SICE Annual Conference 2010
August 18-21, 2010, The Grand Hotel, Taipei, Taiwan

Web Mining using Genetic Relation Algorithm

Eloy Gonzales*, Shingo Mabu*, Karla Taboada* and Kotaro Hirasawa*
*Graduate School of Information, Production and Systems. Waseda University.
2-7 Hibikino, Wakamatsu-ku, Kitakyushu-shi, Fukuoka 808-0135, Japan
E-mail: egonzale@asagi.waseda.jp,
mabu@aoni.waseda.jp,
karla_taboada@asagi.waseda.jp and
hirasawa@waseda.jp

***Abstract*—Web searching is one of the most universal and influential applications on the Internet and it is becoming increasingly a dominant information seeking method. The World Wide Web (WWW) contains a huge amount of information available on-line and it is continuously growing in size in such a way that searching tools are becoming more valuable especially to improve the relative coverage of web search engines. In this paper, a framework for web mining has been proposed using an evolutionary algorithm named Genetic Relation Algorithm (GRA) that performs additional searching for documents according to users interests. GRA can optimize the important relationships between web hyperlinks in web pages by evolution. Querying the standard search engine performs the creation of the GRA individuals. The proposed evolutionary method leads to pages of qualities that are significantly better than those of the standard search engines.**

Keywords *evolutionary computation, genetic relation algorithm, web mining*

I. Introduction

World Wide Web (WWW) continues to grow at a very high speed containing a huge amount of information available online as an information gateway and also a medium for conducting business. Due to its convenience and rich information, the WWW is a fertile area for data mining research[1]. Along with the amount of information available on the WWW, the number of people connected to the Internet and the number of web pages accessed have also increased exponentially over the past years. There are a great variety of resources and information available on the web for people with the most diverse background and interests. However, one of the major problems is the poor quality of information retrieved. Thus, in most of the cases, users have to search for and filter out the relevant information by themselves. Even qualified users, such as students, researchers and lecturers, do spend time searching and filtering the information retrieved from the Web. This is because of the lacking for performing additional computation over their results in standard search engines.

Therefore, performing additional information filtering becomes a necessary and challenging task. In this paper, a method to perform additional search has been proposed in order to find the most interesting pages from an enormous amount of information for users. This work proposes the use of Genetic Relation Algorithm (GRA) to filter the information retrieved by a conventional web search engine.

Genetic Relation Algorithm can positively contribute to the problem of finding an efficient search strategy on the web as an evolutionary algorithm.

II. Related Work

There are some other works to improve the results of web searching using different techniques. In [2] a clustering method is used exploring co-citation and coupling based on common links shared by web pages, however, the quality and structure of the final clusters are still poor, since there are many uncategorized results. A method using user's bookmarks is also proposed in [3], but the amount of bookmarks is too small to produce useful information. Contextual retrieval [4] is another method for improving web searching which modifies the query according to the user's activities and responses, however a lot of manual intervention is needed to acquire, mantain and represent accurate information.

Some approaches use other evolutionary methods such as Genetic Algorithms in web mining. In [5], the authors proposed a method for guiding genetic algorithms to perform information retrieval by fuzzy classification and genetic feature selection of terms from documents evaluated by the user. GeniMiner [6] is a genetic algorithm that manages a population of pages and aims at maximizing a fitness function that is mathematically based on the user query. In [7] a personal agent that mines web information sources and retrieves documents according to users interests was developed using classical information retrieval techniques and a genetic algorithm to learn and adapt the changes in the users interests. However, most of these methods are suitable only for simple queries since the processing time increases exponentially for complex queries.

The advantages of the proposed method are as follows:

- GRA is an evolutionary web pages selection method, since a small number of the most interesting hyperlinks is obtained after an evolutionary process.
- GRA is complementary to standard search engines because its input is a large number of hyperlinks obtained by querying standard search engines. Therefore, GRA does not replace the conventional web search engines.
- The reduced number of hyperlinks provided by GRA in the final generation consist of only the most similar hyperlinks with respect to the query.

PR0001/10/0000-1622 ¥400

- The structure and properties of GRA ensure that redundant and irrelevant pages are removed without data cleaning and query modifications. Therefore, the results can be obtained with minimal user's intervention.
- The number of hyperlinks in the final evolutionary process, is defined flexibly by users because the hyperlinks correspond to nodes in the GRA individuals.
- GRA is a very powerful tool for analyzing and evaluating the results of standard search engines especially for complex queries since GRA provides a good balance between exploration and exploitation of hyperlinks in web pages.

III. Genetic Relation Algorithm

Genetic Relation Algorithm is one of the evolutionary optimization techniques, which evolves both the directed and indirected graphs, where the best relations between events are obtained. Therefore, it is used to extract a fairly small number of events from a large set of events. An event is an abstract concept being represented and encoded by nodes in GRA. The events are defined depending on the problem to solve, for instance, in the stock market, the nodes represent stock brands in a portfolio, in data mining, the nodes represent association rules and in web mining, the nodes represent hyperlinks of a web page. The strength is defined between the nodes and it is used in the fitness function.

There are two kinds of GRA: GRA with directed branches and GRA with indirected branches. The difference between them is the representation of the connection between nodes. Thus, in GRA with directed branches, the connection is represented by a directed branch and the strength is defined depending on the direction of the edge. On the other hand, in GRA with indirected branches, the connection is represented by an indirected branch and the strength is defined without considering the direction.

In this paper, GRA with indirected branches is used because the strength in the relation between nodes is based on their similarity.

Fig. 1 is an example of GRA with indirected branches. The relation between node i and node j has a strength of $S_{ij} = S_{ji}$ in GRA with indirected branches.

Fig. 2 describes the gene of node i, and the set of these genes represents the genotype of GRA individuals. ID_i is an identification number, for example, $ID_i = 1$ means node i has the directed branches to other nodes, while $ID_i = 2$ means node i has the indirected branches to other nodes. F_i denotes the function of node i. C_{i1}, C_{i2}, ..., C_{ik} denote the nodes which are connected from node i, firstly, secondly, ..., and S_{i1}, S_{i2}, ..., S_{ik} denote the strength from node i to node C_{i1}, C_{i2}, ..., C_{ik} or the strength between node i and node C_{i1}, C_{i2}, ..., C_{ik} depending on the arguments of node i.

The properties of GRA are as follows:

- All individuals in a population have the same number of nodes.
- All node functions in a GRA individual are different.

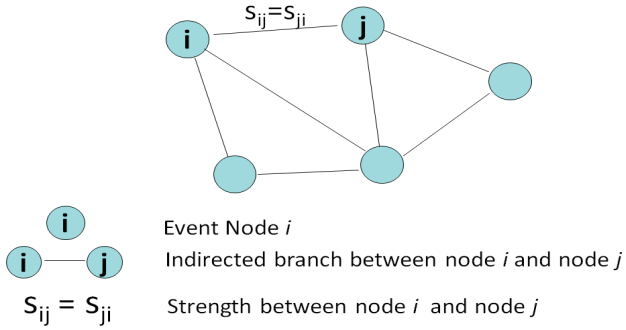

Fig. 1. Genetic Relation Algorithm with indirected branches.

ID_i	F_i	C_{i1}	C_{i2}	.....	C_{ik}
		S_{i1}	S_{i2}	.....	S_{ik}

Fig. 2. Genotype expression of Genetic Relation Algorithm.

- GRA can be evolved in order to optimize the fitness function defined by the strength S_{ij}, in other words, the events and their relations can be evolved by using the strength S_{ij}.
- The point of GRA is that all the connections between nodes do not have to be defined, but the connection itself could be evolved.

The following genetic operators are used in GRA: *Crossover operator* affects two parent individuals. All the connections or contents of the selected nodes in two parents are swapped each other by crossover rate of P_c. *Mutation operator* affects one individual. The connections or contents of each node are changed randomly by mutation rate of P_m.

IV. GRA and Web Mining

A web page is a document or resource of information that is suitable for the World Wide Web and can be accessed through a web browser and displayed on a computer screen. This information is usually in HTML or XHTML format, and may provide navigation to other web pages via hyperlinks[8][9].

Hyperlinks (or links) are the structural components of the web pages which connect a web page to another. This connection is either (1) to other web pages on the same web site, or (2) to web pages located on another web site. Normally, a web page contains many hyperlinks. The typical strategy for accessing information on the WWW is to navigate cross documents through hyperlinks, retrieving the interesting information along the way.

In this paper, a GRA individual represents a Web Page and its nodes represent the hyperlinks existing in the Web Page.

Fig. 3 is an example of GRA for web mining. The cosine similarity between nodes is used as the strength.

The genotype of the GRA individual in the example of Fig. 3 is described in Table I. Notice that $ID_i = 2$ is for GRA with indirected branches. In Table I, *Node Number* is shown for illustration purposes only.

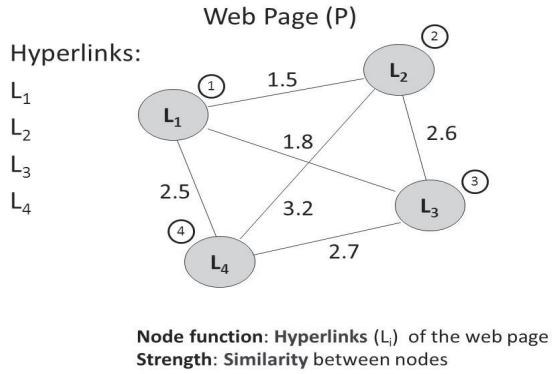

Fig. 3. Genetic Relation Algorithm for Web Mining.

TABLE I
GENOTYPE OF GRA FOR WEB MINING

$NodeNo.$	$NodeFunction$	C_{i1}	C_{i2}	C_{i3}	S_{i1}	S_{i2}	S_{i3}
1	L_1	2	3	4	1.5	1.8	2.5
2	L_2	1	3	4	1.5	2.6	3.2
3	L_3	1	2	4	1.8	2.6	2.7
4	L_4	1	2	3	2.5	3.2	2.7

The query (q) can be words, word pairs or phrases given by users and it is defined as a list of keywords (or terms). L is the set of suffixes of input keywords in query.

Node i is represented as a term vector as follows:

Node $i = (d_{i1}, d_{i2}, \ldots, d_{il}, \ldots, d_{i|L|})$

where, d_{il} represents the occurrence of term l in node i.

$$d_{il} = \begin{cases} 1 & \text{if node } i \text{ contains term } l \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

The quality of node i is calculated as follows:

$$F(i) = \sum_{l \in L} d_{il}. \quad (2)$$

The cosine similarity [9] between node i and node j is calculated as follows:

$$D(i,j) = \frac{\sum_{l \in L} f_{il} f_{jl}}{\sum_{l \in L} f_{il} \sum_{l \in L} f_{jl}}, \quad (3)$$

where, f_{il} is the frequency of keyword l in node i,

Cosine measure determines the similarity between two vectors (nodes) independently of their magnitude. Eq. 3 returns the angle between these two vectors. It is equal to 1 when the vectors point in the same direction, and zero when they form a 90 degrees angle.

Using GRA, an efficient searching strategy can be defined. That is, GRA is used to find the most interesting pages for users.

Since the number of nodes in GRA defines the number of hyperlinks, GRA selects a reduced number of hyperlinks which best match the query generated by users.

A. Fitness of GRA

The fitness function of GRA individuals is defined as follows:

$$Fitness = \frac{1}{|R|} \sum_{i \in R} \frac{1}{|R(i)|} \sum_{j \in R(i)} F(i,j), \quad (4)$$

$$F(i,j) = D(i,j) * F(i) * F(j),$$

where,

$D(i,j)$: cosine similarity between node i and node j described in the previous section.
$F(i)$: quality of node i.
R: set of suffixes of nodes (hyperlinks) in GRA.
$R(i)$: set of suffixes of nodes whose similarity is defined between node i in GRA.

The fitness function evaluates the GRA individuals so that the similarities between hyperlinks are maximized.

The proposed method starts with a query request generated by users. Then, using standard searching engines (ex. Google [10],etc.), the initial population of GRA individuals is created randomly from the results obtained by the search engines. That is, GRA nodes encode the hyperlinks generated by any standard search engine.

Then, the GRA individuals are evaluated with the fitness function and genetic operations (elite selection, tournament selection, crossover and mutation) are carried out. Finally, the final elite individual is the output given to the user.

B. Evaluation Measures

The contents of the function nodes of the final elite individual are listed from the first to the last node to form a rank of hyperlinks.

The evaluation of the final elite individual is realized using the conventional evaluation measures, i. e. , *precision*, *recall* and *F-score*[8] for all nodes as follows:

Let N be the set of hyperlinks retrieved by any standard search engine. Let D_q be the set of actual relevant hyperlinks of query q in N.

Recall of node i is defined as:

$$r(i) = \frac{s_i}{|D_q|}, \quad (5)$$

where s_i is the number of relevant hyperlinks from node 1 to node i. $s_i \leq |D_q|$

Precision of node i is defined as:

$$p(i) = \frac{s_i}{pos(i)}, \tag{6}$$

where, $pos(i)$ is the position of node i, referred as ranked node number.

Average precision is computed based on the precision of each node which contains a relevant document as follows:

$$p_{avg} = \frac{\sum_{i \in D_q} p(i)}{|D_q|} \tag{7}$$

F-score of node i is defined as follows:

$$FS(i) = \frac{2p(i)r(i)}{p(i) + r(i)} \tag{8}$$

V. Simulation Results

The simulations are done using 5 complex queries from different domains and keywords. The average number of terms in queries is 7.2.

In the experiment, 10 runs of GRA are carried out for every query and the results are given as their average.

For each query, the top 400 hyperlinks obtained by a conventional search engine (SE) are used to create the initial population of GRA individuals, ($|N| = 400$).

The number of relevant hyperlinks is obtained from Yahoo directory [11] for every query. The number of nodes of the GRA individual is defined for every query and it has to be at least the same number of the relevant hyperlinks.

All algorithms were coded in Java language. Experiments were performed on a 1GHz Intel Xeon PC with 1G bytes of main memory, running Linux RedHat v4 64bits.

Table II shows the parameters for the evolution of GRA for all queries.

TABLE II
Parameters for Genetic Relation Algorithm.

Parameter	*Value*
Number of GRA individuals	240
Number of generations	100
Number of nodes in each individual	variable
Crossover probability	0.1
Mutation probability	0.01

Fig.4 shows the evolution of the fitness values of the elite GRA individual and the average fitness values of all GRA individuals for q_1. Similar behavior of fitness is found in the rest of the queries.

It can be observed from Fig. 4 that GRA is capable of improving the quality (fitness) of the best individual by about 65%, from the first to the last generation. It can also be seen that the diversity of the population is high at the end of the evolutionary process, which is indicated by the values of the average fitness of the population. This is because the use of genetic operators (crossover and mutation) are helpful to insert and maintain the diversity of the population.

Table III shows the number of terms and the average precision of the proposed evolutionary method and the conventional search engine in every query.

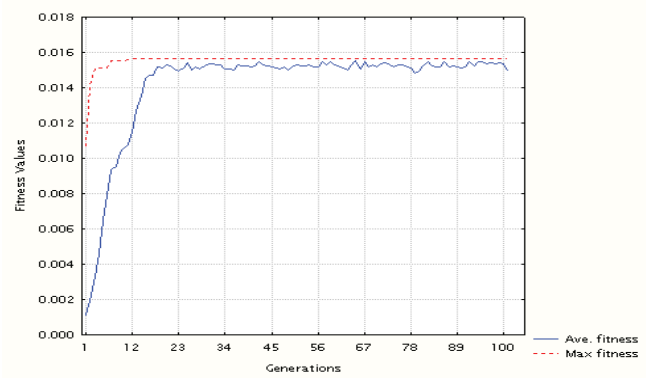

Fig. 4. Evolution of fitness values of elite GRA individual (max. fitness) and average fitness of the population.

TABLE III
Comparison of average precision between GRA and SE.

Query	*No. terms*	*Average Precision GRA*	*Average Precision SE*
q_1	9	72.68%	50.73%
q_2	8	42.22%	32.04%
q_3	9	64.75%	51.83%
q_4	7	60.94%	50.11%
q_5	3	59.55%	59.87%

It is found from Table III that the proposed method outperforms the conventional web search engine in four out of five queries.

Tabla IV shows the average of F-score values for all queries. Since F-score is a harmonic mean of precision and recall, it

TABLE IV
Comparison of average F-score.

Query	*F-score GRA*	*F-score SE*
q_1	24.71%	20.03%
q_2	43.70%	38.06%
q_3	56.64%	49.15%
q_4	52.11%	46.88%
q_5	15.15%	13.98%

is found from Table IV that the proposed method outperforms the conventional web search engine in all cases.

Figure 5, Figure 6, Figure 7, Figure 8 and Figure 9 show the precision-recall curves using the proposed evolutionary method along with the results of standard search engine for q_1, q_2, q_3, q_4 and q_5, respectively.

Fig. 5 shows that the proposed method obtains high precision values even at low recall values.

Fig. 6, Fig. 7, and Fig. 8 show that the proposed method outperforms the conventional web searching engine, which means additional searching would be beneficial for users especially for complex queries. Evaluation functions used in standard search engines are designed for giving a very quick answer to simple queries. Fig. 9 shows that conventional web search engine outperforms a little the evolutionary method in a simple query.

Figure 10 shows the average precision-recall curve over all queries.

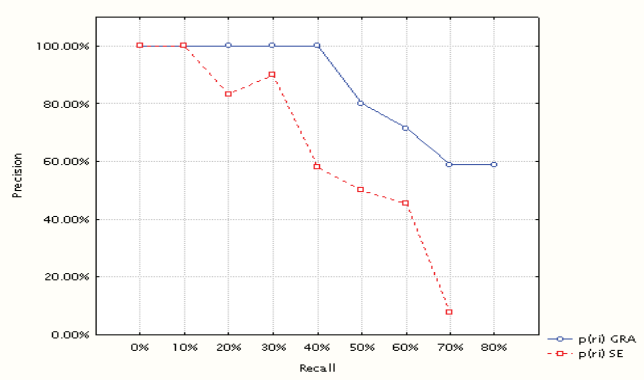

Fig. 5. Precision-recall curve for q_1.

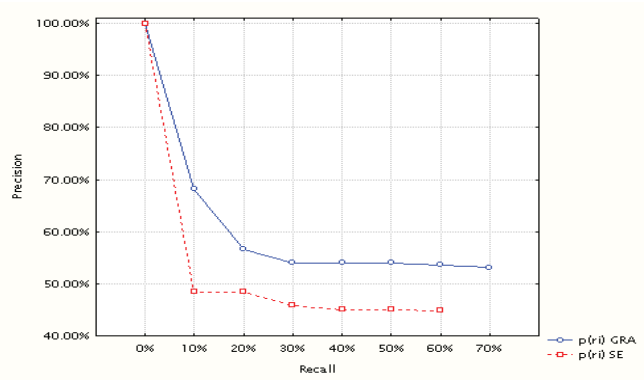

Fig. 6. Precision-recall curve for q_2.

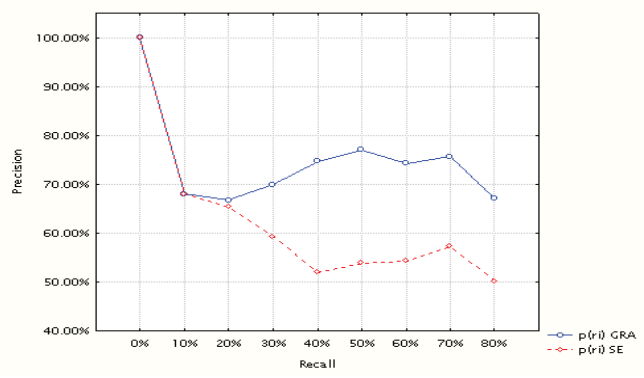

Fig. 7. Precision-recall curve for q_3.

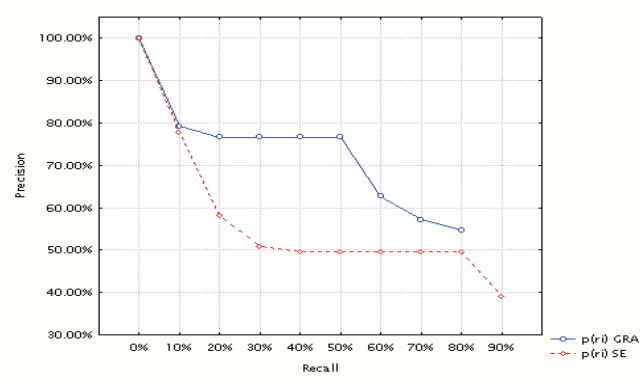

Fig. 8. Precision-recall curve for q_4.

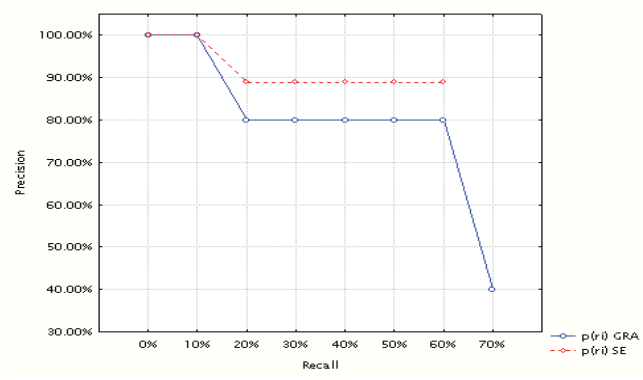

Fig. 9. Precision-recall curve for q_5.

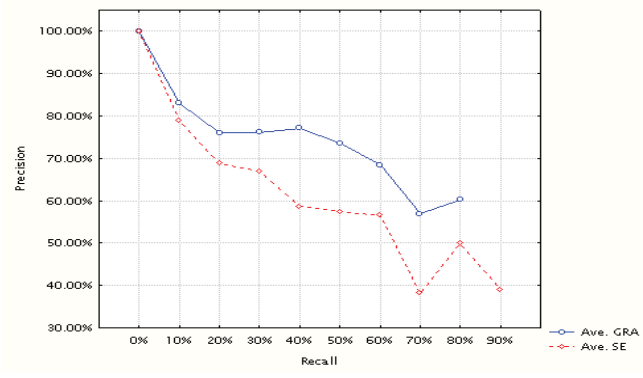

Fig. 10. Average precision-recall curve

It is clear from Fig. 10 that GRA outperforms the conventional web search engine especially between the recall levels of 40% and 50%.

The experiment has demonstrated that the GRA provides a convenient method and retrieves more accurate information by optimizing the web search in several domains. These results show that the GRA system increases the precision and recall of the users' query with information gathering using web search engine.

The preliminary results demonstrated that the system is capable of selecting the most interesting hyperlinks from the web related to a query. It is important to remark that the usefulness of an important hyperlink is related to all relevant words contained in the query.

VI. Conclusions

- The reduced number of interesting hyperlinks is obtained via genetic selection using GRA from the results of a

web search engine and therefore, improve the user's web searching for real and complex queries.

- GRA method can effectively provide a search strategy with minimal user's intervention.
- GRA selects the hyperlinks considering their similarity and quality. GRA provides the most similar hyperlinks with respect to the query.
- The quality of the resultant pages is improved, when GRA is used, compared to the standard search engines.

REFERENCES

[1] R. Kosala and H. Blockeel, "Web Mining Research: A Survey", *Journal of ACM SIGKDD Explorations*, Vol. 2, Issue 1, 2000.

[2] R. Weiss, "Hypursuit: A Hierarchical Network Search Engine that Exploits Content-Link Hypertext Clustering" . *In Proc. of Hypertext'96*, Washigton, USA, 1996.

[3] K. S. Jeong, H. R. Park and S. Y. Kim, "Improving the Performance of Web Search Using User's Bookmarks". *In Proc. of International Symposium on Information Technology Convergence*, Korea, 2007.

[4] D. K. Limbu, A. M. Connor, R. Pears and S. G. MacDonell, "Improving Web Search Using Contextual Retrieval". *In Proc. of Sixth International Conference on Information Technology: New Generations*, Australia, 2009.

[5] M. J. M. Bautista, M. Amparo, V. Henrik and L. Larsen, "A genetic fuzzy classifier to adaptive user interest profiles with feature selection", *In Proc. of the European Society for Fuzzy Logic and Technology (Eusflat-Estylf)*, Joint Conference, pp. 327-330, Palma, Spain, 1999.

[6] F. Picarougne, N. Monmarch, A. Olivier and G. Venturini, "Web mining with a genetic algorithm", *In Proc. of the 11th International World Wide Web Conference*, WWW-2002, Honolulu, Hawaii, USA, May 2002.

[7] M. S. Valim and J. M. A. Coello. "An Agent for Web Information Dissemination Based on a Genetic Algorithm". *In Proc. of IEEE International Conference on Systems, Man and Cybernetics*, IEEE SMC'03. IEEE Press, p.3834 3839, Washington, USA, 2003. Cybernetics- IEEE SMC'03.

[8] B. Liu. Web Data Mining: Exploring Hyperlinks, Contents and Usage Data, Springer-Verlag, Germany, 2007.

[9] G. Salton and M. J. McGill, The SMART and SIRE Experimental Retrieval Systems in Readings in Information Retrieval. Morgan Kaufmann Publishers Inc. USA, 1997.

[10] http://www.google.com

[11] http://www.yahoo.com/directory