International Journal of Control and Automation
Vol.8, No.11 (2015), pp.361-372
http://dx.doi.org/10.14257/ijca.2015.8.11.34

A Method of Collecting Mongolian Web page Based on Hyperlink Correlation Degree

Zhiqiang Ma, Rui Yan, Zeguang Zhang and Shuangtao Yang

(School of Information Engineering, Inner Mongolia University of Technology, Hohhot, 010080, China)
mzq_bim@163.com

Abstract

Since the encoding of Mongolian web pages is not unified and the amount of web pages are is fewer, a method to unify linguistic model and hyperlink analysis is designed to solve the problem. Firstly the web page language identification is carried on by the N-Gram language model, as well as the average distance of language identification is a part of the hyperlink correlation degree. Secondly the hyperlink correlation degree is calculated based on the anchor text, hyperlink increasing and hyperlink depth. Finally the hyperlinks which are sorted by the hyperlink correlation degree become the collecting seeds of the next web page. The experimental results show that the method of collecting Mongolian web page based on hyperlink correlation degree can effectively enhance the information sum, collection speed and the accuracy rate.

Keywords: *Focused Crawler, Topic Collection Strategy, N-Gram Language Model, Mean Distance Algorithm, Hyperlink Correlation Degree*

1. Introduction

Web crawler can be divided into general crawler and focused crawler. The general crawler is a tool which collects all the information on the Internet. On the contrary, the focused crawler collects the information by an assigned topic on the Internet to satisfy the different professions and background people's demand. The colleting process of focused crawler needs to be defined a topic by the user in advance, and then to discover and download the web page relevant to the topic, and finally to extract and store the contents of the same topic. Therefore, the content of the focused crawler research mainly includes Collection frame, computation model, collection strategy and other contents [1-3]. One of the major challenges is to enhance the efficiency of the focused crawler. Thus, in this paper, we address the importance of the hyperlink correlation degree to the topic collection strategy. So, much research work of the topic collection strategy has been done in accordance with the rule theory， probability theory and so on. The concrete method includes Content-based heuristic method [4-7], Appraisal-based Web hyperlink chart method [8-9] and the union of content and hyperlink chart method [11].

Mongolian crawler is a kind of focused crawler, so its topic is to judge whether a web page is in Mongolian writing. The Mongolian writing here refers to the Ancient Mongolian which is used in Inner Mongolia and other places by Mongolians. The collection of Mongolian web page is faced with two key challenges [12-14]: (1) For the Mongolian crawler, it is difficult to how to judge whether the encoding of collecting web page is Mongolian language. Because Mongolian information processing of Inner Mongolia Autonomous Region is relatively backward, many kinds of Mongolian encoding have been designed at present. For instance, MengKeli, Sai Yin and Unicode are commonly used in the network. There is an overlapped phenomenon between Mongolian encodings, and thus it is difficult for the collection strategy of focused crawler to judge

ISSN: 2005-4297 IJCA

whether web page encoding is Mongolian language. Moreover, the research work on language identification of Mongolian text is fewer and the method of encoding sector is mainly used in the identification process. Therefore, it leads to the high error rate of Mongolian language identification. (2) For the collection strategy of the Mongolia crawler, another challenge is how to fast accurately forecast next hyperlink in massive Internet hyperlinks. By the end of July, 2014, there are a total of 2.73 million websites in China, but the websites of minority language roughly accounts for 0.1% of the Chinese websites. The Mongolian website is one kind of minority language websites, so the amount of the Mongolian website is fewer. When the Mongolian crawler collects the web pages of Mongolian language from the Internet, it must avoid analyzing and downloading the non-correlated hyperlink.

Therefore, in the process of the language identification of Mongolian text, we have given the language identification algorithm for Mongolian text based on N-Gram linguistic model. And we have designed the method to calculate the priority of hyperlink which is based on the linguistic model and collection strategy. So, we propose a core concept of the hyperlink correlation degree in Mongolian crawler, which is mainly composed of the anchor text correlation degree, the increasing of parent node hyperlink correlation degree and the hyperlink depth. The language identification of web page is the foundation of hyperlink correlation degree, which is based on the N-Gram linguistic model. The experimental results show that the accurate rate and acquisition rate is better than the baseline algorithm in gathering information. So, it can provide the reference for the focused crawler of other languages.

2. N-Gram-Based Mongolian Topic Recognition

Definition 1: The web page text consists of a sequence of byte, letter or character, the formalization for $C_1C_2\cdots C_{i-1}C_iC_{i+1}\cdots C_{m-1}C_m(1\leqslant i\leqslant m)$, C_i is called the smallest division unit. The N meta segmentation model of text is a set G, G= $\{g_i|\ g_i=C_iC_{i+1}\cdots C_{i+N-1},1\leqslant i\leqslant m-N+1,\ N\leqslant m\}$, N is called the sliding window length. L=|G|, indicates the total of element in the set. The N meta segmentation model is shown in figure 1.

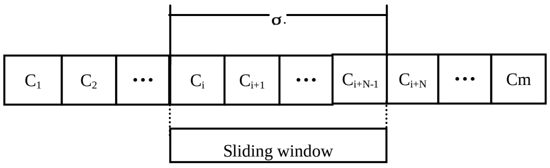

Figure 1. N Meta Segmentation Model

When N is 2, 3 and 4, the model is called dual (bigram), three meta (tri-gram) and four meta (quad-gram) segmentation model. In Figure 2, the Mongolian text of “Inner Mongolia” is carried on bigrams, tri-grams and quad-grams segmentation.

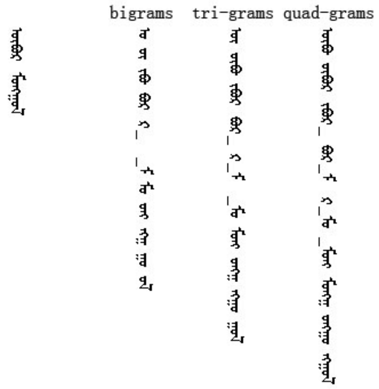

Figure 2. Segmentation Model of Mongolian Text

Definition 2: Count g_i, arrange it in descending order, and obtain N meta model. The training model is called TM, and the testing model is called IM. The index of g_i in the model is called the position, expressed as pos(g_i). The position in TM is expressed as posTM(g_i), as well as the position in IM is expressed as posIM(g_i).

Definition 3: The absolute value of g_i position difference in IM and TM is called the g_i distance, namely dis(g_i), see formula 1.

$$\mathrm{dis}\left(g_i\right)=\begin{bmatrix}\left|pos_{IM}(g_i)-pos_{TM}(g_i)\right| & (\text{find=1}) \\ MAX & (\text{find=0})\end{bmatrix} \tag{1}$$

And "find" describe whether g_i is in the TM model. The find=1 shows that g_i is in TM, otherwise, find=0 shows that g_i is not in TM.

Definition 4: The sum of dis(g_i) of testing text is called text distance, DIS. The mean distance of testing text is expressed as $\overline{DIS}$, see formula 2.

$$\overline{DIS}=\frac{\sum_{i=1}^{i=L_{IM}} dis(g_i)}{L_{IM}} \quad (L_{IM}\neq 0, 1\leq i\leq L_{IM}) \tag{2}$$

There are two steps in the Mongolian identification: First the linguistic model is obtained through the N-Gram model training algorithm, see the algorithm 1, Then the similarity of forecasting text and goal text is obtained through the mean distance identification algorithm, see the algorithm 2.

Algorithm 1. The N-Gram Model Training Algorithm (Train_N_Gram)

```
Input: tests, N
//tests is expressed as the training text, N is expressed as the size of sliding window
Output: N meta model
S[] = split(tests),
for(i = 0, i ≤ S.length(), i++)
  Begin
      for (k = 0, k ≤ L(S[i]) – N , k++) //L(S[i]) is expressed as the length of $S_i$ sentence
          Begin
              $g_k = C_kC_{k+1}…C_{k+N-1}$,
              count[$g_k$]++,//count $g_k$
          End,
      End,
sort(count[g]), // the g is sorted in descending order according
```

```
to counting
    return  g,
```

Algorithm 2. The Mean Distance Identification Algorithm (P_Dis_Ngram)

```
Input: TM, text, N
//TM is expressed as the training model, text is expressed as the test text
output: Mean distance
//Computing IM of testing model, see the algorithm 1
IM = train_N_Gram (text,N),
for ($i = 1, i \le L_{IM}$, i++)
    for ($j = 1, j \le L_{TM}$, j++)
    Begin
        if (find == 1)
            $dis(g_i) = |pos_{IM}(g_i) – pos_{TM}(g_i)|$,
        else
            $dis(g_i)$ = MAX,
        DIS = DIS + $dis(g_i)$,
    End,
return $DIS/L_{IM}$,
```

3. Hyperlink Correlation Degree of Node

Another problem of focused crawler is to forecast the priority sequence of the hyperlink. In order to increase the accurate rate of predicting and computing speed，the algorithm of hyperlink correlation degree is constructed by combining of the topic recognition of N-Gram model with the relevant computation of hyperlink.

Pages and links on the Internet form a directed graph, see figure 3. The collecting process is described as G={H, L, f}, the node in digraph is the web page, H is the node set, namely web page set, the directed edge expresses the hyperlink, L is the directed edge set, namely hyperlink set, The node $a_i(1 \leqslant i \leqslant m)$ is parent node of the node b, f = (H, R), f is the collecting strategy, namely a ordering problem, which predicts node sequence in the digraph on the basis of the H and L, R is the hyperlink correlation degree.

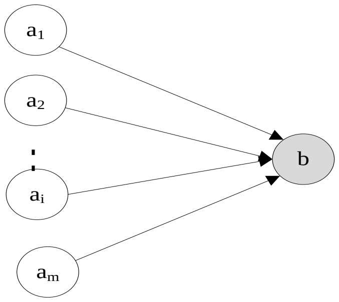

Figure 3. Hyperlink Graph

The web page node includes: the anchor text degree of correlation, the hyperlink increasing and hyperlink depth. Three corresponding parameters are established, shown in Table 1.

Table 1. Corresponding Parameter List

Symbol	Definition
R_t	The correlation degree of anchor text
R_s	Hyperlink increasing
R_d	Hyperlink depth

The calculating process of hyperlink correlation degree is in formula 3.

$$R = w_t R_t + w_s R_s + w_d R_d (w_t, w_s, w_d \geq 0) \tag{3}$$

And, w_t, w_s and w_d are used to standardize R_t, R_s and R_d. They are the weight of priority, and $w_t+w_s+w_d = 1$. w_t, w_s and w_d are obtained through the collecting experiment.

(1) The correlation degree of anchor text (R_t)

Hyperlink correlation degree is related to the language of hyperlink anchor text in parent page. If the anchor text is the Mongolian text, the content of the target page is far more likely Mongolian. Therefore, we have established the correlation degree of anchor text. In this paper, we use the mean distance of N-Gram model to represent the correlation degree of anchor text.

(2) Hyperlink increasing (R_s)

Node hyperlink increasing R_s is related to node in-degree and parent node correlation degree. R^i is hyperlink degree of node a_i ($1 \leqslant i \leqslant m$). In the beginning, it is obtained by the mean distance algorithm of N-Gram model from the extracted web page text, n_i is out-degree of a_i. The computation of hyperlink increasing is in formula 4.

$$R_s^b = \sum\nolimits_{i=1}^{m} R^i / n_i (1 \leq i \leq m) \tag{4}$$

(3) Hyperlink depth (R_d)

It is possible for the collecting algorithm to degenerate into the depth first strategy in the collecting process. Therefore, the attribute of hyperlink depth is defined. The grammar of hyperlink is defined as protocol: // hostname [: port]/path/[, parameters] [? query] #fragment， and the [] is the option. The hyperlink depth is the inverse of the size of path. The size of path is equal to count the separator '/' in the path, namely size(path). See formula 5.

$$R_d = 1/size(path) \tag{5}$$

According to the URL and the W parameters, hyperlinks of page are extracted, the priority of every hyperlink is computed, which consists of the anchor text degree of correlation, the hyperlink increasing and hyperlink depth, the hyperlinks are sorted in waiting queue, see the algorithm 3.

Algorithm 3. Hyperlink Correlation Degree (Rel_Hyperlink)

```
Input: url, W
output: waitQueue//waiting queue
page =Download(url), //Download Web page
links = Jsoup(page), //Analyzes page, hyperlinks to put in the Hyperlinks array
$R_p$ = p_Dis_NGram(page),
for(j=0, j<links.length, j++)
    Begin
        $R_t$ = p_Dis_NGram(links[j].text)； // Links[j].text is the anchor text
    for(k=i, k<waitQueue.length, k++) //waitQueue is waiting queue of URL
        Begin
            if(isParent(page,waitQueue[k]))// page is waitQueue[k] parent node
                $R_s[k]= R_s[k] + R_p/$ links.length,//Update queue hyperlink increasing
        End
        $Rs= R_p/$ links.length；//Calculates this node the hyperlink increasing
        Rd=1/size(links[j]),//Computation depth degree of correlation
        $R = w_tR_t+w_sR_s+w_dR_d$,
    End
return R,
```

4. The Mongolian Web Page Collection Model

The collection model is made up of parameters setting, web page downloading, language identification, Mongolian page content storing, hyperlink duplicate removal, hyperlink analysis and hyperlink correlation degree calculation. Detailed workflow is shown in Figure 4.

(1) Set up a seed queue and parameters,

(2) Run for the first time, put the hyperlink in the seed queue into the priority queue,

(3) Take out the biggest priority hyperlink from the priority queue, download the URL pointing at the web page,

(4) Extract the context from the web page, and calculate the mean distance by N-Gram's mean distance identification algorithm. If the mean distance is greater than a certain threshold in the experiment, the context is judged as Mongolia. If the context is Mongolia, it is stored. Otherwise it is not stored,

(5) Extract all hyperlinks from the web page, and retain the unvisited hyperlinks,

(6) If the hyperlink is non-repeating, then skip to the step (7), otherwise skip to the step (8),

(7) Run the hyperlink correlation degree algorithm, if it is bigger than the threshold, put the hyperlink into the priority queue, update the hyperlink correlation degree in priority queue, and skip to the step (9). Otherwise skip to the step (8),

(8) Abandon the hyperlink,

(9) Judge whether the extracted hyperlinks are completely compared. If they are finished, go to step (10), otherwise go to step (6),

(10) Judge whether the priority queue is empty. If it is non-empty, then jump to step (3), otherwise skip to step (11),

(11) Collection is ended.

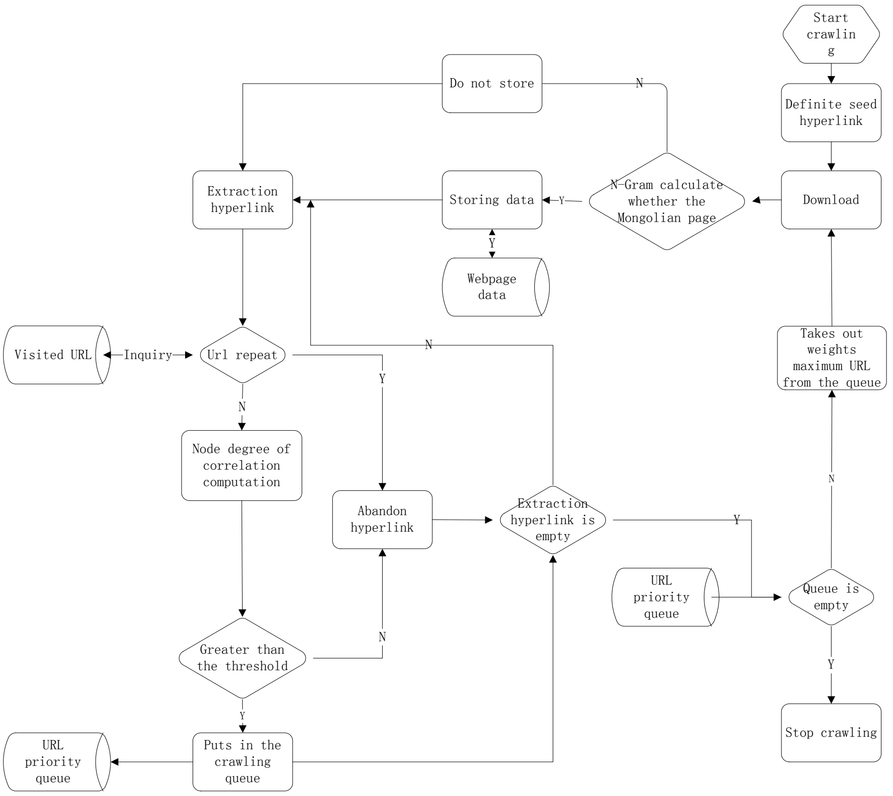

Figure 4. Mongolian Language Web Page Collection Model

5. Experiment Design and Result Analysis

In order to appraise the experimental model, we define the information sum, collection speed and the accuracy rate.

(1)The collection information sum (IS) refers to the disk space of collecting context per unit time, and the unit is Mb. IS can effectively reflect the amount of the collected context, but it cannot reflect whether the context is consistent with the collection topic. In the Mongolian collection, IS includes both the Mongolian context and the non-Mongolian. So, IS is divided into the accurate information sum (AIS) and the fault information sum (FIS). The relation among IS, AIS and FIS is in formula 6.

$$IS = AIS + FIS \tag{6}$$

(2)The collection accuracy rate (AR) is the proportion of the number of Mongolian web page in the number of collected web page. AR can weigh the rate of accuracy of Mongolian collection model data, and also reflect the quality of data source. Accuracy Rate is higher, then it can symbolize that more Mongolian page is collected, and vice versa. The computational method is in formula 7.

$$AR = {}^{n}\!/_{N} \tag{7}$$

And n stands for the number of topic pages which have been collected, N is the number of pages which have been collected.

(3)The collection speed (Collection Speed, CS) is the accurate information sum per unit time. The collection speed can effectively measure the average speed of the collection web pages. In the situation where network bandwidth and the computer hardware resources are invariable, the collection speed is higher, and then the performance of collection system is more outstanding. CS is in formula 8.

$$CS = {}^{AIS}\!\big/_{T} \tag{8}$$

AIS is the Accurate Information Sum, T is the running time of collection system.

In order to confirm the performance of node correlation collection model based on N-Gram model, we have done three experiments. The experiment definitions are shown in Table 2. Mongolian in the web page is commonly encoded by MengKeli and SaiYin. MengKeli code interval is [e264, e34f], namely the coded-decimal interval [57956, 58191], SaiYin code interval is [e244, e293], namely the coded-decimal interval [57924, 58003]. The code interval in table 2 is the Union of two code interval above, namely [57924, 58191].

Table 2. Experiment Definitions

Experiment name	Parameter settings
B_Crawl（Abbreviation B_C）	Collection strategy: Breadth first Mongolian recognition: Code interval recognition
G_Crawl（Abbreviation G_C）	Collection strategy: Content analysis Mongolian recognition: Code interval recognition
N_Crawl（Abbreviation N_C）	Collection strategy: Node degree of correlation computation Mongolian recognition: N-Gram recognition

We have recorded the collection time, collection information sum, the number of web pages and correct web pages. The experimental data is shown in Table 3 in 12 working hours.

Table 3. Experimental Data

Time	Collection information sum			The number of web pages(number)			The number of correct web pages(number)		
	B_C	G_C	N_ C	B_C	G_C	N_ C	B_C	G_C	N_ C
1	99	102	88	9900	10200	8800	2900	3801	8000
2	145	158	132	14498	15801	13197	4400	8011	10956
3	225	245	210	22501	24499	21091	6975	11036	16500
4	300	380	260	30004	38003	25991	8800	19006	22011
5	301	450	316	30099	44991	31611	9250	21009	26021
6	314	681	368	31401	68094	36821	9331	33021	29913
7	323	700	405	32302	70012	40506	10013	34612	34043
8	336	721	460	33598	72111	46018	10441	34802	39005
9	347	734	541	34697	73404	53989	10459	36045	45013
10	354	746	589	35396	74616	59023	11031	36230	49002
11	355	750	610	35503	75019	61007	11038	38021	50010
12	357	772	685	35702	77199	68504	11321	39033	57001

The contrast of the collection information sum is shown in Figure 5. The collection information sum(IS) of G_Crawl which is composed of the Content analysis and the code interval recognition is the biggest, but the collection information sum(IS) of the B_Crawl which consists of the breadth first and the code interval recognition is the lowest. The breadth first does not forecast next hyperlink in close relation with Mongolian, therefore it results in the low collection information sum, On the contrary，the content analysis forecasts the next hyperlink in relation with Mongolian, the collection information sum is high.

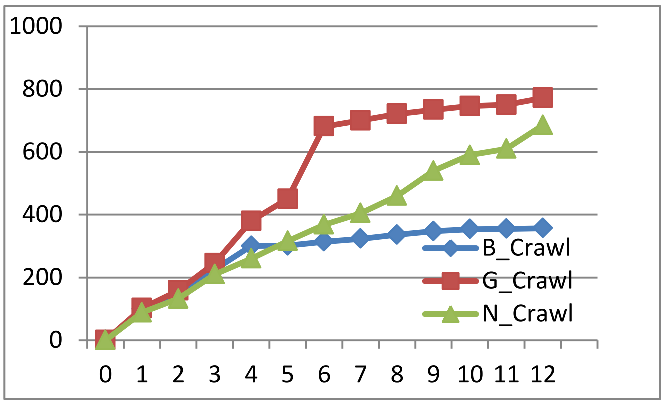

Figure 5. Collecting Information Sum

Because the collection information sum is composed of accurate information sum and fault information sum, the accurate information sum is able to reflect the performance the collection model. The performance is appraised by the accurate information sum, and the statistical data is shown in Table 4.

Table 4. The Accuracy Rate

Time	B_C	G_C	N_ C
1	0.29	0.37	0.91
2	0.30	0.51	0.83
3	0.31	0.45	0.79
4	0.29	0.50	0.85
5	0.31	0.47	0.82
6	0.30	0.48	0.81
7	0.31	0.49	0.84
8	0.31	0.48	0.85
9	0.30	0.49	0.83
10	0.31	0.49	0.83
11	0.31	0.51	0.82
12	0.32	0.51	0.83

The accurate rate of comparison in three experiments is shown in Figure 6.

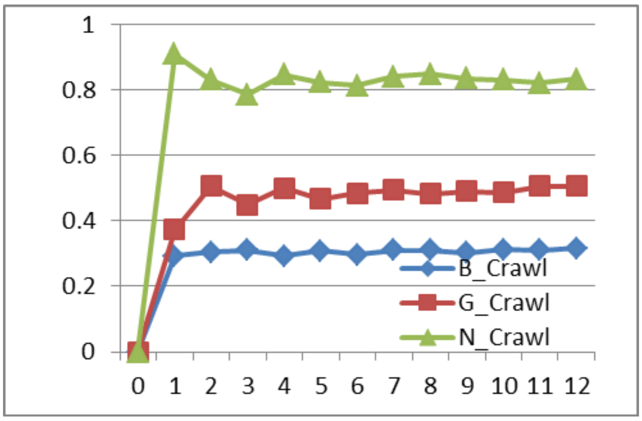

Figure 6. The Accuracy Rate

Finally, according to the accurate rate, the collection information sum and collection speed, the statistics data are shown in Table 5.

Table 5. Statistical Data

Collection project	The Accurate Rate	Collection Information Sum (Mb)	Collection Speed (Mb/h)
B_Crawl	0.31	357	9.22
G_Crawl	0.48	772	31.52
N_Crawl	0.83	685	47.38

The experimental results indicated that the collection information sum of G_Crawl model is bigger than N_Crawl, which consists of N-Gram Mongolian identification and hyperlink correlation degree. But N_Crawl saves the storage space, and enhances the Mongolian identification rate of webpage, collection accuracy rate and collection speed.

6. Conclusion

Through the research of Mongolian webpage encoding on the Internet, we give the mean distance algorithm based on the N-Gram model to identify Mongolian web page, and we give the method of hyperlink correlation degree which combines Mongolian web page identification and hyperlink analysis to order the URL waiting queue. Finally, the collection model has been established on the basis of hyperlink correlation degree. And the comparative experiment has been conducted according to the breadth first strategy and the content analysis strategy. The experimental results show that it obviously surpasses other models in collecting accuracy rate and collection speed. So, it can provide the reference for other minority languages in the information collection system. We plan to do parallel-computing of hyperlink correlation degree to further improve collection speed.

Acknowledgements

Funding project: Inner Mongolia Autonomous Region Natural Sciences Foundation project (2014MS0608). Inner Mongolia Autonomous Region college science research project (NJZY12052). Inner Mongolian university of Technology key Fund (ZD201118).

References

[1] D. Shestakov, “Current challenges in web crawling[C]”, Web Engineering. Springer Berlin Heidelberg, 2013: 518-521.

[2] P. Tadapak, T. Suebchua and A. Rungsawang, "A machine learning based language specific web site crawler[C]", Network-Based Information Systems (NBiS), 2010 13th International Conference on. IEEE, 2010: 155-161.

[3] S. Chakrabarti and M. Van den Berg, "Dom B. Focused crawling: a new approach to topic-specific Web resource discovery[J]", Computer Networks, 1999, 31(11): 1623-1640.

[4] C.C. Aggarwal, F. Al-Garawiand P.S. Yu, "Intelligent crawling on the World Wide Web with arbitrary predicates[C]", Proceedings of the 10th international conference on World Wide Web. ACM, 2001: 96-105.

[5] I. Avraam and I. Anagnostopoulos, "A comparison over focused web crawling strategies[C]", Informatics (PCI), 2011 15th Panhellenic Conference on. IEEE, 2011: 245-249.

[6] L. Kozanidis, "An ontology-based focused crawler[M]", "Natural Language and Information Systems", Springer Berlin Heidelberg, 2008: 376-379.

[7] J. Zong-li, L. Xian-lei and X. Xue-ke, "Based on the theme of the Hub value meta search[J]", Journal of Beijing University of Technology, 2009,35(3):397-342.

[8] S. Brin and L. Page, "Anatomy of a large-scale hypertextual web search engine", Proc. 7th Intl. World-Wide-Web Conference, ,1998:107–117.

[9] J. Kleinberg, "Authoritative sources in a hyperlinked environment[C]", Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998:668-677.

[10] E. Srisukha, S. Jinarat and C. Haruechaiyasak, "Naïve bayes based language-specific web crawling[C]", Electrical Engineering Electronics, Computer, Telecommunications and Information Technology, 2008, ECTI-CON 2008. 5th International Conference on. IEEE, 2008, 1: 113-116.

[11] China Internet Network Information Center. 34th Chinese Internet development situation statistical reports[R],2014,http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201407/P020140721507223212132.pdf.

[12] "Culture China- Chinese net. Chinese minority language present situation [BO/L]", http://www.china.com.cn/culture/2010-07/25/content_20569542.htm.

[13] N. Yao-qun, C. Peng, X. Hong-bo, T. Hui-feng and C. Xue-qi, "Network Uygur language distinguishes and text size world of mortals's discussion [J]", Chinese information journal, 2012,06:109-115.

[14] L. Wei-wei, Z. Wei-qiang and L. Jia. Language identification based on the identification of the vector space model[J]. Journal of Tsinghua University (natural sciences version), 2013,06:796-799.

Authors

Zhiqiang Ma (1972-), male (Hui nationality), Inner Mongolia Hohhot people. He received a B.S. degree in computer application technology from Hohai University in China in 1995. He worked in Inner Mongolia University of Technology. He received a M.S. degree in computer application technology from Beijing Information Science & Technology University in 2007. He became associate professor and graduate students advisor in 2010. His research interests include the search engine, speech recognition, and machine learning.

Rui Yan (1988-), male (Han nationality), Inner Mongolia Erdos people. He received a B.S. degree from Inner Mongolia University of Technology in China in 2012. He is a graduate student in Inner Mongolia University of Technology. His research interests include the speech recognition, data mining.

Zeguang Zhang (1988-), male (Han nationality), Inner Mongolian Tongliao people. He received a B.S. degree from Inner Mongolia University of Technology in China in 2012. He received a M.S. degree in computer application technology from Inner Mongolia University of Technology in China in 2014. His research interests include the search engine, cloud computation.

Shuangtao Yang (1990-), male (Han nationality), Henan Zhoukou people. He received a B.S. degree from Inner Mongolia University of Technology in China in 2013. He is a graduate student in Inner Mongolia University of Technology. His research interests include the deep learning, cloud computation.