

AN INTELLIGENT GIS SEARCH ENGINE TO RETRIEVE INFORMATION FROM INTERNET

Z. Liu, Y. Gao

The University of Calgary

Geomatics Engineering, 2500 University Drive NW., Calgary, AB, CANADA, T2N 1N4.

Fax: 403-2841980; Tel: 403-2208230, 403-2206174

Email: ZHELIU@UCALGARY.CA, GAO@GEOMATICS.UCALGARY.CA

Keywords: Intelligent GIS Search Engine, Information Retrieval, Fuzzy Logic, GIS Thesauri

BIOGRAPH

Zhe Liu is currently a M.Sc. student in Geomatics Engineering at the University of Calgary. He received a ME and a BE from Beijing University of Aeronautics and Astronautics. He currently conducts research on GPS/GIS/Wireless Integration. Dr. Yang Gao is an Assistant Professor whose research focuses on satellite navigation, wireless mobile information management systems and Geographic Information System.

ABSTRACT

There are a lot of search engines available to retrieve GIS related information on Internet, such as the general search engines like Yahoo, AlltheWeb, or the specially designed search engines for GIS Community like GeoCommunity. It is general for these current search engines to provide irrelevant results and miss relevant results when they are used to retrieve GIS related information. No-Name, an intelligent search engine designed by author, is developed intentionally to overcome the disadvantages and inability of current search engines to serve GIS community. No-Name builds a reasonable size keywords database by carefully choosing about 2000 GIS terms to construct a GIS thesauri. Since the terms are usually relevant to each other, like a document represented by "Oracle" and "SQL Server" has an uncertain relevance to a query represented by "Database", the term connection value that represents the term relevance degree can be determined by fuzzy logic. No-Name ranks the websites and documents according to both the frequency of keywords occurring in the text including the heads, titles and bodies with different weights and the relevance level

between the keywords according to the term connection strength. No-Name supports two independent layers, named term layer and location layer. The term layer has a tree-like structure with about 2000 GIS terms to construct GIS thesauri. The location layer has a tree-like structure with about 6 continents and 100 countries that are considered active in GIS field. Moreover, No-Name has limited power to support nature language input and Linguistic Analysis. Several experiments were carried out to assess the performance of No-Name and some famous current search engines. The test results show that No-Name is better than current search engines in three respects: Relevant Sites Retrieval, Irrelevant Sites Dodge, and Natural Language Identification. The test results also show that No-Name can revise the term connection values by learning from users' feedback.

INTRODUCTION

The Internet is an extremely vast source of information and its contents are growing explosively. Geographic Information Systems (GIS), which deals with the management of geospatial data, has become one of the most dynamic area recently. The GIS information is now experiencing an explosive growth on the Internet. Today, it has become extremely difficult for GIS professionals to select qualitative data of the required relevance in a reasonable amount of time. To save the GIS users from the mess information on Internet, search engines with high performance are needed.

Search engine is one of the most essential tools on the Internet. They help you find web sites relating to a particular subject or the email

address of someone you know or articles posted to a newsgroup or even companies which have a presence on the Internet. The search engines are basically huge databases containing millions of records that include the URL of a particular web page along with information relating to the content of the web page which is supplied in the HTML by the author. The search engine obtains this information via a submission from the author or by the search engines searching the Internet for information [28].

Any search engine will fall into one of two categories: for general purpose or for a special community. A general search engine is designed for general populations. Their service covers a wide variety of subjects, and it is the reason why most of current search engines are now used for general purpose. Yahoo and AlltheWeb are good examples of General purpose search engine. However, with width they can not go depth further. On the contrary, GIS search engine is one application of the specially designed search engines for one community. It focuses its service on providing exhaustive GIS information, and its customers are mainly from GIS community. Example for GIS search engine is Geo-Community, founded by GeoComm International Corporation in 1995.

Generally speaking, GIS search engine can provide better service than General purpose search engine for GIS community. By constructing a more compact structure than General purpose search engine, The GIS search engine can retrieve more complete, relevant and balanced GIS information with much lower cost. The work is easier for GIS search engine to increase its performance by retrieving relevant information and dodging irrelevant information than that for General purpose search engine.

However, current GIS search engines like Geo-Community can improve their performance further by applying fuzzy logic and text analysis technology. Fuzzy logic can be used to solve the relevance degree between terms. Current information retrieval method like Bayesian network model [1] tends to assume the terms are atom-like and independent, but the independence assumption is not realistic: Document cannot be represented by a set of independent terms. Terms are inter-independent in most application areas [2]. For example, a document represented by "Oracle" and "SQL Server" is usually relevant to a query represented by "Database" to some

extent. The thesaurus *Wordnet* [3] simplify the term relevance relations by dividing them into several types. Table 1 shows the types of relations. In fact, the relevance degree of every term pair is different from each other and the simplification will impair the relationship between the query and documents. For a small size thesaurus, like a GIS thesaurus containing about 2000 keywords, it is possible to construct a term relevance factor table to quantify the relevance degree of every term-pair. In practice, the term relevance factor table can be first set by experts in GIS fields and then adjusted by the feedback from users.

Relation	Example
Synonymy	Computer – data processor
Antonymy	Big – small
Hyponymy	Tree - maple
Hypernymy	Maple – tree
Meronymy	Computer - processor
Holonymy	Processor - computer

Table 1: Some relations offered by Wordnet

Most of current search engines classify the website by analyzing the heads and titles. Text analysis technology can be used to analyze the whole text of a web page. The concept of text analysis technology was first introduced by IBM in 1999 and was applied in IBM Text Search Engine. By recording the frequency of keywords occurring in the text and rendering weight when the keywords occurring at the heads and titles, search engine applying text analysis technology can rank the website according to the correlation level to the keywords more sophisticatedly and reliably. By integrating text analysis technology and fuzzy logic, the accuracy and integrity of search engine are enhanced and can supply high-quality information retrieval.

This paper is organized as follows: First the information collection methods of No-Name are introduced. Second the basic of information

retrieval, both traditional and recent approaches, are reviewed briefly. Then previous work on thesaurus-based information retrieval, a new method using human-defined knowledge, is described. An implementation of this new method on GIS thesauri is described to show the information retrieval strategies of No-Name, GIS intelligent search engine. Then the unique structure of No-Name to support multiple independent layers is shown and analyzed. Finally, concluding remarks are given.

SPIDER AND ROBOT: COLLECTING INFORMATION ON INTERNET

No-Name is an intelligent GIS search engine to retrieve information from the Internet developed by the author. No-Name, a funny name, comes from misunderstanding between the author and a audience. When the search engine prototype was first demonstrated in 1999, one audience asked a question of "What is the name of it?". The author answered with "No name". Although the message that the author wanted to convey is "It has not gotten a name yet", the audience took "No-Name" as the name of the search engine and the new name was getting popular soon.

No-Name employs spider technology and robot technology to search, analyze and collect GIS related information on Internet. A spider is a program operated by a search engine that surfs the web automatically. As it visits each web site, it records all the words on each site and notes each link to other sites. It then "clicks" on each link and off it goes to read and record another web site. No-Name employs multiple threads to create spiders, so different spiders can search the web parallel at the same time. The web sites of famous enterprises, academics, organizations and technology forums with activities in GIS field are chosen as the start-off points for spiders since there are plenty of links at their web sites. When a spider finds more than one link at a web site, more spiders will be generated to follow each link. If a spider finds that a website contains no GIS information, the spider will kill itself and the search process will stop there. Since the spiders will consume the computer resources, the search engine has to control the number of spiders, or threads, to avoid resource exhaustion.

A robot is a program operated by a search engine that can address other search engines as a user. The current search engines available, both general search engine and GIS search engine,

can response to robot's query represented by GIS terms. However, the results from current search engines are usually considered rough and can not be used directly by No-Name. The rough results will be purified by means of text analysis technology.

Both spiders and robots adopt text analysis technology to analyze and extract the useful keywords from the texts and documents at web sites. The concept of text analysis technology was first introduced by IBM in 1999 and was applied in IBM Text Search Engine. Text analysis technology can be used to analyze the whole text of a web page. By recording the frequency of keywords occurring in the text and rendering weight when the keywords occurring at the heads and titles, search engine applying text analysis technology can extract the feature from a web site and save it to database. The feature information will be used to rank the relevance level of web sites to the keywords by integrating term connection value information. Two terms are considered to have relation with each other if the documents represented by one term are usually considered to have relation with a query represented by the other term. Term connection value is a fuzzy logic approach to determine the connection strength between terms. The concepts and methods to determine term connection value are described in the following sections.

BASIC OF INFORMATION RETRIEVAL

The goal of an Information Retrieval (IR) system is to select the documents relevant to a given information need out of a document database. The present information explosion increases the importance of this area. It is difficult to find out relevant information from a huge information mass like Internet [2].

Traditional approaches to IR, which are popular in current search engines, use direct keyword matching between documents and query representations in order to select relevant documents. The most critical point goes as follows: if a document is described by a keyword different from those given in a query, then the document cannot be selected although it may be highly related. This situation often occurs in real cases as documents are written and sought by different persons [2].

In recent work, there is common agreement that more adequate relevance estimation should be based on inference rather than direct keyword matching [[4], [5], [6], [7]]. That is, the relevance relationship between a document and a query should be inferred using available knowledge. This inference, however, cannot be performed with complete certainty as in classical logic due to the uncertainty inherent in the concept of relevance: one often cannot determine with complete certainty if a document is relevant or not. In IR, uncertainty is always associated to the inference process [2].

In order to deal with this uncertainty, probability theory has been a commonly used tool in IR [[8], [9], [10], [11], [12]]. Probabilistic models usually attempt to determine the relationship between a document and a query through a set of terms that are considered as features. Within the strict probabilistic framework, inferential approaches are often confined to using only statistical relations among terms. The main method adopted by probability theory to determine the relevance degree among terms is by considering term co-occurrences in the document collection [13]. In this case, two terms which often co-occur are considered strongly related. The problem stands out in this method because relations obtained from statistics may be very different from the genuine relations: truly connected terms may be overlooked [14] whereas truly independent terms may be put in relation [15].

A new method using human-defined knowledge like a thesauri to establish the relationship among terms is now getting popular in IR. With the recent development of large thesauri (for example, Wordnet [3]), these relations have quite a good coverage of application areas. A manual thesaurus is then a valuable source of knowledge for IR. However, due to the lack of strict quantitative values of such relations in thesauri, the quantitative values have to be determined by user relevance feedback or expert training.

PREVIOUS WORK ON THESAURUS-BASED INFORMATION RETRIEVAL

Thesauri used in IR may be divided into two categories according to their construction: automatically or manually constructed. The former are usually based on statistics on word (co-)occurrences. While this kind of thesaurus

may help users to some extent, their utilization in early systems shows that their impact on the global effectiveness is limited [16]. The main reason is that real relations (e.g. synonymy) can hardly be identified statistically. In fact, words very similar in meaning tend to repulse from each other in continuous portions of text [14]. For example, "document retrieval", "text retrieval" and "information retrieval" are rarely used simultaneously.

Recent work pays more and more attention to manually constructed thesauri [[17], [18]]. Initiated by Rada [19], a great deal of efforts have been spent in defining IR suited semantic networks based on manually constructed thesauri [[20], [21], [22], [23]]. Metric over semantic networks is determined by measuring the similarity between two terms mainly according to the topography of the thesaurus (the number and length of links). Two problems may occur in these systems. First, the estimation of the strength of term connections which is based heavily (if not only) on the use of thesaurus topography may fail to reflect the real strength of the connections. This strength also depends on the nature of the relations between them which affects their relevance to some application area. Second, the metrics used to measure term connection are often symmetric: for a metric m , we have $m(a,b) = m(b,a)$ for any pair of terms a and b . This property is obviously counterintuitive. For example, a document about object-oriented languages should be more relevant to a query on programming languages than in the reverse situation.

A new method was recently proposed to revise the strength of term connections according to users' feedback using fuzzy logic [[2],[24]]. The core idea can be represented as follows. Assuming one term is relevant to another if a document represented by the first term is relevant to the query represented by the second term alone. The term relevance can be represented with a fuzzy implication relation such as $a \supset_{\mathbf{b}} b$ where $\mathbf{b} \in [0,1]$. In this way, the entire thesaurus may be represented as a set of fuzzy term relevance relations:

$$\{ \{a \supset_{\mathbf{b}} b\}, \dots \}$$

The key problem lies in the estimation of term relevance strength \mathbf{b} given a thesaurus relation between two terms according to the users' judge on if the documents represented by a is related to query represented by b only.

The principle goes as follows. The system gives a tentative query evaluation and provides an answer (a set of ordered documents). Then the user is required to give his or her own relevance evaluation of the retrieved documents. The user's evaluation is used by the system to revise the strength of term relevance relation in order to better fit the user's evaluation.

This new method is right in concept but is hard to be realized on a large size thesauri like WordNet, which contains about words and phrases. To determine the strength of each term connection, in other word, a node of the semantic networks which contains enumerable nodes, needs infinite users' feedback, which will rapidly increase the cost and time. Moreover, biased users' feedback, for example, users' feedback focuses on some nodes heavily but scares on other nodes, will deteriorate the final results. In practice, the full thesauri is divided into a few groups, a group of thesaurus relations are adopted instead of individual relations among terms [2]. However, the results will be definitely deteriorated since the group classification is very coarse and can not represent the true relations among terms.

Although this new method, to revise the strength of term connections according to users' feedback using fuzzy logic, is hard or impossible to be realized in a large size thesauri, it is very possible to be realized in a small size thesauri. The following section describes how this method is improved and realized on a GIS thesauri, which contains about 2000 terms.

INFORMATION RETRIEVAL BASED ON GIS THESAURUS

The tree-like GIS thesaurus is constructed by adding a few hundred GIS terms into an online GIS dictionary [25]. With the fast advancement of GIS technology, the GIS terms of the GIS thesaurus have to be updated frequently, even the structure of the GIS thesaurus might need to be renewed in a few years. To enhance the reliability and provide an adequate service, any

update on the GIS thesaurus should be approved by the GIS expert board.

The GIS thesaurus prototype adopted in No-Name, the GIS intelligent search engine, are composed of about 2000 GIS terms and further divided into 16 categories. Each category contains about a few tens to a few hundreds terms. Figure 1 shows the structure of the GIS thesaurus prototype.

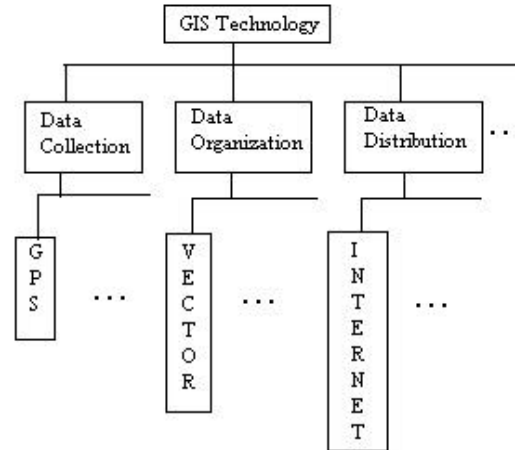


Figure 1: GIS thesaurus structure

The possible combination for 2000 terms can be as much as $2000 * 2000 = 4,000,000$. Notice $\mathbf{b}(a,b)$ is different from $\mathbf{b}(b,a)$ where a, b represent any two keywords. $\mathbf{b}(a,b)$ stands for term connection value from a to b. Two assumptions are made to simplify the computation for term connection strength.

Assumption1: $\mathbf{b}(A_0, A_i) = 0$; $\mathbf{b}(A_i, A_{i,j}) = 0$

where A_0 stands for "GIS technology", A_i ($i=1, \dots, 16$) stands for the categories like "Data Collection" and "Data organization" (notice the categories are also terms), and $A_{i,j}$ ($j=1, \dots$) stands for the terms under category i .

Assumption2: $\mathbf{b}(A_i, A_j) = 0$ ($i \neq j$); $\mathbf{b}(A_i, A_j) = 1$ ($i = j$); $\mathbf{b}(A_{i,j}, A_{p,q}) = 0$ ($i \neq p$ or $j \neq q$); $\mathbf{b}(A_{i,j}, A_{p,q}) = 1$ ($i = p$ or $j = q$).

The first assumption shows that a document represented only by term "Data collecting" has no relation with a query represented only by "GPS", but a document represented only by term "GPS" relates to a query represented only by "Data collecting" to some degree. The second assumption shows that a document represented

only by term “GPS” has no relation to a query represented only by “Digitizing” or “Vector” and vice versa.

The number of term connection with unknown values is compressed from 4,000,000 to around 2,000 under these two assumptions. It is now a reasonable work for the expert board to set a prior value for each term connection and it can be realized to modify the prior values for term connection by users’ feedback.

More specifically, for a given term a , set B contains all terms with unknown connection value to a . $B = \{b_i\}, (i=1, \dots)$. Assume a document Γ is one of the query results of a , and Γ contains a term set C . $C = \{c_j\}, (j=1, \dots)$. Set D is the intersection of B and C . $D = \{d_k\}, (k=1, \dots)$. b_k represents the corresponding relevance relation of d_k and a ,

$m(d_k, a)$. b_k is determined as follows:

1. When a user input a query that contains only term a , a list of documents including Γ are given as query results. The decision on the degree how the document Γ is related to the query is based on b_k with prior values .

2. The user examine the document Γ and indicates whether it is relevant to a or not.

3. Then the value b_k is modified to either

$$b_k' = \min[1, b_k * (1 + e)]$$

or

$$b_k'' = \max[0, b_k * (1 - e)]$$

where $e \in [0, 1]$ is the change scale. The query is evaluated again with b_k' and b_k'' .

4. The fuzzy value for this relation is adjusted to the value which leads to the best answer.

The approach used to adjust the strength of a relevance relation is simple and efficient. However, the performance is possible to be further improved if the users' feedback can be classified. For example, the experts' feedback should have more weight than those of amateur users'. It is an interesting topic to identify the experts' feedback from the amateur users', or more practically, to divide the feedback into groups, then identify one group containing more feedback from experts than other groups. The analysis of the feedback shows that the feedback from experts has some different characteristics from that from the amateur users on statistics. The results might be used in No-Name in the future.

MULTIPLE INDEPENDENT LAYERS STRUCTURE

Most search engines provide one textbox to users. Users input all the keywords and organize them with some simple binary combination like "And" and "Or". Some search engines divides the keywords into several overlapped categories like Yahoo. Some search engines has begun to divide the keyword into independent categories. However, the importance for search engines to divide keywords into independent categories has not been fully recognized. In fact, by dividing keywords into independent categories, or called multiple independent layers, the users can operate a query easier and get better results close to the desire. Moreover, the query speed will be improved.

The rule to divide keywords into multiple independent layers is described as follows:

Given A is a set of terms, B and C are subset of A, and $A = B \cup C$. $B = \{b_j\}$

($j=1, \dots$). $C = \{c_k\}$ ($k=1, \dots$). If for any j

and k , $b(b_j, c_k) = 0$ and

$b(c_k, b_j) = 0$, then A can be divided into two independent layer B and C.

No-Name currently support two independent layer: Term layer and Location layer (Figure 2). Term layer has a tree-like structure containing about 2,000 terms. Location layer also has a tree-like structure containing 7 continents and about 200 countries that are involved into GIS. The tree-like structure of term layer and location layer makes them easy to expand and be organized. Term layer and Location layer meet the requirement to be independent. The terms in term layer has no connection strength with the terms in location layer, and it accords to the intuition: A document represented by term "USA" has no "cause and effect" relation with a query represented by "GIS", and vice versa.

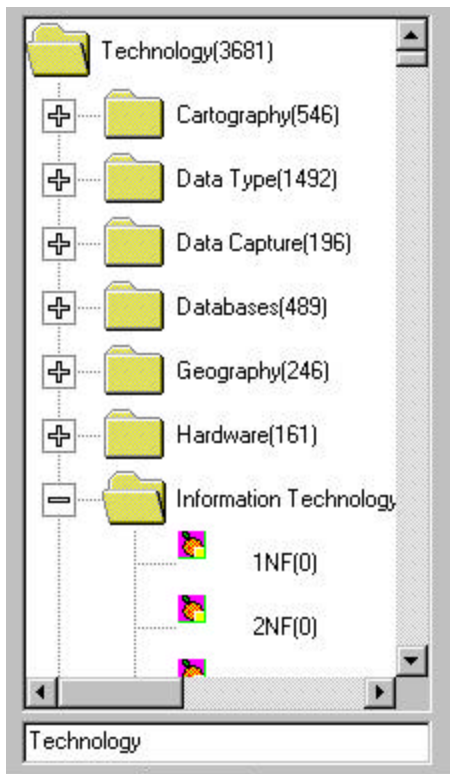


Figure 2 : Term Layer

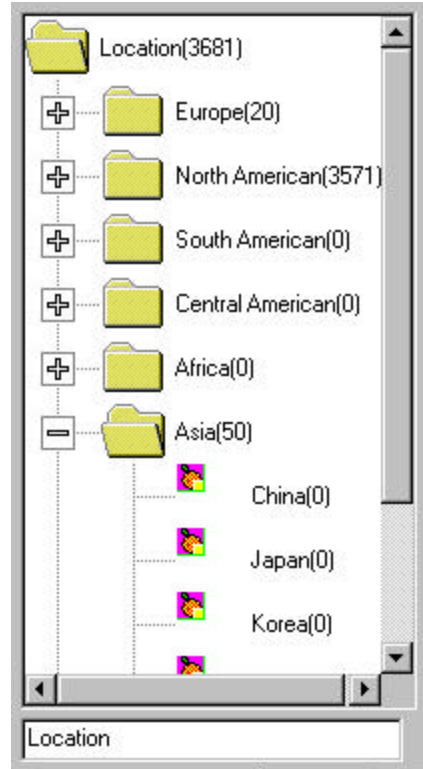


Figure 3 : Location Layer

The independent layers structure can help the users to query useful results. If a user wants to find out all documents about GIS activities in China, he/ she can choose "China" in location layer and perform the query. The result set will contain the all the address of web sites that are involved into the GIS activities in China. If a user wants to know the activities and advancement about data capture technology in China, he/she can choose "Data Capture" in term layer, and perform the query in former result set. Users can also get a rough sketch about the knowledge distribution in countries, or even cities in the future, or in a special technology field. The information is very useful for researchers.

The work to add a new independent layer, Application layer, is now undergoing (Figure 3). Application layer contains about a few hundred existing or potential applications for GIS technology, such as Government, Education, Business, Public Service etc. Application layer has a tree-like structure, and the terms are organized mainly according to careers.

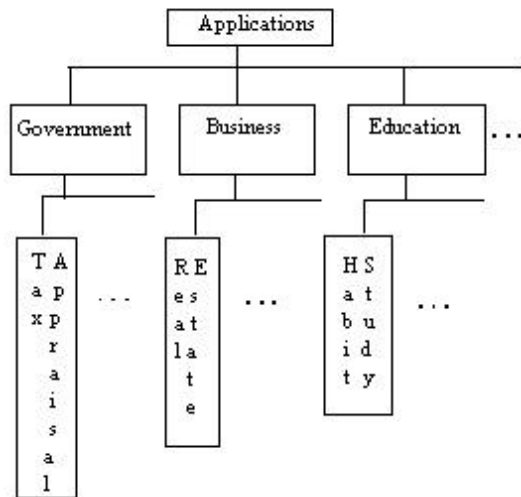


Figure 3 : Application Layer

No-Name has limited power to support Natural Language input and Linguistic Analysis. No-Name matches the natural language input with the terms and some common words in database. The common words includes "all", "some", "somehow", "somewhat", "precise", "precisely", "accurate" ... The common words are used in No-Name to set a threshold for query results. A query needs all related documents will generate more results than a query needs somehow related documents, and a query needs somehow related documents will generate more results than a query needs only precisely matched documents. Conversely speaking, a query needs precisely matched documents will generate results with higher quality than a query needs somehow documents. No-Name achieves the goal by assigning the words different fuzzy values, like "all" equals to 0.1, "somewhat" equals to 0.5, "precisely" equals to 1.0. The fuzzy values are then used to set the threshold for query results., and a document can not be output if its relevance degree to a query is less than the threshold.

Natural language support and multiple independent layer structure enable No-Name to support complicate queries from users. For example, when a user input "I want to know something about remote sensing used in Canadian government". No-Name will first identify "Canadian", "Government", "Remote sensing", and "Something" from the query, the term "Canadian" will be mapped to the term "Canada" automatically, and then find the terms "Canada", "Government", and "Remote sensing" in the independent layers, and perform a query.

"Something" will be assigned a fuzzy value as threshold, and the final results will be achieved by filtering the prior query results according to the threshold.

EXPERIMENTS

Introduction

Due to the limitation of free disk space available at server, the records number in Database are limited to 20,000 with a total size up to 300M. Some general search engine usually generate more than thousands of results toward one query. Generally speaking, the results after the first 200 are usually not relevant to the query according to our experience. To avoid wasting time and speeding our test, only the first 200 results are accepted and analyzed.

Assessment on the search engines

While Yahoo and AlltheWeb are chosen to represent the general-purpose search engines, Geo-Community is chosen to represent the current GIS search engine. Their performances are compared with that of No-Name by studying their retrieval results according to same queries. An expert board with 3 persons is responsible to evaluate the query results.

Two term sets with 10 and 50 terms respectively are adopted to represent the queries. Each query can only contain one term. The results are shown in Table 2 and Table 3.

	Relevance results	Irrelevance results
No-Name	443	172
Yahoo	391	677
AlltheWeb	227	985
Geo-Community	36	12

Table 2 : Query results of term set with 10 terms

	Relevance results	Irrelevance results
No-Name	1951	547
Yahoo	1593	2640
AlltheWeb	1145	3256
Geo-Community	125	23

Table 3 : Query results of term set with 50 terms

Natural Language Support and Linguistic Analysis

Two sentences, "Tell me all about GPS activities in China" and "I want to know something about AM/FM applications in Canada", are selected as queries to test the ability of No-Name and current search engines to support natural language input and linguistic analysis. The results are shown in Table 4 and Table 5. While "GPS" stands for "Global Positioning Satellite System", "AM/FM" stands for "Automated Mapping and Facilities Management".

	Relevance results	Irrelevance results
No-Name	261	43
Yahoo	132	68
AlltheWeb	1	199
Geo-Community	0	0

Table 4 : Query results of the first sentence

	Relevance results	Irrelevance results
No-Name	42	7
Yahoo	10	190
AlltheWeb	0	200
Geo-Community	0	11

Table 5 : Query results of the second sentence

Comparison of the system performances before learning and after learning

No-Name can revise the term connection value by learning from the users' feedback. A category name "Database" is chosen as termX, and 5 terms, ("Binary Large Object", "Conceptual Model", "Data Definition Language", "Georelational Model", and "Spatial Database") in this category are chosen to represent the queries. The initial term connection values between termX and the 5 terms are set to 0.5. Each query can only contain one term. The expert board is responsible to judge if the query results are relevant to termX. The results of the evaluation from expert board are used as users' feedback to train the search engine. 100 results are chosen as feedback from the total 408 evaluation results randomly. The query results before learning and after learning are shown in Table 6 and the term connection values before learning and after learning are shown in Table 7.

\	Relevance results	Irrelevance results
No-Name \	271	137

After learning	256	92
----------------	-----	----

Table 6 : Query results before learning and after learning

Term Connection value	TermX (before learning)	TermX (after learning)
Term1	0.5	0.7
Term2	0.5	0.3
Term3	0.5	0.8
Term4	0.5	0.5
Term5	0.5	0.9

Table 7 : Term connection values before learning and after learning

CONCLUSIONS

Compared to current search engines, No-Name, an Intelligent GIS search engine, has a higher performance in Relevant Sites Retrieval, Irrelevant Sites Dodge, and Natural Language Identification.

Spider, robot, and full text analysis technology are applied in No-Name successfully. The information collecting from Internet is efficient and adequate.

It is realized to determine the strength of term connections in small size thesauri according to users' feedback using fuzzy logic. The experiment results support the search engine can learn from the users' feedback. The relevance results are almost kept as same, but the irrelevance results are highly reduced after training.

The test results show No-Name has the best performance in natural language supporting among the four chosen search engines. Anyway, Yahoo also shows now it can support natural language identification somehow. The improvement should be attributed to the partnership between Yahoo and Google that began in June of 2000 [27].

No-Name's unique multiple independent layers structure can speed the query and help the users to find better matched results.

Although the experimental results shows No-Name is an Intelligent GIS search engine to retrieve GIS information from Internet with high performance, there are still some room left for improvement. First, the expert board should contain more experts to avoid misjudgment. The decisions made by only a few experts are usually fatal to misjudgment and thus become doubtful. The noise, or misjudgment by one expert is expected to be weakened by recruiting more experts to the expert board. Second, The assumption 2, that any term has no connection with its peers in the same category, is too strong and might be not the truth. The term connection value among terms in the same category should be determined in the future which means the number of term connection will be increased from a few thousand to several tens of thousand. Finally, GIS is an explosive interdisciplinary application science, and the terms that GIS involves with are increasing dramatically. It is a challenging problem to maintain a reasonable size GIS thesauri for the efficiency's sake, and at the same time, to keep up with the advancement of GIS technologies.

ACKNOWLEDGE

The authors wish to thank and acknowledge the encouragement and support of Dr. C. Vincent Tao, Dr. Chaowei Yang, Dr. Quanke Wang, and Dr. Yong Hu. Dr. C. Vincent Tao contributed a lot of good ideas at the beginning of our research. Dr. Chaowei Yang, Dr. Quanke Wang, and Dr. Yong Hu helped us a great deal by forming the expert board to evaluate the query results.

REFERENCES

- 1 J. Pearl. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. *Morgan Kaufmann, San Mateo CA, 1988.*
- 2 J. Y. Nie and M. Brisebois. An inferential approach to information retrieval and its implementation using a manual thesaurus. *Artificial Intelligence Review, 10:1-31, 1996.*
- 3 G. Miller. Wordnet: an on-line lexical database. *International Journal of Lexicography, 3, 1990.*
- 4 W. B. Croft. Approaches to intelligent information retrieval. *Information Processing & Magement, 23:249-254, 1987.*
- 5 X. Lu. Document retrieval: A structure approach. *Information Processing & Magement, 26(2):209-218, 1990.*
- 6 C.J. van Rijsbergen. A non-classical logic for information retrieval. *The Computer Journal, 29(6):481-485, 1986.*
- 7 H. Turtle and W. B. Croft. Inference network for document retrieval. *In Proceedings of 13th ACM-SIGIR Conference, Brussels, 1990.*
- 8 A. Bookstein. Outline of a general probabilistic retrieval model. *Journal of Documentation, 39:63-72, 1983.*
- 9 Norbert Fuhr. Probabilistic models in information retrieval. *The Computer Journal, 35(3):243-255, 1992.*
- 10 M.E. Maron and J.K. Kuhns. On relevance, probabilistic indexing and information retrieval. *Journal of the ACM, 7:216-244, 1960.*
- 11 C. J. van Rijsbergen. Information Retrieval. *Butterworths, London, 2 edition, 1979.*
- 12 S. Robertson, M. Maron, and W. Cooper. Probability of relevance: a unification of two competing models for document retrieval. *Information Technology: Research and Development, 1:1-21, 1982.*
- 13 Tadeusz Radecki. Fuzzy set theoretical approach to document retrieval. *Information Processing & Magement, 15:247-259, 1979.*
- 14 J. Sinclair. Corpus, concordance, collocation. *Oxford University Press, 1991.*
- 15 Helen J. Peat and Peter Willett. The limitation of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science, 42(5):378-383, 1991.*

- 16 K Sparck-Jones. Notes and references on early automatic classification work. *SIGIR Forum*, 25(1):10-17, 1991.
- 17 Martha Evens, Yih-Chen Wang, and James Vanderdorpe. Relational thesauri in information retrieval. *Journal of the American Society for Information Science*, 36(1):15-27, 1985.
- 18 E.A. Fox. Lexical relations: Enhancing effectiveness of information retrieval systems. *SIGIR Forum*, 15(3):6-35, 1980
- 19 Roy Rada, Hafedh Mili, Ellen Bicknell, and Maria Blettner. Development and application of a metric on semantic nets. *IEEE Transaction on Systems, Man, and Cybernetics*, 19(1):17-30, 1989.
- 20 Hsinchun Chen and Vasant Dhar. Cognitive process as a basis for intelligent retrieval system design. *Information Processing & Magement*, 27(5):405-432, 1991.
- 21 Young Whang Kim and Jin H. Kim. A model of knowledge based information retrieval with hierarchical concept graph. *Journal of Documentation*, 46(2):113-136, 1990.
- 22 Joon Ho Lee, Myoung Ho Kim, and Yoon Joon Lee. Information retrieval based on conceptual distance in IS-A hierarchies. *Journal of Documentation*, 49:188-207, 1993.
- 23 Joon Ho Lee, Myoung Ho Kim, and Yoon Joon Lee. Ranking documents in thesaurus-based boolean retrieval systems. *Information Processing & Magement*, 30(1):79-91, 1994.
- 24 Shehu S. Farinwata, Dimitar Filev, Reza Langari, Fuzzy control, ayntesis and analysis, p47-70. *John Wiley & Sons Ltd*, 2000.
- 25 AGI on-line GIS dictionary, Association for Geographic *Information and the University Of Edinburgh Department of Geography*.
- 26 *GeoCommunity, GeoSpecific Search Engine!* <http://search.geocomm.com>
- 27 CNN, partnership between Google andYahoo, <http://www.cnn.com/2001/TECH/internet/04/12/wireless.google.idg/index.html>
- 28 WinStar Communications, Inc. of Mansfield support page, <http://support.ici.net/browsers/searcheng.html>