UTMCrawler: Crawling the E-business Social Network Using Genetic Algorithm for Relevant Document Searching

Siti Nurkhadijah Aishah Ibrahim, Ali Selamat, Mohd Hafiz Selamat
Faculty of Computer Science and Information Systems,
Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia.
echoas1306@yahoo.com, aselamat@utm.my, mhafiz@utm.my

ABSTRACT

The increasing of online social network in the Internet has caused the explosion of search results from the search engines. According to the Google search engine statistics, until 2008 almost 1 trillion web pages have been indexing including the online social network website. Thus, how can we retrieve the massive online social network information with the exploded information accessible in the web? In this paper, we have designed the internet agent/ crawler based genetic algorithm to retrieve the e-business web pages from the lelong.com.my, the Malaysia online auction website. We used genetic operation in order to retrieve the information connected between the users by expanding the keywords. Our result shows that the genetic algorithm can be a promising technique in terms of accuracy of the retrieval results.

1. INTRODUCTION

The rapid growth of text documents in digital form in the Internet has increased the importance of using methods to analyze the content of text documents. Current search engines that rely on keyword matching usually return too many low quality matches. To make matters worse, some advertisers attempt to gain people's attention by taking measures meant to mislead search engines. As a result, the identification and classification of text documents based on their contents are becoming very necessary (Page, 1998).

Nowadays, there has been tremendous expansion of online auction activities over the last several years with millions of people buying and selling goods online (Beyene, 2008). eBay, Lelong and Yahoo are some of the major well-known online auction sites. On any given day, eBay has more than a 100 million items available for sale, with 6.4 million new items added every day. eBay users worldwide trade more than $1,812 worth of goods on the site every second.

In order to prevent auction fraud, reduce the buyers' costs and increase sellers' competence, increasing numbers of researchers have begun to study issues surrounding on-line auctions (Beyene, 2008) (Chau, 2007) (Hsien, 2008). In recent years, empirical research on on-line auction has been flourishing because of the availability of large amounts of high-quality bid data from on-line auction sites. Researchers want to study price, bid behavior and auction market characteristics to locate suggestions for buyers and sellers participating in on-line auctions (Hsien, 2008). However, the increasingly large information of bid data has made data collection increasingly complex and time consuming, and there are no effective resources that can support this work. So, in this paper, we focus on the multi crawling and capturing of on-line auctions from the automatic agent perspective to help researchers collect auction data more effectively.

This paper is organized as follows. Section 2 provides the related works on e-business social network. In Section 3, we discuss experimental setting and follow by result analysis and discussion in Section 4. The conclusion is summarized in Section 5.

2. E-BUSINESS SOCIAL NETWORKS

An online social network consists of individuals who are linked to each other in the same network. Some well-known examples of online social network include Facebook and Friendster which helps build personal and professional relationship. Facebook is also inserting advertisements into its social graph (the feed of activities from friends) attaching ads related to activity information from them. Online auction sites, in fact also social network where the auction users take part in buying and selling activities (Fig. 1). Indeed, Lelong (lelong.com.my) which is the Malaysia's largest and most visited trading and auction portal has many registered users for online social network.

Online social networks are part of the web, where a lot of interesting phenomena take place; and many of them are worth studying or need to be considered. For example, on an online auction, we may want to find out the patterns of fraudulent or suspicious transactions among users (Beyene, 2008).

However, before any analysis can be done, we need to gather the data that describes the social networks. These networks are often huge, therefore crawling these networks could be both challenging and interesting. However, there

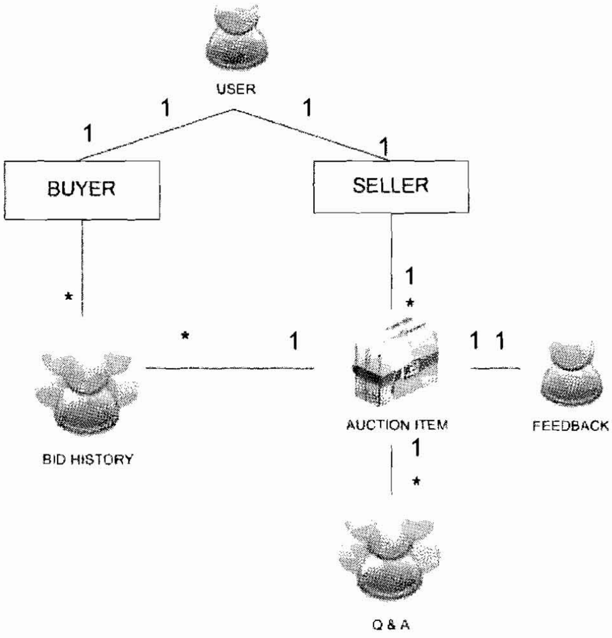


Fig. 1 Online auction structure

has been little documented work on the crawling of these data.

2.1 Crawling *lelong.com.my* Social Network

Lelong (lelong.com.my), which is the Malaysia's largest and most visited trading and auction portal, has over 30 Million page views per month. It also has over a million visitors per month and has been established more than 10 years. At lelong.com.my, more than 70,000 products traded in all stores. Moreover, they gained almost a total of RM 10 million transactions each month.

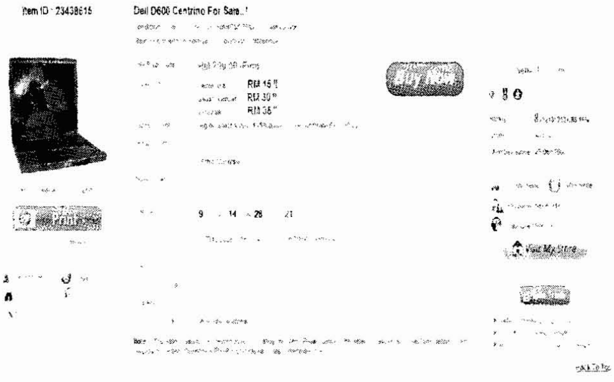
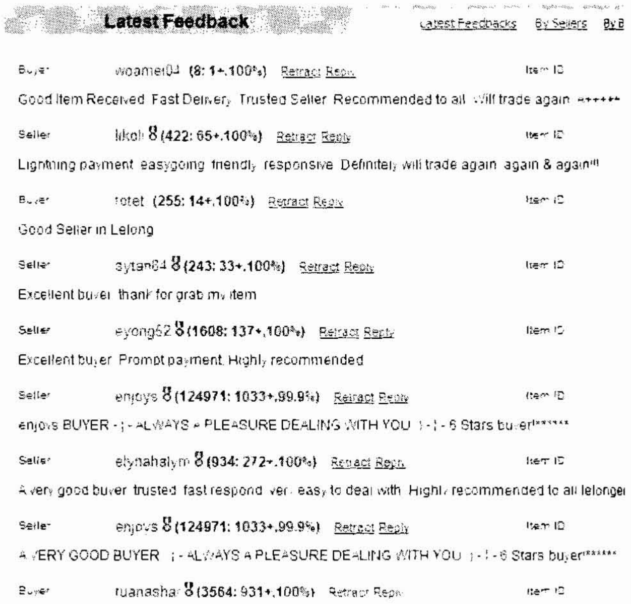


Fig. 2 Lelong user's profiles

We assumed that this portal can be benefit to spreading the Malaysia products to the whole world and at same time making the social network with users. As we concerned, we believed that the data represented in the online social network are very different from general web pages. The web pages that describe an individual in an auction social network are typically well-structured, as they are usually automatically generated, unlike general web pages which could be authored by any person. Therefore, we can be very certain about what pieces of data we can obtain after crawling a particular individual's web pages. In this paper, we use Lelong as an example of online social network.



Latest Feedback   Latest Feedbacks   By Sellers   By B

Buyer   woamei04 (8: 1+,100%)   Retract Reply   Item ID
Good Item Received Fast Delivery Trusted Seller Recommended to all Will trade again ++++++

Seller   likoli 8 (422: 65+,100%)   Retract Reply   Item ID
Lightning payment easygoing friendly responsive Definitely will trade again again & again!!!

Buyer   fotet (255: 14+,100%)   Retract Reply   Item ID
Good Seller in Lelong

Seller   sytan64 8 (243: 33+,100%)   Retract Reply   Item ID
Excellent buyer thank for grab my item

Seller   eyong62 8 (1608: 137+,100%)   Retract Reply   Item ID
Excellent buyer Prompt payment. Highly recommended

Seller   enjoys 8 (124971: 1033+,99.9%)   Retract Reply   Item ID
enjoys BUYER - ; - ALWAYS A PLEASURE DEALING WITH YOU) - ! - 6 Stars buyer!******

Seller   elynahalym 8 (934: 272+,100%)   Retract Reply   Item ID
A very good buyer trusted fast respond very easy to deal with Highly recommended to all lelonger

Seller   enjoys 8 (124971: 1033+,99.9%)   Retract Reply   Item ID
A VERY GOOD BUYER ; - ALWAYS A PLEASURE DEALING WITH YOU) - ! - 6 Stars buyer******

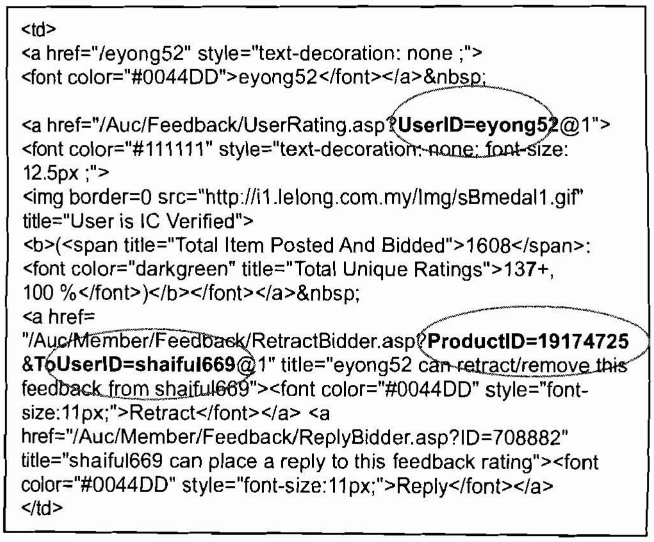
Buyer   ruanasha 8 (3564: 931+,100%)   Retract Reply   Item ID

Fig. 3 List of user feedback in user profiles

Fig. 2 shows the profile of a user in Lelong. From the page, we can obtain the user's ID, date of membership, location, item's offered by users ad other personal details. A part from the personal local information, there are usually explicit links (e.g.; ratings, bidding, etc.) that we can use to trace the user's connections to the others.

Referring to Fig. 3, the list of feedback received by the user is shown, including the IDs of the users who left the feedback, which are hyperlinked to those users' profile pages. Thus, by crawling these hyperlinks, we can construct the graph of connections between all the users in the social network. In addition, the users can add their own favorite seller whereby this function can increase their social network between the buyers / seller. From those hyperlinks, we also considered the explicit hyperlinks whereby it can provide us with related promotion products between users.

Web crawlers are one of the most crucial components in search engines and their optimization would have a great effect on improving the searching efficiency.



```
<td>
<a href="/eyong52" style="text-decoration: none ;">
<font color="#0044DD">eyong52</font></a>&nbsp;

<a href="/Auc/Feedback/UserRating.asp?UserID=eyong52@1">
<font color="#111111" style="text-decoration: none; font-size:
12.5px ;">
<img border=0 src="http://i1.lelong.com.my/Img/sBmedal1.gif"
title="User is IC Verified">
<b>(<span title="Total Item Posted And Bidded">1608</span>:
<font color="darkgreen" title="Total Unique Ratings">137+,
100 %</font>)</b></font></a>&nbsp;
<a href=
"/Auc/Member/Feedback/RetractBidder.asp?ProductID=19174725
&ToUserID=shaiful669@1" title="eyong52 can retract/remove this
feedback from shaiful669"><font color="#0044DD" style="font-
size:11px;">Retract</font></a> <a
href="/Auc/Member/Feedback/ReplyBidder.asp?ID=708882"
title="shaiful669 can place a reply to this feedback rating"><font
color="#0044DD" style="font-size:11px;">Reply</font></a>
</td>
```

Fig. 4 User information in Lelong website that will be use in the crawling process

We used crawler for web retrieval to make a fast crawling process yet gather the web documents as many as it can. Our crawler consists of a user interface for URLs submission and a database to store the information retrieved by the crawlers (Ibrahim, 2008). From the user interface, the seed URLs is distribute to multiple crawler agent to crawl and retrieve the e-business information from the Internet. The crawlers are using the breadth-first search technique so that it can fetch as many web pages as it can and will not be stuck in blind alleys, and will always find the shortest path first (Pant, 2005). Fig. 5 shows our proposed crawling design for an online auction social network.

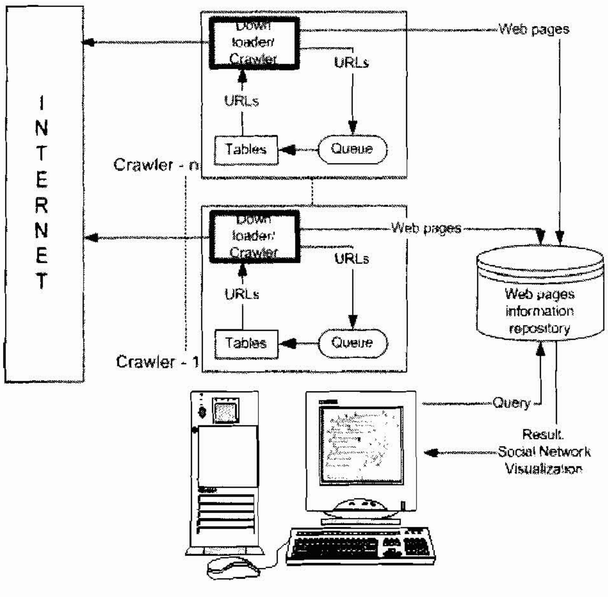


Fig. 5 Our proposed online auction crawler

2.2 Genetic Algorithm as Crawling Search Algorithm

We crawled these user data in a breadth-first search. We used a queue data structure to store the list of pending users which had been seen but not yet crawled. Initially, a seed set of users was inserted into the queue. Then at each step, the first entry of the queue is popped and all feedbacks for that user were crawled. Every user who had left a feedback was crawled based on queue. Once all of the user's feedbacks were crawled, the user was marked as visited and stored in a separate queue.

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 widely 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,

$$\text{let } V = \{t1, t2, \ldots t_{|V|}\};\ \text{V= terms}$$

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

$$d_j,\ wij = 0;\ d_j = (w_{1j}, w_{2j}, \ldots, w_{|V|j})$$

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.

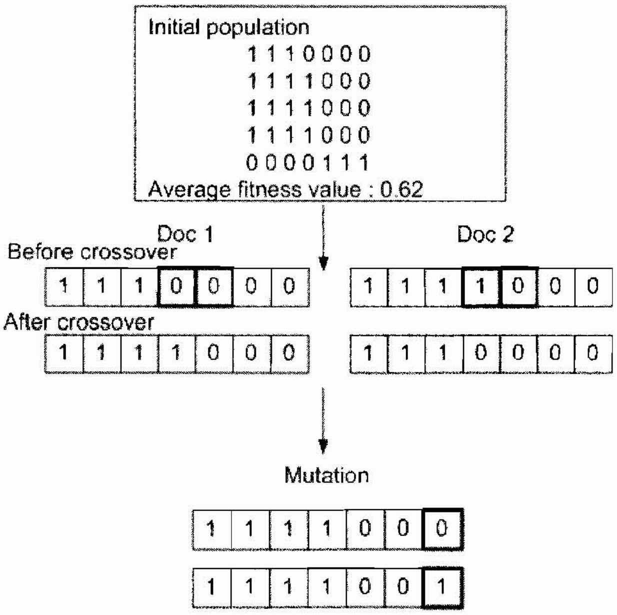


Fig. 6 GA process in web crawling

Crossover is the genetic operators that mix two chromosomes together to form new offspring. Genetic algorithm constructs a better solution by mixture good characteristics of chromosomes together. Higher fitness chromosomes have an opportunity to be selected more than the lower ones, so good solution always alive to the next generation. Crossover technique includes one point crossover, two point crossovers and multiple point crossovers. However, we are choosing the point at random for crossover because the structure is represented as binary strings.

After the selection and crossover process, we implemented the mutation process whereby it involves the modification of the values of each gene of a solution with some probability. In accordance with changing some bit values of chromosomes, it gives the different breeds. Chromosomes may be better or poorer than old chromosomes. If they are poorer than old chromosomes, they will be eliminated in selection step. The main objective of mutation process is restoring lost and exploring variety of data.

3. EXPERIMENTAL SETUP

In this work, we have applied the proposed crawling framework as shown in Fig. 5. Data preparation is a step that data of interest or raw data are identified and collected.

1. Data preparation

All related data of user's profile from lelong web pages are retrieved. The data that is known as the initial seed set will be used to form the social network. Fig. 4 shows some useful and relevant information that will be use in the crawling process

2. Preprocessing

In the preprocessing stage, the html syntax is removed. For instance <html></html>, <head></head> and <title></title>. We used stopping function to eliminate the most using keywords such as 'a', 'to', 'the', 'she' and etc. Then, the Porter stemmer (Porter, 1980) library is used in order to stem the long words into their root word such as 'enjoyable' = 'enjoy', 'supplying' = 'supply' and so on. Finally, the output from the preprocessing data is stored in the database for later processing.

3. Processing

GA was used to expand the keywords retrieved from the lelong user's profile. GA will automatically generate the new keywords based on the parameter shows in Table 1. The GA process is shows as in Fig. 6. The next step is to crawl the lelong web pages based on the new keywords generates by GA.

Table 1 Parameter use in GA Process

parameter	value
Total chromosomes in a population	5
Total generations in GA process	5
Probability of crossover, *Pc*	0.4
Probability of mutation, *Pm*	0.005

3.1 Results and Discussion

Based on our findings (Fig. 7), the expansion of keywords found by the crawler can obtain such a relevant or online social network between users. However, the results is not really extensive and convincing where the results only relevant to a small corpus of web documents. Thus, we are encouraged to do extra experiment of more web documents to extend the results.

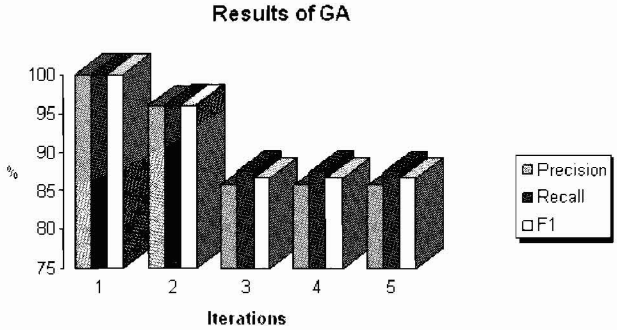


Fig. 7 Results for crawling results

Table 2 shows some sample results that can be obtain in the Lelong website. The connection between each user can be shown to analyze how the seller and buyer can interact with each other in order to successfully achieve what they aim for.

Table 2 Sample results

user	link with
e-Nabill	ssadikun nee29 akuamuscute csyedz tuanashar enghua icellular Mastereos
joow	gn175546 newspeed bryanwern8
shaiful669	woamei04 likoli totet sytan84 eyong52 enjoys elynahalym

4. CONCLUSION AND FUTURE DIRECTION

In this paper, we investigated the use of keywords in web pages retrieval using GA for keywords expansion. We proposed a new approach to crawl the interest and connection between web-user in online auction social network. In the future, we intend to extend our findings into the other types of recognizing the online auction social network. We hope that our works can contribute something to the social communities whereby this research can be more benefit to the online social network users.

5. REFERENCES

Beyene, Y.; Faloutsos, M.; Duen Horng Chau; Faloutsos, C., "The eBay graph: How do online auction users interact?," *Computer Communications Workshops, 2008. INFOCOM. IEEE Conference on* , vol., no., pp.1-6, 13-18 April 2008.

Brin, S. and Page, L., The anatomy of a large-scale hypertextual Web search engine. *Comput. Netw. ISDN Syst.* 30, 1-7 (Apr. 1998), 107-117, 1998.

Chau, D. H., Pandit, S., Wang, S., and Faloutsos, C., Parallel crawling for online social networks. In *Proceedings of the 16th international Conference on World Wide Web* (Banff, Alberta, Canada, May 08 - 12, 2007). WWW '07. ACM, New York, NY, 1283-1284, 2007.

Cheng-Hsien Yu, Shi-Jen Lin: Parallel Crawling and Capturing for On-Line Auction. *ISI Workshops 2008*: 455-466, 2008.

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

G. Pant, P. Srinivasan. Learning to Crawl: "Comparing Classification Schemes". *ACM Transactions on Information Systems*, Oct 2005.

M.F. Porter, An algorithm for suffix stripping, *Program*, 14(3) pp 130−137, 1980.

Pandit, S., Chau, D. H., Wang, S., and Faloutsos, C., Netprobe: a fast and scalable system for fraud detection in online auction networks. In *Proceedings of the 16th international Conference on World Wide Web* (Banff, Alberta, Canada, May 08 - 12, 2007). WWW '07. ACM, New York, NY, 201-210, 2007.

S. N. A, Ibrahim and A. Selamat. "Multi-agent Crawling System for Effective Web Retrieval". *3rd Postgraduate Annual Seminar (PARS' 2007)*, FSKSM, UTM, July 4-5 2007.

Facebook. http://www.facebook.com. 2008.

Friendster. http://www.friendster.com. 2008.

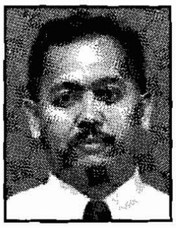
Lelong. http://www.lelong.com.my. 2008.

6. ACKNOWLEDGMENT

We appreciate SEATUC'09 reviewers for their helpful suggestion which improve the quality of this work. We also wish to acknowledge that this work is supported by the Ministry of Science & Technology and Innovation (MOSTI), Malaysia and Research Management Center, Universiti Teknologi Malaysia (UTM) under the Vot. 79267.

S.N.A. Ibrahim has received a B.Sc. (Hons.) in Computer Science from Universiti Teknologi Malaysia (UTM) and currently in M.Sc. by research in computer science under Assoc. Prof. Dr Ali Selamat supervision's at the same university since 2007 until now. She has been published several IEEE conferences paper and attending related courses in computer science to enhance her research skills. Her research interests include software engineering, crawler agents, information retrievals, social network, e-business and genetic algorithms.

Ali Selamat has received a B.Sc. (Hons.) in IT from Teesside University, U.K. and M.Sc. in Distributed Multimedia Interactive Systems from Lancaster University, U.K. in 1997 and 1998, respectively. He has received a Ph.D. degree from Osaka Prefecture University, Japan in 2003. Currently, he is an associate professor and IT Manager at the School of Postgraduate Studies (SPS), UTM. His research interests include software engineering, software agents, web engineering, information retrievals, genetic algorithms, neural networks and soft computing.



Md. Hafiz Selamat has received B.Sc. (Hons.) and M. Sc. in Computer Science from Universiti Teknologi Malaysia (UTM) in 1988 and 1998, respectively. Currently, he is a lecturer at the Department of Information Systems, Faculty of Computer Science & Information Systems, Universiti Teknologi Malaysia (UTM) since 2000. Currently, he is pursuing a PhD degree at International Islamic University, Malaysia. His research interests include mobile agent, information retrieval, and knowledge management.