FOCUSED WEB CRAWLING USING DECAY CONCEPT AND GENETIC PROGRAMMING

Mahdi Bazarganigilani[1], Ali Syed[2] and Sandid Burki[3]

[1] Faculty of Business, Charles Sturt University, Melbourne, Australia
mahdi62b@yahoo.com

[2] Faculty of Business, Charles Sturt University, Melbourne, Australia
ASyed@studygroup.com

[3] Faculty of Business, Charles Sturt University, Melbourne, Australia
sandid.burki@gmail.com

ABSTRACT

The ongoing rapid growth of web information is a theme of research in many papers. In this paper, we introduce a new optimized method for web crawling. Using genetic programming enhances the accuracy of simialrity measurement. This measurement applies to different parts of the web pages including the title and the body. Consequently, the crawler uses such optimized similarity measurement to traverse the pages .To enhance the accuracy of crawling, we use the decay concept to limit the crawler to the effective web pages in accordance to search criteria. The decay measurements give every page a score according to the search criteria. It decreases while traversing in more depth. This value could be revised according to the similarity of the page to the search criteria. In such case, we use three kinds of measurement to set the thresholds. The results show using Genetic programming along the dynamic decay thresholds leads to the best accuracy.

KEYWORDS-COMPONENT;

Focused Web Crawler; Genetic Programming; Decay Concept; Similarity Space Model.

1. INTRODUCTION

The World Wide Web contains more than 11.5 million pages and is growing rapidly. According to recent statistics, 60 percent of the users search for a special theme and often use popular and commercial search engines to obtain their results [2,3]. Such users do not have the capability to use a generic search engine.

In reality, many search engines do not cover all the visible pages [1]. Therefore, there is a need for a more effective crawling method to collect more accurate data. One of the most common approaches is to limit the crawler to a few specified subjects. In this way, the crawler is able to retrieve the most common pages. This method called Focused crawling [4]. The following figure illustrates the difference between regular crawling and focused crawling.

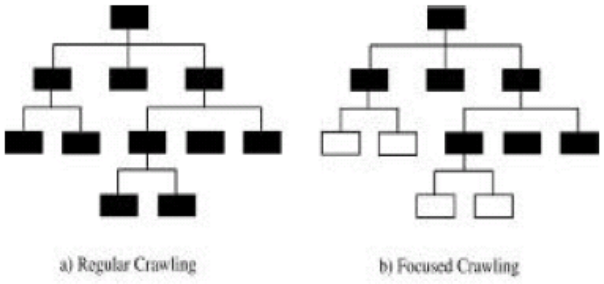

Figure 1. Comparision between focused and regular cralwers.

Web is similar to a social network. In such a network, the links among pages are not meaningless. Each link in every page denotes a semantic relation among the pages. The crawlers use such links to effectively fetch the correlated pages. We can proceed in two ways:

One way is using content based similarity. In such a way, the crawler estimates the similarity of the current page and the subject. The similarity above a predefined threshold leads to fetching entire page's links.

Another approach is to rank the page's links. We fetch the links with desired rank score.

Since, we use a predefined threshold in our crawler, it leads to unpredictable noises. In this paper, we use a Decay concept to lessen the effect of inaccuracy incurred by the noises [5].

2. PREVIOUS WORKS

One of the first focused Web crawlers is discussed in [6]. Experiences with a focused crawler implementation were described by Chakrabarti in 1999 [7]. Focused crawlers contain two types of algorithms to keep the crawling scope within the desired domain [8]: (1) Web analysis algorithms are used to judge the relevance and quality of the Web pages pointed to by target URLs; and (2) Web search algorithms determine the optimal order in which the target URLs are visited [9].

One the fundamental basis of any crawler is to determine the similarity of the pages according to the query. In this way, the pages are mapped to vector space for computing the similarity. This method is still regarded as an effective solution [6]. In another works, there is a use of implicit ontology concepts to establish the similarity measurement more accurately. They use ontology to enrich the text with similar word concepts .Such works have been proven to be more accurate than pure use of similarity concept [7,8].

In the structural approach, link analysis of pages is vastly used building a relation graph. For example, the Page Rank algorithm can be used [9,10]. The Page Rank method computes the score of the URL not fetched yet as well as the URL with best scores firstly retrieved. The Page Rank algorithm ranks them according to the frequency of repetitions and the importance of these in other pages. Under this approach, the crawling would not stop upon reaching an irrelevant page [11]. As a result, there may be pages which are not completely coherent to the main search theme.

Some of focused crawlers rely on the use of the database of previous crawlers. Such databases contain the relevancy information of the links and URLs. There has been some research into the combined use of such content and link based algorithms [12]. Some researchers have used a

concept graph to build link based graphs [11]. In such cases, they use Meta searches to establish the pages linked to a particular page.

In some instances, Genetic Programming has been used to improve the accuracy of the content based classification. As an extension of Genetic Algorithms, Genetic Programming has also been applied to data classification. Castillo [14] developed a multi-strategy classifier system for document classification. Different types of classifiers (e.g., Naïve Bayes, Decision Trees) were applied to different parts of the document (e.g., titles, references). Genetic algorithms were applied for feature selection as well as for combining the output of the different classifiers. In her recent studies, Zhang [15-18] proposed a GP-based classification framework to intelligently fuse evidence from multiple sources in order to improve classification of text documents into predefined categories [19].

2.1. SIMILARITY MEASURES

In his research, Chen [19] had introduced three methods of similarity measurements; bag-of-words, cosine, and Okapi. Such content-based similarity measures have been applied to the content of Web. These three similarity measures have been widely used in scientific research activities especially in the text classification field [15-18]. In order to compute these similarity measures, the documents are required to be represented as vectors, as in the Vector Space Model [20]. Suppose we have a collection with t distinct index terms t_j, a document d_i can be represented as follows: $d_i = (w_{i1}, w_{i2}, ..., w_{it})$, where w_{ij} represents the weight assigned to term t_j in document d_i.

For the bag-of-words measure, the similarity between two documents d_1 and d_2 can be calculated as:

$$bag-of-words(d_1, d_2) = \frac{|\{d_1\} \cap \{d_2\}|}{|d_1|}$$

where $\{d_i\}$ corresponds to the set of terms occurring in document d_i.

For the cosine measure, the similarity between two documents can be calculated as [21]:

$$\cos ine(d_1, d_2) = \frac{\sum_{i=1}^{t} w_{1i} * w_{2i}}{\sqrt{\left(\sum_{i=1}^{t} w_{1i}^2\right) * \left(\sum_{i=1}^{t} w_{2i}^2\right)}}$$

For the Okapi measure, the similarity between two documents can be calculated as:

$$Okapi(d_1, d_2) = \sum_{t \in d_1 \cap d_2} \frac{3 + tf_{d2}}{0.5 + 1.5 * \frac{len_{d2}}{len_{avg}} + tf_{d2}}$$

$$* \log \frac{N - df + 0.5}{df + 0.5} * tf_{d1}$$

where tf is the term frequency in a document, df is the document frequency of the term in the whole collection, N is the number of documents in the whole collection, len is the length of a document, and $lenavg$ is the average length of all documents in the collection. [19].

From these equations, we can deduce that the cosine similarity measure is symmetric, while the bag-of-words and Okapi similarity measures are not [19].

2.2. EVIDENCE COMBINATION

Chen has used different combination of above similarity measurements on different parts of web pages including body and title [19].

TABLE I. DIFFERENT EVIDENCE

Content-Based Evidence	Bag of words Using Title
	Cosine Using Title
	Okapi Using Title
	Bag of words Using Body
	Cosine Using Body
	Okapi Using Body

2.3. GP BASED CLASSIFICATION

Each type of evidence shown in the previous section is represented as a document × document matrix, and serves as the input matrix to the GP-based classification framework [19]. The matrix is defined as:

$$M_k = \begin{pmatrix} a_{11} & .. & a_{1n} \\ .. & .. & .. \\ a_{n1} & .. & a_{nm} \end{pmatrix}$$

In this matrix, a_{ij} is the similarity value between the two documents d_i and d_j based on one type of similarity measure mentioned in the previous section. GP will try to find a best non-linear function f to combine the matrices M_1, M_2,…, M_n, where n is the number of types of evidence. The computational output of the combination through such a non-linear function f is an output matrix defined as M_{GP}:

$$M_{GP} = f(M_1, M_2, ..., M_n)$$

M_{GP} is a matrix of similarities between pairs of documents. In order to take advantage of information represented in M_{GP} to predict the class label for a document in the classification process, we introduced a method based on a nearest neighbour classifier KNN [22].Comparing with Mk, MGP is denser, more accurate, and can produce better classification results [19].

3. GENETIC PROGRAMMING

Based on the principle of biological inheritance and evolution, Genetic Programming (GP) is an extension of Genetic Algorithms (GAs), It is a set of artificial intelligence search algorithms which have strong ability to traverse a very large search space efficiently and find approximate global optimal solutions instead of local optimal solutions. Genetic Programming has been widely used and proved to be effective in solving optimization problems, such as financial forecasting, engineering design, data mining, and operations management [23]. GP is capable of solving complex problems for which conventional methods cannot find an answer easily [19].

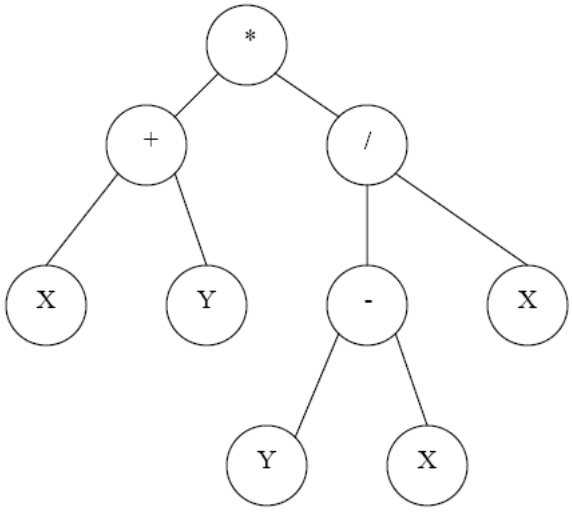

Figure 1. A sample GP tree Structure [19].

GP searches for an "optimal" solution by evolving the population, generation after generation. It works by iteratively applying genetic transformations, such as reproduction, crossover, and mutation, to a population of individuals, to create more diverse and better performing individuals in subsequent generations. The reproduction operator directly copies, or using a more appropriate term, clones some individuals into the next generation. The probability for an individual to be selected for reproduction should be proportional to its fitness. Therefore, the better solution for solving the problem, the higher probability it has to enter the next generation. While reproduction keeps the best individuals in the population, crossover and mutation introduce transformations, and provide variations to enter into the new generation. The crossover operator randomly picks two groups of individuals, selects the best (according to the fitness) individual in each of the two groups as parent, exchanges a randomly selected gene fragment of each parent, and produces two "children". Thus, a "child" may obtain the best fragments of its excellent parents and so may surpass them, providing a better solution to the problem. Since parents are selected from a "competition", good individuals are more likely to be used to generate offspring. The mutation operator randomly changes a gene code of an individual. Using these genetic operators, subsequent generations keep individuals with the best fitness in the last generation, and take in "fresher air", providing creative solutions to the target problem. Better solutions are obtained either by inheriting and reorganizing old ones or by lucky mutation, simulating Darwinian Evolution [19].

In order to apply GP to the Web classification problem, several required key components of a GP system need to be defined. Table 2 lists these essential components along with their descriptions.

TABLE II. APPLYING GP TO CLASSIFICATION

Terminals	Evidence Types
Functions	+,/,*, sqrt
Fitness Function	Macro F1
Genetic Operators	Reproduction, cross over, mutation

3.1 FITNESS FUNCTION

The choice of fitness function can have a huge impact on the final classification results [24]. A good similarity function is defined as a similarity function with a high value. When it is applied to a document y of class C, it ranks documents from class C in such a way that those documents with greater similarities to y are top-ranked. The higher the fitness value, the better the fitness function. The fitness function we select to use is Macro F1. This algorithm uses kNN to predict the category label for a document. Its detailed procedure is shown below [19].

Let R = 0, P = 0, T = 0

for each document y in test collection do

Find the k documents most similar to y

Predict a category for y according to the kNN algorithm using the k documents as the k nearest neighbours

if this is a correct prediction then

R = R + 1, P = P + 1

end if

T = T +1

end for

Let p = P/T, r = R/|C|

F=2p*r/(p+r) (F stands for Macro F1)

4. PROPOSED METHOD

As we mentioned, the application of estimating similarity of content between the desired subject and the page is the essential part of each focused crawler. One of drawbacks of such methods is how to determine the best threshold. Considering low threshold may lead to entering huge numbers of irrelevant pages .Such ineffective pages results in less efficiency. On the other hand, considering such threshold as high to comply with user's criteria may result in loosing so many effective pages.

By considering the best possible threshold, we may face another problem. In such case, there may be huge amount of ineffective pages due to little difference with the desired threshold [5].

We propose a decay concept to overcome the drawbacks of thresholds. For each page, we set a variable among 0 and 1.This variables shows the decay concept. Each page with value near 1 shows better similarity. Each child pages inherit the decay of the parent with a percentage of reduction. We set such a value to half of its parent decay [5].

The use of decay concept is applied when the decay value reduces less than a threshold T_d .In such cases the crawler stops traversing new pages. Actually, we use two threshold values. In such way, the decays of child pages always decrease. Such assumption is not proper since we may face a page completely relevant to the desired search query. In such case, if the similarity is greater than a threshold T_r ,we reset the decay to 1.In other words, for each page, we have two values. One is similarity between the page and the query and the other is the decay value which put among 1,1/2,1/4,1/8,…Moreover, such approach has three thresholds. Main Threshold or T_m represents the similarity between the query and the page. Decay Threshold or T_d which the decay value less than threshold results in stoppage of the crawling. Reset Threshold or T_r causes the decay to reset to 1 for the pages which have similarity above [5].

4.1 FOCUSED CRAWLING FRAMEWORK

In this section, we present high level steps of our focused crawler.

Step 1: Discovery of best similarity function. A data collection with both computing and non-computing documents obtained from DMOZ [25] will be used as the training and validation collections. Contents of these Web pages will be analyzed, and similarities based on different measures such as bag-of-Words, cosine [21], and Okapi [26] will be calculated. GP will be used to discover the best similarity function which is a combination of these similarity measures. This newly discovered similarity function can represent the similarity relationship among these Web pages more accurately. The discovered best similarity function will be used in classification step [19].

Step 2: Initialization. In this step, the Web pages pointed to by the starting URLs are fetched by the crawler to form the base set. This is usually the first results from a meta search engine.

Step 3: Classification. For each fetched Web page, the GP discovered best similarity function will be used by a kNN classifier to decide if this is a computing-related Web page. If yes, this Web page will survive and be put into the collection. Otherwise, this Web page will be discarded [19].

Step 4: Breadth-first search. The breadth-first search algorithm will be used to fetch new Web pages. The outgoing links of the surviving relevant Web pages will be collected and put into the crawling queue. The reason to choose breadth-first search is that it is not a local search algorithm (like best-first search), and it does not share the natural limitations of the local search algorithms. Although breadth-first search may increase the crawling time, it is still considered as a good method to solve the problems caused by local search, since crawling time is not a crucial factor when building a domain-specific collection.[19].we apply our decay concept to each page .Each page which dos not comply with predefined threshold would cause to stoppage of the crawling.

Step 5: Meta-search. Top 10 results from some search engines are combined and put into the crawling queue. The meta-search step will try to obtain diverse relevant URLs globally from the whole search space, and it will not be limited by the boundaries between relevant Web communities because it does not have to follow hyperlinks to find relevant pages [19].

Step 6: Termination. Steps 3 - 5 are repeated until the number of Web pages in a local collection repository reaches.

5. RESULTS AND EXPERIMENTS

To show the strength of our GP based classifier, we used 30% and 10% of DMOZ [25] resources. In 30% dataset, we considered 25% for validation and the rest for testing. We

compare our GP based classifier with simple SVM based and combination based SVM. For 10%,we set 7% for validation.

The difference between content-based SVM and combination-based SVM lies in the kernel combination. Joachims et al. [27] have indicated how to combine different similarity measures by means of composite kernels in their research work. Their approach contains a combination of simple and well-understood kernels by a series of "kernel preserving" operations, to construct an increasingly matching feature space S. In order to apply their method to our application domain, we started by finding that each one of our types of evidence can be represented as positive document × document matrices, as described in Section 5. Therefore, we can represent each type of evidence as a kernel matrix. The kernel matrix for our final feature space S is obtained by means of a linear combination of our initial kernel matrices.

For the evaluation of the algorithm, factors including Precision, Recall and F are computed based on the following parameters [28].

TABLE III. EVALUATING PARAMETERS

Pages	**Assigned to** c_i	**Not Assigned to** c_i
Belonging to c_i	tp	fn
Belonging to other cluster than c_i	fp	tn

The Precision shows the accuracy of the algorithm while the recall represents the integrity of search algorithm.

$$\Pr ecision = \frac{tp}{tp + fp}$$

$$\mathrm{Re}\, call = \frac{tp}{tp + fn}$$

F Also can be computed as follows:

$$F = \frac{2 * \Pr ecision * \mathrm{Re}\, call}{\Pr ecision + \mathrm{Re}\, call}$$

We have shown the $F1$ of both 10 and 30 % datasets. We set the GP properties as set in [19].for three thresholds we set the Tr =0.3 and Tm=0.15 and Td=0.3.

TABLE IV. COMPARISION EVALUATION

Class	GP +meta search+ decay	GP + meta search	GP	Content-based SVM	Combination-based SVM
10%	65%	64%	62%	54%	55%
30%	72%	69%	66%	58%	56%

As sown, the promising method is using GP along with the meta search and limiting the crawled pages with decay concepts.

In another experiments to determine different thresholds for our decay limitation, we inspired from [5].To consider different levels of rigidness on crawler. They set Tr with three values.0.1,0.2 and 0.3 which are consecutively lenient and average and rigid.

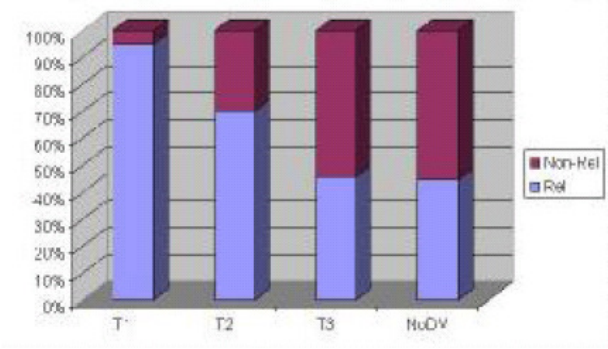

Figure 2. Crawling Accuracy for different Tr values [5].

As shown, in rigid approach there are the least irrelevant pages. it has been shown, in lenient way the half of the fetched page were irrelevant. In another experiments they have shown how much stack memory consumed in each approach.

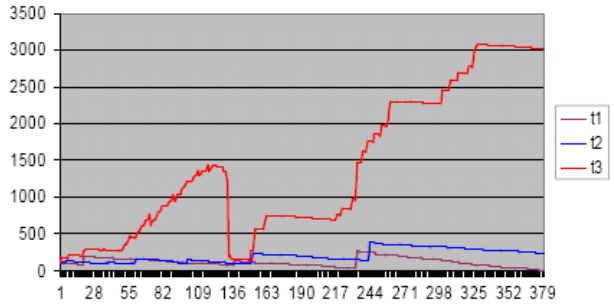

Figure 4. Stack Size in different approches [5].

Obviously the rapid incline in the drawing shows the discovery of a relevant page. While a decline represents some irrelevant pages in cache set. In such datagram it is obvious the rigid approach saved the most memory [5].

6. CONCLUSIONS

After a comparison of generic and focused web crawling, we introduced a new method for focused crawlers. With a goal of reducing inefficiencies, a novel method using Genetic Programming is proposed to most efficiently discover the best combination for estimation the similarity evaluation among pages. Such genetic programming approach gained the best possible similarity measurement among web pages. The method applied to both the title and body of the web pages according to the different similarity measurement. The results showed GP approaches surpass the SVM based approach. Even the combination based SVM has the least accuracy. Furthermore, we use a decay concept approach to limit the crawler from fetching irrelevant pages. Our decay methods dynamically score the pages while traversing and could eliminate or revise them for traversing. In such way the crawler picked the most relevant pages.

The results of research demonstrate that the proposed method produces better accuracy and efficiency as compared to other algorithms.

7. REFERENCES

[1] A. Gulli, A. Signorini, "The Indexable web is more than 11.5 billion pages", In Proceedings of the 14th international conference on World Wide Web, pp. 902- 903, ACM Press, 2005.

[2] Internet Metrics and Statistics Guide: Size and Shape, http://caslon.com.au/metricsguide13.htm, Version of 2003 "Evaluating Crawling Efficiency Using Different Weighting Schemes with Regional Crawler", P. Chubak,

[3] M. Shokouhi, Proceedings of IEEE 4th International Conference o Intelligent Systems Design and Applications (ISDA2004), Budapest, Hugary, 2004.

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

[5] M. Khalilian, K. SheikhEsmaili, M. Neshati, H. Abolhassani, "Boundy Threshold Controling Using Decay Concept In Focused Crawling ", Thirteenth National CSI Computer Conference, Kish Islan, Iran, March 2008.

[6] Bra, P.D., Houben, G., Kornatzky, Y., and Post, R., Information Retrieval in Distributed Hypertexts. in Proceedings of the 4th RIAO Conference. p. 481-491. 1994. New York.

[7] M. Ehrig, A. Maedche. Ontology-focused Crawling of Web Documents, In Proceedings of the 2003 ACM symposium on Applied computing.

[8] Yuxin Chen. A novel hybrid focused crawling algorithm to build domain-specific collections. PhD thesis, United States – Virginia, 2007.

[9] Qin, J. and Chen, H., Using Genetic Algorithm in Building Domain-Specific Collections: An Experiment in the Nanotechnology Domain. in Proceedings of the 38th Annual Hawaii International Conference on System Sciences - HICSS 05. p. 102.2. 2005. Hawaii, USA.

[10] F. Menczer and G. Pant and P. Srinivasan. Topic-driven crawlers: Machine learning issues, ACMTOIT, Submitted, 2002.

[11] S. Chakrabarti,M. van den Berg and B. Dom. Focused Crawling: A New Approach to Topic- Specific Web Resource Discovery, In Proceedings of the 8th International WWW Conference, Toronto, Canada, May 1999.

[12] D.Maghesh Kumar,, (2010) "Automatic Induction of Rule Based Text Categorization", International Journal of Computer Science & Information Technology (IJCSIT), Vol 2, No 6, December 2010.

[13] J. Cho, H. Garcia-Molina, L. Efficient Crawling Through URL Ordering, Page. In Proceedings of the 7th International WWW Conference, Brisbane, Australia, April 1998.

[14] L. Page, S. Brin, R. Motwani, T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web, Stanford Digital Library Technologies Project.

[15] D. Bergmark and C. Lagoze and A. Sbityakov. Focused Crawls, Tunneling, and Digital Libraries.

[16] M. Jamali, H. Sayyadi, B. Bagheri Hariri and H. Abolhassani. A Method for Focused Crawling Using Combination of Link Structure and Content Similarity, WI/IEEE/ACM 2006. Hong Kong, 2006.

[17] M. Diligenti, F. Coetzee, S. Lawrence, C. Giles and M. Gori. Focused Crawling Using Context Graphs, In Proceedings of the 26th International Conference on Very Large Databases (VLDB 2000), Cairo, Egypt, September2000.

[18] Castillo, M.D.D. and Serrano, J.I., A multistrategy approach for digital text categorization from imbalanced documents. SIGKDD, 2004. 6(1): p. 70-79.

[19] Zhang, B., Chen, Y., Fan, W., Fox, E.A., Gonçalves, M.A., Cristo, M., and Calado, P., Intelligent Fusion of Structural and Citation-Based Evidence for Text Classification. in Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 667-668. 2005. Salvador, Brazil.

[20] Zhang, B., Chen, Y., Fan, W., Fox, E.A., Gonçalves, M.A., Cristo, M., and Calado, P., Intelligent GP Fusion from Multiple Sources for Text Classification. in Proceedings of the 14th Conference on Information and Knowledge Management. p. 477-484. 2005. Bremen, Germany.

[21] Zhang, B., Gonçalves, M.A., Fan, W., Chen, Y., Fox, E.A., Calado, P., and Cristo, M., Combining Structure and Citation-Based Evidence for Text Classification. in Proceedings of the 13th Conference on Information and Knowledge Management. p. 162-163. 2004. Washington D.C., USA.

[22] Zhang, B., Gonçalves, M.A., Fan, W., Chen, Y., Fox, E.A., Calado, P., and Cristo, M., A Genetic Programming Approach for Combining Structural and Citation-Based Evidence for Text Classification in Web Digital Libraries, in Soft Computing in Web Information Retrieval: Models and Applications. 2006: p. 65-83.

[23] Salton, G., Automatic Text Processing. 1989, Boston, Massachusetts, USA: Addison-Wesley.

[24] Kleinberg, J.M., Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM), 1999. 46(5): p. 604-632.

[25] Bergmark, D., Collection Synthesis. in Proceedings of the 2nd ACM/IEEE-CS joint conference on digital libraries p. 253-262. 2002. Portland, Oregon, USA.

[26] Dean, J. and Henzinger, M.R., Finding Related Pages in the World Wide Web. in Proceedings of the 8th International WWW Conference. p. 1467-1479. 1999. Toronto, Canada.

[27] Kitsuregawa, M., Toyoda, M., and Pramudiono, I., WEB Community Mining and WEB Log Mining: Commodity Cluster Based Execution. in Proceedings of the 13th Australasian Database Conference. p. 3-10. 2002. Melbourne, Australia.

[28] Salton, G. and Buckley, C., Term-weighting approaches in automatic text retrieval. IPM, 1988. 24(5): p. 513-523.

[29] Yang, Y., Expert network: effective and efficient learning from human decisions in text categorization and retrieval. in Proceedings of 17th ACM International Conference on Research and Development in Information Retrieval - SIGIR 94. p. 13-22. 1994. Dublin, Ireland.

[30] Koza, J.R., Genetic programming: On the programming of computers by natural selection. 1992, Cambridge, MA, USA: MIT Press.

[31] Fan, W., Fox, E.A., Pathak, P., and Wu, H., The effects of fitness functions on genetic programming-based ranking discovery for web search. JASIST, 2004. 55(7): p. 628-636.

[32] DMOZ, Directory Mozilla, http://www.dmoz.org,.

[33] Robertson, S.E., Walker, S., and Beaulieu, M.M., Okapi at TREC-4. in TREC-4. p. 73-96. 1995.

[34] Joachims, T., Cristianini, N., and Shawe-Taylor, J., Composite kernels for hypertext categorisation. in Proceedings of 18th International Conference on Machine Learning - ICML 01. p. 250-257. 2001. Williams College, USA.

[35] Cleverdon, C. W. and Mills, J. The testing of index language devices. Aslib Proceeding. 15, 4, 106–130, 1963

Authors

1-Mahdi Bazarganigilani: He finished his bachelor in Computer Science, Software Engineering. Shahid Beheshti University, Tehran. Iran in 2008. He is currently studying Master of Information Systems at Charles Sturt University, Melbourne. His research interests are in the areas of Artificial Intelligence, Intelligent Data Mining, Information Security and Image Processing.

2- Ali Syed : Is the Deputy Academic Director and Adjunct Senior Lecturer at Charles Sturt University, Study Centre Melbourne. With over 12 years of Academic experience, his research interests are generically in the areas of IS management and specifically in knowledge management, data mining and their applications.

3- Sandid Burki : He is currently a Lecturer at Charles Sturt University, Study Centre Melbourne. His areas of interest include Information Security and Online Information Systems.