Key Spatial Relations-based Focused Crawling (KSRs-FC) for Borderlands Situation Analysis

Dongyang Hou [a,b,] *, Hao Wu [b], Jun Chen [b], Ran Li[b]

[a] School of Environment Science and Spatial Informatics , China University of Mining and Technology, Xuzhou,Jiangsu,221116, China-houdongyang1986@163.com
[b] National Geomatics Center of China, 28 Lianhuachi West Road, Beijing 100830, China

KEY WORDS: Focused Crawling, Place Names, Web Information Collection, Borderlands Situation Analysis, Relevance Calculation, Spatial Relations

ABSTRACT:
Place names play an important role in Borderlands Situation topics, while current focused crawling methods treat them in the same way as other common keywords, which may lead to the omission of many useful web pages. In the paper, place names in web pages and their spatial relations were firstly discussed. Then, a focused crawling method named KSRs-FC was proposed to deal with the collection of situation information about borderlands. In this method, place names and common keywords were represented separately, and some of the spatial relations related to web pages crawling were used in the relevance calculation between the given topic and web pages. Furthermore, an information collection system for borderlands situation analysis was developed based on KSRs-FC. Finally, F-Score method was adopted to quantitatively evaluate this method by comparing with traditional method. Experimental results showed that the F-Score value of the proposed method increased by 11% compared to traditional method with the same sample data. Obviously, KSRs-FC method can effectively reduce the misjudgement of relevant webpages.

1. INTRODUCTION

Borderlands Situation refers to an overall description and tendency of a borderlands entity, including its location, activity and events etc. (Chen et al., 2013a). Borderlands situation analysis is one of the significant components for risk assessment and emergency response, The process of borderlands situation analysis can be divided into four steps, which are information collection, comprehension, projection and monitoring (Chen et al., 2013a; Roy, 2001). While, the collection of borderlands situation information is a tedious and complicated work. Officers always have to browse thousands of web pages and then select some newest information related to a certain borderlands entity. Therefore, focused crawling, which can download the information relevant to a given topic from web pages automatically, may be a good choice (Almpanidis et al., 2007; Siemiński, 2009; Wu et al., 2012).

As shown in Figure 1, focused crawling needs some seed URLs as its beginning. By requesting the URLs, it can extract and parse the web contents and sub-links in the web pages. Furthermore, it will calculate the relevance value between web contents and a given topic that can be defined as a series of keywords. If the relevance value is above a certain threshold, the crawling will store the web page into a repository. Otherwise, it will drop the page. After the relevance calculation, the crawling will turn to the sub-links, and circle the above operations. To advance the efficiency of the whole crawling process, it also has some sorting algorithms to assign the priority of sub-links, which can control the request list.

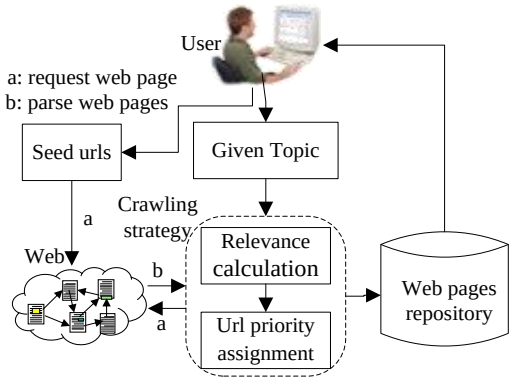

Figure 1. Workflow of the focused crawling

For the definition of a given topic in Figure1, there are mainly three methods for the moment. The first one represents a given topic as a series of keywords that are independent with each other. The keywords can be formulized as a vector based on VSM (vector space model), and a weight value will be predefined for every element in the vector (Batsakis et al., 2009; Hersovici et al., 1998; Wu et al., 2012). This method can make the formulization of a topic easier and intuitive, but it cannot express the semantic information of keywords. The second method uses an existing classified catalogue (such as Open Directory Project and Yahoo Directory) to define a given topic (Pant and Menczer, 2003; Srinivasan et al., 2005). The method could describe not only detailed information on topic itself but also some semantic information. However, the classified catalogues usually serve the common users. So they lack of some domain topics, such as Borderlands Situation topics and spatial topics. The third one depends on domain ontology to depict a given topic. It mainly maps a given topic to the corresponding classes and properties through their interrelationships organized in domain ontology (Ehrig and Maedche, 2003; Ye and Ouyang, 2011). The method can consider polysemy, synonyms and other semantic information of a given topic, but different topic corresponds to different domain ontology, whose construction is a complex and

* Corresponding author.
Email address:houdongyang1986@163.com.

time-consuming task. What's more, all above methods are unable to discriminate common keywords and place names. They treat place names in the same way as other common keywords, and the spatial characteristics of place names are ignored, which may reduce the precision of the focused crawling methods.

Another important process of focused crawling is the relevance calculation, which is shown in Figure 1. In general, the relevance calculation is performed based on vector space model. In the model, topic and web page is formalized as weight vectors of keywords, and the cosine of the two vectors indicate the relevance. In order to improve the accuracy of the relevance, term frequency (tf), term frequency- inverted document frequency (tf-idf) and positions (title and content) of keywords are general used to represent weight of keywords (Hersovici et al., 1998; Yang et al., 2010). Besides, polysemy and synonyms of keywords are also considered through WordNet and ontology (LIU and DU, 2012; Liu et al., 2011; Siemiński, 2009; Ye and Ouyang, 2011). These methods can improve the accuracy of the relevance to some extent. However, they do not consider spatial relations of place names, such as overlap and adjacency relations etc., which may affect relevance between web pages and the given topic.

While, geographic information retrieval (GIR) considers place names and theirs spatial relations, but GIR is a retrieval technology, which can limit space scope of queries and handle queries in the form of the triplet of <theme><spatial relationship><location> (Frontiera et al., 2008; Jones and Purves, 2008; Purves et al., 2007). However, focused crawling is a crawling technology, which can provide web pages repository for GIR, as shown in Figure 1.

According to the discussion above, we could find that traditional focused crawling methods lack the ability to deal with place names and their spatial relations, which may lead to the omission of many useful web pages and the acquisition of many useless web pages. Besides, place names play an important role in Borderlands Situation topics, such as the topic of North Korea Nuclear Issue. Therefore, we develop a new focused crawling method considering place names and theirs key spatial relations to improve the precision and recall for borderlands situation analysis in this paper. In the method, place names and common keywords are represented separately, and some key spatial relations are used in the relevance calculation between the given topic and web pages. Besides, based on the proposed approach, we develop an information collection system for borderlands situation analysis (ICS-BSA), which can download, index and search borderlands situation information automatically.

The rest of the paper is set out as follows. In Section 2, the proposed focused crawling method is described in detail. In Section 3, the ICS-BSA is designed and implemented to validate our proposed method. Finally, the conclusion is discussed in Section 4.

2. KEY SPATIAL RELATION-BASED FOCUSED CRAWLING METHOD

2.1 Spatial characteristic of place names in web pages

Each geographic entity is associated with a specific location on the Earth's surface and has spatial relations with other geographic entities (Zhang et al., 2008). And place names are the proper names of geographic entities, such as areas, regions, localities, cities, suburbs and towns etc. (Commission, 2007). Therefore, place names also can be used to express locations and be qualified by spatial relations. So locations and spatial relations are the two main spatial characteristics of place names, and they are also the two main distinctions between place names and common keywords.

As discussed above, each place name can be assigned a specific location. Therefore, the given topic including place names can be related to certain locations through the spatial characteristic of location. For instance, the topic of North Korea Nuclear Issue is associated with the location of North Korea. We call this the geographic scope of the given topic. Similarly, web pages whose contents contain place names also can be related to certain locations, which are called the geographic scope of web pages.

Intuitively, when web page is relevant to the given topic, their geographic scopes should be equal or have overlapped part. For example, the web page about the third nuclear test in North Korea(http://news.163.com/13/0124/11/8LVU9J3J0001121M.html) is relevant to the given topic of North Korea Nuclear Issue, and they have equal geographic scope of North Korea. Another web page about Iran nuclear issue (http://news.xinhuanet.com/2013-10/17/c_117761284.htm) is relevant to the topic of nuclear issue, but it is not relevant with the given topic of North Korea Nuclear Issue, because their geographic scopes of Iran and North Korea do not have overlapped part. Therefore, the geographic scope of web pages can be individually used to filter out the irrelevant web pages. In other words, place names can be served as an independent factor in the focused crawling method.

Similar to geographic entities, place names have three commonly used types, which are directional, distance and topological relations. The directional and distance relationships both normally refine the disjoint topological relationship (Zhang et al., 2008). Therefore, we put the emphasis on topological relationship, which is invariant under topological transformations (Chen et al., 2001). There are some key topological relations, including equal, overlap, contain, contained and disjoint etc. Intuitively, when place name B in web page is equal with place name A in the given topic, place name B must affect the relevance between the given topic and web page. When place name B in web page contains, is contained by and overlaps place name A in the given topic, place name B may affect the relevance because of their common geographic scopes. When they are disjoint with each other, it is complex to judge whether place name B impact the relevance, because some domain knowledge may change disjoint relation between A and B in social meaning. For instance, North Korea is disjoint with America, but America play an important role in North Korea Nuclear Issue. Therefore, we will consider equal, contains, contained, overlap and disjoint to compute the relevance between the given topic and web pages.

In the paper, place names and their spatial relations are organized in ontology, which we call toponym ontology. Although construction of toponym ontology is also a complex and time-consuming task, it can be used in other domains, such as Geographic information services and navigation. In the toponym ontology, we consider Chinese place names only, and assume that place names are well-defined.

2.2 Topic representation considered with place names

Place names are not considered especially in the three topic representation methods mentioned in Section 1. For example, assume given a topic T and a web page D, they will be represented as a series of keywords and weight values of keywords by utilizing the first method, as shown in equations (1) and (2),

$$V_T = \{(kw_1, w_{t1}), (kw_2, w_{t2}), \dots, (kw_n, w_{tn})\} \quad (1)$$
$$V_D = \{(kw_1, w_{d1}), (kw_2, w_{d2}), \dots, (kw_n, w_{dn})\} \quad (2)$$

Where V_T and V_D represent topic vector and web page vector, kw_i denotes the i-th keyword, w_{ti} and w_{di} represent the weight of kw_i in topic T and web page D respectively, and n depict the number of keywords in topic T. Place names can not be identified in V_T and V_D, because the method treats place names in the same way as other common keywords.

However, place names can be used as an independent factor to filter out irrelevant web pages according to their spatial characteristic of location. Therefore, place names and the rest common keywords in given topic and web pages will be represented separately in the paper.

Common keywords in topic T and web page D will be represented as common keywords vector CKV_T and CKV_D respectively, as shown in equation (3) and (4),

$$CKV_T = \{(kw_1, w_{t1}), (kw_2, w_{t2}), \dots, (kw_k, w_{tk})\} \quad (3)$$
$$CKV_D = \{(kw_1, w_{d1}), (kw_2, w_{d2}), \dots, (kw_k, w_{dk})\} \quad (4)$$

Where w_{ti} and w_{di} indict the weight of i-th common keyword kw_i in topic T and web page D separately, and k represents the number of common keywords in topic T.

In addition to place names in topic T and web page D, some other place names are also selected from the toponym ontology to extend the topic T and web page D through spatial relations. Then they can be represented as place names vector TVT and TVD separately, as shown in equation (5) and (6),

$$TV_T = \{(pl_1, w_{t1}), (pl_2, w_{t2}), \dots, (pl_k, w_{tm})\} \quad (5)$$
$$TV_D = \{(pl_1, w_{d1}), (pl_2, w_{d2}), \dots, (pl_q, w_{dq})\} \quad (6)$$

Where w_{ti} and w_{di} represent the weight of i-th place name pl_i in topic T and web page D respectively, and m and q indict the number of place names in topic T and web page D respectively, m and q may not be equal.

The weight w_{tk} and w_{tm} can be set by experts or be calculated in predefined corpus. The weight w_{dk} and w_{dq} are often calculated through tf or tf-idf algorithm. Because computing inverse document frequency (idf) weights during crawling may be problematic (Batsakis et al., 2009), we will utilize the tf algorithm to compute them in the paper. Therefore, $w_{dk} = tf_{dk}$ and $w_{dq} = tf_{dq}$ where tf_{dk} and tf_{dq} represent the occurrence of k-th common keyword and q-th place name in the web page D.

2.3 Relevance calculation using key spatial relations

In section 2.2, the given topic and web pages both are represented as common keywords vector and place names vector. Therefore, we should calculate two relevance values, which are common keywords relevance and place names relevance.

The common keywords relevance $Sim(CKV_T, CKV_D)$ between common keywords in topic T and web page D is calculated by traditional cosine similarity formula (Batsakis et al., 2009), as shown in equation (7).

$$Sim(CKV_T, CKV_D) = \frac{\sum_{i=1}^{k} w_{ti} * w_{di}}{\sqrt{\sum_{i=1}^{k} w_{ti}^2} * \sum_{i=1}^{n} w_{di}^2} \quad (7)$$

If $Sim(CKV_T, CKV_D)$ is greater than the specific threshold σ_1, the web page D will be stored in web pages repository and the place names relevance will be implemented. If otherwise, the web page D will be abandoned.

As discussed in Section 2.1, we find that spatial relations affect the relevance calculation. Therefore, the place name relevance between the topic and web page will be considered some key spatial relations. The place names relevance $Sim(TV_T, TV_D)$ is computed through the equation (8),

$$Sim(TV_T, TV_D) = \frac{\sum_{x=1}^{m} \sum_{y=1}^{q} [w_{tx} * w_{ty} * R(tx,ty)]}{\sqrt{\sum_{x=1}^{m} w_{tx}^2} * \sqrt{\sum_{y=1}^{q} w_{ty}^2}} \quad (8)$$

Where $R(tx, ty)$ represents spatial relevant factor between places names tx and ty. The spatial relevant factor reflects the influence of key spatial relations on place names relevance. It is computed by equation (9),

$$R(tx, ty) = \begin{cases} 1 & \text{equal} \\ R_1(tp, tq) & \text{contain} \\ 0.4 & \text{overlap} \\ f(tx, ty) & \text{adjacency} \end{cases} \quad (9)$$

Where $R_1(tp, tq)$ means the spatial relevant value when tp contains tq and $f(tx, ty)$ represents the spatial relevant value when tx is adjacency with tx. $f(tx, ty)$ Value is different with domain knowledge. Therefore, we set $f(tx, ty)$ as 0 in the paper. $R_1(tp, tq)$is computed by equation (10),

$$R_1(tp, tq) = \begin{cases} 0.2^{(4-x)} & x = 1,2,3 \\ 0 & x > 3 \end{cases} \quad (10)$$

Where x denote hierarchical distance in toponym ontology.

If $Sim(TV_T, TV_D)$ is greater than the specific thresholdσ_2, the document is judged to be relevant with the given topic. If otherwise, the web page D will be abandoned.

At last, the finial relevance $Sim(V_T, V_D)$, which will be utilized in assigning URL queue priority, is computed by the weighted average of $Sim(CKV_T, CKV_D)$ and $Sim(TV_T, TV_D)$:

$$Sim(V_T, V_D) = a * Sim(CKV_T, CKV_D) + b * Sim(TV_T, TV_D) \quad (11)$$

Where a and b are weighted factors, and they satisfy $a + b = 1$. In the paper, a is assigned to 0.4 and b is assigned to 0.6.

Note that if the specific topic does not contain place names, the method will be the same to the traditional method that is only the first step is implemented.

2.4 The workflow of KFRs-FC

The proposed focused crawling method KFRs-FC is based on place names and theirs key spatial relations. Figure 2 shows the workflow of the KFRs-FC method. First, user assigns ***Given Topic*** and ***Seed URLs***. The ***Given Topic*** is represented as

common keywords vector and place names vector which is complemented by ***Toponym ontology***, as shown in section 2.2. Then, the KFR-FC method is beginning with Seed URLs through ***requesting web pages and parsing web pages***. After parsing web pages, ***Relevance calculation*** is implemented as shown in section 2.3. If common relevance and toponym relevance both are greater than the given threshold, the web page will be stored in ***web pages repository*** and the two values will be utilized into ***URL priority assignment***. At last, URLs in ***URL priority assignment*** will continue to being submitted for request web pages until ***URL priority assignment*** is empty or other conditions is fulfilled.

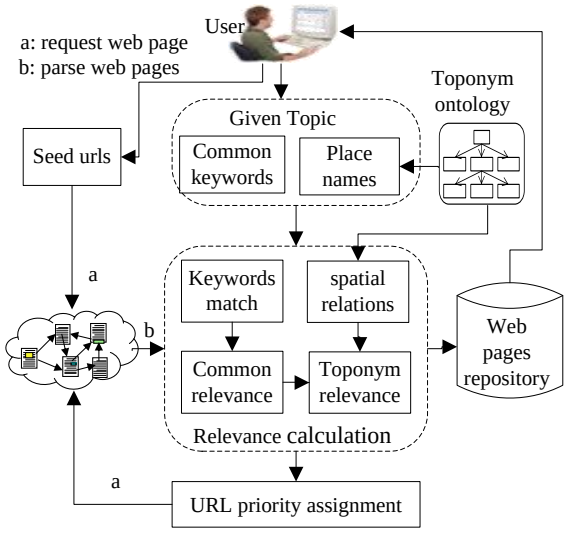

Figure 2. Workflow of focused crawling based on place names and spatial relations

There are two main differences between the KFR-FC method and traditional focused crawling methods. One is topic representation method. The given topic is divided into common keywords and place names in KFR-FC method, while the given topic is represented as one keywords vector in traditional methods. The other is relevance calculation method. The relevance will be computed step by step, where common relevance is firstly computed through keywords match method and then place names (toponyms) relevance is calculated through considering spatial relations.

3. IMPLEMENTATION AND ANAYSIS

3.1 Develop ICS-BSA System by using KFR-FC

We implement proposed focused crawling method into an information collection system for borderlands situation analysis (ICS-BSA) based on the Microsoft .NET framework 3.5. The prototype system can download and index the information relevant to given topic automatically. Moreover, it can provide query service in the form of text list and map. The main goal of the system is to enable officers to customize the process of crawling for borderlands situation topic and to retrieval relevant information from web pages repository.

3.1.1 Modules of the ICS-BSA

The prototype system contains the ***toponym ontology*** and four main modules of information agents, including ***focused crawling***, ***information indexing***, ***information retrieval*** and ***user query interfac***e, as shown in Figure 3. The focused crawling and information indexing module is a desktop application based on C# win form. The information retrieval module and user query interface is a web application based on ASP.net in Browser/Server architecture.

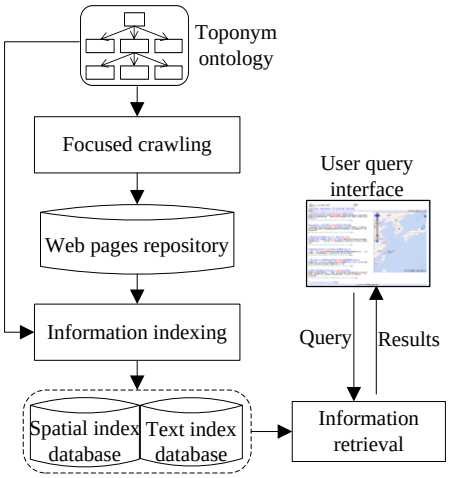

Figure 3. Design of the IGS-BAS system

The ***toponym ontology*** is applied not only in focused crawling module for topic representation and relevance calculation as discussed in Section 2, but also in information indexing module for constructing spatial index. The ontology is edited by Protégé Ontology Editor (see http://protege.stanford.edu/ for more information) and is integrated into the ICS-BSA system through an open source soft dotNetRDF (see http://www.dotnetrdf.org/default.asp for more information).

The ***focused crawling module*** is implemented based on the proposed focused crawling method as discussed in Section 2. The module is responsible for downloading web pages relevant with borderlands situation topic to serve for the information indexing module. The focused crawling module can run periodically and allow multitask simultaneous operation. The main ***graphic user interface*** (GUI) of the module is shown in Figure 4. Through the GUI, Borderlands workers can set task parameters such as topic, timing parameters and thread number etc. after clicking new task button, and they can start the task after clicking run task button. Besides, Borderlands workers also can monitor the process of the crawling.

The ***information indexing module*** is responsible for indexing the downloaded web pages to serve for the information retrieval module. The module is implemented by Lucenc.net API (see http://blogs.apache.org/lucenenet/ for more information). In the module, text information about web pages is indexed as an inverted file structure. Besides, in order to locate and visualize web pages on map, spatial location of web page is also indexed as an inverted file structure. In the paper, the highest occurrence frequency of place name in the web page is simply considered as its spatial location. Longitude and latitude of the place name are obtained from toponym ontology.

The ***information retrieval module*** is responsible for searching and ranking the information from the index database. The module is also implemented by Lucenc.net API. In the module, we just adopt keywords matching method to search relevant results. The results are ranked at the descending order of the relevance value, and also can be ranked at the descending order of publish time of web pages.

The ***user query interface*** is a bridge connecting users and the information retrieval module, the aim of which is to submit

query term into the system and display the returned results. It comprises the five parts including query term input box, search button, statistics button, textual list and map visualization, as shown in Figure 5. When user input query term and click search button, relevant results will be displayed in textual list and map visualization. When user input query term in form of <inclass:topic name> and click statistics button, a time trend figure of the topic will be displayed in another interface. The textual list part contains title, abstract, URL, publish time and a full text link. The map visualization part which is implemented by OpenLayers API (see http://www.openlayers.org/ for more information) contains number icons corresponding to ranked textual information and some simple map tool such as pan, zoom and modify etc. When user click the number icon, a label box which contains place name, longitude and latitude will be displayed and the corresponding result in textual list will be highlight.

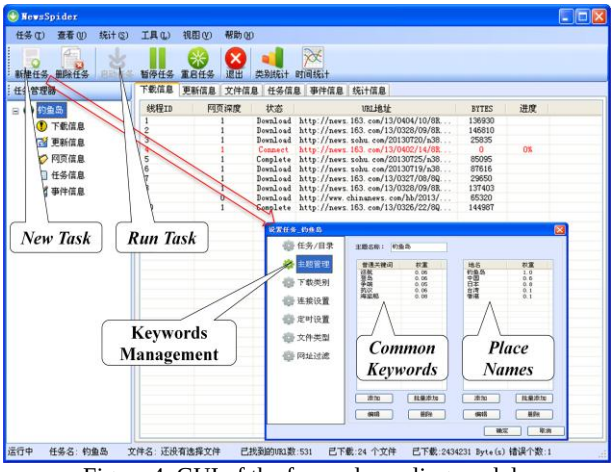

Figure 4. GUI of the focused crawling module

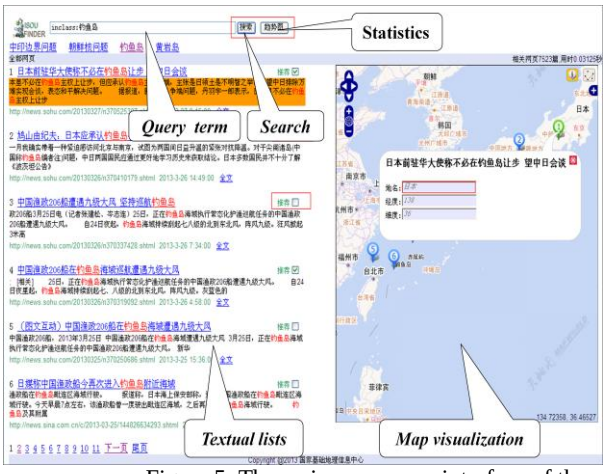

Figure 5. The main user query interface of the system

3.2 Experimental analysis

3.2.1 Evaluation metrics

The two most frequent and basic evaluation metrics for focused crawling method effectiveness are precision and recall (Manning et al., 2008; Peng and Liu, 2013). Precision represents the fraction of relevant web pages in web pages crawled (Chen et al., 2013b; Peng and Liu, 2013), as shown in equation (12), where p denotes precision, CR represents the number of crawled relevant web pages and TC represents the total number of crawled web pages. The higher precision value implies the focused crawling method has better ability to filter out irrelevant web pages. Recall is the fraction of relevant web pages crawled in total relevant web pages (Peng and Liu, 2013), as shown in equation (13), where r represents recall and TR denotes the total number of relevant web pages in the whole web. The higher recall value means the focused crawling method has better capacity to obtain relevant web pages. However, the total number of relevant web pages for the given topic is unknown, which means that the true recall is hardly to measure. Therefore, in the paper a web pages collection in which TC and TR are determinate is constructed to compute precision and recall.

$$p = \frac{CR}{TC} \tag{12}$$

$$r = \frac{CR}{TR} \tag{13}$$

Although in theory precision and recall are not related, in practice high precision is achieved almost always at the expense of recall and high recall is achieved at the expense of precision (Liu, 2010). Thus, a trade-off metric F-score which is the harmonic mean of precision and recall is adopted in the paper (Liu, 2010; Manning et al., 2008), as shown in equation (14). Because the harmonic mean of two numbers tends to be closer to the smaller of the two, the high F-score value means that both precision and recall must be high (Liu, 2010).

$$F = \frac{2pr}{p+r} * 100\% \tag{14}$$

3.2.2 Sample data

In order to validate the proposed focused crawling method, a simple experiment is implemented. In the experiment, the given topic is North Korea Nuclear Issue, which is an international hotspot issue and takes place in borderlands of China. According to high term frequency, we select ten meaningful keywords to represent the North Korea Nuclear Issue topic from 100 sample web pages which are relevant to North Korea Nuclear Issue and are dealt with Chinese word segmentation, deleting stop word and other process. The ten keywords contain five common keywords and five place names, which represent generalized scope of the given topic, as shown in equation (15) and (16).

$$CKV_T = \begin{Bmatrix} (\text{nuclear test}, 1), (\text{nuclear gallery}, 0.29), (\text{resolution}, 0.26), \\ (\text{launch}, 0.24), (\text{sanction}, 0.21) \end{Bmatrix} \tag{15}$$

$$TV_T = \begin{Bmatrix} (\text{North Korea}, 1), (\text{South Korea}, 0.34), \\ (\text{American}, 0.2), (\text{Japan}, 0.07), (\text{China}, 0.06), \end{Bmatrix} \tag{16}$$

Where CKV_T represent common keywords and TV_T denotes place names. The keywords weight is computed in corpus through normalized tf algorithm (Liu, 2010).

To facilitate recall computation, we construct a corpus by manual, which includes 100 web pages relevant to North Korea Nuclear Issue topic and different with above 100 web pages for topic, 50 web pages relevant to Iran Nuclear Issue topic and 50 web pages relevant to other topics. In the corpus, the total number of relevant web pages for the given topic is known in advance.

3.2.3 Comparison with traditional method

In the experiment, traditional focused crawling method based on vector space model is also implemented as a comparison. The Figure 6 shows the results of the relevance through the traditional focused crawling method. Figure 7 and Figure 8 represent common relevance value and toponym relevance value through the proposed focused crawling method. In the three figures, x axis represents the web page number, where web pages number 1 to 100 means irrelevant and number 101 to 200 represent relevant web pages, y axis denotes the relevance value, and the dashed line is the dividing line between actual irrelevant and relevant web pages. And we set 0.5 as the threshold value.

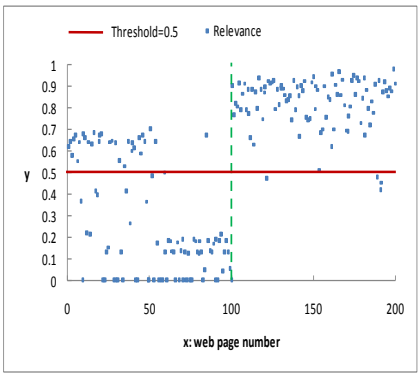

Figure 6 Relevance in traditional method

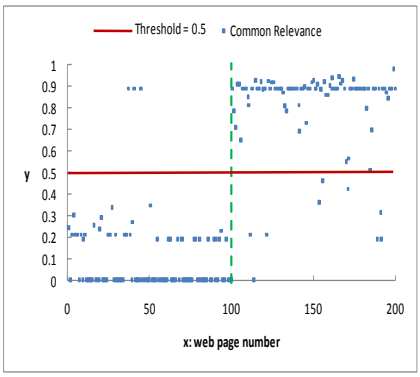

Figure 7 Common Relevance in proposed method

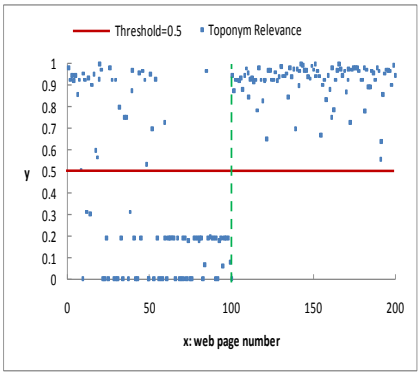

Figure 8 Toponym Relevance in proposed method

In the Figure 6, there are 128 web pages and 96 actual relevant web pages whose relevance values are greater than threshold value 0.5. Therefore, the total number of crawled web pages TC is 128 and the number of crawled relevant web pages CR is 96. In the Figure 7, there are 3 web pages numbered 37, 41 and 45 whose common relevance values are greater than threshold value 0.5. In the Figure 8, the toponym relevance values of 37, 41 and 45 web pages are smaller than threshold 0.5. And there are 91 web pages in number 101 to 200 whose common relevance value and toponym relevance value are greater than threshold value 0.5. Thus, the total number of crawled web pages TC is 91 and the number of crawled relevant web pages CR is also 91.

According to equations (12), (13) and (14), we can compute the precision, recall and F-score values. Table 1 shows the final results.

Table 1 the precision, recall and F-score values

	CR	TC	TR	Precision	Recall	F-score
Traditional method	96	128	100	75%	96%	84.2%
proposed method	91	91	100	100%	91%	95.3%

In table 1, we find that precision of the traditional method is lower than precision and F-score value of the proposed method, while recall of the traditional method is higher than recall of the proposed method. Compared with traditional method, the higher precision value implies the proposed focused crawling method has better ability to filter out irrelevant web pages and the lower recall value means the focused crawling method has poor capacity to obtain relevant web pages. However, F-score value of the proposed method is higher 11.1% than traditional method, which means the proposed focused crawling method is better than traditional method.

4. CONCLUSIONS

Traditional focused crawling methods treat place names in the same way as other common keywords. In order to solve the problem, we proposed a key spatial relations-based focused crawling method according to the spatial characteristics of place names in the paper. The proposed focused crawling method has the ability to deal with place names and theirs spatial relations. In the proposed method, the borderlands situation topic is represented as common keywords and place names separately. And some of the spatial relations related to web pages crawling were used in the relevance calculation between the given topic and web pages. Besides, we developed an information collection system for borderlands situation analysis based on the method. The system can download, index and search borderlands situation information automatically. As shown in the experiment, the F-Score value of the proposed method is increased by 11% compared with traditional focused crawling method. It means that the proposed method has better performance on filtering out the irrelevant web pages than the traditional method.

REFERENCES

Almpanidis, G., Kotropoulos, C., Pitas, I., 2007. Combining text and link analysis for focused crawling—An application for vertical search engines. Information Systems 32, 886-908.

Batsakis, S., Petrakis, E.G., Milios, E., 2009. Improving the performance of focused web crawlers. Data & Knowledge Engineering 68, 1001-1013.

Chen, J., Ge, Y., Hua, Y., Wang, F., Yang, S., Qu, B., Li, R., 2013a. Digital Border-land: Conceptual FrameworkandResearchAgenda. ActaGeodaetica et CartographicaSinica 40, 502-508.

Chen, J., Li, C., Li, Z., Gold, C., 2001. A Voronoi-based 9-intersection model for spatial relations. International Journal of Geographical Information Science 15, 201-220.

Chen, J., Wu, H., Li, S., Liao, A., He, C., Peng, S., 2013b. Temporal logic and operation relations based knowledge representation for land cover change web services. ISPRS Journal of Photogrammetry and Remote Sensing.

Commission, E., 2007. Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). Official Journal of the European Union 50, 1-14.

Ehrig, M., Maedche, A., 2003. Ontology-focused crawling of Web documents, in: BG, L. (Ed.), the 2003 ACM symposium on Applied computing. ACM Press, New York, pp. 1174-1178.

Frontiera, P., Larson, R., Radke, J., 2008. A comparison of geometric approaches to assessing spatial similarity for GIR. International Journal of Geographical Information Science 22, 337-360.

Hersovici, M., Jacovi, M., Maarek, Y.S., Pelleg, D., Shtalhaim, M., Ur, S., 1998. The shark-search algorithm. An application: tailored Web site mapping. Computer Networks and ISDN Systems 30, 317-326.

Jones, C.B., Purves, R.S., 2008. Geographical information retrieval. International Journal of Geographical Information Science 22, 219-228.

Liu, B., 2010. Web data mining: exploring hyperlinks, contents, and usage data, Second Edition ed. Springer-Verlag New York Incorporated, New York.

LIU, W., DU, Y., 2012. An Improved Topic-specific Crawling Approach Based on Semantic Similarity Vector Space Model. Journal of Computational Information Systems 8, 8605-8612.

Liu, Z., Du, Y., Zhao, Y., 2011. Focused crawler based on domain ontology and fca. Journal of Information & Computational Science 8, 1909-1917.

Manning, C.D., Raghavan, P., Schütze, H., 2008. Introduction to information retrieval. Cambridge University Press Cambridge.

Pant, G., Menczer, F., 2003. Topical crawling for business intelligence, Research and Advanced Technology for Digital Libraries. Springer, pp. 233-244.

Peng, T., Liu, L., 2013. Focused crawling enhanced by CBP–SLC. Knowledge-Based Systems 51, 15-26.

Purves, R.S., Clough, P., Jones, C.B., Arampatzis, A., Bucher, B., Finch, D., Fu, G., Joho, H., Syed, A.K., Vaid, S., 2007. The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet. International Journal of Geographical Information Science 21, 717-745.

Roy, J., 2001. From data fusion to situation analysis, Proceedings of Fourth International Conference on Information Fusion.

Siemiński, A., 2009. Using WordNet to measure the similarity of link texts, in: T, N.N., R, K., M, C.S. (Eds.), First International Conference ICCCI. Springer, Wroclaw, Poland, pp. 720-731.

Srinivasan, P., Menczer, F., Pant, G., 2005. A general evaluation framework for topical crawlers. Inf Retrieval 8, 417-447.

Wu, H., Liao, A., He, C., Hou, D., 2012. Topic-Relevance Based Crawler for Geographic Information Web Services. Geography and Geo-Information Science 28, 27-30.

Yang, X., Sui, A., Tang, Z., 2010. Topical Crawler based on multi-level vector space model and optimized hyperlink chosen strategy, in: Sun F, Wang Y, Lu J, Zhang B, W, K., A, Z.L. (Eds.), Cognitive Informatics (ICCI), 2010 9th IEEE International Conference on. IEEE, pp. 430-435.

Ye, Y., Ouyang, D., 2011. Semantic-Based Focused Crawling Approach. Journal of Software 22, 2075-2088.

Zhang, Y., Gao, Y., Xue, L., Shen, S., Chen, K., 2008. A common sense geographic knowledge base for GIR. Science in China Series E: Technological Sciences 51, 26-37.