Improving Retrieval by Querying and Examining Prestige

Jesús Ubaldo Quevedo T.
Department of Computer Science
University of Houston
Houston, TX 77204-3010
Phone: 713-743-3350
Fax: 713-743-3335
jquevedo@uh.edu

Ana Gabriela Covarrubias S.
Maestría: Ciencias Computacionales
Univ. Autónoma de Guadalajara
Guadalajara, Jal. 44100 México
Phone: +52 3110 1131
anagabriela@smartware.com.mx

S. H. Stephen Huang
Department of Computer Science
University of Houston
Houston, TX 77204-3010
Phone: 713-743-3338
Fax: 713-743-3335
shuang@uh.edu

ABSTRACT

The increasing amount of data stored in the World Wide Web demands efficient techniques for information retrieval. When consulting a regular search engine, it is very common to receive millions of documents as an answer. In this paper, we present a web query language capable of extracting additional knowledge from the response provided by the searcher to prune undesired documents. Moreover, we use this language to define queries that explore the theory of "prestige". This concepts states that the prestige of a web page that treats a particular topic "t", depends on the number of documents referencing it and treating the same topic "t". Various experiments confirm the efficiency of our approach.

Keywords

Web Query Language, World Wide Web, Information Retrieval.

1. INTRODUCTION

Nowadays, there are several well known and widely used search engines [6, 8, and 9] to retrieve information from the World Wide Web; however, the quality of the output of these searchers has been questioned in several publications [12]. Each system applies different techniques to find and rank pages according with their assumed relevance with a particular topic. Frequency of the term in a document is one of those techniques, which has the problem that quantity not always leads us to quality. In fact, there are some companies that for marketing purposes hide in their html source multiple copies of possible keywords so that the searcher will rank them among the top of the list. This phenomenon is called Search Engine Persuasion (SEP) or Web Spamming [10, 11].This obviously has nothing to do with significance of a topic. Another technique involves human experts called themselves editors [8] processing and classifying pages in directories which so far have classified two million web sites [8] from more than one billion pages [13] available through all the web. This task of classifying that huge number of pages goes beyond human capacity. There are other efforts that involve Boolean queries, proximity queries and several other advanced algorithms used in information retrieval [14, 15]. Web query systems have been very popular during the past years. Systems like WebSQL[1], WOQL[2][3], WebSSQL [4] and WebLog[5] retrieve information from the World Wide Web. WebSQL, for instance, is a high level declarative query language for extracting information from the Web, it is written in Java, so the queries can be embedded into Java programs; however, our system uses primitives implemented in an Interbase Database, so they can be called from any application that can access a Interbase Server independently of the programming language.

2. IMPLEMENTATION

Our system uses its own web query language to identify pages that are considered relevant by others, and, therefore, prestigious.

2.1 Web Query Language

The web query language has three main components: metadata, primitives and interface.

The metadata is the knowledge representation of all documents retrieved by the search engine. The search engine is prompted with a keyword. Then, our system collects all the attributes needed to fill the relations illustrated in figure 1 for every page. This approach transforms the result of the searcher into a relational database.

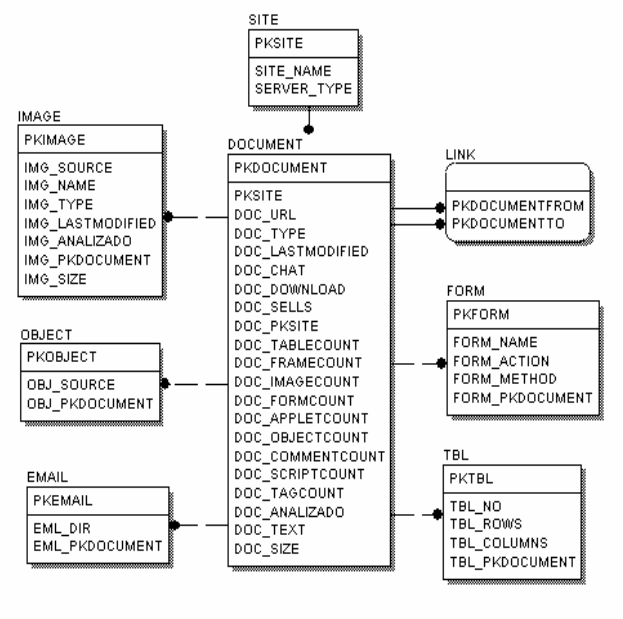

Figure 1. Relations needed to query documents

The primitives are the set of operations that can be performed over the relations. Table 1 shows all primitives, where *Q*, *P* and *R* are sets of urls, *x* and *y* represent one url, and *n* is an integer greater or equal t0 1.

PRIMITIVE	**SYNTAX**
Mentions	Mentions (Q, q)
Link_to	Link_to(x,y,n)
References_to	References_to(Q, x, n)
Length	Length(Q, n)
Intersect	Intersect(P,Q, R)
Outgoing_links	Outgoing_links(Q, q, n)
Sites	Sites(P,Q)

Table 1: List of Primitives

The interface to execute queries can be either a QBE (Query by Example) or a SQL style interface.

2.2 Querying Capability

Due to the relational representation illustrated in figure 1, we are capable of using a query for further pruning the result provided by the searcher. For instance, we may want to restrict our search to documents containing certain number of images and having a chat option. This subsection briefly illustrates the way queries can be defined.

Query 1: Find documents containing keyword “travel” and selling something

```
SELECT * FROM MENTIONS(‘travel’) WHERE DOC_SELLS ='Y'.
```

Query 2: Find the sites that contain documents containing the keyword “hotel”

```
SELECT SITE_NAME FROM MENTIONS('hotel') M
```

```
INNER JOIN SITE S ON M.DOC_PKSITE = S.PKSITE
```

2.3 Prestige

Prestige is the level of recognition at which one is regarded by others. When looking for a physician in a big city, you may go and consult the yellow pages and find thousands of names. It would help to have some references from somebody that used their services to select the best one. However, it would be even better to know those doctors that have recognition from their own colleagues. Hence, we could measure the prestige of a doctor by his/her references. The level of prestige is higher when somebody from the same profession recommends him/her than just a regular patient. Our system locates documents related to a specific topic "t" that are referenced by others documents containing the same topic "t". This guaranties a higher level of prestige. The following query attempts to computes a set of prestigious pages that mentions keyword "q". This query is expressed in relational calculus style to illustrate the concept. The implementation uses several primitives, and prestige is actually saved as a procedure or complex primitive in our system.

$$\text{Prestige} = \{t \mid \text{mentions}(Q,q),\ t \in Q,\ \text{References\_to}(R, t, n),\ n=1,\ \text{Intersect}(P,Q,R),\ \text{Length}(P, m),\ m>=1 \}$$

Since, "prestige " is considered as procedure, we could further query the prestigious pages . A simple implemention is:

```
SELECT doc_url FROM prestige('swimming',1)
```

where 'swimming' represents the topic and "1" indicates "direct" reference".

3. RESULTS AND CONCLUSIONS

There is an initial version of the system working under windows. There were several experiments conducted that verify that our system eliminates irrelevant pages that searcher like google [6] missed. We have a system that can be used on top of any search engine to improve retrieval and perform more advanced and detailed search through its query capacity. For additional explanations please refer to [7]. There are still other aspects that should be considered for future research in this area, like quality of the content in relation to prestige. We are now working on adding a quality-checker for the reference, considering two cases. Case 1: there is a known authority or famous page related to the topic, and then its references to another page in the same topic should have a higher weight in prestige. Case 2: an authority is selected by user.

4. REFERENCES

[1] G. Arocena, A. Mendelzon and G. Mihaila. Applications of a Web Query language, Proceedings of the 6th International WWW Conference, Santa Clara, California, April 1997. HTML version

[2] Gustavo Arocena and Alberto Mendelzon. WebOQL: Restructuring Documents, Databases, and the Web, In Proceedings of ICDE, 1998, Orlando, Florida

[3] G. Arocena, WebOQL: Exploiting Document Structure in Web Queries Master's Thesis, University of Toronto, 1997

[4] C. Zhang, W. Meng, Z. Zhang and Z. Wu. WebSSQL - A Query Language for Multimedia Web Documents. IEEE Conference on Advances in Digital Libraries (ADL'00), Washington, D.C., May 2000.

[5] L. Lakshmanan, F. Sadri, and I. Subramanian. A Declarative Language for Querying and Restructuring the Web. In Proc. 6th Int. workshop on research Issues in Data Engineering, New Orleans, 1996.

[6] Google. http://www.google.com.

[7] J. U. Quevedo, S-H S. Huang, and A. G. Covarrubias: Improving Retrieval using Prestige, technical report, Computer Science Department -University of Houston 2002.

[8] Altavista. http://www.altavista.com.

[9] Yahoo. http://www.yahoo.com.

[10] M. Marchiori. The quest for correct information on the Web: hyper search engines. Computer Networks and ISDN Systems, 29(1997) 1225-1235

[11] G. Pringle, L. Allison and D.L. Dowe. What is a tall poppy among Web pages? Computer Networks and ISDN Systmes, 30 (1998) 369-377

[12] An efficient algorithm to rank Web resources Proceedings of the 9th international WWW conference on Computer networks, 2000 Dell Zhang, Yisheng Dong Pags. 449 – 455

[13] J. Carriere, and R. Kazman, WebQuery: Searching and visualizing the Web through connectivity, in: Proc. 6th International World Wide Web Conference, 1997.

[14] Grossman D.A. and Frieder O, "Information Retrieval: Algorithms and Heuristics", Kluwer, August 1998.

[15] Baeza-Yates R. and Ribeiro-Neto B., "Modern Information Retrieval", Addison Wesley 1999.