search answer-page.

CONCLUSIONS AND FUTURE WORK

This paper has presented "U-Search", a new meta engine that allows the creation of knowledge paths on the Web based on specific user requirements and knowledge levels. We have implemented a prototype and are presently running some experiments to refine the algorithms used to place words in the different correlation categories and to validate the hypotheses made on the existence of those kinds of correlations.

As a future work, we plan to use semantic analysis to eliminate similar words and other words (beside common words and similar ones) that may have no relevance to the user navigation path. Moreover, we want to use the U-Search methodology together with some "machine learning algorithms" for directory classification and ontologies creation. Finally, we plan to analyze the fourth word category (words that appear in a few documents and with a low number of occurrences) to understand whether and how to correlate it to the user learning paths. This is a very critical issue because we would like to facilitate, if possible, the user towards serendipity discoveries and so we must be very careful in not eliminating words that can lead to serendipity even though they are apparently uncorrelated to the initial keyword(s).

REFERENCES

[1] Alfano M. and Lenzitti B. A web search methodology for different user typologies. Proc. of ACM International Conference on Computer Systems and Technologies (CompSysTech' 2009). 2009.

[2] Barker J. and Kupersmith J. Recommended Search Strategy: Analyze your topic & Search with peripheral vision. http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/Strategies.html. 2009.

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

[4] Barroso L. A. et al. Web search for a planet: The Google cluster architecture. IEEE Micro, Vol. 2,No.23, pp.22–28, 2003.

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

[6] Hotho A. Information Retrieval in Folksonomies: Search and Ranking. L.N.C.S., Vol. 4011. Springer Berlin , 2006.

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

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

[9] Manning C.D. et al. An Introduction to Information Retrieval. Cambridge University Press, UK. 2009.

[10] O'Reilly T. What is Web 2.0 Design Patterns and Business Models for the Next Generation of Software. <http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html>. 2005.

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 Informatica & CITC, University of Palermo, Phone: +39 091 238 91101, E-mail: lenzitti@math.unipa.it.