International Journal of Hybrid Information Technology
Vol. 10, No.3 (2017), pp. 57-66
http://dx.doi.org/10.14257/ijhit.2017.10.3.06

Research on Knowledge Acquisition and Knowledge Extraction of Forest Management

LiuJiancheng[1], Wu Baoguo[1*] and Dong Chen [2]

[1]*School of Information Science and Technology, Beijing Forestry University, Beijing, 100083 China*
[2]*School of Information Engineering, Zhejiang Agriculture and Forestry University Hangzhou, 311300 China*
**Corresponding Author: wubg@bjfu.edu.cn*

Abstract

The abundance of knowledge seriously influence the ability of expert decision support system deal with problems, it makes the problem that how to find the reliable knowledge on the Internet. A method that use web crawler crawl forest management web and extract expert knowledge in them is put forward in this paper, to solve the problem. This method limit the web crawler, and it is adopted to grasp the knowledge related to forest management and to extract the expert knowledge therein, enriching knowledge base of expert decision support system. Crawling web priority is calculated by reliability value, and Pearson correlation is used to make sure that the collected knowledge is related to forest management. The knowledge of forest management is extracted and repetitive knowledge is removed on the basis of collecting knowledge. Included angle cosine and Euclidean distance are used in this method to avoid acquit duplicate content. Experimental results show that this method has good accuracy.

Keywords*: knowledge acquisition; knowledge extraction; web crawler; DSS*

1. Introduction

Being interested only in afforestation and exploitation, Chinese forestry has ignored tending management process and technology in- between for quite a long time. Forest management lagging behind seriously is caused by lack of forest management planning system, backward theory and technology system, scarcity of management experts [1]. Serving production units at the grass-roots level with the utilization of expert decision support system can solve the problem of production units at the grass-roots level lacking management experts.

Knowledge collection is a combination of knowledge management and knowledge engineering, also belonging to the scope of artificial intelligence. The core of problem-handling capability of expert decision support system is the richness of knowledge. Knowledge collection is a "bottleneck problem" of building expert decision support system, which also a main factor to determine the superior performance of an expert decision support system performance as well one of the key technologies for developing the expert decision support system [2]. Its task is to extract information needed by the expert decision support system from information source for forming knowledge which later is converted into a form of easy storage and expression for computer. Finally, knowledge base is formed.

Search engine is a typical representative of the knowledge collection, which has a wide coverage, nearly traversing all the corners on the Internet by means of powerful crawlers.

However, its information search results are often dissatisfied due to the poor professional classification [3]. Taking the commonly-use word "sub compartment" as an example, search results of two general search engine "Google" and "Baidu" are always

ISSN: 1738-9968 IJHIT

information related to Chinese word "Group study in kindergarten", which is in poor forestry specialty.

On the other hand, document research on knowledge collection and knowledge extraction is rare. There are more researches using XML related technologies to extract knowledge in current research literatures, and also researches adopting DOM- based technology [2-6]. In addition to this, knowledge collection related research hotspots have always focused on study on topic crawler and topic-specific search engine [7], so do the related researches on forestry. Targeted at poor coverage rate and low accuracy ratio of general search engines on forestry topic, an overview of topic information collection strategy and a study on design scheme of forestry topic crawler have been conducted to improve coverage and accuracy [8,9]; forestry information can be divided into seven categories: forestry science and technology information, forestry production data, forest products market supply and demand information, flowers information, forestry policies and regulations, forestry labor services, meteorology and environment for realizing rapid search on dynamic information of forestry. It can achieve collection and classification of dynamic forestry information through extracting domain names and identifying categories of collected webpages [10]. These documents were mainly studies on collection technology, collection method and search technology as well as search algorithm related to forestry information, mainly solving the problems of information collection to improve the coverage rate of forestry topic related information and converting deficiency in low accuracy of traditional search engines forestry topic related information. However, without combination of expert decision support system, the above studies have no researched on converting the extracted information related to forestry management into management knowledge of knowledge base of expert decision support system.

This article studies that the web crawler is adopted to grasp the knowledge related to forest management and to extract the expert knowledge therein, enriching knowledge base of expert decision support system. It also studies on carrying structural analysis on topic pages related to forest management knowledge, focusing on qualification rules of crawler, credibility- based grasp algorithm, filtering of grasp information, extraction of webpage knowledge information and avoiding of extracting repeated knowledge into the base, *etc*. Compared to manual knowledge collection, computer technology can solve the problem of collecting forest management knowledge to greatly improve the efficiency of knowledge collection, to enrich the knowledge base of forest management decision support system and to enhance the handling capability of expert decision support system on forest management. At the same time, due to the connectivity of computer technology in various industries, this study can also provide references for building knowledge collection models of decision support system for other industries.

2. Key technology and Algorithm

2.1. Web Crawler Algorithm

(1) Qualification rules of crawler

Based on system structure and implementation technique, web crawler [11] can be roughly divided into: general web crawler, topic web crawler, incremental web crawler, deep web crawler, qualified web crawler. As current studies not considering the reliable requirements on forest management knowledge, this study combines the characteristics of forest management knowledge to qualify the web crawler modules. It only processes the internal links of collection links and the links containing inside the domain in index database with adoption of method of priority weights, grasp in depth, so as to ensure all sources of crawler collection knowledge are from identified webpages.

Forestry in China undertakes mainly a task of public welfare, development and research of which are depended on financial input of government and guided by national

policies. Regarding forest management and research, it mainly depends on the supports from government or national projects that are studied by universities and research institutes. Research results are reported back to the government or fund supporting units in the form of academic papers, monographs and research reports. Therefore, in terms of overall authority in the forestry industry, government, scientific research units, colleges and universities as well as central units are higher than that of enterprises and social organizations as well as local units, respectively. Based on this kind of authority, webpages are ranked for rating based on the division of administrative ranks of nation, province (autonomous region, municipality directly under the central government), city and county (district). The higher administrative levels of unit, the higher site weight and the larger credibility will be.

(2) Grasp algorithm of crawler

Before crawler grasp, it shall firstly compute the URL priority scores, then according to descending order to grasp the priority scores. Weight of topic relevance of father node is greater than that of rank rating of father node. This article chooses formula (1) to carry out priority scores on the father node for computing.

$$Score_{potential}(Node_{current}) = \begin{cases} \eta * Rank(Node_{current}) + (1-\eta) Score_{relativity}(Node_{current}) \\ Score_{relativity}(Node_{current}) = \sin(q, Node_{current}) \end{cases} \quad (1)$$

Where $\boldsymbol{Score_{potential}}$ represents potential score computing, $\boldsymbol{Rank()}$ represents node rating method, $\boldsymbol{Score_{relevance}}$ represents topic relevance between $\boldsymbol{Node_{current}}$ and topic q. $\boldsymbol{Node_{child}}$ under the father node will inherit potential scores from the father node. Yet as each child node will inherit the father node, influences among adjacent children nodes will also exist. Priority scores of each child node can be obtained via formula (2).

$$Score_{potential}(Node_{child}) = \begin{cases} \lambda * Score_{inherited}(Node_{child}) + (1-\lambda) * Score_{inherited}(Node_{others}) \\ Score_{inherited}(Node_{child}) = \delta * Score_{potential}(Node_{current}) \end{cases} \quad (2)$$

$\boldsymbol{Score_{inherited}}$ of $\boldsymbol{Node_{child}}$ is obtained from $\boldsymbol{Score_{potential}}$ and decay factor $\boldsymbol{\delta}$. In the formula, $\boldsymbol{\lambda} \in \{0,1\}$, while decay factor $\boldsymbol{\delta} \in \{0,1\}$.

URL array is sorted based on priority scores after computing priority scores of URL, grasping is conducted based on descending order of the priority scores. Web crawler algorithm is:

Step 1. Pre- read data and compute priority scores on the URL array.

Step 2. Resort based on the URL priority score.

Step 3. Obtain source documents of corresponding webpages based on the URL.

Step 4. Extract source document URL link sets from the webpage to enter into the Todo work array.

Step 5. Obtain the URL of No. i data of Todo work array, namely Todo[i]. Url to judge whether the URL is within the hostname scope of URL index database.

Step 6. If yes, extract the webpage information and carry out further information on formatting process, etc. After that, i=i+1, and skip to Step 3; if no, bypass the webpage information processing, i=i+1, and skip to Step 5.

Step 7. Judge whether Todo array has URL that is not processed; if yes, i=i+1, skip to Step 5; If no, skip to Step 8.

Step 8. End of the work.

Crawler module only process internal links, ensuring information collected is from webpages provided by the index base and improving reliability of the information. Rank rating on webpages and computing priority scores for URL can be applied to not only the grasping order of web crawl, but also confirmation of credibility of knowledge collected.

(3) Filtering of grasp information

A webpage with the topic of forest management mainly consists of tree species information, forestry knowledge, operation mode information, technical measures information as well as forest pest and disease information that is generally mixed with species name, Latin name, type of tree species, characteristics of tree species, seed production technique, seedling raising technique, afforestation technology, tending technology, cutting operation regeneration technology, cultivating operation technology, forest cleaning technology, pests and disease information and other features. Knowledge data collected by the crawler are not surely the forest management knowledge, which should be filtrated to retain knowledge related to forest management.

Based on VSM (Vector Space Model)[12], it has established a n-dimensional feature vector (incl. species name, Latin name, type of tree species, characteristics of tree species and seed production technique, etc.) for identifying whether the targeted webpage has a topic relevance with forest management knowledge. The feature vector is denoted as $\boldsymbol{v^T}$=((t1,ω1),(t2,ω2)……(tn,ωn)), ti(i=1,2,3, …,n) to represent different attributes of the topic webpage. ωi(i=1,2,3,…,n) represents the corresponding weights of different attributes. In this study, $\boldsymbol{t_i}$ is defined as the function of frequency of occurrence $\boldsymbol{tf_i}$ of the topic webpage, obtained by the formula (3).

$$\omega_i = \frac{tf_i \times \log\left(\frac{N}{n_i} + 0.1\right)}{\sqrt{\sum_{i=1}^{n} (tf_i)^2 \times \log^2\left(\frac{N}{n_i} + 0.1\right)}} \tag{3}$$

Where, $\boldsymbol{N}$ represents total number of documents; $\boldsymbol{n_i}$ represents the number of documents containing property $\boldsymbol{t_i}$. Assuming that the target webpage $\boldsymbol{Q}$ is collected currently, its vector established is denoted as $\boldsymbol{Q^T}$. And then, compute whether its topic is relevant to the forest management knowledge. This study computes the topic relevance with the help of Person related coefficients, while Person related coefficient $\boldsymbol{\varepsilon}$ of webpage vector $\boldsymbol{Q^T}$ and feature vector $\boldsymbol{v^T}$ are obtained by formula (4).

$$\varepsilon = Pearson(Q^T, \upsilon^T) = \frac{n\sum_{i=1}^{n}(Q_i d_i) - \sum_{i=1}^{n} Q_i \bullet \sum_{i=1}^{n} d_i}{\sqrt{n\sum_{i=1}^{n} Q_i - \left(\sum_{i=1}^{n} Q_i\right)^2} \bullet \sqrt{n\sum_{i=1}^{n} d_i - \left(\sum_{i=1}^{n} d_i\right)^2}} \tag{4}$$

Where, $\boldsymbol{Q_i}$ represents the value of different elements of n- dimensional vector $\boldsymbol{Q^T}$ of webpage $\boldsymbol{Q}$; $\boldsymbol{d_i}$ represents the value of different elements of feature vector of forest management knowledge. When $\boldsymbol{\varepsilon}$ value approaches 0, the relevance is lower; otherwise, the relevance is higher.

After obtaining the results, judge the relationship between $\boldsymbol{\varepsilon}$ and threshold value $\boldsymbol{\theta}$. If $|\boldsymbol{\varepsilon}| \in \{0, \boldsymbol{\theta} | 0<\boldsymbol{\theta}<1\}$, it considers that the targeted webpage $\boldsymbol{Q}$ conforms to features of forest management knowledge, extracting forest management knowledge herein for processing; If $|\boldsymbol{\varepsilon}| \notin \{0, \boldsymbol{\theta} | 0<\boldsymbol{\theta}<1\}$, it considers that the targeted webpage $\boldsymbol{Q}$ is not in conformity with features of forest management knowledge, no data processing. As feature vector $\boldsymbol{v^T}$ of forest management knowledge shall try to cover main features of forest management knowledge, there is no vector $\boldsymbol{Q^T}$ of the targeted webpage that is completely relevant to the feature vector $\boldsymbol{v^T}$ with an absolute value of 1 in the practical application.

2.2. Webpage Information Processing

The web crawler processes the URL array to obtain webpage documents with a set of HTML source code lines [13]. It can be seen from structure figure of tree form node of webpages collected, the webpages have the structures, as shown in Figure 1. Tree form structure of webpage figure can be given by:

(html (head (meta(title , keywords ,description), style , script))，(body (table (tr (td (text))))),(div (ul (li (a))),(span (text))), script..).

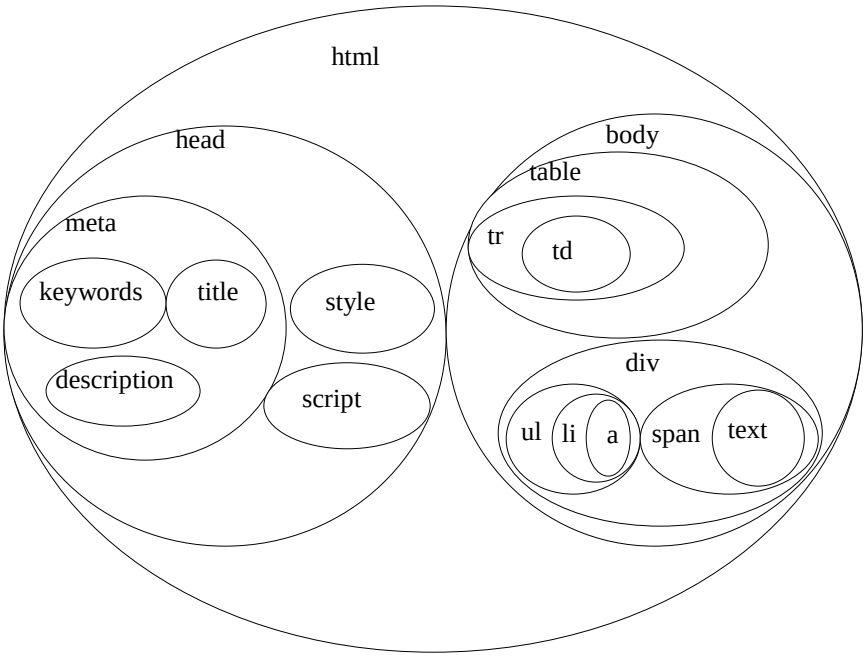

Figure 1. The Web Page Structure

According to the tree representation, distinguish the node positions of the forest management knowledge information and the node positions of noise. As for the webpages with the topic of forest management, title, keywords, description part marked under the head are the direct description of the webpage or webpage, generally presenting introduction of technical information related to forest management that can be used for establishment of vector to compute the relevance with the feature vector of forest management knowledge. If the relevance scores greater than or equal to the threshold $\boldsymbol{\theta}$, it is considered that it's useful; if the relevance scores less than the threshold $\boldsymbol{\theta}$, then it's considered as noise. Although script part of all webpages styles are not the noise information, this study after analyzing major topic webpages of forest management knowledge, found that a majority of style and script parts of the webpages are used for beautifying the display effect, which is irrelevant to the topic content of forest management knowledge, causing interference with the collection works of the crawler. Also, efficiency for the crawler analyzing these contents is low, so if the crawler analyzes the contents one by one, it will affect the work efficiency of the crawler. In the study, it will not process this information, but regarding it as noise. Generally, it can extract precise information of forest management knowledge in the tag of table <td></td>; Partial contents in the mark of have larger possibility to become the targeted content. Anchor text information in the marks of and <a> can be the basis for judging the relevance, while URL linked can generally be extracted to the crawl array.

The study selects a regular expression to extract information on the basis of analyzing the web page structure of forest management knowledge related topic webpage. Different HTML webpage contents match with different regular expressions. Therefore, besides customizing the regular expressions extracted by the information, the study also specifies HTML tag analysis rules and generates the regular expressions used by knowledge extraction through analysis. Analysis rules mainly contain webpage properties (including title, keywords, description and other Meta information of the webpages) and knowledge contents (title, source and release time, knowledge content and technical measures), *etc*. Different formulas will match with different parts of the webpage, generally divided into monolayer or multi-layer tag. Contents of monolayer tag like < h1 > < / h1 >, *etc*., are uniline or with a consecutive character string that can be matched directly; Contents of

multiple tags like < table > < / table >, *etc.*, are relatively complex and nest single tag, which can be analyzed after extraction. Matching formula is as shown in formula (5).

$$Z = \begin{cases} split(Tag).[0] + Wildcards(.*?) + split(Tag).[1] & Tag.layerCount = 1 \\ split(Tag).[0] + Wildcards([\hat{}\hat{}]*?) + split(Tag).[1] & Tag.layerCount > 1 \end{cases} \quad (5)$$

Where, ***Tag*** represents the HTML tags that require to carry out matching rules; ***Split (Tag).[i]*** represents different segments of character strings; ***Wildcards()*** represents the method of adding wildcard character; and ***layerCount*** represents layer amount of ***Tag***.

2.3. Processing of Similar Contents

Two aspects are conducted to avoid duplicate contents. On the one hand is to control the crawler to process only the uncollected URL. The URL should be confirmed whether it's collected before the crawler collection of URL; on the other hand is to control the contents avoiding duplicate contents or high- similarity knowledge into the knowledge base, which guarantees the performance of expert decision support system. It shall use the detection algorithm of duplicate webpages for content control, which firstly generates a fingerprint value for each webpage, and then computes the similarities of fingerprint values of two webpages. If the similarity is greater than the threshold value θ, it's considered that the two webpages are repeated; otherwise, the two are different. Content-based detection algorithm of duplicate webpage can be divided into: the algorithm based on word frequency statistics and the algorithm based on string comparison [14], according to the particle size of generated fingerprints. The study adopts the detection algorithm based on the string comparison.

It shall firstly establish vector υ_d of the webpage d and vector υ_q of the webpage q according to spatial vector establishing algorithm of forest management knowledge topic webpage introduced in Section 2.1.3, before carrying out similarity algorithm on forest management knowledge topic webpages. Both υ_d and υ_q are multidimensional vector. To solve the ***sim(q,d)*** of webpage d and q is equal to solving the angular separation and spatial distance of the multidimensional vector υ_d and υ_q. Therefore, this study uses respectively included angle cosine and Euclidean distance to determine the similarity.

Included angle cosine: the values obtained by the included angle cosine are in- between 0 and 1. If the vector is one consistently, it indicates the highest similarity of documents; if the value obtained is 0, the vector is orthogonal; it indicates the lowest similarity of documents. Value of Angle cosine of webpage d and webpage q is computed by formula (6).

$$sim(q,d) = sim(\upsilon_q, \upsilon_d) = \cos\theta = \frac{\sum_{i=1}^{n}(\omega_i(q) \times \omega_i(d))}{\sqrt{\left(\sum_{i=1}^{n}\omega_i^2(q)\right) \times \left(\sum_{i=1}^{n}\omega_i^2(d)\right)}} \quad (6)$$

Euclidean distance: it's an algorithm commonly used in the space to compute the distance of two n- dimensional vectors. Differing from the included angle cosine, the larger value of Euclidean distances, it has the farther distance of vectors and the lower similarity of documents; computing formula of Euclidean distance is as shown in formula (7).

$$sim(q,d) = sim(\upsilon_q, \upsilon_d) = ED_{qd} = \sqrt{\sum_{i=1}^{n}((tq_i, \omega q_i) - (td_i, \omega d_i))^2} \quad (7)$$

Multiplicity of text is obtained by computing, result of which is used for determining the relationship between ***sim (q, d)*** and ***Field*** convention. If ***sim (q, d)*** $\in$ ***Field***, then it's too high similarity to delete the webpage; if ***sim (q, d)*** $\notin$ ***Field***, the similarity meets the requirements, URL and text information of the webpage will be stored in the database.

3. Test Results

This study chooses the master station, station group, sub- station and other authority webpages of Chinese forestry web to be the main testing webpages. It sets grasping time to be 20 hours, during which, 273 M URL sample data containing 2, 130,000 URL data are collected. Two-thirds of the data collected will be used as training samples, the remaining one-third data as testing samples. In order to identify the forestry vocabularies contained in the webpage accurately during establishment of index base, following specific word entries are added: afforestation model, suitable land for forest, splash restoration, sub compartment, region compartment, subplot, glyptostrobus pensilis family, Masson pine family, stock map, brown patch, grey speck disease, iron- deficiency disease and other nearly 2000 data of entries. Samples are analyzed and computed, following results are obtained.

3.1. URL Extraction Results

URL extraction results are mainly to judge whether the crawler extraction capability can meet the requirements of link coverage required by this study. Table 1 is URL Extract Test Result Data. Of which, the extraction percentage is obtained through comparing the extraction URL amount of test samples divided by the URL amount of the webpage.

Table 1. URL Extract Test Result Data

Site name	Site URL	Actual amount of URL	Extract amount of URL	Extract percentage
Chinese forestry information webpage	http://www.forestry.gov.cn/	1104	1011	91.6%
Academy of Forestry Sciences (China)	http://www.caf.ac.cn/	292	241	82.5%
Management Division of Forest resources	http://slzy.forestry.gov.cn/	83	76	91.6%
Management Division of afforestation	http://zls.forestry.gov.cn/	123	121	98.4%
Chinese Society of Forestry	http://lxh.forestry.gov.cn/	45	43	95.6%
Chinese Eucalyptus Site	http://eucalypt.forestry.gov.cn/	107	105	98.1%
Chinese Cedar Site	http://cedar.forestry.gov.cn/	81	79	97.5%
Chinese Robur Site	http://robur.forestry.gov.cn/	84	82	97.6%
Chinese Fir Site	http://fir.forestry.gov.cn/	72	70	97.2%
Chinese Birch Site	http://birch.forestry.gov.cn/	71	69	97.2%

From Table 1, it's observed that extracting URL is slightly less than the actual one. Of which, webpage of Academy of Forestry Sciences (China) has the lowest accuracy, where has a larger gap between extract amount of URL and the actual amount. The main cause of generating the problem is extracted URL used in the study have been filtrated, removing useless and wrong URL, document URL and external URL of knowledge collection. Different amount of these URL can be existed in different sites, different accuracy will be presented. But for most forestry webpages, extract percentages are above 90%, satisfying requirements of the study on URL coverage.

3.2. Topic Relevance Analysis

The study randomly selects forest management knowledge grasped by the crawler modules for topic relevance verification, which aims to test and study whether the topic vector established for forest management knowledge can meet the requirements. 300 webpage data grasped by the crawler module are selected by the study for verifying samples. Sample vector $\boldsymbol{Q^T}$ is established based on the method described in the Section 2.1.3 of the paper. And computational algorithm of Person correlation coefficient is employed to compute targeted webpage vector $\boldsymbol{Q^T}$ and Person correlation coefficient ε of feature vector $\boldsymbol{v^T}$ of forest management knowledge. Method described in the Section 2.2 of the paper is employed to extract knowledge from the targeted webpage and make statistics on knowledge words extracted. The correlation coefficient and alphanumeric data obtained are randomly divided into A, B, C and D, each of which has 75 data. Extracting word is regarded as a vertical coordinate, while the Person correlation coefficients as the horizontal coordinate. And then draw scatter diagram based on four data separately, as shown in Figure 2.

Figure 2 shows that related correlation coefficients are negatively correlated to the number of words. Under the situation of the more words, it has lower topic related coefficients, and the largest differences between vector $\boldsymbol{Q^T}$ of the targeted site and the feature vector $\boldsymbol{v^T}$ of forest management knowledge. Sample values are mostly concentrated in- between 0.2 and 0.5, which accounts for about 85% of the total samples. From manual data calibration, it found that not all webpages are in the situation of the more words, the lower the topic relevance. Special cases exist. Main cause of which is there is a certain deviation on feature vector of the forest management knowledge established by the study and the vector of the targeted webpage. Target of the feature vector is to try to identify a majority of webpages as much as possible, not all. When the threshold $\boldsymbol{\theta}$ is set to 0.15, the sample coverage rate reaches 94.86%, reaching 82.13% of sample fitness. It indicates that most of the samples can be identified accurately, and accuracy of which accords with the study requirements.

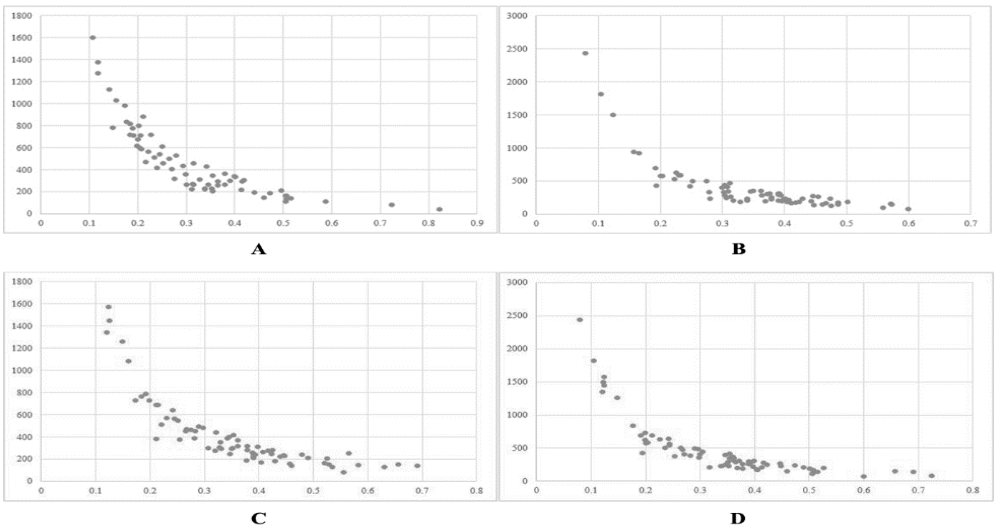

Figure 2. The Result of Topic Correlevance Value

3.3. Accuracy Verification

The ultimate goal of this study is to enrich forest management knowledge base. Finally, knowledge base is presented in the form of knowledge. Specialty, accuracy, usefulness

and credibility of forest management knowledge are influenced by subjective factors. Therefore, 30 experts in the field of forestry related are invited to randomly select 50 knowledge data to rate via Delphi method for further verifying the accuracy of the testing results. Rating is divided into rating of relevance, rating of specialty and rating of credibility, and weight factors are 3, 3 and 4, respectively. After arranging the result, it shows that sample standard deviation is 0.7818 with an average rate of 7.8. Overall average score is higher due to small deviations of the samples. Good accuracy is presented in the overall study.

4. Conclusion and Discussion

This study proposes that the web crawler is employed to grab forest management information and to extract forest management knowledge for finding a solution to the richness of decision support system of forest management, which highlights on qualification rules of crawler, credibility- based and use of the Person related coefficient for filtrating information grasped, etc. structure of forest management topic webpages is also analyzed on the basis of crawler modules studied. Information extraction is conducted on the forest management knowledge with the combination of the algorithm of regular expression. Also, the included angle cosine and Euclidean distance are employed to compute the similarity for avoiding duplicate knowledge into the base.

The core module of this study is the web crawler module that is depended by the overall works. Main consideration when designing crawler module is the requirements of specialty and accuracy of forest management knowledge collected, but lacking consideration of efficiency of crawler collection. Therefore, certain problems are existed in the study, which should be further verified and improved in the future study. For example:

(1) Heavy workloads borne by the crawler module leads to higher time complexity of the module. Besides computation of link grasping priority is needed, it also needs to compute Person correlation coefficient, consuming larger time. Crawler module has low efficiency.

(2) Content is in accurate fitness with the use of the regular expression, but different webpage structures need different matching rules. Tags are used in this study to match and extract knowledge combined with the regular expression. However, it has a fair matching accuracy in certain webpages with complex structures.

Acknowledgments

This work is partially supported by the 863 Program (the National High Technology Research and Development Program of China, Project NO. 2012AA102003.) and National Forest Management Science and Technology Support Research Projects (Project NO. 169201531).

References

[1] Z. Tie, “Academician YIN W L: Forest Quality Improvement Related to the Ecology Society”, Journal of Green China, no. 7, **(2016)**, pp. 46-50.
[2] D. C. Wimalasuriya and D. Dou D, “Ontology-based information extraction: An introduction and a survey of current approaches”, Journal of Information Science, vol. 36, no. 36, **(2010)**, pp. 306-323.
[3] J. H. Zhang, “Application of professional intelligent search system in Veterinary Medicine”, Journal of Northeast Agricultural University, vol. 40, no. 9, **(2009)**, pp. 141-144.
[4] N. Esfandiari, M. R. Babavalian, A. M. E. Moghadam and V. K. Tabar, “Knowledge discovery in medicine: Current issue and future trend”, Journal of Expert Systems with Applications, vol. 41, no. 9, **(2014)**, pp. 4434-4463.
[5] Y. F. Jin, J. Y. Fan and Y. Feng, “Design and Realization of Distributed Web Crawler”, Journal of Harbin University of Science and Technology, vol. 15, no. 1, **(2010)**, pp. 116-119.

[6] Z. Chen and D. M. Zhang, "Survey of Web information extraction technologies", Journal of Application Research of Computers, vol. 27, no. 12, **(2010)**, pp. 4401-4405.
[7] B. M. Shi, Y. X. He and C. Z. Wu, "Research on search strategy of web spider in topic-oriented search engines", Journal of Computer Engineering and Applications, vol. 50, no. 2, **(2014)**, pp. 116-119.
[8] F. Zhang, X. L. Feng and J. S. Yuan, "Vertical Search Engines on Forestry", C. Advanced Materials Research. Trans Tech Publications, vol. 143, **(2011)**, pp. 321-323.
[9] J. S. Yuan and Y. F. Guo, "Algorithm Research and Design of Forestry Focused Web Crawler", Journal of Computer Engineering and Design, vol. 32, no. 6, **(2011)**, pp. 2003-2006.
[10] L. S. Zhang, G. Zhang, C. X. Long and S. Zhang, "Search and Integration of Thematic Dynamic Information on Forestry", Journal of Central South University of Forestry and Technology, vol. 33, no. 5, **(2011)**, pp. 47-51.
[11] S. Xu, H. J. Yoon and G. Tourassi, "A user-oriented web crawler for selectively acquiring online content in e-health research", Journal of Bioinformatics, vol. 30, no. 1, **(2014)**, pp. 104-114.
[12] H. P. Deng and G. Wu, "Discovery of topic-specific information source based on web crawler and website classification", Journal of Computer Engineering and Applications, vol. 52, no. 3, **(2016)**, pp. 59-65.
[13] Y. C. Wu, "Language independent web news extraction system based on text detection framework", Journal of Information Sciences, vol. 342, **(2016)**, pp. 132-149.
[14] J. Qin, F. L. Yan, H. F. Zhu, Q. Si and H. Xie, "A Webpage Classification Algorithm Based on Link Information", Journal of Microelectronics and Computer, vol. 29, no. 6, **(2012)**, pp. 108-112.

Authors

Liu Jiancheng, born in 1989, Ph. D. candidate. His research interests include Forestry Information Technology, Forestry Decision Support System.

Wu Baoguo, born in 1955, professor. His research interests include Foestry Information Technology, Forestry Decision Support System.

Dong Chen, born in 1989, Ph. D. She is a lecturer worked in Zhejiang Agriculture and Forestry University. Her research interest is Forestry Information Technology.

