Enhancing Web Searches from Concept Map-based Knowledge Models[i]

Marco Carvalho, Rattikorn Hewett & Alberto J. Cañas

Institute for Human & Machine Cognition
40 South Alcaniz, Pensacola, FL. 32502
www.ihmc.us

ABSTRACT

Although many publicly available search engines can retrieve relevant information reasonably well, the list of the retrieved web pages is still often too large or contains information that has no relevance to the query. Our goal is to improve the results of these search engines for queries generated by users while constructing and/or browsing concept map-based knowledge models. By exploiting the propositional and hierarchical nature of concept maps, we have developed two algorithms, SAgent and WAgent, for filtering and ranking the results obtained by the search engines. The algorithms were implemented via mobile agents and evaluated empirically. In our experiments, six subjects submitted queries based on a concept map to publicly available search engines (Google, Altavista, Yahoo, Excite), and were asked to rank the relevance of the results; the agents also filtered and ranked the engines' output. The results indicated that the proposed algorithms are capable of identifying pages that the subjects considered are relevant to the context on the map

Keywords

Information filters, concept maps, searching, knowledge models, information retrieval, mobile agents

INTRODUCTION

It's a well-known fact that online information is growing at an exponential rate and each day more and more data is made available on diverse formats like text documents, images, movies, sounds, etc. Some reports [1] have estimated a growth on the number of Internet hosts going from about 93 million in July 2000 to about 109.5 million on January 2001. All this information is available in a matter of seconds, and to be useful in addressing our daily questions, it must be properly organized, searched, filtered, sorted and displayed.

This paper presents the preliminary results from an ongoing research effort on the development of an information software agent aimed at improving the filtering and ranking of results obtained from conventional search engines, displaying a reduced set of documents with high relevance to a user who is constructing or navigating through a graphical knowledge model based on a concept map.

CONCEPT MAPS AS GRAPHICAL REPRESENTATIONS OF KNOWLEDGE MODELS

Concept Maps have been widely used to represent knowledge in all domains since their inception in the 70's by Novak [2]. Recently, software packages like CmapTools[1], which facilitate the online manipulation of concept maps, have extended their use and applicability to knowledge sharing, organization and browsing [3][4]. Electronic concept maps provide an elegant representation of an expert's domain knowledge in a browsable, sharable form, easily understood by others. For example, the CmapTools have been used for applications such as the creation at NASA of a large-scale multimedia CD and web site on Mars

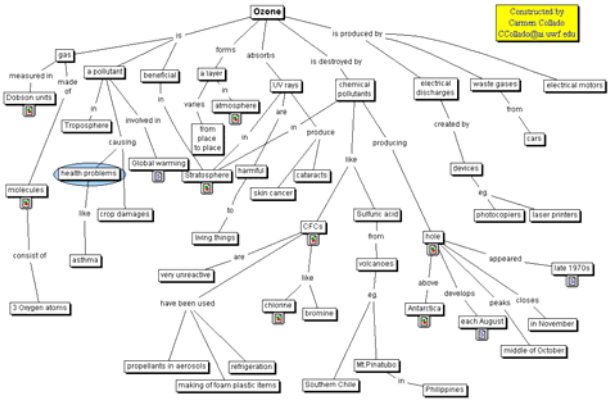

Figure 1. Concept map about ozone. Available online at http://public-cmaps.coginst.uwf.edu/cmaps/Ozone/

[1] The CmapTools software package is available free for non-profit use at http://cmap.ihmc.us.

(http://cmex.arc.nasa.gov).

This effort aims at developing a search-enhancer module for the CmapTools, which aids the user – whether an expert, a student, or any other type of user – who is constructing or browsing through a concept map, on his task to search on the web for additional information related to the map. When the user requests to perform a search on a particular part of the map, a mobile agent [5][6] is created that analyzes the content of the particular map, and moves through a set of meta-search servers to carry out the searching, filtering and ranking, and bringing back to the client only those links considered to be of high relevance.

The agent takes advantage of the nature and topology of concept maps. A concept map is a two-dimensional representation of a set of concepts that is constructed so that the interrelationships among them are evident (see Figure 1). Concept maps represent meaningful relationships between concepts in the form of propositions. Propositions are two or more concepts linked by words to form a semantic unit. In the simplest form, a concept map would contain just two concepts connected by a linking word to form a single proposition. For example, "ozone absorbs UV Rays" would represent a simple map forming a valid proposition about the concepts "ozone" and "UV Rays." A concept acquires additional meaning as more propositions include the concept. The vertical axis expresses a hierarchical framework for the concepts. More general, inclusive concepts are found at the highest levels, with progressively more specific, less inclusive concepts arranged below them. These maps emphasize the most general concepts by linking them to supporting ideas with propositions. The agent takes the set of concepts in the concept map and the links between them as the context to rank the information.

THE META-SEARCH SERVER

The filtering agent takes the query from the user, examines the concept map for contextual information, and moves to one or more meta-search servers. At the server, the agent queries the publicly available search engines. For each result obtained from the engines, it retrieves the corresponding web page, parses it to remove HTML tags, scripts and stop words, and adds it to a reverse index stored at the server with a time stamp that defines it's time-to-live or expiration date. (The time stamp should save time from repeated retrievals of the same page, which is one of the most expensive steps of performed by the server. It is expected that when working on an specific concept map, the client will perform several queries under the same context which would result on repeated URLs returning from the search engines that won't have to be retrieved, parsed and indexed again. It is expected that with continuous use of the tool over the same map or set, or related maps its overall performance will improve.) The agent then applies the context information it extracted from the concept map to the indexed information, ranks the pages, moves back to the client machine and presents the ranking of results to the user.

DOCUMENT RETRIEVAL

Traditional information retrieval relies mostly on Salton's theory of the vector space model [7]. TF-IDF vector analysis is based on the determination and geometric evaluation of two n-dimensional vectors, one representing the document and the other representing the query. Such vectors are usually weighted through several different techniques [8][9] and then compared through different "operations", e.g., inner product or cosine coefficient [10][11].

Although such approach has provided good results on text retrieval applications, concept maps by their nature and topology provide more information than just a set of weighted words. From the map structure we can extract propositions that can provide strong context correlation between the retrieved text and the map. The use of phrases to evaluate similarity has been presented elsewhere [7]. We take advantage of the propositional nature of concept maps to provide equally strong metrics with better coverage. Such features can be measured with the help of what we refer to as the distance matrix.

THE DISTANCE MATRIX

A distance matrix is used to match the propositions from the concept map with the text of each web page retrieved. It contains information about proximity of each concept found along the document as well as a count of each word. Propositions extracted from the concept map form phrases like:

[Plants have roots]

The distance matrix is used to locate and capture text information, similar to the proposition above, that might exist in the text, for example in the form:

[Plants are flourishing all over the place and their roots go deep into the ground…]

This phrase, although written on a much more discursive way does contain the somewhat concealed notion that plants have roots, which would make it similar to the proposition first presented.

This approach behaves in a similar way to the vector space similarity proposed by Salton (except for the lack of the inverse document frequency term) but permits the simultaneous evaluation of one more metric: the proximity measure between terms that constitute a proposition from the map.

On a concept map with N concepts, the distance matrix will be an NxN matrix, where each element will contain the inverse of the distance between terms. From the list of concepts from the map, the distance matrix can be calculated by the following equation:

$$E_{i,j} = 1 \Big/ abs(PW_j - PW_i)$$

Where, for concepts W_i and W_j, E_{ij} is the element of the distance matrix and PW_j and PW_i are respectively the position of the words W_i. and W_j in the document, expressed as the number of words from the beginning of the page.

Additional metrics could be added to improve the ranking process [10][11], such as relative position of the word within the document, it's HTML format, its appearance in meta-tags, link reference [12], anchor text, etc.

RANKING ALGORITHM

Two different ranking algorithms were implemented as agents (SAgent, for simple agent and WAgent, for weighted agent), both making use of the information on the distance matrix.

The first algorithm, utilized by SAgent, is based on a straightforward evaluation of the distance matrix: all elements of the matrix are summed and the resultant value is used for the ranking. The higher the resultant value, the better the rank.

WAgent, based on the second algorithm, attempts to highlight significant features of the concept map and the distance matrix before summing its elements to obtain the ranking coefficient. Based on the hierarchical nature of concept maps, each concept C_i is assigned a weight of 5–*n*, where *n* is the shortest path from the concept to the root of the map, with 1 being the minimum weight. (The root concept has a weight of 5). The algorithm uses a second matrix, derived from the concept map, that is multiplied (element-wise) by the distance matrix to force on it information about the concept weights. This NxN weight matrix, where N is the number of concepts on the concept map, is calculated by the following equation:

$$WE_{i,j} = \left(\frac{WC_i + WC_j}{2} \right) \bullet \delta_{i,j}$$

where $WE_{i,j}$ is the element (i, j) of the weight matrix, WC_i and WC_j are respectively the weights of the concepts C_i and C_j, and $\delta_{i,j}$ is a function that can assume three values: 1.0, 0.5, and 0, if the shortest path length between C_i and C_j is 1, 2, and otherwise, respectively.

Each element of this weight matrix will intensify or lower the importance of the correspondent element of the distance matrix. The goal is to reinforce possible propositions, captured by the distance matrix, that appear on the map and have a high weight for their concepts and at the same time lower the importance of the remaining elements.

After multiplying each element of these two matrices, WAgent then performs the same similarly measure as SAgent, that is, it calculates the sum of all elements of the matrix, and uses the resultant value as the ranking coefficient.

EXPERIMENT METHODOLOGY

To evaluate whether the proposed agents improved on the ranking of results provided by the publicly available search engines, we proceeded with an experiment where subjects were provided with a concept map as a basis to perform a search, were asked to execute a query to public search engines looking for information relevant to a piece of the map, and were then asked to rank the results. Separately, both agents also ranked the web pages returned by the search engines. The objective was to determine whether the agents improved the ranking provided by the search engines, coming closer to the users' ranking of documents.

Six subjects were presented with the concept map on the Ozone shown in Figure 1. By printed and read instructions, each subject was separately asked to submit one query that they believed would provide the most relevant information related to the "health problems" concept in the map.

After submission, the query string was used to retrieve 15 URLs from each of four search engines (Yahoo, AltaVista, Google and Excite) amounting

to a total of 60 URLs. This list of URLs was then checked for repetitions, randomized, and presented back to the subject for ranking. Classification for each URL was done by the subjects to indicate if each URL was "Highly Relevant", "Relevant" or "Not Relevant" to the context presented by the map and the highlighted (health problems) concept.

There were no time constrains for classification and each subject was allowed to visit all pages and sub-pages to review their content. On average, each subject spent about 23 minutes classifying all the URLs. The concept map was also available at all times to be consulted. Users were allowed to change their minds on the ranking of the pages as they went along the classification.

Given the 60 URLs returned by the search engines, the content of each page was retrieved to be processed by SAgent and WAgent. Each agent then displayed its own ranking of the pages based on its algorithm.

EXPERIMENTAL RESULTS

The experiment involved 6 subjects. For each subject, we obtained the top five documents as ranked by each of six different ranking schemes (four publicly available search engines and the two agents). The subjects rated each document as 1 (no relevance), 2 (moderate relevance), or 3 (high relevance). An example of a subject's ranking is shown in Figure 2. In this example, the subject rated none of the top five documents ranked by Yahoo as highly relevant.

Figure 3 (a) shows the number of documents, of the top five produced by each ranking scheme, considered highly relevant by each subject. Averaging over the six different subjects, the documents found by WAgent and by Google have the highest number of "high relevant" hits, whereas those found by Yahoo have the lowest. Similarly, we obtain ratings by the six subjects for the top eight documents found by the six ranking schemes. This is shown in Figure 3 (b). WAgent gives the highest average number of "high relevance" hits, followed by SAgent and Google. .

Yahoo	Google	AltaVis	Excite	WAgent	SAgent
1	3	1	3	3	3
1	3	1	3	1	1
1	1	3	1	3	3
2	3	1	2	2	2
2	3	3	2	3	3

Figure 2: An example of results for one subject.

To check whether the above averages of ordinal data

Subject	Yahoo	Google	AltaVis	Excite	WAgent	SAgent
1	0	4	2	2	3	3
2	0	3	2	3	3	3
3	0	5	2	5	4	3
4	0	4	3	0	5	4
5	3	5	1	0	5	5
6	0	1	1	0	2	1
Avg. Top5	**0.50**	**3.67**	**1.83**	**1.67**	**3.67**	**3.17**

(a)

Subject	Yahoo	Google	AltaVis	Excite	WAgent	SAgent
1	0	7	3	5	3	5
2	0	4	1	4	5	6
3	0	8	3	8	6	6
4	0	4	4	3	8	7
5	3	7	1	0	8	7
6	0	2	2	2	3	2
Avg. Top8	**0.50**	**5.33**	**2.33**	**3.67**	**6.00**	**5.50**

(b)

Figure 3: Number of high relevance hits.

are statistically significant, we apply the non-parametric Kruskal-Wallis Test [13] to the results for each subject. The null hypothesis is H_0: there are no differences between the ranking schemes.

For the results using the top five ranks, we must accept the null hypothesis for three of the six subjects. The ranking schemes are significantly different (*0.05*) only for subjects 3, 4 and 5. For the top eight ranks, the Kruskal-Wallis Test shows that the ranking schemes are statistically (*0.05*) for all six subjects. As shown in Figure 3 (b), WAgent and SAgent have the highest number of "high relevance" hits for four of the six subjects. For the other two subjects, Google and Excite outperformed WAgent and SAgent by a small margin.

If we treat the subjects' ratings of documents (which we will refer to as relevance score) as continuous values rather than ordinal values, we can obtain a mean relevance score for each ranking scheme for each subject. Assuming that the distribution of mean sample is normal, we apply ANOVA [13] to each subject to test the null hypothesis, H_0: all the sample means of relevance scores of each ranking schemes are the same. Based on ANOVA, the null hypothesis is rejected, and thus, all the means are significantly different (0.05). The same results are obtained for both the top five and top eight cases.

Figure 4 summarizes the average relevance scores over the six subjects for the top five and top eight cases. The top three performers, in order, are

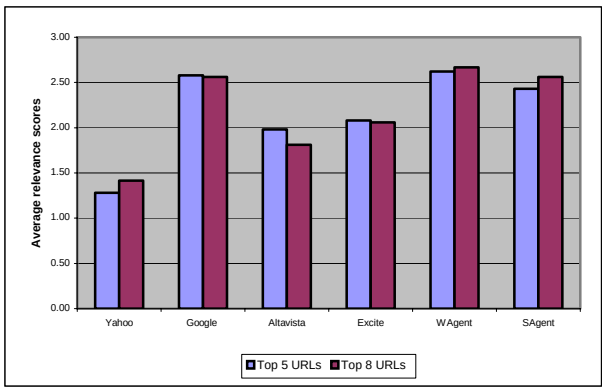

Figure 4 Average relevance scores by different ranking schemes.

WAgent, Google and SAgent in the top five case and WAgent, SAgent and Google in the top eight case.

DISCUSSION OF RESULTS

The results above seem to indicate that the proposed algorithms, particularly that of WAgent, are capable of identifying pages that the subjects considered are relevant to the context on the map. The concise and precise propositions extracted from the concept maps provided strong contextual information for ranking to SAgent. WAgent, taking advantage of the concepts' weights based on the map's topology, further improved on the results of SAgent, performing in most cases close to or better than the best of the four search engines, Google. We are investigating various ways to refine the metrics in order to improve the Agents' performance, among them using the structural information provided by the retrieved pages themselves, as Google does.

CONCLUSIONS

The proposed algorithms scored similarly or better than the best of the four search engines in the ranking of retrieved documents for relevance to the concept map according to the subjects' criteria, and clearly performed better that the other three. This seems to indicate that, given the results from several search engines, there is room for improvement over the original ranking provided by the engines. The propositional nature of the concept maps, together with their hierarchical topology, seem to provide enough contextual information to identify and rank those documents that are more relevant to the map. At a minimum, our results justify further work in the area.

ACKNOWLEDGEMENTS

We would like to thank Dale Welch and Mary Jo Carnot from IHMC for the outstanding technical support, database expertise, and data analysis help.

REFERENCES

[1] Internet Domain Survey (2000). Internet Software Consortium (http://www.isc.org).

[2] Novak, J. D. and D. B. Gowin. (1984). *Learning How to Learn*. Cambridge University Press, N.Y.

[3] Cañas, A. J., K. M. Ford, J. Brennan, T. Reichherzer, and P. Hayes. (1995). *Knowledge Construction and Sharing in Quorum*. In Proc. of AI in Education, Washington D.C., pp. 218-225.

[4] Coffey, J. W and A. J. Cañas (2000). *A Learning Environment Organizer for Asynchronous Distance Learning Systems*. Twelfth IASTED Intern. Conf. Parallel and Distributed Computing and Systems (PDCS 2000), Las Vegas, Nevada.

[5] Suri, N., J. M. Bradshaw, M. R Breedy, P. T. Groth, G. A. Hill, R. Jeffers, and T. S. Mitrovich. (2000). *An Overview of the NOMADS Mobile Agent System*, Sixth ECOOP Workshop on Mobile Object Systems.

[6] Suri, N., J. M. Bradshaw, M. R. Breedy, P. T. Groth, G. A. Hill, & R. Jeffers. (2000). *Strong Mobility and Fine-Grained Resource Control in NOMADS.* Proc. of the 2nd Intern. Symp. on Agents Systems and Applications and the 4th Intern. Symp. on Mobile Agents (ASA/MA 2000). Springer-Verlag.

[7] Salton, G. and M. J. McGill. (1983). *Introduction to Modern Information Retrieval.* McGraw-Hill, New York.

[8] Pazzani, M., J. Muramatsu and D. Billsus, (1996). *Syskill & Webert: Identifying Interesting Web Sites*. Proc. of the National Conference on Artificial Intelligence, Portland, OR.

[9] Pazzani, M., J. Muramatsu & D. Billsus. (1996). *Syskill & Webert: Identifying Interesting Web Sites*. Proc. of the National Conf. on Artificial Intelligence, Portland, OR.

[10] Fireder, O. (2000). "*On the Integration of Structured and Semi-Structured Data: From Concept to Prototype to Deployment*", Illinois Institute of Technology.

[11] Brin, S. and P. Lawrence (1998). *The Anatomy of a Large-Scale Hypertextual Web Search" Web*. 7th WWW Conference.

[12] Bharat, B. et al. (1998). *The Connectivity Server: Fast Access to Linkage Information on the Web*. 7th WWW Conference.

[13] Lapin, L. L. (1973). *Statistics for Modern Business Decisions*, Harcourt Brace Jovanovich, Inc.

---

[i] Proceedings of SCI 2001: Fifth Multi-Conference on Systems, Cybernetics and Informatics, Orlando, FL (July 2001).