International Journal of Smart Home
Vol. 9, No. 8 (2015), pp. 187-196
http://dx.doi.org/10.14257/ijsh.2015.9.8.20

A Focused Crawling Method Based on Detecting Communities in Complex Networks

ShenGui-lan, [1, 2]Sun Jie, [2] and Yang Xiao-ping[1*]

[1]*School of Information Renmin University of China, Beijing 100872;*
[2]*Business Collage of Beijing Union University, Beijing 100025*
guilan.shen@buu.edu.cn; jie.sun@buu.edu.cn; yang@ruc.edu.cn

Abstract

The rapid growth of the large-scale World-Wide Web poses great challenge to existing focused crawling methods. Whetheranalyzing text content or link structure, traditional focused crawler were mainly based on the page granularity. Random walking in the network composed of a large number of pages, the focused crawler is easy to get lost. Obviously, narrowing the focused crawling range from the entireWeb can improve the precision and efficiency. A focused crawling method based on the twogranularitiesis put forward. Firstly, using detectingcommunity algorithm to analyze the link structure of the network composed of websites, a given topic web sites group is built up. It contributes to narrow the crawling range. Secondly, all topic relevant analysis for web pages and link prediction are performed inside this generated group. Topic relevant analysis is implemented through calculating the topic similarity for title and content separately. The similarity of father pages, anchor texts and the string text for URL all are considered to predict the topic relevance for unknown links.The experimental results suggest that this method is very effective for given topic, and it can improve the precision.

Key word: *detecting community, focused crawling, web site granularity,similarity analysis, link precision*

1. Introduction

With the explosive growth of web pages, fetching information about a particular topic is becoming more and more important. As a web crawler always downloads just a fraction of web pages, it is highly desirable that the downloaded fraction contains the most relevant pages and not just a random sample[1]. Focused crawling is a kind of technology that try to download only those web pages that are relevant to predefined topic or several topics rather than crawling the entire WWW. In fact, a variety of techniques are required to avoid gathering non-topical resources when web resources are very huge.In order to improve the efficiency of large-scale automatic crawling, we argue that we should be more concerned crawling range instead of the content analysis, and link prediction, which were paid attention to traditionally. The recent studies[2][3] have shown the link distribution of web resources with special topic has the topic locality phenomenon, that is, the similar topical web sites or web pages often are closely linked together. It was obvious that predetermined the overall distribution of the similar topical web sites that may enable web crawlers to effectively narrow crawling range. Therefore, we summarize the focused crawling as two key issues: the first issue is that how to narrow the crawling range from the entire WWW, and the second one is content analysis and link prediction within the narrowed range.

We propose a new approach that is based on detecting local communities in complex networks whereby to narrow the crawling range. Through this method, we can find those web sites related to closely which are relevant to predefining topic or topics, and then we perform topic relevant analysis for web pages and link prediction within those web sites.

ISSN: 1975-4094 IJSH

The results of our experiment demonstrate that the proposed approach is helpful for improving the precision.

We proceed to report our work in the rest of the paper as follows. We precede the literature review about focused web crawlingin Section 2. In section 3, we propose our approach for focused crawling based on detecting local communities. In order to verify our approach, we conducted extensive experiments. The experimental design and results analysis are given in Section 4. Finally, we make a conclusion in Section 5.

2. Related Works

A wide range of approaches have been proposed to fetching topical resources from the entire web. Fish Search algorithm [4], Shark-Search algorithm [5] based on web content heuristic method put forward by De Bra and M. Hersovic respectively, which assign priority values to candidate pages using simple keyword matching or Vector Space Model. Classifier prediction methods [6] represented by S.C hakrabarti, network link structure analysis method based on Page Rank algorithm [7] designed by Cho. There are emerging semantic focused crawling and ontology-learning-based focused crawling since semantic technologies provide shared knowledge for enhancing the interoperability between heterogeneous components. Such as, Zhang et al [8] proposed a supervised ontology-learning-based focused crawler; Dong et al[9] present a novel framework of a self-adaptive semantic focused crawler with the purpose of discovering, formatting, and indexing mining topical information over the Internet. Although these ideas use the different analysis methods, such as web content, Hyper Text Marked Language or link graph structure respectively to determine the best subsequent links, but they all inherit the general crawling technology, that is, starting the given seeds, crawler travel all nodes on the entire web as long as there is a connection exists between two nodes. The essence of focused crawling is to download as much as possible relevant network resources depending on link structure from the entire web.

The most available focused crawling technology is in the form of roaming, in crawling granularity, focusing only on web page characteristics such as web content, anchor text. In fact, only by means of them, we can't grasp crawl range in difference perspectives. It is easy to get lost and waste time and network bandwidth. So how to crawl large-scale topical resources effectively has become difficult to realize.

Zhu-min Chen [10] enlarges the traditional crawling granularity, promoted to site granularity. She focuses on URLs priority calculation of site granularity and parallels processing a page and site granularity. At the same time she judges URLs crawl priority of site granularity by the relevant web pages size of the total in the dynamic process. Obviously, she does not distinguish the computing order between site granularity and page granularity. It is unable to narrow the range for crawler to crawl page efficiently.

Flake et al [11-12] first introduce a definition of a web community that may enable web crawlers to effectively focus on narrow but topically related subsets of the web. They adopt a maximum flow/minimum cut framework to identify members of a community, and argue that the identified members are the topic related pages. This method only uses link analysis while ignoring the content feature of web pages. Ding [2] conducted systematic analysis of applying a topology-based community detection approach and a topic-based community detection approach to several networks which nodes contain information and found communities detected by the topology-based community approach tend to contain different topics with each community.

3. Designing Focused Crawling Approach

In this section we present a focused crawling method based on community detection for websites and content analysis and link prediction for webpages. This method can narrow the crawling range in website granularity and calculate the topical similarity in

webpage granularity. We first give the framework of our method in Section 3.1.We describe the topic-oriented community detection method based on website granularity in Section 3.2. Content analysis and link prediction based on web page granularity are presented in Section 3.3

3.1. Framework

The framework for focused Crawling method is illustrated in Figure 1.The method includes two phase crawling tasks. The purpose of the first round of crawling is to get the URL queue based on website granularity, and the second round is to analyze and download the topical resources according to the topic similarity based on webpage granularity. The whole process for focused crawling is divided into 4 key modules.

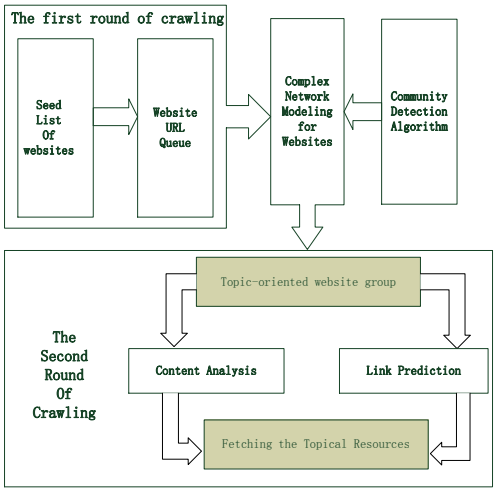

Figure 1. The Framework for the Focused Crawling Method

(1) Complex Network Modeling for Websites: This module aims to structure the crawled website URLs into formal complex network model for processing.

(2) Community Detection Algorithm: According to the link analysis of the websites, we detect the community related to the given topic. The numbers of detected community constitute the topic-oriented website group.

(3) Content Analysis: Based on the vector space model, and taking the different role of webpage content and page title into account, we carry the content analysis for each pages

(4) Link Prediction: We perform the link prediction by means of calculating the similarity of father pages, anchor texts and the string text for the unknown URLs respectively.

3.2. Community Detection based on Website Granularity

Community in complex network is a set of nodes, which has more similarity with each other, but has more different with other parts in the network, that is, the internal nodes in a community link together tightly; however, the connection between the communities is relatively sparse [12]. Topic locality in the specified topical resources is similar to the community structure of complex networks. Local community discovery algorithm regards web site as the basic granularity, it can not only avoid localization of resources acquisition by using Fish Search algorithms, and help to detect a topical information web site group.

3.2.1. Network Model based on Website: In our method, the web sites on the Internet are regarded as the node set V; links among web sites are represented by the edge set E. So,

all web resources on the Internet are represented by a graph G = (V, E). Let Number of nodes n be | V |, number of edges m be | E |. Edge$< i, j >$indicates that site i has a link to site j, here, we ignore the direction of links, that is, G is an undirected graph. See Figure 2.

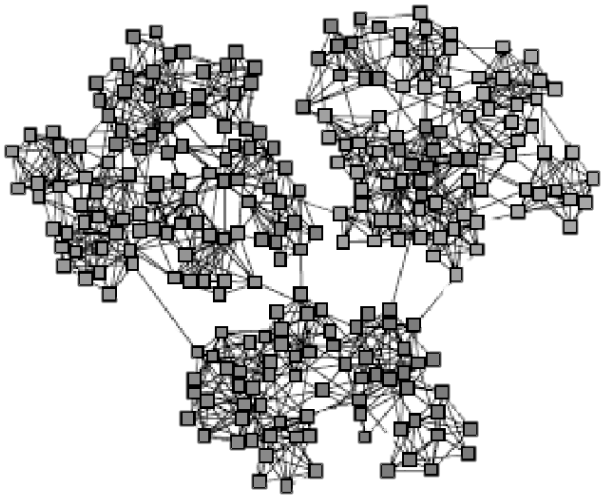

Figure 2. A Complex Network Model based on Websites

3.1.2. Building the Topology of Websites: It contributes to promote the efficiency that if we can grasp the entire topology of web sites on the Internet. However, it is not feasible to build the topology of the entire Web sites at present, and it is also meaning less referred to the focused crawling range. We observe that the relationships among those web sites with the same topic are usually tightly, which are realized by link exchange. In a given web site, link exchanges usually are those web site URLs which have the same topic with the site. But it also includes some web site URLs, for example, advertising links, which have no relationship with this site. So we need select the link exchanges relevant to the website topic.

Starting from the seed site, we extract link exchanges from the crawled home page, and then save the crawled URLs to an unvisited URLs queue. Repeat the same step to handle the home page of unvisited URLs until crawl queue is empty or the number of sites is satisfied.

We call the above processing as the first round of crawl. A complex network consisting of Web sites has been constructed. By extracting link exchanges, we record the parent nodewhich the URL is recommended and content. When extracting only link exchanges of first page of each site, we will analysis of the site in a shorter time and can test the validity of the link at the same time. It can also avoid link cost of late in valid URLs.

3.1.3. Discovering Local Community Structure: In order to further discover those members that linked together tightly in the network G which was built up before, we adopt a detecting community method, and we call it FRTW. At first, we select site V which has high relevance to the given topic in G as the website seed. In FRTW, We perform the following algorithm steps to find the community C.

step1: Select site V and its adjacent node as the initial community members.

step2: Judge whether each initial community members meet the following conditions:$k_v^{in}(C)$which is degree with the current community must be greater than $k_v^{out}(C)$ which is degree with the outside society nodes or the value of $k_v^{in}(C)$is not less than the current corporate scale$\varphi * Size(C)$.

step3: For the node V which does not meet the above conditions, we further to find its

neighbor node F which owing the most edges with the initial community. Condition judgment in thestep2 should be carried out. $k_v^{in}(C)$Mustbe greater than $k_v^{out}(C)$, or the value of $k_v^{in}(C)$is not less than the current corporate scale$\varphi * Size(C)$.

step4: If the node V can be found, we sent node V to Linked C link queue.

step5: Use step 2 to check every neighbor nodes outside community in Linked C. All the nodes satisfying the condition are sent to Linked C. If the condition will check the node into the community link queue Linked C.

step6: Repeat step 5 until no node joins the community.

Where φ is the scale factor, it obtains the value between 0 and 1.A series of tests are carried out to determine the value range of scale factor. Community structure of the node V is composed of nodes stored in the queue Linked C. They are linked to each other and constitute the local community C closely related to topics. According to the topic locality phenomenon, Community C constitutes the desired focused crawling website group.

3.2. Content Analysis and Link Prediction Based On Webpage Granularity

Specifying the crawling range where web resources are collected automatically is the essential condition and important factor in the process of focused crawling. Even had detected the topic-related website group, we cannot download all pages from these websites owing to there are a large number of pages in a target site. Some pages content are relevant to the topic; more over some others are irrelevant to the topic. Irrelevant information on the target site should be filtered for effective collection theme resources. This filtering is based on topic related to web content analysis and topic similarity prediction of link URLs.

3.2.1. Correlation Analysis of Webpage Content: The page theme is expressed by using vector space model based on principles in the page content relevant analysis. The basic idea is to represent vector components as feature item weight. In the vector space model, each document d_i is represented as an n-dimensional feature vectors, namely $< w_{i1}, w_{i2}, \cdots, w_{in} >$. w_{ij}Is denoted as a document weighting factor, considering word frequency, inverse document frequency and the fact that frequency high matching words will drown other matches term, we use the TF-IDF weighting method [13].w_{ij}Is best calculated by the following equation:

$$w_{ij} = \frac{(lgtf_{ij}+1.0)\times idf_j}{\sum_{j=1}^{t}[(lgtf_{ij}+1.0)\times idf_i]^2} \quad \text{(Formula 1)}$$

Where$tf_{ij}$is the number of lexical itemt_j appearing in the documentd_i, $df_j$is denoted as the number of documents containing lexical items t_j, idf_j is expressed as$\lg\left(\frac{d}{df_j}\right)$, d represents the number of crawling document.

As far as web pages are concerned, similarity degree of titles plays an important role on determining whether it is similar to themes. If you ignore the title text, or only put its lexical item and these of the page content in the same processing, it will affect the accuracy of correlation analysis. Vectors subject headingsT_T, content vectorD_T, crawling document title vectorD_C, content of the crawling document vector$D_C$were respectively constructed. By calculating angle cosine between T_T andD_T, $T_C$and D_C,similarity value is obtained .It can be calculated in the following equation .

$$SC(D,T) = \theta * SC(D_T, T_T) + (1-\theta) * SC(D_C, T_C) \quad \text{(Formula 2)}$$

where：

$$SC(D_T, T_T) = \frac{\sum_{j=1}^{t} D_{Ti} \times T_{Ti}}{\sqrt{\sum_{j=1}^{t}(D_{Ti})^2 \sum_{j=1}^{t}(T_{Ti})^2}}$$

$$SC(D_C, T_C) = \frac{\sum_{j=1}^{t} D_{Ci} \times T_{Ci}}{\sqrt{\sum_{j=1}^{t}(D_{Ci})^2 \sum_{j=1}^{t}(T_{Ci})^2}}$$

Where θ is the impact factor, its value is [0, 1].It is used to adjust right factor. Where θ is large, title has more influence on topic; where θ is small, it tends to be affected by the text content.

3.2.2. Link Predicting: There are normally a large number of hyperlinks in the crawled pages. We need to predict whether these hyperlinks are related to topics. Predicting link is to calculate priority the extracted URL.

According Shark Search [5] algorithm idea proposed by Hersovici, child nodes can inherit page topic similarity from parent node page. Anchor text, that the hyperlink text, usually represent what web designers describe on the hyperlinks Web page. So it can provide important information of theme forecasting. And the URL address itself is a string text; it carries meaningful information to predict the target.

Considering the above three factors, that is, the correlation parent page, anchor text and URL strings, we can use the following formula to evaluate priority of the URL and predict the next crawling target.

$$Score(Url) = \alpha * \frac{SC(D,T)}{n} + \beta * SC(anchor, T) + \gamma * SC(Url, W_{url}) \text{(Formula 3)}$$

Here, SC (D, T) represents the value of topic similarity of the parent web pages; nis denoted as the degree of the parent web page. It is the number of pages contained in the parent link.

SC (anchor, T) is described as the value of topic similarity of the anchor text. $SC(Url, W_{url})$Is the value of URL, URL is composed of character strings. We should take feature words into account when evaluate the topical similarity of the provided URLs. For example, feature words of “http://digi.it.sohu.com/mobile.shtml”, are "digi", "it", "sohu", "mobile" respectively. Considering the characteristic of URL, which is composed of short, independent words, we adopt common lexicon matching method to calculate the similarity. The lexicon is build up based on the work in section 3.1. At first we extract meaningful words from each URL in addition to host name, that is $W_{url} = (term_1, term_2, \cdots, term_n)$, and then combine all W_{url} into a vocabulary $C_{url} = (word_1, word_2, \cdots, word_m)$.We eliminate the repeat words during the process. When we get the lexicon, the similarity of URLs can be calculated as Formula 4:

$$SC(Url, W_{url}) = \sum_{i=1}^{m} \frac{t_i}{n} \text{ (Formula 4)}$$

Where n is the number of words that W_{url}collected, m is the number of characteristic words of Curl. If $Word_i \in W_{url}$, then$t_i = 1$, otherwise$t_i = 0$.

Where α, β, γ represents weighting factors of the parent page correlation, the anchor text, and URL respectively in predicting link, Here $\alpha + \beta + \gamma = 1$

4. Experiment

4.1. Dataset and Parameter Settings

We perform two related experiments about sports resources and Chinese boxing resources. Starting from the seed list, we perform the first round crawling and get the website URL queue. And then pruning is done before modeling the website network. Firstly, we exclude the website URLs having more than 1000 in-links or out-links, such as

Sohu, Baidu et al from the queue. Secondly, we prune the websites that cannot be visited. Finally, merging the mirrors, we only select unrepeated website URLs. The results of the preprocessing are illustrated in Table 1.

Table 1. Seed List Used For Experiments, the Number of Crawled Website Links after Being Pruned, the Number of Crawled Page Links After Being Pruned

Seed List	Topic	Website Link number	Page Link number
http://www.sports.cn **http://www.sport.gov.cn**	Sports	494	62367
http://www.xingyiquan.cn **http://cnw5.cn** **http://www.bjvingtsun.com** **http://www.cntjq.net**	Chinese Boxing	4714	832489

Firstly, we extracted size 100, size 500, and size 1000and size 5000 pages from above two datasets respectively to conduct different group experiments to verify the values of parameters in Formula 2.Here we used the precision as the metric. In order to calculate the precision, we think the tested web page content ***pc*** is same as the topic ***T*** when *SC(pc,T)*>0.7.

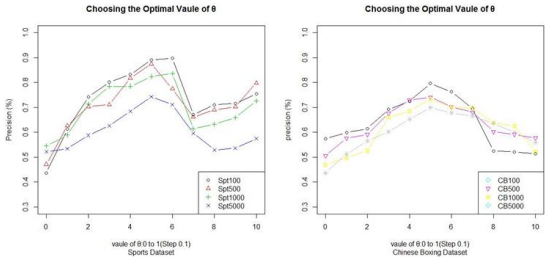

Figure 3. Precision for Two Given Datasets Conducted By Choosing Different Value of Θ When Analysis Page Content

Secondly, we conducted the experiments to choosing the optimal value α, β in Formula 3, we also took the precision as the metric, and $Score(Url_{pc})$>0.6 is the right case. The result shows in Figure 4.

The results show that θ is set to 0.5 for page content analysis, and set α = 0.3, β = 0.3, γ = 0.4 for link prediction. It can obtain the best performance both in two experiments.

4.2. Evaluation

There are many indicators to measure the performance of a focused crawling method. The most commonly used indicators are precision and recall. Here, precision is the fraction of crawled page resources that are relevant to the given topic, while recall is the faction of relevant page resources that are crawled. However, it is very difficult to measure for a focused crawling method, because we have a rather incomplete and subjective notion of what is "good coverage" on a topic [3].So we adopt the precision as the evaluation metric.

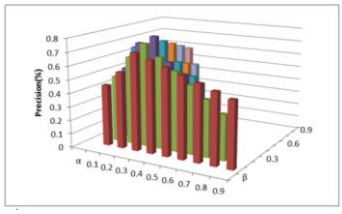

Figure 4. Precision for Given Datasets Conducted by Choosing Different Value of α, β When Predict Links

4.3. Results and Analysis

We choose MyEclipse8.5 as development tools. Hardware environment is designed as follows: CPU is Intel Core i7 2.7GHZ, memory is 8GB, and the operating system is Win7.

After the first round of crawling, we build a complex network model with websites, this is the initial topology. For Sports resources, the network model is composed of 494 nodes, and 2341 edges. For Chinese Boxing resources, the network model has 4714 nodes, and 65234 edges. Based on the complex network models, we use FRTW algorithm to detection the topic-oriented community. Finally we get the topic-oriented website groups respectively. There are 72 websites related to the sports topic and 983 websites related to the China Boxing topic.

Then we conducted the second round of crawling. For Sports resources, we choose these 72 websites as crawling range to collect pages, and 983 websites for Chinese Boxing resources. In the process of crawling, all links beyond the range are discarded. In the focused crawling for sports resources, we have 1000 pages increments in ten times crawling, crawling a total of 10,000 pages. Accounting for the larger scale, we have 2000 pages increments and crawl 20000 pages when crawl the Chinese Boxing resources. We select the Best First algorithm, Fish Search algorithm and Shark Search algorithm as the bench mark.

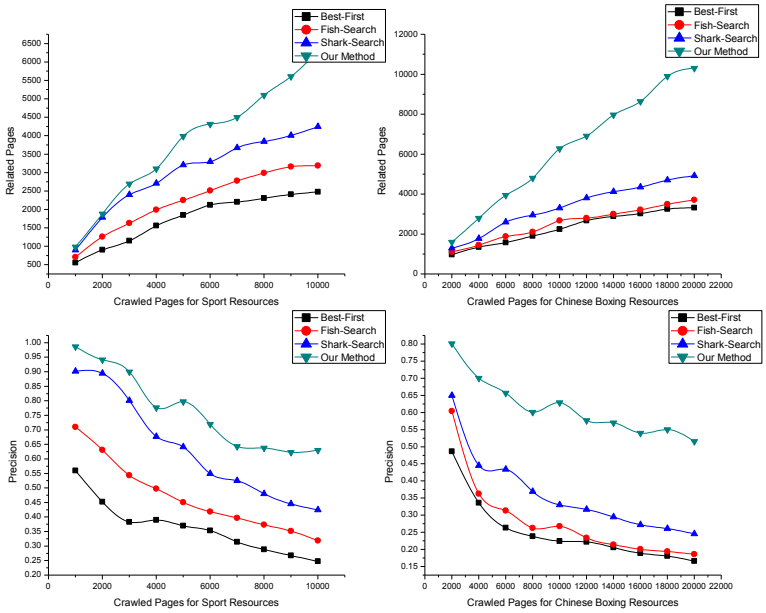

Figure 5. Related Page Numbers and Precision for Two Given Topic Resources Conducted By Four Algorithms

The comparisons are illustrated in Figure 5.It can be seen our approach is significantly better than the other three algorithms owing to we have narrowed the range of the group's

website. Even increasing the number of pages in the crawling, precision of other algorithms significantly decline, this algorithm still maintains a high precision, such as when crawling 10,000 pages, which is 63% precision in sports resources. We observe that the number of the crawled pages is more, the performance of our approach is better than other algorithms.

5. Conclusion

In this paper, we propose a new focused crawling approach based on community detection of complex network. It can avoid crawling and analyzing a large number of unrelated pages due to narrowing the crawled range in website granularity. Considering the important role of page title, we conduct similarity calculation involving page title and page body when analyze the content in web page granularity. Meanwhile, we argue that the link priority becomes more reasonable by comprehensively considering the relevance degree of parent page, anchor text and URL string text when we take link prediction. This paper presents a method which has been tested experimentally. Compared with the benchmark Best First algorithm, Fish Search algorithm and Shark Search algorithms, our approachhas significantly improved on the precision of the focused crawling especially in large-scale web resources.

Acknowledgement

This paper is supported by New Start Project of Beijing Union University (NoZk10201506), Key Scientific and Technological Project of Beijing Municipal Education Commission (No.SZ201311417001), Scientific Research Base Project of Beijing Municipal Education Commission (No.pxm2014_014209_07_000076)

References

[1] B. Gangly and R. Sheikh, "A Review of Focused Web Crawling Strategies. International Journal of Advanced Computer Research", vol. 2, no. 4, **(2012)**.
[2] Y. Ding, "Community detection, topological vs. topical", Journal of Informatics, vol. 5, no. 4, **(2011)**.
[3] B. D. Davison, "Topical Locality in the Web", the 23rdAnnual International Conference on Research and Development in Information Retrieval **(2000)** July 24- 28, Athens, Greece.
[4] P. D. Bra, R. Post, "Searching for Arbitrary Information in the World-Wide Web", the Fish-Search for Mosaicl, Proceedings of the Second International World Wide Web Conference **(1994)** October 17-19, Chicago, US.
[5] M. Hersovici, A. Heydon, M. Mitzenmacher and D. Peeleg, "The Shark-Search Algorithm, An application Tailored website mapping", Proceedings of the 7th International World Wide Web Conference **(1998)** April 14-18, Brisbane, Australia.
[6] S. Chakrabarti and M. V. D. Berg, "Focused crawling, a new approach to topic-specific Web resource discovery", Proceedings of the 8th International World Wide Web Conference **(1999)** May 11-14, Toronto, Canada.
[7] S. Brin and L. Page, "The Anatomy of a Large-Scale Hyper textual Web Search Engine", Computer Networks, vol. 1, no. 30, **(1998)**.
[8] H. T. Zheng, B. Y. Kang and H. G. Kim, "An ontology-based approach to learnable focused crawling", Information Sciences, vol. 23, no. 178, **(2008)**.
[9] H. Dong and F. K. Hussain, "Self-Adaptive Semantic Focused Crawler for Mining Services Information Discovery", IEEE Trans, Industrial Informatics, vol. 10, no. 2, **(2014)**.
[10] C. Zhumin, "Research on focused crawling technique for vertical search engine", Shandong University, **(2009)**.
[11] G. W. Flake, S. Lawrence, C. L. Giles, "Self-organization and identification of web communities", Computer, vol. 3, no. 35, **(2002)**.
[12] L. Yafei, "Research on Algorithms for Detecting Community Structure in Complex Networks", Beijing Jiao tong University, **(2011)**.
[13] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval", Information Processing and Management, vol. 5, no. 24, **(1988)**.

Authors

ShenGuilan, is an associate professor in Business College of Beijing Union University. She got M. D. of Computer Science from Information and Computer College, Beijing University of Chemical Technology (BUCT) in 2006. She is currently a Ph.D. candidate of Information College in People's University of China (RUC), She is interested in the following fields: Communities detection for information network; topical modeling for short-text. She has published articles in several professional journals and international conferences, participated in two textbook's writing, hosted and participated in several externally funded research projects.

Sun jie, is an associate professor in Business College of Beijing Union University. She got M. D. of Management Science and Engineering from Beijing JiaoTong University in 2006.She is currently a Ph.D. candidate at Beijing Telecommunication University. She is interested in fields such as knowledge management and data mining. Her publications include three textbooks and several scholarly papers in journals and international conferences. She was hosted and participated in several externally funded research projects.

Yang Xiao-ping, is aprofessor, doctoral supervisor of Information College in People's University of China (RUC), the major field is web data mining, information system engineering. Prof. Yang is Vice President of Association of Fundamental Computing Education in Chinese Universities (AFCEC), Committee of China Computer Users Association and Vice President of Systems Engineering Society of Beijing