9

Malaysian Business Community Social Network Mapping on the Web Based on Improved Genetic Algorithm

Siti Nurkhadijah Aishah Ibrahim, Ali Selamat and Mohd Hafiz Selamat
Universiti Teknologi Malaysia
Malaysia

1. Introduction

The issues of community social network mapping on the web have been intensively studied in recent years. Basically, we found that social networking among communities has become a popular issue within the virtual sphere. It relates to the practice of interacting with others online via blogsphere, forums, social media sites and other outlets. Surprisingly, Internet has caused great changes to the way people do business. In this chapter, we are focusing on the networks of business in the Internet since it has become an important way of spreading the information of a business via online. Business networking is a marketing method by which business opportunities are created through networks of like-minded business people. There are several popular businesses networking organization that create models of networking activity that, when followed, allow the business person to build new business relationship and generate business opportunities at the same time. Business that increased using the business social networks as a means of growing their circle of business contacts and promoting themselves online and at the same time develop such a "territory" in several regions in the country. Since businesses are expanding globally, social networks make it easier to keep in touch with other contacts around the world.

Currently, searching and finding the relevant information become a high demand from the users. However, due to the rapid expansion of web pages available in the Internet lately, searching the relevant and up-to-date information has become an important issue especially for the industrial and business firms. Conventional search engines use heuristics to decide which web pages are the best match for the keyword. Results are retrieved from the repository which located at their local server to provide fast searched. As we know, search engine is an important component in searching information worldwide. However, the user is often facing an enormous result that inaccurate or not up-to-date. Sometimes, the conventional search engine typically returned the long lists of results that saddle the user to find the most relevant information needs. Google, Yahoo! and AltaVista are the examples of available search engine used by the users. However, the results obtain from the search engines sometimes misrelated to the users query. Moreover, 68% of the search engine users will click a search result within the first page of results and 92% of them will click a result

within the first three pages of search results (iProspect, 2008). This statistic concluded that the users need to view page by pages to get the relevant result. Thus, this will consume the time to go through all the result provides by search engine. From our experienced, the relevant result also will not always promise found even after looking at page 5 and above. Internet also can create the abundance problem such as; limited coverage of the Web (hidden Web sources), limited query interface: keyword-oriented search and also a limited customisation to individual users. Thus, the result must be organized so that them looks more in effective and adapted way. In previous research, we present the model to evaluate the searched results using genetic algorithm (GA). In GA, we considered the user profiles and keywords of the web pages accessed by the crawler agents. Then we used the information in GA for retrieving the best web pages related to the business communities to invest at the Iskandar Malaysia in various sectors such as education, entertainment, medical healthcare etc.

The main objective of this chapter is to provide the user with a searching interface that enabling them to quickly find the relevant information. In addition, we are using the crawler agent to make a fast crawling process and retrieve the web documents as many as it can and scalable. In the previous paper, we also using genetic algorithm (GA) to optimize the result search by the crawlers to overcome the problem mention above. We further improve the GA with relevance feedback to enhance the capabilities of the search system and to find more relevant results. From the experiments, we have found that a feedback mechanism will give the search system the user's suggestions about the found documents, which leads to a new query using the proposed GA. In the new search stage, more relevant documents are retrieved by the agents to be judged by the user. From the experiments, the improved GA (IGA) has given a significant improvement in finding the related business communities to potentially invest at Iskandar Malaysia in comparison with the traditional GA model.

This chapter is organized as follows. Section 2 defined the problem that related to this chapter. Section 3 is details on improved genetic algorithm and section 4 are the results and discussion. Section 5 explains the results and discussion of this chapter and Section 6 presented the case study. Finally, section 7 describes the conclusion.

2. Problem Definition

In this chapter, we define the business networks as βƉ whereby it will be represent as a graph G= (V, E) where V is a set of vertices (URL or nodes) and E is a set of links (URLs) that link two elements of V. Fig. 1 shows the networks that represent as a graph. As explained in (Pizutti, 2008), a networks of community is a group of vertices that have a high density of edges among them but have lower density of edges between groups. The problem with the community network is when the total of group, g is unknown how can the related g' can be found? Basically, adjacency matrix is used to find the connection between g. For instance, if the networks consist of V nodes then the networks can be represented as $N \times N$ adjacency matrix (Pizutti, 2008). Nevertheless, we used the binary coding [0, 1] to represent the occurrence of terms in the network or each web page so that we can find the related networks. In the results section, we will show how the searching technique using genetic algorithm and improved genetic algorithm works in order to get the most related information to the V.

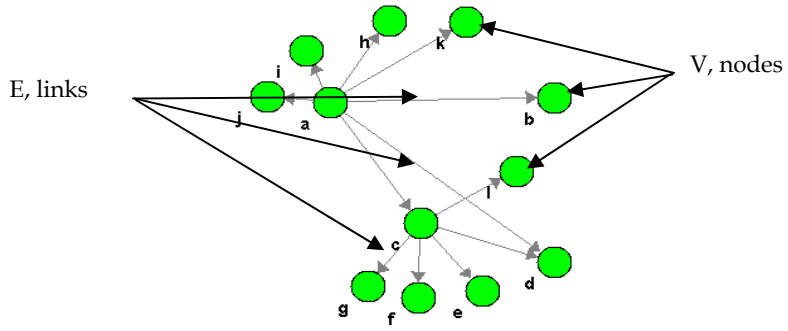

Fig. 1. Networks that represent as a graph

3. Improved Genetic Algorithm

As claim by Zhu (Zhu et al., 2007), a traditional and very important technique in evolutionary computing (EC) is genetic algorithm (GA). GA are not particularly a learning algorithms but they offer a powerful and domain-independent search capability that can be used in many learning tasks, since learning and self-organization can be considered as optimization problems in many cases. Nowadays, GA have been applied to various domain, including timetable, scheduling, robot control, signature verification, image processing, packing, routing (Selamat, 2005), pipeline control systems, machine learning (Bies, 2006) (Goldberg, 1989) and information retrieval (Zhu, 2007) (Selamat, 2005) (Koorangi).

Genetic algorithms (GA) are not new to information retrieval. So, it is not surprising that there have recently appeared many applications of GA's to IR. Genetic algorithm (GA) is an evolutionary algorithm that used for many functions such as optimization and evolves the problem solutions (Luger, 2002). GA used fitness function to evaluate each solution to decide whether it will contribute to the next generation of solutions. Then, through operations analogous to gene transfer in sexual reproduction, the algorithm creates a new population of candidate solutions (Luger, 2002). Figure 2 shows the basic flow of genetic algorithm process.

Fitness function evaluates the feature of an individual. It should be designed to provide assessment of the performance of an individual in the current population. In the application of a genetic algorithm to information retrieval, one has to provide an evaluation or fitness function for each problem to be solved. The fitness function must be suited to the problem at hand because its choice is crucial for the genetic algorithm to function well.

Jaccard coefficient is used in this research to measure the fitness of a given representation. The total fitness for a given representation is computed as the average of the similarity coefficient for each of the training queries against a given document representation (David, 1998). Document representation evolves as described above by genetic operators (e.g. crossover and mutation). Basically, the average similarity coefficient of all queries and all document representations should increase.

Text-based search system is used for constructing root set about user query. However, the root set from text-based search system does not contain all authoritative and hub sources about user query (Kim, 2007). In order to optimize the result, we are using the genetic

algorithm that works as a keyword expansion whereby it expends the initial keywords to certain appropriate threshold.

```
Input: N, size of population;
Pc, Pm: ratio of crossover and mutation.
Output: an optimization solution

Procedure:
Begin
Initialize with population of size N at generation t=0;

Repeat while (|Pop (t+1)|<N)
Select two parent solutions from Pop (t) by fitness function
Copy the selected parents into child solutions (Cloning process)
Randomly generated a number, r, 0≤r≤1

If(r<Pc)
        Carry out crossover operation on the child solution
        Randomly generated a number, r, 0≤r≤1
If(r<Pm) // in this process mutation are set to 0 to prevent changes, Pm=0;
        Carry out mutation operation on the child solution
        Place the child solution in the new population, denoted by Pop (t+1)
End While

Until termination conditions are satisfied return the best solution in Population
End.
```

Fig. 3. Improved genetic algorithm pseudocode

3.1 Process in Improved Genetic Algorithm (IGA)

The main difference between GA and IGA is how to generate new individuals in the next population. We combine two mechanisms to generate new individuals. IGA used the Jaccard coefficient (formula 1) since the vector space model (VSM) has been used in this research.

$$\frac{1}{n} \bullet \sum_{k=1}^{n} \frac{\left| d_j \cap d_q \right|}{\left| d_j \cup d_q \right|} \tag{1}$$

Then, we implement the elitism process to the selected best chromosomes (parents) and clone them into the appropriate size of population. The main purpose of using the elitism is to maintain the best parents and keep the population in the best solution until the end of the optimization process.

We proceed to the cloning process to keep the child as same as the best parents. After that, we used two point crossover and mutation to prevent the solution stuck at the local optimum. The process is repeated until the stopping conditional is fulfilled.

In addition, relevance feedback is used because it is one of the techniques for improving retrieval effectiveness. The user first identifies some relevant (D_r) and irrelevant documents (D_{ir}) in the initial list of retrieved documents and then the system expands the query, q by extracting some additional terms from the sample relevant and irrelevant documents to produce q_e.

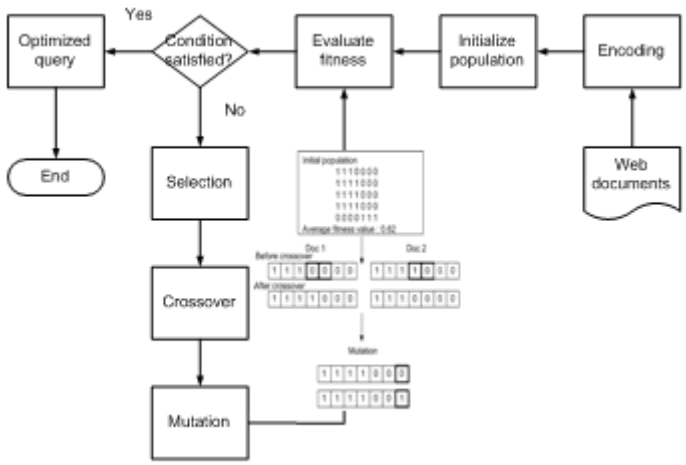

Fig. 4. Improved genetic algorithm flow chart design

4. Experimental Setup

We retrieved the web pages of business networks that related to Iskandar Malaysia (Table 1). The seed URLs are retrieved from the website and several URLs need to be retrieved from each of the URL. The related web pages can be defined in many categories such as ICT or computers, government, bank and etc. There are several processes involve in this research such as initialization, web crawling, optimization and visualization. Below are the details about the processes:

4.1 Initialization

Crawling process start with defines the initial seed URLs to explore the related business web pages from the Internet. The list of URLs is obtained from the Iskandar Malaysia website. The business web pages can be defined in many categories such as ICT or computers, government, universities, bank and etc. Table 1 shows some examples of related URLs from Iskandar Malaysia's web pages.

No	Categories	URLs
1	ICT/ computer / information technology	http://www.msc.com.my
2	Government/ business areas	http://www.iskandarjohoropen.com http://www.khazanah.com.my http://www.epu.jpm.my http://www.kpdnhep.gov.my http://www.mida.gov.my http://www.kpkt.gov.my http://www.imi.gov.my http://www.customs.gov.my http://www.jpj.gov.my http://www.jkr.gov.my http://www.marine.gov.my http://www.rmp.gov.my http://www.nusajayacity.com http://www.ptp.com.my http://www.iskandarinvestment.com http://www.cyberport.my http://www.royaljohorcountryclub.com
3	Bank	http://www.bnm.gov.my
4	Tourism	http://www.tourismjohor.com http://www.dangabay.com

Table 1. Related URLs from Iskandar Malaysia's web pages

4.2 Web crawling

Crawler will take place on retrieved the related business web pages after initialized the seed URLs. The crawler will use the breadth-first search technique.

4.3 Optimization

Optimization is the process of making something better. The advantages of optimization are to save the building time and memory. In this phase, GA is used to select the best result in the searching process whereby the keyword entered by the user will be expanded to produce the new keyword. In the improved genetic algorithm we set the parameter slightly different from the conventional genetic algorithm. Table 2 is details on paramater setting for improved genetic algorithm compared to previous genetic algorithm and Table 3 shows some example of user queries.

Techniques	Population	Generation	Crossover rate, Pc	Mutation rate, Pm	Elitism
GA	5	5	0.4	0.005	No
IGA	16	100	0	0	Yes

Table 2. Setting paramaters for improved genetic algorithm

Queries	Information	Expanded queries
Q1	iskandar malaysia development	IRDA
Q2	iskandar	IRDA, Malaysia, johor
Q3	iskandar malaysia	IRDA, development
Q4	iskandar johor open	Johor, Iskandar
Q5	IRDA iskandar johor	IRDA

Table 3. Example of user queries and expanded queries found by the system

The detail processes in the system are as below:

1. User enter query into the system.
2. Match the user query with list of keywords in the database.
3. Results without GA are represented to the users.
4. Used user profiles when selecting the relevant results found by the system.
5. Encode the documents retrieved by user selected query to chromosomes (initial population).
6. Population feed into genetic operator process such as selection, crossover and mutation.
7. Repeat Step 5 until maximum generation is reached. Then, get an optimize query chromosome for document retrieval.
8. Decode optimize query chromosome to query and retrieve new document (with GA process) from database.

Most of the information in the Internet is in the form of web texts. How to express this semi-structured and unstructured information of Web texts is the basic preparatory work of web mining (Song, 2007). Vector space model (VSM) is one of the most widely used model in the application of GAs to information retrieval. In this research, VSM has been chosen as a model to describe documents and queries in the test collections. We collect the data from the (Iskandar Malaysia, 2009) to retrieve the related web pages link to it.

4.5 Term Vectorization and Document Representation

Before any process can be done, we first implement the pre-processing to the retrieve data. To determine the documents terms, we used procedure as shows in Fig. 4. Vector space model (VSM) is one of the most widely used models in the application of GAs into information retrieval. Thus, VSM has been chosen as a model to describe documents and queries in the test collections. Let say, we have a dictionary, D;

$$D = \left(t_1, t_2, ..., t_i\right) \tag{2}$$

where i is the number of distinguished keywords in the dictionary. Each document in the collection is described as i-dimensional weight vector;

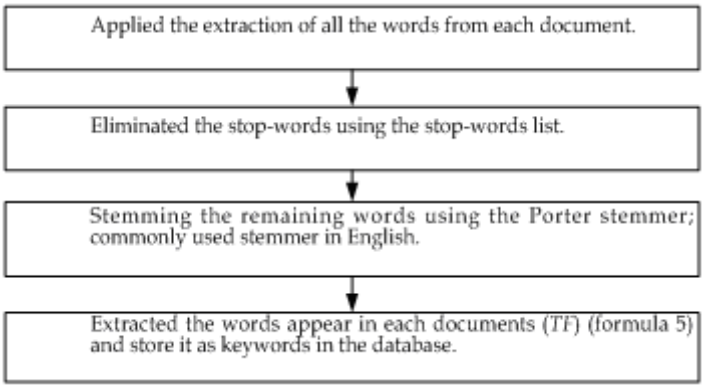

Fig. 4. Pre-processing procedure

$$d = (w_1, w_2, .., w_i) \tag{3}$$

where w_j represents the weight of the j^{th} keyword t_j for $j = 1,2,...i$ and is calculated by the Term Frequency Inverse Document Frequency (TF.IDF) method. Each query in the collection is also described as a weight vector, *q*;

$$q = (u_1, u_2, .., u_i) \tag{4}$$

where u_j represents the weight of the j^{th} keyword t_j for $j = 1,2,...i$ and is calculated by the Term Frequency (TF) method.

$$tf_{ij} = \frac{f_{ij}}{\max\{f_1, f_2, .., f_{|v|j}\}} \tag{5}$$

Based on natural selection in environments and natural genetics in biology, the GA is evolved according to the principle of survival of the fittest and is mostly applied in many optimization problems. When applying the binary GA to the document classification, most research uses gene positions in the chromosome to represent candidate keywords. In this paper, we used GA as an optimization method to expanding the keywords to form new queries. Basically, genetic algorithm is the heuristics search that previously applied in data mining (Chou et al., 2008). Each term is represented as a vector. Given a collection of documents *D*,

$$letV = \{t_1, t_2, .., t|V|\}; V = terms \tag{6}$$

be the set of distinctive words/terms in the collection. A weight $w_{ij} > 0$ is associated with each term t_i of a document $d_j \in D$. For a term that does not appear in document

$$d_j, w_{ij} = 0; d_j = \left(w_{1j}, w_{2j}, .., w|V|_j\right) \tag{7}$$

Then, the terms are encoded as chromosome such as:

Doc1 = 0000001000001000

Doc2 = 0110000110100000

Doc3 = 1000001000000100

Doc4 = 0001100011010111

Doc5 = 0000011000000100

These chromosomes are called the initial population that will be feed into genetic operator process. The length of chromosome depends on number of keywords of documents retrieved from user query. From our example the length of each chromosome is 16. We used fitness function to evaluate how good the solution will be. After evaluated the population's fitness, the next step is the chromosome selection. Satisfied fitness chromosomes are selected for reproduction. Poor chromosomes or lower fitness chromosomes may be selected a few or not at all.

4. Results and discussion

From Table 4, we can see how slightly the results between the conventional GA and IGA. This is because of the small improvement made to the conventional GA. 0.85% of the precision average increase and 0.37% of the recall average increase from the conventional GA to IGA. However Q2 and Q5 do not shows any improvement. These results shows that the IGA can perform better than conventional GA even the results are slightly different. From the results also, we can detected the community of the Iskandar Malaysia since the results can be expanded to many other terms that related to Iskandar Malaysia.

Queries	GA			IGA		
	P	R	F1	P	R	F1
Q1	98.01	49.50	65.78	99.34	50.17	66.67
Q2	100.00	100.00	100.00	100.00	100.00	100.00
Q3	95.86	50.00	65.72	96.05	50.15	65.89
Q4	51.22	33.87	40.78	53.95	34.91	42.39
Q5	100.00	100.00	100.00	100.00	100.00	100.00

Table 4. Results on conventional genetic algorithm and improved genetic algorithm

5. Case Study: Iskandar Malaysia

Iskandar Malaysia (formerly known as Wilayah Pembangunan Iskandar) is one of the well-known community company that closely related to social communities in Malaysia. There are five existing clusters within Iskandar Malaysia that are mostly not fully developed. The most developed is the electrical and electronics (E&E) cluster which as noted, is actually an integral part of the Singapore E&E cluster, though the part within Iskandar Malaysia occupies the lower end of the value chain.
Fig. 5 shows an example of the main networks of Iskandar Malaysia whereby this network will brings many advantages to the country in order to improve the quality of the social networking in business matter.

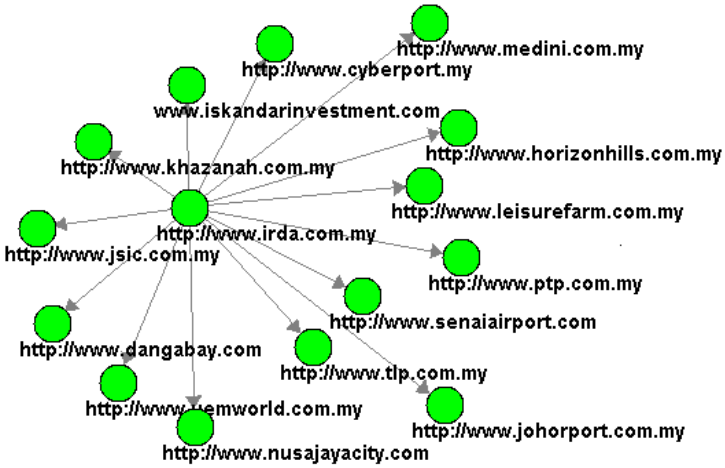

Fig. 5. Business networks of Iskandar Malaysia

Table 1 shows several numbers of related URLs for Iskandar Malaysia. Most of the URLs are connected to the governments and business areas web pages. So, logically we found that the business networks can be spreading into several fields that can attract investors to come and invest in Malaysia. As stated before, our main objective is to show the related industries or links connected to this web page so that we can see the total or statistics of the investors and profits of the business areas.
From the results, we can see the expansion of the related web pages from the first URL. In Fig. 6 we show that the networks can be spreading into many categories and influence the business industries improvement. This result clearly explained that the objective to see the related industries can be achieved by using the mention technique.

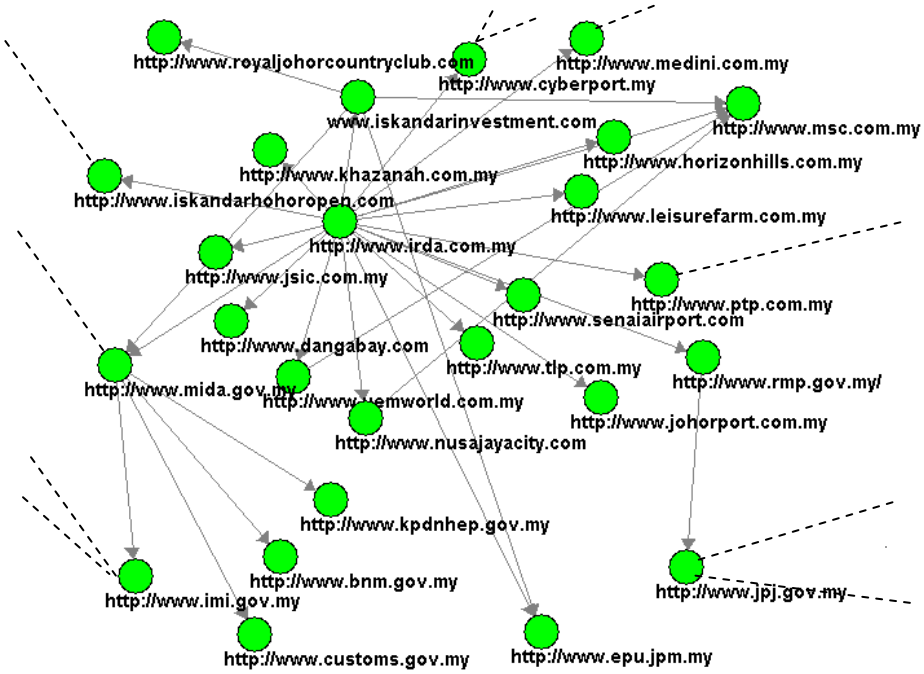

Fig. 6. Expansion of business networks of Iskandar Malaysia

6. Conclusion

In this chapter, we have shown how an evolutionary algorithm can help to reformulate a user query to improve the results of the corresponding search. The evolutionary algorithm is in charge of selecting the appropriate combination of terms for the new query. To do this, the algorithm uses fitness function a measure of the proximity between the query terms selected in the considered individual. Then, the top ranked documents are retrieved using these terms. We have carried out some experiments to have an idea of the possible improvement that the IGA can achieve. In these experiments, we have used the precision obtained from the user relevance judgments as fitness function. Results have shown that in this case, the IGA achieve a slight improvement compared to the conventional GA. However, we want to emphasize that this feedback mechanism improves the search system by considering users suggestions concerning the found documents, which leads to a new query using IGA. In the new search stage, more relevant documents are given to the user. As a conclusion, the conventional genetic algorithm model was improved. In the future, we hope that the system can be improved further and the results can achieve higher accuracy rate in solving the data mining problems.

7. References

Bies, Robert R; Muldoon, Matthew F; Pollock, Bruce G; Manuck, Steven; Smith, Gwenn and Sale, Mark E (2006). A Genetic Algorithm-Based, Hybrid Machine Learning Approach to Model Selection. *Journal of Pharmacokinetics and Pharmacodynamics*: 196-221. Netherlands: Springer.

Chih-Hsun Chou , Chang-Hsing Lee , and Ya-Hui Chen (2008). GA-Based Keyword Selection for the Design of an Intelligent Web Document Search System. The Computer Journal Advance Access published on October 14.

David A. Grossman, Ophir Frieder (1998). Information Retrieval: Algorithms and Heuristics. *Springer*.

iProspect.http://www.midiaclick.com.br/wp-content/uploads/2008/10/researchstudy_apr2008_blendedsearchresults.pdf. Accessed on October 2008.

Goldberg, David E (1989), *Genetic Algorithms in Search, Optimization and Machine Learning*, Kluwer Academic Publishers, Boston, MA.

Iskandar Malaysia Development. *www.idr.com.my*. Accessed on January 2009.

Kim, K. and Cho, S. (2007). Personalized mining of web documents using link structures and fuzzy concept networks. *Appl. Soft Comput.* 7, 1 (Jan. 2007), 398-410.

Luger, F. G. (2002). Artificial Intelligent Structures and Strategies for Complex Problem Solving. 4th Ed. Edinburgh Gate, England: Eddison Wesley.

M. Koorangi, K. Zamanifar, A distributed agent based web search using a genetic algorithm, IJCSNS, *International journal of computer science*

Pizzuti, C. (2008). GA-Net: A Genetic Algorithm for Community Detection in Social Networks. *Springer,* 2008, 5199, 1081-1090.

Selamat, A. and Selamat, M. H. (2005). Analysis on the performance of mobile agents for query retrieval. *Inf. Sci. Inf. Comput. Sci*. 172, 3-4 (Jun. 2005), 281-307.

Song Liangtu, Zhang Xiaoming (2007). Web Text Feature Extraction with Particle Swarm Optimization. *IJCSNS International Journal of Computer Science and Network Security*, Vol. 7 No. 6 pp. 132-136.

Zhengyu Zhu, Xinghuan Chen, Qingsheng Zhu, Qihong Xie (2007): A GA-based query optimization method for web information retrieval. *Applied Mathematics and Computation* 185(2): 919-930.

Knowledge Management

Edited by Pasi Virtanen and Nina Helander

ISBN 978-953-7619-94-7
Hard cover, 272 pages
Publisher InTech
Published online 01, March, 2010
Published in print edition March, 2010

This book is a compilation of writings handpicked in esteemed scientific conferences that present the variety of ways to approach this multifaceted phenomenon. In this book, knowledge management is seen as an integral part of information and communications technology (ICT). The topic is first approached from the more general perspective, starting with discussing knowledge management's role as a medium towards increasing productivity in organizations. In the starting chapters of the book, the duality between technology and humans is also taken into account. In the following chapters, one may see the essence and multifaceted nature of knowledge management through branch-specific observations and studies. Towards the end of the book the ontological side of knowledge management is illuminated. The book ends with two special applications of knowledge management.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821