Published in Procs. Tercera Conferencia de Ingeniería Eléctrica, CIE 97, Vol. I, pp. 263-266, September 1997, Cinvestav-IPN, México, D.F.

An Improved Search Tool for the WWW

Moisés Homero Sánchez López, Aurelio López López
INAOE, Electronics
Luis Enrique Erro No. 1
Tonantzintla, Puebla, 72840 México.
Tel. (22) 472-011 Fax (22) 472-940
e-mail: mhosl@tulum.inaoep.mx, allopez@gisc1.inaoep.mx

Summary: The need to find documents with specific information at a site has been satisfied with the creation of automatic search engines. Although several search engines with different features have been developed, we embarked in the creation of a search engine that brings together the basic search options scattered across several of the already existing public available engines. It is important to remark that the proposed engine was not intended to improve performance or the administration of the computer resources. Instead, it was intended to increase the basic search options, i.e., to provide a more flexible search form that allows users to do an easy search according to their needs.

I. INTRODUCTION.

Nowadays, computers store huge quantities of documents dealing with different information; in the past, when a computer user wanted to find documents dealing with a specific topic, he/she had to scan the content of every document. When a collection of documents is small, manual searches are not as difficult; however, in today's huge collections of documents, the use of automatic search engines is essential [1].

Among the most powerful available automatic search engines that have been developed are FFW [2], Harvest [3], Glimpse [4], and CNIDR [5]. Two of the most complete engines that have been developed using C++ language are HARVEST and FFW. HARVEST is an integrated set of tools to gather, extract, organize, search and replicate information across the Internet, and FFW, a software package that can be used to build free-text search facilities on a World Wide Web server.

Besides HARVEST and FFW, there are others search engines having very complex features and very good performance [6]; however, none of them were found neither having all the desirable basic features nor allowing term-weighted searches, that is, search engines in which the user can quantify the importance of terms. With this in mind, we embarked on creating a search engine that provides more search options (including term weighting) than existing tools.

Improved Search Engine (ISE) is a software package made to incorporate most of the basic search features contemplated in the today's search engines. ISE is meant to cover the search needs for single users who want to be able to provide complex search queries and to change the queries' structure easily.

II. IMPROVED SEARCH ENGINE, ISE.

A. Main Features

ISE is a complete indexing and query system that allows a user to search through the files in a server, and includes several features. The indexing system allows the user to supply a stop list containing words he/she does not want to include in the index, supports ISO-Latin-1 or 8 bit characters, and does not index HTML tags. During indexing, ISE tries to find the summary of each document and saves it if found.

ISE allows the user to do very flexible searches once the index has been built. ISE supports Boolean queries, case sensitive/insensitive searches, and weighted-term searches, and permits the use of synonyms, wild cards, and word stemming. Furthermore, the user has the possibility of choosing between two formulas for weighting forms: The Inverse Document Frequency (IDF) [1], and the Signal Noise Ratio [1]. In the first formula, term importance is proportional to the standard occurrence frequency of each term k in each document i (that is $FREQ_{ik}$) and inversely proportional to the total frequency in all the documents in which the term occurs. On the other hand, the Signal-Noise Ratio takes into account the concentration of a term in a document collection. That is, for perfectly even distributions, when a term occurs in every document of the collection an identical number of times, the noise is maximized. Contrariwise, for perfectly concentrated distributions, when a term appears in only one document, the noise is zero.

Published in Procs. Tercera Conferencia de Ingeniería Eléctrica, CIE 97, Vol. I, pp. 263-266, September 1997, Cinvestav-IPN, México, D.F.

ISE was developed in C++ language in such a way that it is easy to modify or extend; we tried to use Object Oriented Technology at all levels. In addition; ISE works with a WWW gateway to provide access to WWW users.

Because ISE is based on FFW engine, it inherited the main index and search structure of FFW. Like FFW, ISE consists of two principal components: the indexing system and the search system.

B. Indexing System

The indexing system builds an index consisting of three files: main index, pointer file and url file. The main index consists of pairs (word, offset) giving a minimum of data for each indexed word. The search system uses the main index, and once the desired words are found, the offsets from the index file are used as addresses into the pointer file. The offsets point to a list (length, pointer1, weight1, pointer2, weight2,...) in the pointer file; there is one such list for each word. These pointer lists are stored sequentially and can be read quickly.

The pointers are offsets into the url file where the records describing the documents are kept. Each record is a quartet (date, url, title, summary). The structure described above is illustrated in Figure 1.

C. Search System

The search system parses the query, carries out the search and returns the matched documents. There are two kinds of queries accepted by the ISE: Boolean queries and weighted queries. An example of each one is shown below:

Boolean: ([bag] or marsupial) and Australian and animals
Weighted: retrieval .3 information .6 automatic .1

Where square brackets around a word indicate that ISE might search also for the synonyms of such word. When ISE has found the synonyms, it displays a list of different families of synonyms to the user, in such a way that the user may choose the most convenient sense. The synonym table of ISE is an extract of WordNet, a lexical database for English [7].

Once these queries are parsed and used for the search, the output is a list of pairs (document, summary) in decreasing order according to their similarity to the query supplied by the user.

The ISE indexing system (ISEindex) indexes items stored all in the same server. This characteristic makes necessary the installation of ISEindex in each server where an index file is created; however, the index file can be transferred toward the ISE gateway location, where the gateway can easily retrieve documents from remote servers. So, if the user wants to see the whole text of an item, he/she just need to click on the title of such item, and the document will be displayed wherever it is.

III. RESULTS.

ISE is part of an information server that allows several ways of searching besides those provided by the engine.

ISE is already on-line; however, it is being tested with an experimental collection of items in order to determine adequate threshold values for the indexing system [8].

Figures 2, 3, and 4 illustrate some stages of ISE's functionality for a Boolean search. Figure 2 shows the first form presented to the user. On this screen, the query "[number] and last" was submitted; then, Figure 3 shows the synonyms screen where the group number 1 was selected. Finally, Figure 4 shows a screen of the resulting items.

CONCLUSIONS

ISE can be useful for user needs at different levels. While some users have problems in formulating a Boolean query, others find Boolean queries simple to use.

Because of the features of ISE, it can be used for organizing several servers; furthermore, it also can be used to do searches in other languages such as Spanish.

The use of Object Oriented Technology made FFW easy to modify; consequently we thought the ISE would also be useful for future improvements.

ACKNOWLEDGEMENTS

This work was supported by a scholarship granted by CONACYT to Moisés Homero Sánchez López, who appreciates this support.

REFERENCES

[1] SALTON Gerard; J. McGILL Michael., Introduction to Modern Information Retrieval, McGraw-Hill, 1983.

[2] STEIN. L.D. How to set up and maintain a world wide web site, the guide for information providers, Addison Wesley, 1995, 496 p.

[3] Information about FFW can be found at: http://www.nta.no/produkter/ffw/ffw.html

[4] Announcements and documentation about HARVEST can be found at: http://harvest.colorado.edu/harvest

[5] Information about GLIMPSE can be found at: http://glimpse.cs.arizona.edu: 1994/

[6] CNIDR Isite is an integrated Internet publishing software package. Information about it can be found at: http://vinca.cnidr.org/software/Isite/Isite.html

[7] MILLER G. A. WordNet: A Lexical Database for English, Communications of the ACM, Vol. 38, No. 11, Nov, 1995, pp. 39-41.

[8] SANCHEZ L. Moisés. An Improved Search Tool for the WWW, INAOE, Master Sc. in Electronics, Thesis in progress.

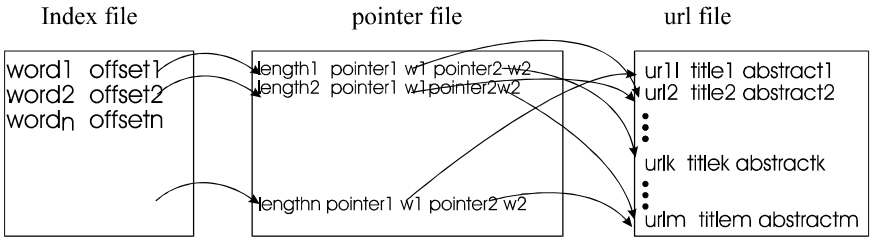

Figure 1. File Structure created by ISE Indexing System.

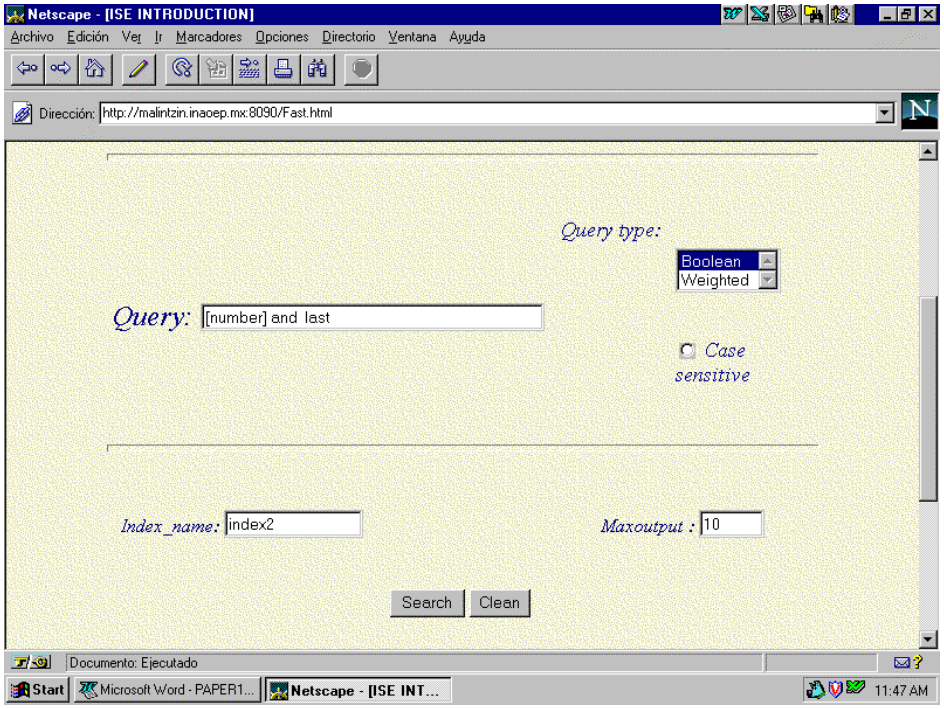

Figure 2. Search Form presented by ISE.

Published in Procs. Tercera Conferencia de Ingeniería Eléctrica, CIE 97, Vol. I, pp. 263-266, September 1997, Cinvestav-IPN, México, D.F.

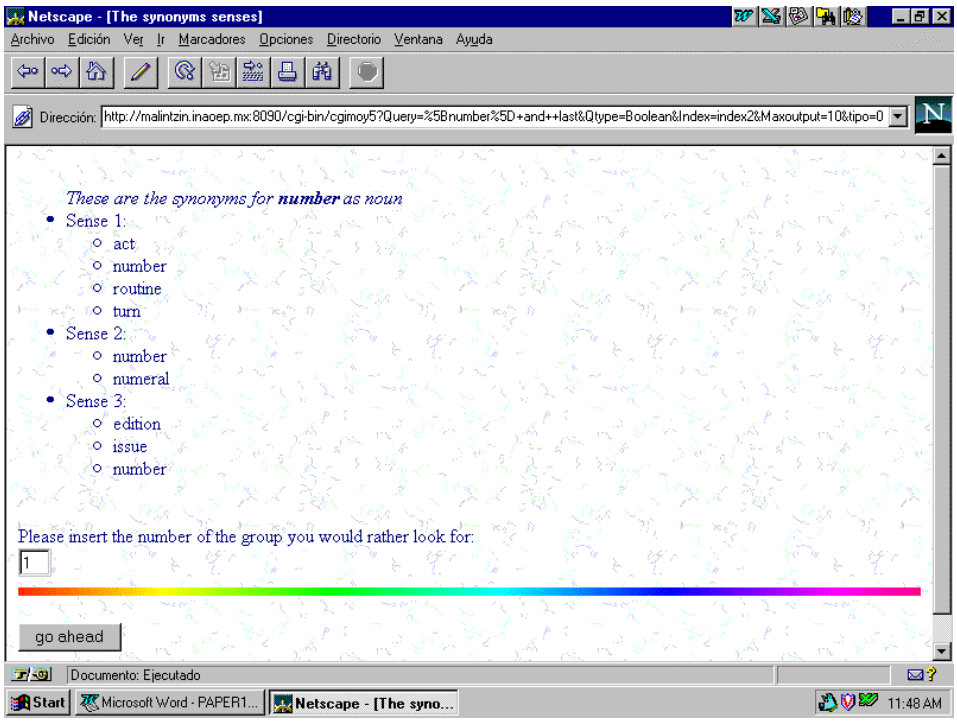

Figure 3. ISE Synonyms Form asking for Selection.

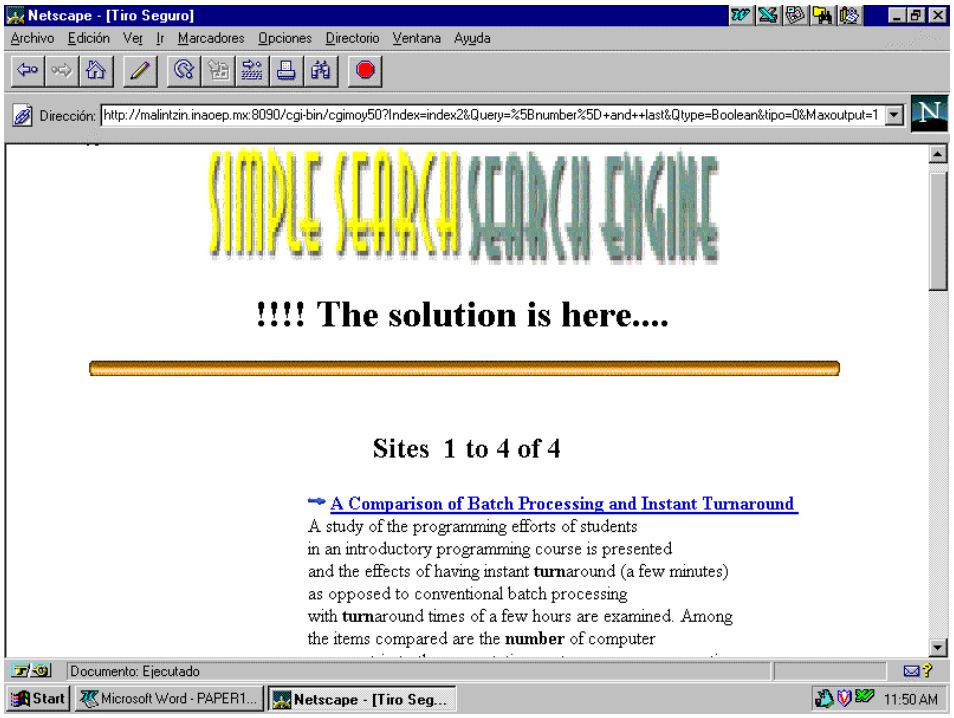

Figure 4. Example of Results produced by ISE.