AN INTELLIGENT QUASI-THESAURUS FOR ACCESSING INFORMATION SYSTEMS

Wieslaw GLINSKI
Warsaw University/E-mail: wglinski@kent.edu

RESUMO

As fontes de informação disponíveis através da Internet são imensas. O volume total de arquivos acessíveis, via Internet, é contado em milhões de *gigabytes*. O principal problema, no entanto, na Internet é a enxurrada de informação e a dificuldade de encontrar itens adequados. Entretanto, a questão em identificar as fontes relevantes e acessá-las é de importância crucial em particular da perspectiva dos usuários casuais. Como norma, o conhecimento específico dos usuários acerca da própria rede, e a distribuição e o conteúdo das fontes de informação, é limitado. Este artigo endereça essa questão à proposta de um método que auxilie os usuários a estabelecer "bons" questionamentos e a submetê-los às fontes de informação (residentes na rede), as quais devem provavelmente conter os documentos relevantes. A idéia central da abordagem proposta é a de criar um mecanismo de auto-aprendizagem, que subsidie o estabelecimento das perguntas e que as encaminhe para o lugar certo dentro da rede. O mecanismo denominado ADRESSER tem sido concebido como um simples quasi-tesauro, o qual é composto por termos, endereços URL relacionados com os termos e as relações ligando os termos. Um protótipo para testar e avaliar a idéia foi implementado. Os experimentos têm provado que a idéia é viável, válida e possível de ser trabalhada.

Palavras chave: Quasi-Tesauro inteligente; Acesso a Sistemas de Informação; Recuperação da Informação.

1. Introduction

At the beginning I would like to offer three observations which prompted the writing of this paper which is the outline of my Ph.D. research conducted in 1995 at the Institute of Library and Information Science of Warsaw University (Poland).

The first observation is the assumption that, it is impossible or at least very difficult to apply oneself to the study of the humanities without knowing the modern tools for information retrieval, processing and presenting the information. The huge amount of information accessible in the modern well developed and democratic countries makes access to it is more and more difficult. The flood of information is as problematic as its lack. In some groups of professionals there is reluctance to use

the modern tools of information retrieval and processing. This reluctance is probably caused by the opinion that using those tools is very difficult and is reserved only for the information professionals and specialists. Of course there is some justification for this, but there is an increasing group of tools which are user friendly. One of my objectives was an attempt to create such a useful tool which could help users in accessing relevant information.

The second observation relates to modern computer networks. The slogan: "the network is a computer" became true, and what is more the opposite saying "the computer is the network"- is now true too. Thanks to the computer networks, and the Internet is the chief example, by means of the ordinary PC equipped with a modem and connected to the computer network one can compute very sophisticated tasks on geographically distributed mainframes. Access to a library which is on another continent can be as easy as access to the local library. We can say that information networks reduced the distances in information reality. The consequences of this fact are tremendous for humanists, technicians, and administration officers. Of course the objective of this paper is not to discuss it. However we can be sure of one thing - the lack of accessibility and/or the difficulty in accessing the proper information limits the professionals for whom the information is the crucial thing.

The final observation relates to the discussion of the new media of information versus traditional ones. In such discussions the computers and computer networks are set against books. Here is what Umberto Echo, Italian writer and semiotician says:

> *"We should learn to use the Internet and the CD-ROM as well as teach our neighbors read books. That is possible. Oncoming times proclaim new duties and new experiences for the educated person. At one time the educated man who could read and write books, but he could have written the books by himself or asked his copyist to do this. Today the educated man is required to know books as well as new forms of writing and processing information. Only by his means one can guarantee that new media will be used in democratic ways, without preventing anyone from accessing information; this is the only way one can teach the others how to choose and evaluate the information received and how to keep alive the books which are necessary to our cultural development."*[1]

1. The quatations is from the lecture "New mass media and the future of the book", by Prof. Umberto Eco PEN Club, 23 Febr. 1996, (translated from Adam Szymanowski).

The same author wrote that we can notice the division of modern societies into three groups: those who have no access to the computers (and also to books) and almost completely dependant on passive audio-visual transmission of information like TV. The second consists of those who know how to use the computer passively (e.g. bank officers, travel agents etc.). The third group consists of those who are using the computer actively, making some analytical work, being able to estimate the value of information, distinguishing between valuable and useless information and being able to use computer and/or information networks.

Easing the access to the computers (which is the network!) which was the main and most general objective of my work can help to increase the size of the third group. Sociologists and researchers of a modern culture appeal to the notion of an *information society*. The most distinctive feature of this society is the constantly growing influence of information and different techniques of processing the information on the functioning of such societies. Today the information is becoming an important and difficult to overestimate factor of developing modern societies in the same way as capital, energy, or natural resources.

But pure information without the inexpensive means of its transmission and access would not play as important a role as it is now in the most advanced technologically societies. We can say that the computer network is the most decisive factor in sending information to any user at any time. Thanks to computer networks a user can access any information even if it is hundreds or thousands miles away.

Computer networks are the most powerful means of sending information. Therefore, they are building new and faster ones, the computer networks become more and more complex, they are merging with other computer networks and going out of the communities in which they were originally created. The best example of this is of course the Internet.

The main problem from the user point of view, created by the growing amount of computer networks and extreme increase of information resources and of information itself is the lack of an easy ability of surfing the computer network, which means an ability of accessing and finding the needed information.

The information resources available through the Internet are immense. The total volume of the files accessible via the Internet is counted in thousands of gigabytes.

In this flood of information it is very difficult to find what one seeks

Another problem is that not all of the information on the Internet is of equal quality or value. This is why, in spite of the tremendous capaci-

ties of the Internet, the skills and knowledge of its users will still remain a decisive factor.

The issue of identifying the relevant resources and accessing them is of crucial importance, in particular from casual users' perspective. As a rule their specific knowledge about the network itself, and the distribution and contents of the information sources is limited. This paper addresses this issue by proposing a method helping the users to establish "good" queries and submitting them to the information resources (residing on the network) which are likely to contain the relevant documents.

Section 2 presents the underlying concept of the front-end processor devised for facilitating the process of building and forwarding the queries. Preparation of a query for searching, which is the main operation executed by the front-end processor, is described in Section 3. The principle of *ADDRESSER*'s functioning is in Section 4. The last Section focuses on the follow-up activities regarding the development of the front-end processor.

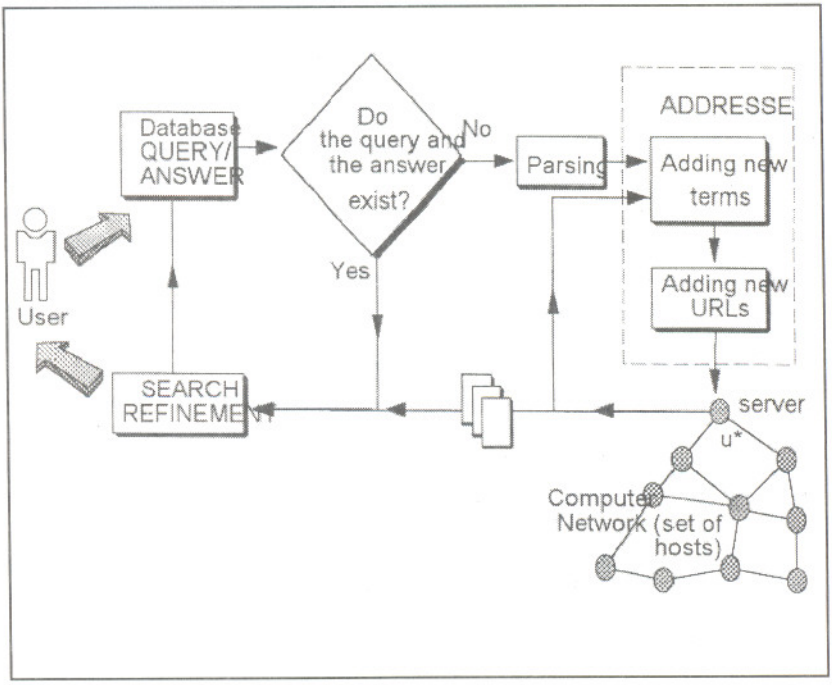

Figura. 1 - Conceptual architecture of NetExp

2. The Concept

As mentioned in the previous Section the problem of formulating an appropriate query and launching it against the Internet resources is not an easy task even to an experienced information officer, not to mention casual users. Besides technical obstacles connected to the use of the Internet client type tools such as Gopher or WWW, there are two significant reasons why the process of establishing good queries is difficult. The first reason is related to the fact that the user has to use some key-words for expressing his/her needs whatever the query language. At this point the user is never sure whether the terms used for setting up the query are actually those which are well "understandable" by the network.

Secondly, the user in general does not know where the information resources are storing the documents which are likely to be relevant to the key-words. To summarize: when sitting in front of the Internet one can hardly formulate a "sharp" query and readily locate the pertinent databases.

The core idea of the proposed approach facilitating the process of establishing queries and efficiently accessing the information resources distributed across the Internet is to create a self-learning mechanism supporting the queries' establishment and forwarding them to the right places within the network. The mechanism, called *ADDRESSER*, has been conceived as a simple quasi-thesaurus which is composed of terms, URL addresses related to the terms and relations linking the terms. We assume that each Internet data server has its *signature* which is simply the sets of key-words characterising the information resources and/or documents placed on the server. The *ADDRESSER* is the main component of the front-end processor, called *NetExp* , whose structure is depicted in Fig.1.

Roughly speaking, the process of preparing and launching the query looks as follows (see Fig.1):

1.A Boolean query is set up by the user. The user can either use key-words or pick up the terms from the quasi thesaurus.

2.The query is compared with the entries in the so-called QUERY database where all the queries which have already been submitted have been registered along with the answers. Should the input query match one of the items kept in the Query database, the answer is picked up and immediately delivered to the user without accessing the Internet. Now, it is up to the user whether the query should be addressed to the Internet, hoping that additional (updated) information might be found, or not.

3.If there is nothing like the input question in the QUERY database, it is parsed and transferred to the ADDRESSER for semantic tuning and determining the set of URL addresses of the resources where the relevant information is expected to be available.

4.The servers indicated by the set of URL addresses are accessed from the user's server (marked as u*), and search is being done by the standard Internet tools. The documents found are moved back to NetExp, and after some refinement (e.g. formatting, editing) transferred to the user.

5.Having identified the pertinent servers on the Internet, any information in form of terms on the contents of the files residing on those servers is picked up and sent back to the ADDRESSER and added to the quasi thesaurus. This information is taken from the servers' signatures. This process can be considered as teaching the thesaurus.

More on the whole process and the modules depicted in Fig.1 will be said in Section 2.

3. Preparation for Searching

The prototype of the front-end processor was implemented by means of the Toolbook ver.3.0 software. In this paper we shall not discuss the *NetExp* architecture in detail. We shall rather focus on how the preparation to the search process is performed by the *NetExp* and how the quasi thesaurus can learn from the network.

The user can formulate a question as a Boolean formula. Let us take an example. The query is:

POLAND AND INFORMATION SCIENCE

The query is made up of two terms:

term 1: POLAND
term 2: INFORMATION SCIENCE

Now, the process goes to the ADDRESSER. If the term which occurs in the query exists in the ADDRESSER, all the URLs related to this term are taken into account. If the NetExp cannot find any URL related to the term, it tries to find all synonyms of this term. Should the synonyms not exist or the URLs are missing, the broader terms are looked for. Again, if these terms cannot be found or URLs are not specified, the system seeks narrower terms. For instance, if the URLs for POLAND do not exist in the quasi-thesaurus, the synonyms are considered.

synonyms:
RZECZYPOSPOLITA POLSKA
POLSKA RZECZYPOSPOLITA LUDOWA
PRL

If there is no any URL for those synonyms, *NetExp* looks for URLs of broader and narrower terms, which in our experimental quasi thesaurus were as follows:

broader terms:
EASTERN EUROPE
WARSAW PACT

narrower terms:
MAZOVIA
SILESIA
WARSAW

Thus, the following terms along with corresponding URL addresses were found:

TERM TEMP	ADRES TEMP
EASTERN EUROPE	http://148.81.213.3/info4.htm
WARSAW PAC	Thttp://148.81.213.3/info2.htm
SILESIA	http://148.81.213.3/info2.htm
WARSAW	http://148.81.213.3/info2.htm
MAZOVIA	http://148.81.213.3/info4.htm

For the term *INFORMATION SCIENCE* the system found 4 URL addresses:

http://148.81.213.3/info2.htm
http://148.81.213.3/info4.htm
http://148.81.213.3/info5.htm
http://148.81.213.5/info6.htm

Since our input query was formulated as a conjunction we get as a result the following table:

http://148.81.213.3/info2.htm
http://148.81.213.3/info4.htm

Now, the front-end processor can access the servers/files specified in the above table and get the relevant documents.

4. Intelligent quasi-thesaurus (*addresser*)

As has been emphasised the intelligent quasi-thesaurus is a core of the *NetExp*. It has the ability for accumulating knowledge about the network. Now, let us define the intelligent quasi-thesaurus in a more formal manner. By SYG(u) we understand the *signature* of the server *u*, which is a set of all the terms indexing the files sitting on this server. Now, we can define the *ADDRESSER* as an ordered quadruple:

ADR=(T, A, {B,N,S}, t)

where:

T - is a set of terms such that $\mathbf{T} \neq \emptyset$;

A - is a set of URL addresses

{B,N,S} -sets of broader and narrower terms, and synonyms, respectively, in **T**

τ - is a relation $\tau \subseteq \mathbf{T} \times \mathbf{2^A}$ such that for the server *u* with which URL addresses **A´** are assigned the following condition holds

$$\forall\, t \in T\ \forall A' \subseteq A\ \forall a \in A'\ (t \tau A' \rightarrow t \in SYG(u))$$

As a spin effect of the searching process, signatures of all the accessed servers are brought back to the front-end processor where the terms are added to the quasi thesaurus along with their URL addresses. The action of incorporating the terms into the quasi thesaurus and establishing the relations between them and the existing terms is assumed to be done by a person who is administering the *NetExp*. This is how the quasi *ADDRESSER* accumulates, or in other words learns, knowledge about the resources available on the network. Note that this process is by its nature monotonic since the knowledge can only increase or remains the same after any search (transaction) within the network. Fig.2 shows the learning curve of the quasi thesaurus; it also displays the fact that the knowledge about the network is finite.

It has to be stressed that before incorporating the quasi thesaurus to the front-end processor in question, a critical mass of knowledge has to be injected into the quasi-thesaurus by the front-end processor administrator. One has to accept that it is impossible to start any kind of searching and learning without certain prior knowledge.

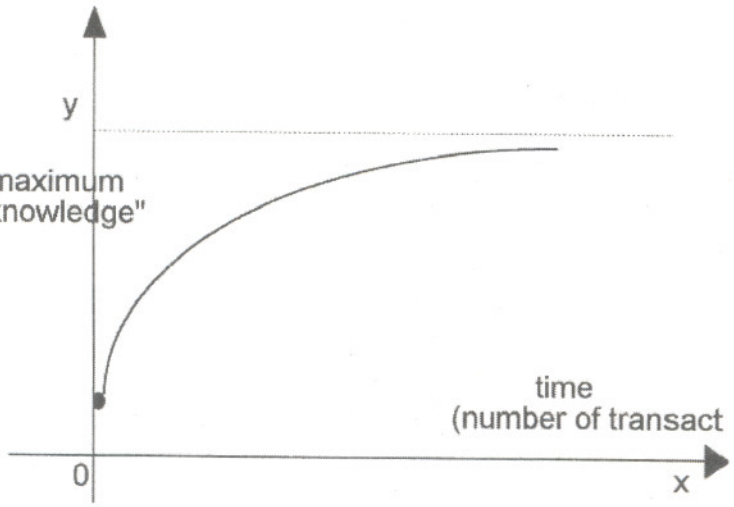

Figura. 2 - Monotonic accumulation of knowledge by the quasi-thesaurus

When designing the *ADDRESSER* we have encountered an interesting quantitative problem related to the approximation of original terms as given in the query submitted by the user. In other words, the issue is: how good is the approximation proposed by the quasi thesaurus. We have tackled this problem in the following manner. Let us denote the accuracy coefficient related to the term *t* as Q_t. For Q_t , the following assumption were adopted:

1. $0 \leq \Theta_t \leq 1$

2. *l* is a number of all URL found of broader and narrower term (related to term *t*)

$$\lim_{l \to \infty} \Theta_t = 1$$

3. $\Theta_t \mid_{t=0} = 0$

At the beginning of our experiment we decided to use the following function:

$$\Theta_t = \frac{\sqrt{(l-1)^2 - 1}}{l-1}$$

However, during the experiment it turned out that the function was not a good one, so we devised another function (see Fig.3):

$$f(i) = 1 - [(1-\varepsilon)e^{\frac{1-i}{T}}]$$

where i is the number of URLs and the parameters take the following values:

ε **-0.1,** it means that for i=1, f(i)=0.123.

T-100 , this factor is responsible for the function growth pace rate.

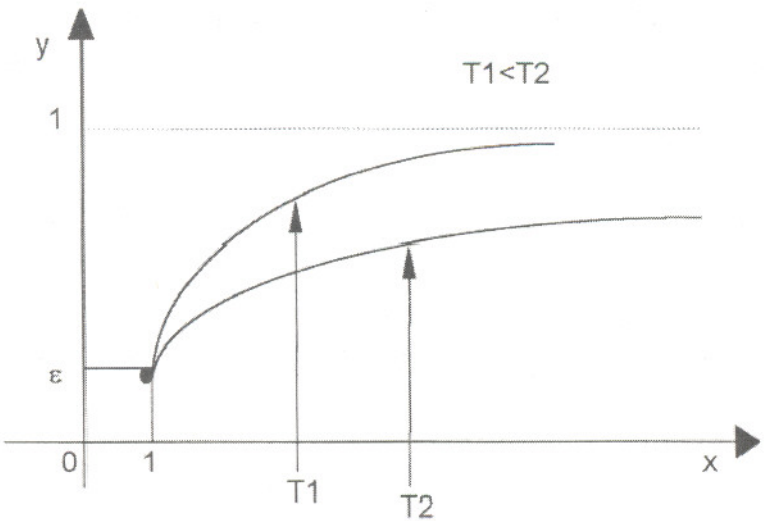

Figura. 3 - Function $f(i) = 1 - [(1-\varepsilon)e^{\frac{1-i}{T}}]$

Thus, the function looks like

$$f(i) = 1 - [(1-0.1)e^{\frac{1-i}{100}}]$$

The coefficient of accuracy for the whole query composed of n terms is defined as follows:

$$\Theta_w = \frac{\sum_{i=1}^{n} \Theta_{t_i}}{n}$$

This measure proved to be a fair tool for estimating the approximation of original queries during our experiments with the front-end processor.

5. Final Remarks

The paper constitutes an attempt to facilitate the process of interacting with large, geographically distributed, heterogeneous networks such as the Internet. In particular, in this context formulating queries is an error prone and difficult exercise. The work on the front-end processor, which has been implemented as an experimental project, proved its usefulness for students and casual users. It is believed it may also be useful to the information professionals.

Two issues deserve further elaboration, namely:

·tuning of the accuracy function which is used when some terms provided in an input query are not covered by the quasi-thesaurus;

·how large should the prior knowledge be to be given to the quasi-thesaurus before starting its operation, and how the amount of "critical mass" of terms depends on the subject matter covered by the front-end processor.

Another important aspect is whether it is better to devise a general quasi thesaurus encompassing many domains or to maintain an array of narrow discipline oriented thesauri governed by a kind of meta-thesaurus. These problems are addressed by the follow-up project that has just been started.

6. References

ENGST A. C, LOW, C. S., SIMON M. A. **Internet Starter Kit**. Hayden Books, 1995.

LIU C., PEEK J., JONES R., BUUS B., NYE A. **Managing INTERNET Information Services**. O'Reilly & Associates, Inc., 1994.

MATHIESEN M. **Marketing on the Internet**. Maximum Press, 1995.

ABSTRACT*

The information resources available through the Internet are immense. The total volume of the files accessible via the Internet is counted in thousands of gigabytes. The main problem however on the Internet is the flood of information and the difficulty of simply finding items. Therefore, the issue of identifying the relevant resources and accessing them is of crucial importance, in particular from the casual users' perspective. As a rule users' specific knowledge about the network itself, and the distribution and contents of the information sources is limited. This paper addresses this issue by proposing a method helping the users to establish "good" queries and submitting them to the information resources (residing on the network) which are likely to contain the relevant documents. The core idea of the proposed approach is to create a self-learning mechanism supporting the queries' establishment and forwarding them to the right places within the network. The mechanism, called *ADDRESSER*, has been conceived as a simple quasi-thesaurus which is composed of terms, URL addresses related to the terms and relations linking the terms. A prototype for testing and evaluating the idea was implemented. The experiments have proved the idea is viable, sound and workable.

Key Words: Intelligent Quasi-Thesaurus; Acessing Information Systems; Information retrieval.

* The work reported in this paper has been jointly supported by: grant nº 8 T11C 038 of the State Committee for Scientiic Research (KBN), Poland, and by grant nº 1146/94/TM of the Foundation of the Polish-German Cooperation, by the grant of the European Community (TEMPUS - Phare - Individual Mobility Grant nº IMP-95-PL-2136), and by Polish-American Fulbright Commision.