*Sustainability* **2014**, *6*, 6529-6552; doi:10.3390/su6106529

OPEN ACCESS

*sustainability*

**ISSN 2071-1050**

www.mdpi.com/journal/sustainability

*Article*

# A Focused Crawler for Borderlands Situation Information with Geographical Properties of Place Names

**Dongyang Hou [1,2], Hao Wu [2,*], Jun Chen [1,2] and Ran Li [2]**

[1] School of Environment Science and Spatial Informatics, China University of Mining and Technology, Xuzhou 221116, China; E-Mails: houdongyang1986@gmail.com (D.H.); chenjun@nsdi.gov.cn (J.C.)

[2] National Geomatics Center of China, 28 Lianhuachi West Road, Beijing 100830, China; E-Mail: liran@nsdi.gov.cn

* Author to whom correspondence should be addressed; E-Mail: wuhao@nsdi.gov.cn; Tel.: + 86-10-6388-0216.

External Editor: Shangyi Zhou

*Received: 3 June 2014; in revised form: 2 September 2014 / Accepted: 5 September 2014 / Published: 29 September 2014*

---

**Abstract:** Place name is an important ingredient of borderlands situation information and plays a significant role in collecting them from the Internet with focused crawlers. However, current focused crawlers treat place name in the same way as any other common keyword, which has no geographical properties. This may reduce the effectiveness of focused crawlers. To solve the problem, this paper firstly discusses the importance of place name in focused crawlers in terms of location and spatial relation, and, then, proposes the two-tuple-based topic representation method to express place name and common keyword, respectively. Afterwards, spatial relations between place names are introduced to calculate the relevance of given topics and webpages, which can make the calculation process more accurately. On the basis of the above, a focused crawler prototype for borderlands situation information collection is designed and implemented. The crawling speed and F-Score are adopted to evaluate its efficiency and effectiveness. Experimental results indicate that the efficiency of our proposed focused crawler is consistent with the polite access interval and it could meet the daily demand of borderlands situation information collection. Additionally, the F-Score value of our proposed focused crawler increases by around 7%, which means that our proposed focused crawler is more effective than the traditional best-first focused crawler.

**Keywords:** focused crawler; place name; web information collection; borderlands situation; relevance calculation; spatial relations

## 1. Introduction

Borderlands situation information refers to the events and tendencies in borderlands regions [1,2]. It is extremely important for emergency risk assessment, geopolitical analysis, decision-making and other borderlands studies [3–5]. With the availability of ever-increasing World Wide Web resources, borderlands situation information can be collected and/or mined from the Internet with search engines, which use crawlers to continuously fetch a large number of webpages and return ranked results with keywords-based similarity comparison [6,7]. In general, borderlands events and tendencies have an explicit or implicit geographical dimension [3,8]. For instance, "North Korea Nuclear Issue" was an international event, which occurred in the Korean Peninsula, with impacts on the surrounding area. Therefore, geographical location needs to be taken into account in search engines during borderlands situation information collection.

Search engines can be categorized into general-purpose search engine and topic-specific search engine [9]. A general-purpose search engine (such as Google or Baidu) uses a general crawler to endlessly obtain a huge number of webpages and put them into an indexed database [6,9]. Since a breadth-first or depth-first strategy is conducted without keywords by a general crawler, the resulted indexed database contains hundreds of millions of webpages with an enormous number of duplicates [10]. The keywords are defined by users and used only for calculating their relevance with the indexed webpages to deliver final searching results with rankings [11]. Taking "North Korea Nuclear Issue" as a querying topic, a search on Google on 21 December 2013, returned as many as 55,700,000 results. Many of the results are unrelated webpages, belonging to diverse topics, such as "North Korean economy", "Iran nuclear issue", "Issue", and "North". Because the indexed database is so large, many webpages about other topics containing one or more keywords of the querying topic are captured [12,13]. In order to reduce the unrelated webpages, topic-specific search engines, targeted for collecting webpages, relevant to a specific topic, have been proposed [9,14]. Unlike the general-purpose search engine, a topic-specific search engine often employs a focused crawler with a given topic to selectively fetch webpages [15]. In other words, a topic-specific search engine utilizes two groups of keywords. The first is the topic keywords with weights and are defined by experts with prior knowledge [9], and they are utilized in the focused crawler to reduce numbers of unrelated webpages in the indexed database. The second are the user-defined querying keywords without weights [11], and they are used for the relevance calculation with the indexed webpages. With the introduction of topic keywords, a large number of unrelated webpages can be filtered out or reduced. Therefore, the utilization of the querying keywords can better satisfy a user's actual search interests in a filtered indexed database. This makes the final results of a topic-specific search engine better than a general-purpose search engine.

However, a traditional topic-specific search engine can only filter out or reduce those webpages about the topics "issue" or "north" in borderlands situation information collection. A number of

unrelated webpages may still remain, such as those about the topics "North Korean economy" and "Iran nuclear issue". The reason is that "North Korea" and "Iran" are treated by a topic-specific search engine in the same way as other common keywords, is that they have no geographical properties. In fact, "North Korea" and "Iran" are place names, referring to specific geographical locations on the Earth [16]. They should be used as special keywords to limit geographical scopes concerned by users. For instance, webpages about the "Iran nuclear issue" describe the events that occurred in the geographical scope of "Iran". Its geographical scope is disjoint with "North Korea", concerned by users. Therefore, with the comparison of two special keywords "Iran" and "North Korea", all the webpages related to "Iran nuclear issue" could be filtered out.

This paper reports on the utilization of place name as special keyword in topic-specific search engine and the development of a new focused crawler for improved searching effectiveness of borderlands situation information. A two-tuples of topic representation method is proposed to represent place names and common keywords, respectively. Some key spatial relations of place names are abstracted and used for relevance calculation. With the use of geographical locations and spatial relations of place names, the new focused crawler has improved its effectiveness in comparison with traditional focused crawler. The rest of the paper is structured as follows. Section 2 gives a review of related works on focused crawlers and geographic information retrieval. Section 3 presents a new focused crawling method by the two-tuples-based topic representation and hierarchical relevance calculation with key spatial relations of place names. The design and implementation of an Information Collection Prototype for Borderlands Situation Information (ICP-BSI) is described in Section 4. Section 5 gives the results of experiments and analysis, including the efficiency and effectiveness analysis. Conclusions and future works are described in Section 6.

## 2. Related Works

In this section, we will review the related works in three parts. The first part is the review of traditional focused crawlers. The second part discusses the geographic information retrieval and then analyzes what we can learn from it. The last one describes "spatial" focused crawlers related to geographical information.

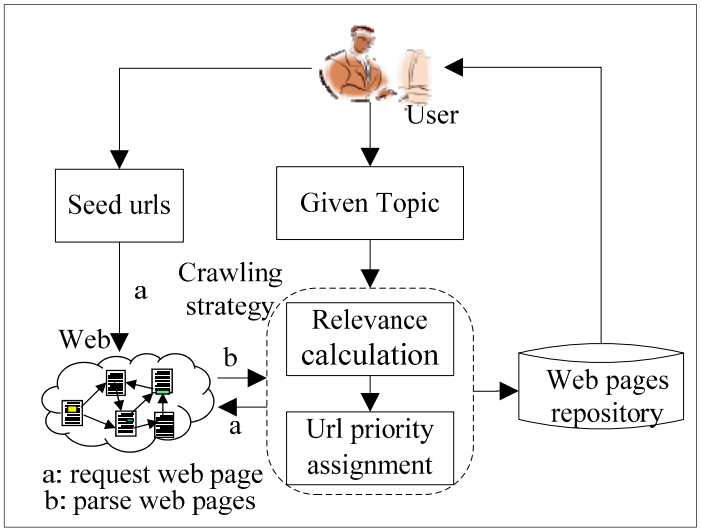
### *2.1. Traditional Focused Crawlers*

A focused crawler, also known as topic or topic-driven crawler, is a program/software or programmed script, which automatically collects webpages relevant to a given topic [9,17,18]. It usually contains four primary modules, including seed Uniform Resource Locators (URLs), given topic, relevance calculation and URL priority assignment, as shown in Figure 1. *Seed URLs* are the entrance for a focused crawler to the World Wide Web. *Given Topic* is used to determine the domain for crawling. Therefore, given topic and its representation can affect the effectiveness of focused crawler. *Relevance Calculation* is responsible for determining whether a webpage is related to given topic. *URL Priority Assignment* can determine the order of crawling, which can also affect the efficiency of focused crawler.

The main process of focused crawler can be divided into four steps. Firstly, the user should select some seed URLs by experts or common search engines and define a given topic as initial value. Then, the focused crawler starts to fetch webpages after requesting these seed URLs [19]. Secondly, the

focused crawler will parse these webpages to obtain their web contents and sub-links. Furthermore, it will calculate the relevance between web contents and given topic. If the relevance value is above a certain threshold, it means that these webpages are relevant to the given topic and will be stored into a webpage repository. Otherwise, these webpages will be discarded immediately. Thirdly, the crawling process will turn to the sub-links and repeat the above operations. Finally, to improve the efficiency of the whole crawling process, some ranking algorithms are implemented to assign the priority of sub-links, which can control the request list.

**Figure 1.** Process of focused crawler.




Currently, various focused crawlers have been designed and have been widely used in many domains, such as search engines, web data mining, business intelligence, social network analysis and many more domains [20–23]. Since the focused crawler was first launched in 1999 [14], the main challenges for designing an effective focused crawler have always in three aspects, which are the representation of given topic, the relevance calculation, and the assignment of URL priority [24].

For the presentation of the given topic, there are mainly three methods. The first one represents the given topic as a series of independent keywords. These keywords can be formalized as a vector based on VSM (Vector Space Model), and a weight value will be predefined for every element in the vector [18,22,25]. This method can make the formalization of given topic easier and intuitive, but it fails to express the semantic information of different keywords. The second method adopts an existing classified catalogue (such as Open Directory Project and Yahoo Directory) to define a given topic [26,27]. This method could describe, not only the detailed information about given topic itself, but also some semantic information. However, some domain topics (e.g., borderlands situation topics and spatial topics) are incomplete or missing in the classified catalogues. The third method depends on domain ontology to depict a given topic. It mainly maps given topic to the corresponding classes, properties and individuals through their interrelations organized in domain ontology [28,29]. This method can consider polysemy, synonyms and other semantic information of given topic, but a different topic corresponds to a different domain ontology, of which construction is a complex and time-consuming task. Furthermore, all the above-mentioned methods fail to discriminate common keywords with place names. They just treat place name in the same way as any other common

keyword, which may weaken the role of the place name. Therefore, this paper will represent place names and common keywords in given topic about borderlands situation respectively.

Generally, the relevance calculation is performed based on a vector space model. In the model, the given topic and webpage are formalized as weight vectors of keywords, and the cosine of two vectors indicates the relevance [17,30]. In this method, the relevance value is just a composite value of common keyword and place name, and focused crawler will use the only relevance value to determine whether webpages are relevant to given topic. The composite value may weaken separate effects of common keyword and place name, which may lead to low effectiveness of focused crawler. Therefore, the topical relevance will be calculated and judged step-by-step based on the relevance of place names and relevance of common keywords. Additionally, in traditional relevance calculation, various characteristics of keywords are commonly used to improve its accuracy, such as term frequency (tf, the number of occurrences of keywords in a webpage), inverted document frequency (idf, a measure of whether the keyword is common or rare across all webpages) and positions (title, anchor text and content in which the keyword occurs) [25,31]. Polysemy and synonyms of keywords are also applied through WordNet or ontology [17,29,30,32]. These characteristics can improve the accuracy of relevance calculation to some extent. However, spatial relations of place names are not considered, such as equal, contain, contained and overlap. Hence, this paper will calculate the relevance of place names considering some key spatial relations.

### *2.2. Geographic Information Retrieval with Place Name*

Geographic Information Retrieval (GIR) can be seen as a specialized branch of traditional Information Retrieval [33]. It focuses on retrieving geographically information from the Internet [34]. In GIR, webpages and user queries are assigned to one or more geographical scopes through place names and their spatial relations [35,36]. The geographic scopes of webpages and user queries are represented as point or polygons in form of geographic coordinates (latitude and longitude) and then utilized to calculate the relevance between them for ranking the search results [34,37,38]. Finally, GIR will return the webpages that meet the relevance of text and relevance of geographical scopes at the same time [37,39]. For instance, the SPIRIT prototype [37,40] is capable of handling queries in the form of the triplet of "theme, spatial relation, location". It assigns geographical scopes to webpages and ranks webpages according to both textual and spatial relevance. Frontiera *et al.* [38] employs the minimum bounding box and the convex hull to represent the geographical scopes and computes a spatial similarity score for a query–document pair based on logistic regression models.

In general, GIR and focused crawler are two different parts in a search engine. GIR puts emphasis on indexing and searching spatial information from a webpages repository [41], while focused crawler focuses on collecting relevant information from the Internet in order to generate the webpage repository [10]. However, the process of judging the relevance of webpages is analogous. Since place name is treated as special keyword to assign webpages and user queries with geographical scopes in GIR and the geographical scopes are used to calculate spatial relevance for ranking the final results. Similarly, we can also treat a place name as special keyword and utilize the geographical scopes to indicate the role of place name in focused crawler.

### *2.3. "Spatial" Focused Crawler*

Recently, some researchers have been involved in focused crawlers, related to geographical information directly or indirectly [41–44]. For example, Li *et al*. [42] and Patil *et al*. [43] design focused crawlers for collecting geospatial web services, such as a web map service and web feature service. The collected results of these crawlers are related to geographical information, but they do not take place names into consider. Kozanidis and Stamou [41] propose a geo-focused crawler for GIR. While, the geo-focused crawler only fetches webpages containing one or more place names. It introduces the average distance between place names of the URL into URL priority queue. However, our focused crawler concentrates on topic representation and relevance calculation towards hybrid topics, which contain, not only place names, but also common keywords. Furthermore, Ahlers and Boll [44] proposed a geospatially focused crawler, which employs a Geoparser, to identify relevant pages containing place names. It puts emphasis on predicating geospatially URL priority with Naive Bayes Classifier. However, there are no details about topic representation and relevance calculation in the crawler, and spatial relations are also not covered.

Additionally, focused crawler also has been used to collect social media data [20,21]. For example, Catanese *et al*. [20] and Gjoka *et al*. [21] both develop a breadth-first-search crawler to collect users in Facebook and analyze their relationships. Although place names are not used in the crawling process for collecting user information, place names could be used to collect social media content for conducting spatial analysis or understanding group change [45,46]. Perhaps, a focused crawler considering breadth-first-search and place names may be applied to collect social media content in the future.

## 3. Focused Crawling with Place Name

### *3.1. Using Place Name in Focused Crawler*

Place name is generally used to represent a specific location on the Earth's surface, and can be related to other place names by their spatial relations. For example, Beijing is located in Northern China and is adjacent to Tianjin. Therefore, place name has two specific properties of location and spatial relation, which are the two main distinctions between place names and common keywords.

In focused crawler, a given topic about borderlands situation information can be reflected to certain locations through locations of place names. For instance, the topic "North Korea Nuclear Issue" is associated with the location of North Korea. Analogous to GIR, it is called geographical scope of given topic. Similarly, webpages also can be related to certain locations using place names' locations [47], which are called the geographical scope of webpages.

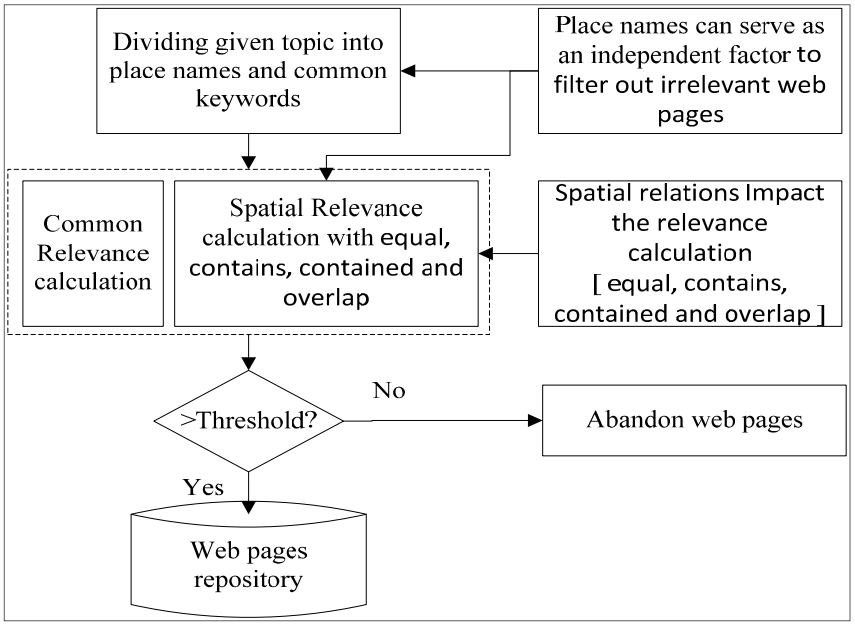
Intuitively, when a webpage is relevant to given topic, their geographic scopes will be equal or overlapped. For example, the webpage about the third nuclear test in North Korea [48] is relevant to given topic "North Korea Nuclear Issue", and they have equal geographic scope of North Korea. Another webpage about "Iran nuclear issue" [49] is relevant to the topic of a nuclear issue, but it is not relevant to given topic "North Korea Nuclear Issue", because the geographical scopes of Iran and North Korea are disjoint with each other. Therefore, when researchers collect borderlands situation information with focused crawler, the geographical scope of webpages can be individually used to filter out the irrelevant webpages. In other words, place name can serve as an independent factor in

focused crawler and the relevance of place names can be utilized to filter out some irrelevant webpages, which have different geographical scopes within given topic. In a similar way, the relevance of common keywords also can be applied to filter out some irrelevant webpages, which have different common keywords within a given topic.

Similar to geographic entities, a place name has three universal types of spatial relations including directional, distance and topological relations [50], which are often applied as terminologies in the discipline of a geographic information system. Both directional and distance relations normally refine the disjoint topological relation [47]. Therefore, attention is paid to topological relation, which is invariant under topological transformations [50]. There are some key topological relations, including equal, overlap, contain, contained and disjoint, *etc.* Intuitively, if a webpage and given topic include the same place name, it must affect their relevance because of the equal or overlapped geographic scopes. If a webpage and given topic have place names with spatial relations of contain, contained or overlap, these place names may impact the relevance, due to their overlapped geographic scopes. If a webpage and given topic have place names with disjoint relations, these place names may have little impact on the relevance because of no common geographic scopes. Therefore, spatial relations of equal, contain, contained, overlap and disjoint should be taken into account to compute the relevance between webpages and given topic.

To sum up, a place name can serve as an independent factor in a focused crawler, and the relevance of place names and relevance of common keywords can be utilized to filter out irrelevant webpages individually. In addition, spatial relations of equal, contains, contained and overlap may impact the relevance calculation. Therefore, we will apply place name to the focused crawler with two major steps, as shown in Figure 2. Firstly, a given topic is divided into place names and common keywords, and it is represented as a two-tuples, because place names can be used alone to filter out irrelevant webpages. Secondly, based on the first step, the spatial relations of equal, contain, contained, overlap and disjoint are introduced into the cosine formula for calculating the relevance of place names.

**Figure 2.** Framework of focused crawler using place name.



### *3.2. Two-Tuple-Based Topic Representation Method*

Place name is not especially considered in the three topic representation methods mentioned in Section 2. For example, assuming given topic $T$ and webpage $D$, 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),

$$V_T = \{(k_1, w_{T1}), (k_2, w_{T2}), ..., (k_n, w_{Tn})\} \quad (1)$$

$$V_D = \{(k_1, w_{D1}), (k_2, w_{D2}), ...(k_n, w_{Dn})\} \quad (2)$$

where, $V_T$ and $V_D$ represent topic vector and webpage vector, respectively, $k_i$ denotes the *i-th* keyword, $w_{Ti}$ and $w_{Di}$ represent the weight of $k_i$ in given topic $T$ and webpage $D$, respectively, and $n$ depict the number of keywords in given topic $T$. In vectors of $V_T$ and $V_D$, place names cannot be identified from the keywords of $k_1$, $k_2$, …, $k_n$, because the traditional method just treats place name in the same way as other common keyword.

However, as mentioned in Section 3.1, place name in a given topic can be used as the independent factor to filter out irrelevant webpages. Therefore, place names, and the rest of the common keywords in given topic and webpages, will be represented separately in the paper. That is to say, a given topic and webpages will be represented as a two-tuple in the form of "common keywords, place names". Given topic $T$ and webpage $D$ can be represented in Equations (3) and (4).

$$T = \langle V_{T-K}, \mathrm{V}_{T-PN} \rangle \quad (3)$$

$$D = \langle V_{D-K}, V_{D-PN} \rangle \quad (4)$$

where <,> represents two tuples, $V_{T\text{-}K}$ and $V_{D\text{-}K}$ denote common keyword vector of given topic $T$ and webpage $D$, respectively. In addition, $V_{T\text{-}PN}$ and $V_{D\text{-}PN}$ represent place name vector of given topic $T$ and webpage $D$, respectively.

To be specific, given topic $T$ is represented as the common keyword vectors of $V_{T\text{-}K}$ and the place name vectors of $V_{T\text{-}PN}$ separately, as shown in Equations (5) and (6).

$$V_{T-K} = \{(k_1, w_{T-k1}), (k_2, w_{T-k2}), ..., (k_s, w_{T-ks})\} \quad (5)$$

$$V_{T-PN} = \{(p_1, w_{T-p1}), (p_2, w_{T-p2}), ..., (p_m, w_{T-pm})\} \quad (6)$$

where $k_i$ represents the *i-th* common keyword and $p_i$ indicates the *i-th* place name in given topic $T$. Variables of $s$ and $m$ denote the number of common keywords and place names in given topic $T$, respectively and $s + m = n$. Variables of $w_{T\text{-}ki}$ and $w_{T\text{-}pi}$ represent the weight of a common keyword $k_i$ and the weight of a place name $p_i$ in a given topic $T$.

In this paper, common keywords of $k_1$, $k_2$, …, $k_s$ and place names $p_1$, $p_2$, …, $p_n$ are extracted from sample corpus through the maximum-frequency methods, where the sample corpus is obtained by submitting simple queries to the search engines Google and Baidu. In the maximum-frequency methods, keywords are selected through the descending order of term frequency in the sample corpus.

Similar to given topic $T$, webpage $D$ is expressed as the common keyword vectors of $V_{D\text{-}K}$ and the place name vectors of $V_{D\text{-}PN}$ respectively, as shown in Equations (7) and (8).

$$V_{D-K} = \{(k_1, w_{D-k1}), (k_2, w_{D-k2}), ..., (k_s, w_{D-ks})\} \quad (7)$$

$$V_{D-PN} = \begin{Bmatrix} (p_1, w_{D-p1}), (p_2, w_{D-p2}), ..., (p_m, w_{D-pm}), \\ (p_{m+1}, w_{D-p(m+1)}), ..., (p_{m+u}, w_{D-p(m+u)}) \end{Bmatrix} \quad (8)$$

where variables of $k_i$,*s* and *m* are the same as Equations (5) and (6). Variable of $p_i$ indicates the *i-th* place name in the webpage *D*. The variable of *u* shows the number of place names in a webpage *D* but not in a given topic *T*. Therefore, the dimension of a place name vector $V_{D\text{-}PN}$ in webpage *D* is greater or equal to the dimension of $V_{T\text{-}PN}$ in given topic *T*. Variables of $w_{D\text{-}ki}$ and $w_{D\text{-}pi}$ represent the weight of a common keyword $k_i$ and the weight of a place name $p_i$ in the webpage *D*.

Weight $w_{T\text{-}pi}$ and $w_{T\text{-}ki}$ can be set by experts or be calculated in predefined corpus. In this paper, $w_{T\text{-}pi}$ and $w_{T\text{-}ki}$ are calculated by the normalized term frequency in a predefined corpus [51], as shown in Equation (9).

$$tf_i = \frac{f_i}{\max\{f_1, f_2, ..., f_{s \text{ or m}}\}} \quad (9)$$

where, variable of $tf_i$ denotes the normalized term frequency. Variables of *s* and m are the same as Equations (5) and (6). Variables of $f_i$ represent term frequency (the number of times that $k_i$ or $p_i$ appears in the corpus) of the *i-th* common keyword or place name. The *max{…}* denotes the maximum value of $f_i$.

Weight $w_{D\text{-}pi}$ and $w_{D\text{-}ki}$ are often calculated through term frequency or term frequency-inverse document frequency algorithm [51]. However, because computing inverse document frequency may be problematic during the focused crawling process [22], the term frequency algorithm is adopted to compute them in this paper. That is to say, $w_{D\text{-}pi} = tf_{D\text{-}pi}$ and $w_{D\text{-}ki} = tf_{D\text{-}ki}$, where $tf_{D\text{-}pi}$ and $tf_{D\text{-}ki}$ represent the occurrence of *i-th* common keyword and the place name in webpage *D*.

### *3.3. Hierarchical Relevance Calculation with Key Spatial Relations*

In traditional methods, the relevance $sim(V_T, V_D)$ between a given topic *T* and a webpage *D* is computed only in one step, which is the cosine of $V_T$ and $V_D$, as shown in Equation (10) [22],

$$sim(V_T, V_D) = \frac{\sum_{i=1}^{n} w_{Ti} \times w_{Di}}{\sqrt{\sum_{i=1}^{n} w_{Ti}^2 \times \sum_{i=1}^{n} w_{Di}^2}} \quad (10)$$

where, variables are the same as Equations (1) and (2). If the relevance value $sim(V_T, V_D)$ is greater than a given threshold, it means that the webpage *D* is relevant to given topic *T* and the focused crawler will store webpage *D* in a webpage repository, otherwise, the webpage *D* is irrelevant to given topic *T* and the focused crawler will abandon webpage *D*. In this method, the relevance value $sim(V_T, V_D)$ is the only standard to determine whether the webpages are relevant to the given topic or not. This may weaken separate effects of common keywords and place names on relevance calculations. For example, if an irrelevant webpage contains most of the common keywords in a given topic and these common keywords account for a big proportion, the relevance value $sim(V_T, V_D)$ may be greater than the given threshold. However, the geographic scope of the webpage is different from the geographic

scope of given topic. Therefore, the relevance value weakens the role of place name that can filter out irrelevant webpages in the focused crawler.

In Section 3.2, given topic and webpages are both represented as a two-tuple in the form of "common keywords, place names". Based on this, and the role of place name, the relevance between a given topic and a webpage is calculated from the hierarchy of common keyword and place name. In the hierarchical method, there are two steps to calculate and judge topic relevance.

Firstly, the relevance $sim\ (V_{D\text{-}K}, V_{T\text{-}K})$ of common keywords between given topic $T$ and webpage $D$ is calculated with Equation (11).

$$sim(V_{D\text{-}K}, V_{T\text{-}K}) = \frac{\sum_{i=1}^{s} w_{T-ki} \times w_{D-ki}}{\sqrt{\sum_{i=1}^{s} w_{T-ki}^2 \times \sum_{i=1}^{s} w_{D-ki}^2}} \tag{11}$$

where, variables are the same as Equations (5) and (7). If $sim\ (V_{D\text{-}K}, V_{T\text{-}K})$ is greater than the given threshold, webpage $D$ will be judged to relevant to given topic $T$ preliminary, and the relevance of place names between given topic $T$ and webpage $D$ will continue to be computed, otherwise, webpage $D$ will be abandoned.

Secondly, the relevance $sim\ (V_{D\text{-}PN}, V_{T\text{-}PN})$ of place names between given topic $T$ and webpage $D$ will be calculated. In the two-tuple, the dimension of place name vector $V_{D\text{-}PN}$ in webpage $D$ is greater or equal to the dimension of $V_{T\text{-}PN}$ in given topic $T$. Therefore, the cosine formula cannot be directly utilized to calculate the relevance $sim\ (V_{D\text{-}PN}, V_{T\text{-}PN})$. This paper will reduce the dimension of place name vector $V_{D\text{-}PN}$ through spatial relations. That is to say, the weights of redundant place names will be transmitted to other place names through the spatial relations of equal, contain, contained and overlap, according to the discussion in Section 3.1. In addition, the changed weight will be used in the cosine formula. At last, the relevance $sim\ (V_{D\text{-}PN}, V_{T\text{-}PN})$ is calculated, as shown in Equation (12).

$$sim(V_{D\text{-}PN}, V_{T\text{-}PN}) = \frac{\sum_{i=1}^{m} (w_{T-pi} \times w_{D-pi}) + \sum_{i=1}^{m} [w_{T-pi} \times \sum_{j=m+1}^{m+u} R(p_i, p_j) \times w_{D-pj}]}{\sqrt{\sum_{i=1}^{m} w_{T-pi}^2 \times \sum_{i=1}^{m} [w_{D-pi} + \sum_{j=m+1}^{m+u} R(p_i, p_j) \times w_{D-pj}]^2}} \tag{12}$$

where, $R\ (p_i, p_j)$ represents spatial relevant factor between places names $p_i$ and $p_j$. The other variables are the same as Equations (6) and (8). If the relevance $sim\ (V_{D\text{-}PN}, V_{T\text{-}PN})$ is greater than or equal to the specific threshold, it means that webpage $D$ is relevant to given topic $T$ and the focused crawler will store webpage $D$ into a webpage repository, otherwise, webpage $D$ will be abandoned.

Spatial relevant factor $R\ (p_i, p_j)$ reflects the influence of key spatial relations on transmitting weights of place names. When spatial relation between places names $p_i$ and $p_j$ is different, the spatial relevant factor is also different. For instance, places names $p_i$ and $p_j$ are independent when they are disjoint. Therefore, spatial relevant factor of disjoint is zero. When places names $p_i$ and $p_j$ overlap, the proportion of the area of the overlapping part and geographic scope of a given topic affect weight transmission. Because it is difficult to obtain the area of the overlapping part and geographic scope of given topic, the spatial relevant factor of overlap is set as 0.4 through many experiments. Finally, the spatial relevant factor $R\ (p_i, p_j)$ is computed by Equation (13),

$$R(p_i, p_j) = \begin{cases} 1 & \text{equal} \\ R_1(p_i, p_j) & \text{contain or contained} \\ 0.4 & \text{overlap} \\ 0 & \text{disjoint} \end{cases} \quad (13)$$

where $R_1$ $(p_i, p_j)$ means the spatial relevant value when place name $p_i$ contains $p_j$ or is contained by $p_j$. The hierarchical distance between places names $p_i$ and $p_j$ mainly affect the spatial relevant value. The relation of contain represents concretization and the relation of contained means generalization, therefore, the relations of contain and contained is processed in different formulas. $R_1$ $(p_i, p_j)$ is calculated with Equation (14),

$$R_1(p_i, p_j) = \begin{cases} 0.2 \times (4 - z) & z = 1, 2, 3 \\ 0.2 \times (3 + z) & z = -1, -2 \\ 0 & z = \text{others} \end{cases} \quad (14)$$

where, z denotes hierarchical distance in toponym ontology. The negative value implies the relation of contain and positive value represents the relation of contained. In order to prevent transmitting weight excessively, when the absolute value of hierarchical distance is greater than a given value (e.g., 2 and 3), the spatial relevant value of contain or contained is zero.

In this step, spatial relations of overlap and contain are obtained by toponym ontology, which will be mentioned in Section 4.2. To be specific, a set $M$ (as shown in Equation 15) of place names, which overlap or belong to five hierarchies of place names in a given topic, are extracted from the toponym ontology before the focused crawler is started.

$$M = \{< p_j, contain, p_i, \text{z} >, < p_j, overlap, p_i >, \ldots\ldots\} \quad (15)$$

where, variable $p_i$ represents place name in given topic and variable $p_j$ represents the place name that has spatial relation of contain or overlap with the place name $p_i$. Variables of *contain* and *overlap* represent corresponding spatial relations. A variable of z is the same as Equation (14), but its value is only one of $\{1, 2, 3, -1, -2\}$. The set $M$ can avoid place names in webpages, searching in the whole toponym ontology online, which can improve the efficiency of a focused crawler.

Then, when a place name in webpages is identified by the Pan Gu Segment [52], three judging rules will be conducted. The first rule is that if the place name is in a given topic, an equal relation can be extracted. The second rule is that if the place name matches with the place name $p_j$ in the set $M$, overlap relation, contain relation and its hierarchical distance z can be obtained. The last rule is if the place name is not in given topic and set $M$, we assume the place name is disjointed with place names in a given topic.

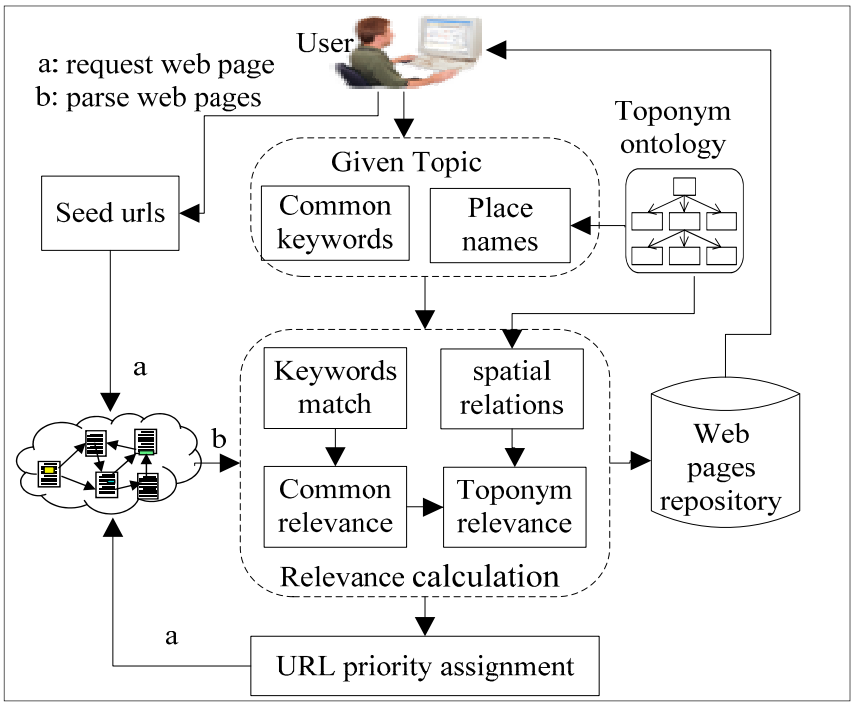
Note: if a given topic does not contain place names, the method will be the same to the traditional method, which means that only the first step is implemented.

## 4. Design and Implementation

### *4.1. Focused Crawler with Geographical Properties of Place Names*

On the basis of two-tuple-based topic representation method and hierarchical relevance calculation algorithm, a focused crawler with place names' geographical properties is developed for borderlands situation information collection. Figure 3 shows the process of our proposed focused crawler. Firstly, the user assigns given topic and seed URLs. Given topic is represented as a two-tuple in the form of "common keywords, place names", as shown in Section 3.2. Then, our proposed focused crawler begins with Seed URLs through requesting webpages and parsing webpages. After parsing webpages, the relevance calculation is implemented, as shown in Section 3.3. If both the relevance of common keywords (named common relevance) and the relevance of place names (named toponym relevance) are both greater than or equal to the given threshold, the webpage will be stored in a webpage repository, and both values will be utilized for URL priority assignment. Otherwise, the webpage will be abandoned. At last, URLs in the URL priority queue will continue to be submitted for requesting webpages until the URL priority queue is empty or other conditions are fulfilled.

**Figure 3.** Process of focused crawler with place names' geographical properties.



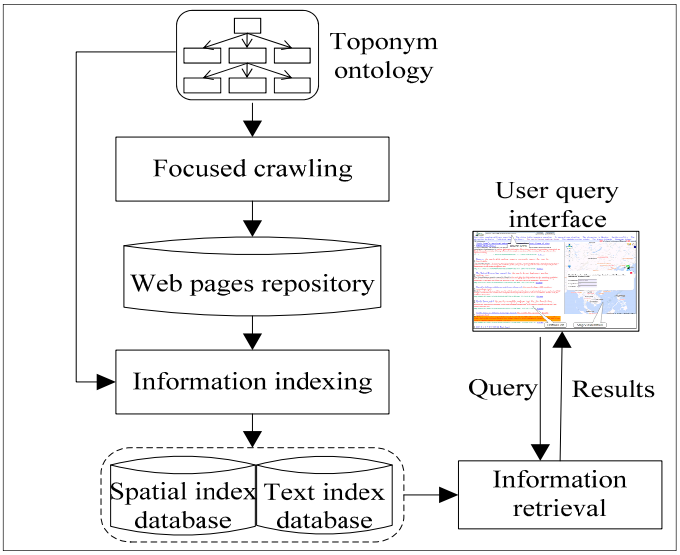

There are two main differences between our proposed focused crawler and traditional best-first focused crawler. One is the topic representation method. The given topic is divided into common keywords and place names in our proposed focused crawler in the form of "common keywords, a place names", while given topic is represented as one keyword vector in the traditional best-first focused crawler. The other is the relevance calculation algorithm. In our proposed focused crawler, the relevance will be calculated by two steps and some key spatial relations are introduced into the relevance of place names.

### *4.2. Borderlands Situation Information Collection Prototype*

Our proposed focused crawler is implemented in an Information Collection Prototype for Borderlands Situation Information (ICP-BSI), based on the Microsoft NET framework 3.5. The prototype can automatically download and index borderlands situation information. Moreover, it can provide a query service in the form of textual list and map. The main goal of the prototype is to enable borderlands researchers to customize the process of crawling for borderland situation topics and to retrieve relevant information from webpage repositories.

The prototype contains toponym ontology and four main modules of information agents, including focused crawling, information indexing, information retrieval, and user query interface, as shown in Figure 4. 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 4.** Design of ICP-BSI.

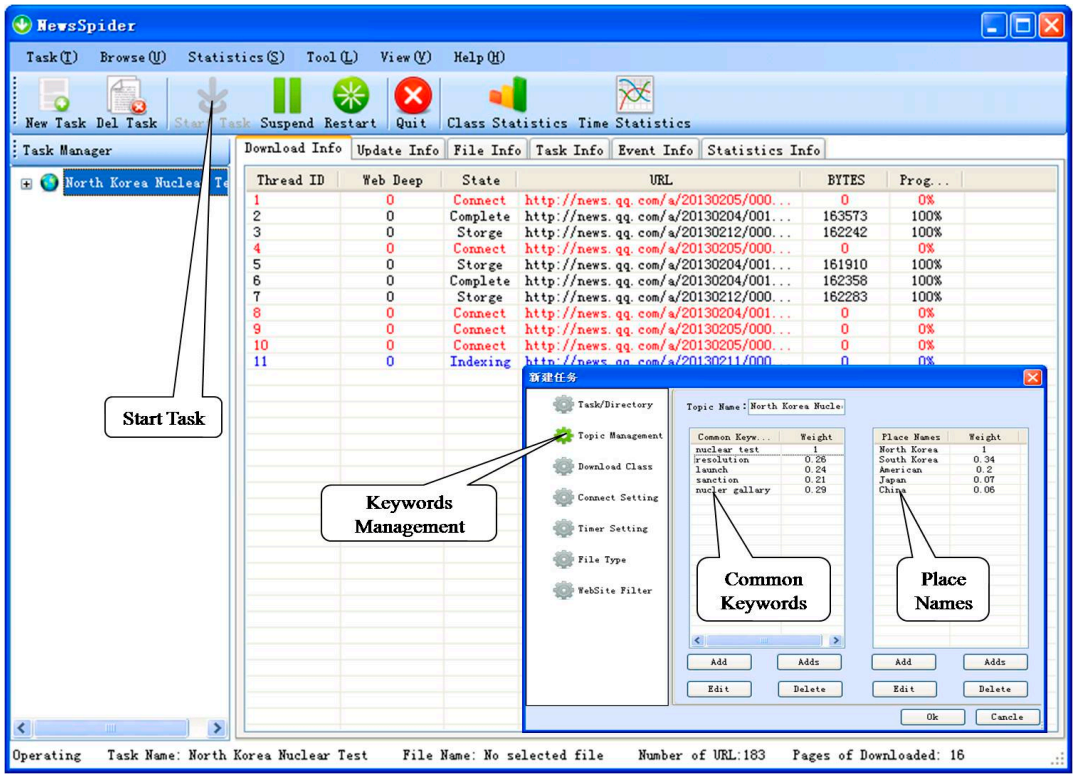



Toponym ontology is applied not only in focused crawling module for topic representation and relevance calculation as discussed in Section 3, but also in information indexing module for constructing a spatial index. The toponym ontology is edited by Protégé Ontology Editor [53] and is integrated into an ICP-BSI system through an open source soft dotNetRDF [54]. The toponym ontology is made up of place names and their spatial relations. In the toponym ontology, there are two types of place names, including names of administrative zones and physical geography. To be specific, places names of administrative zones are divided into the seven hierarchies of world, continent, country, province/state, city, county, and town. In addition, these place names mainly concern China and its neighboring countries in the Chinese language. Currently, the toponym ontology has about 46,000 place names, associated with a point coordinate (longitude and latitude), for visualization, 45,000 "direct contain" relations and only 113 overlap relations. 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 addition, place names in webpages are identified by open source software of Pan Gu Segment [52], in which we add place names from the toponym

ontology. However, there are many place names shared by different places [34]. To deal with this situation, the place names are simply disambiguated by their co-occurrence place names in given topic and webpages in this paper.

The focused crawling module is implemented based on our proposed focused crawler as discussed in Section 4.1. The module is responsible for downloading webpages relevant to borderlands situation to serve 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 5. Through the GUI, borderland researchers can set task parameters, such as topic, timing parameters and thread number, *etc.* After clicking the new task button, they can start the task and also monitor the process of the crawling.

**Figure 5.** GUI of focused crawling module.

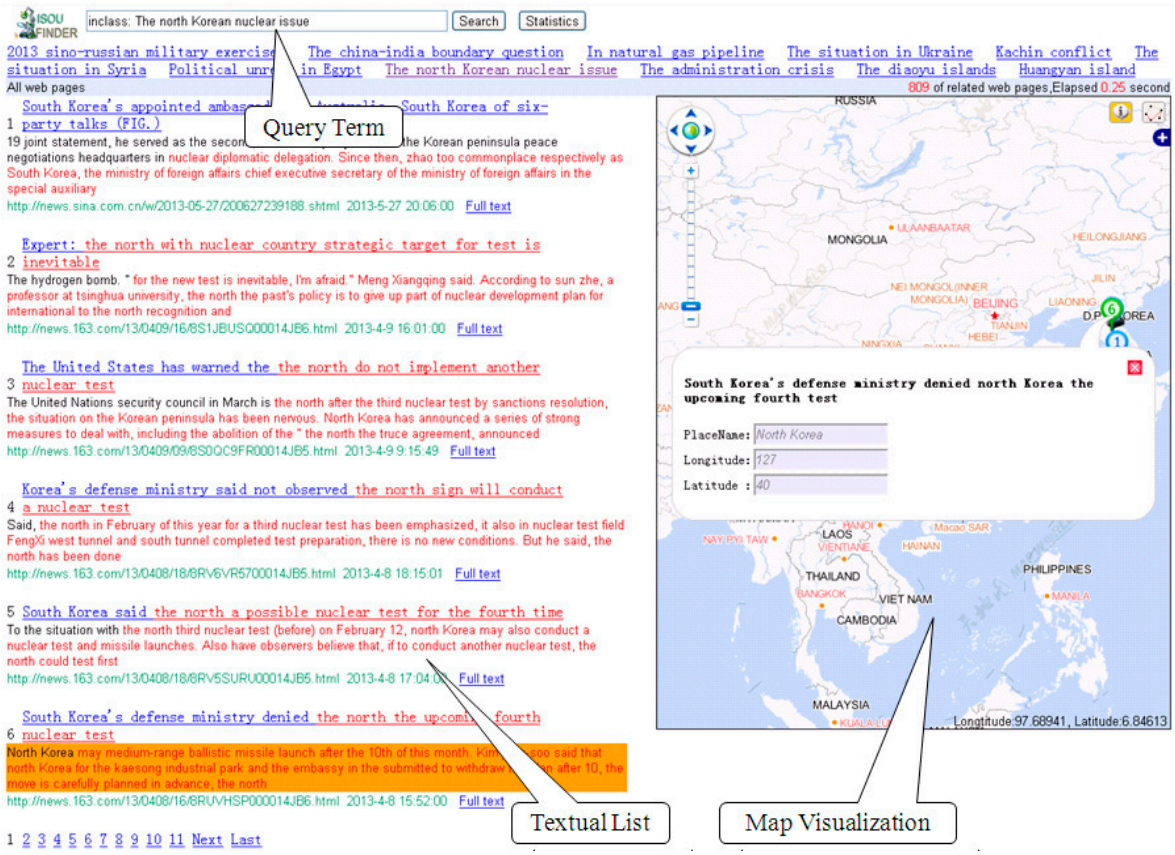



The information-indexing module is responsible for indexing the downloaded webpages to serve the information retrieval module. The module is implemented by the Lucenc.net API [55]. In the module, text information about webpages is indexed as an inverted file structure. In addition, in order to locate and visualize webpages on the map, the spatial locations of webpages are also indexed as an inverted file structure. In this paper, the highest occurrence frequency of a place name in the webpage 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 with the Lucenc.net API. In the module, only the keyword-matching method is adopted to search for relevant results. The results are ranked in descending order of the relevance value, and also can be ranked in descending order of publication time of the webpages.

The user query interface is a bridge connecting users and the information retrieval module, with the goal of submitting query term to the system and display the returned results. It is comprised of five parts, including query term input box, search button, statistics button, textual list, and map visualization, as shown in Figure 6. When a user inputs a query term and click the search button, relevant results will be displayed in textual list and map visualization. When a user inputs the query term in the form of "inclass: topic name" and click the 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 implemented by OpenLayers API [56] contains number icons, corresponding to ranked textual information and some simple map tools, such as pan, zoom, modification, *etc.* When a user clicks the number icon, a label box containing place name, longitude and latitude will be displayed, and the corresponding result in the textual list will be highlighted.

**Figure 6.** The main user query interface.




## 5. Experiments and Analysis

This section compares the performance of our proposed focused crawler with the traditional best-first focused crawler. The experiments are carried out in an environment with an Intel Pentium 4 CPU 3.20 GHZ, 1 GB of RAM, and 6 M bandwidth. The evaluation involves efficiency and effectiveness.

### *5.1. Preparation of Experiments*

In experiments, the topic of "North Korea Nuclear Issue" is used as an example to evaluate the efficiency and effectiveness of our proposed focused crawler. The topic is an international hotspot

issue about borderland situations. It is made up of common keywords and place names, limited in borderland regions. In addition, it is consistent with the definition of borderland situation information. Therefore, the topic of "North Korea Nuclear Issue" can be used as a representative to evaluate the efficiency and effectiveness of our proposed focused crawler for borderland situation information collection.

In the experiments, ten keywords are selected to represent the topic of the North Korea Nuclear Issue through a descending order of term frequency. Firstly, 100 relevant webpages are manually fetched from ifeng.com, qq.com, 163.com, souhu.com, *etc.* Then, these webpages are segmented into keywords by Chinese word segmentation using the Pan Gu Segment. Next, the term frequency of keywords are recorded and ranked in descending order. Finally, five meaningful common keywords and five place names with high term frequency are selected. The weight of the ten keywords is calculated with Equation (9). On the basis of two-tuple-based topic representation method in Section 3.2, the topic of "North Korea Nuclear Issue" is represented as two vectors, which are shown in Equations (16) and (17).

$$V_{T-K} = \begin{Bmatrix} (\text{nuclear test}, 1), (\text{nuclear gallery}, 0.29), \\ (\text{resolution}, 0.26), (\text{launch}, 0.24), (\text{sanction}, 0.21) \end{Bmatrix} \tag{16}$$

$$V_{T-PN} = \begin{Bmatrix} (\text{North Korea}, 1), (\text{South Korea}, 0.34), \\ (\text{American}, 0.2), (\text{Japan}, 0.07), (\text{China}, 0.06) \end{Bmatrix} \tag{17}$$

To facilitate effectiveness evaluation, 200 test webpages are manually constructed from ifeng.com, qq.com, 163.com, sohu.com, *etc.* These webpages include 100 webpages relevant to the topic of North Korea Nuclear Issue and 100 webpages irrelevant to the topic. These relevant webpages are different from above 100 webpages for determining given topic and these irrelevant webpages are about the Iran nuclear issue, the North Korean Nuclear economy, *etc.*, which are easily confused with North Korea Nuclear Issue.

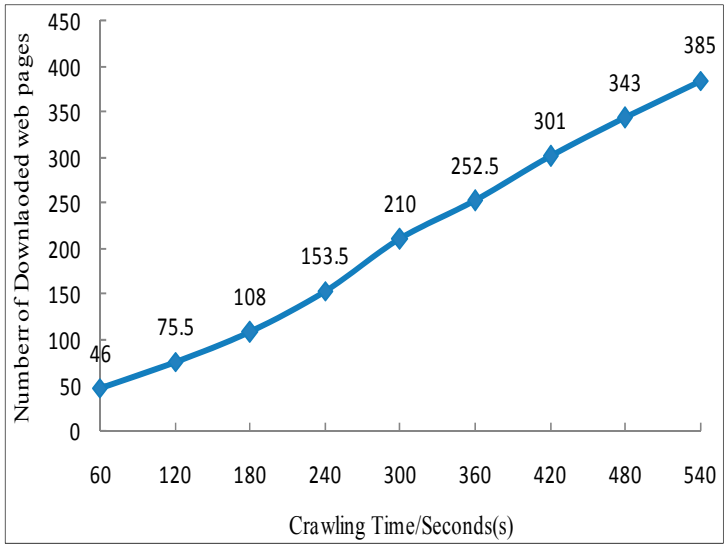
### *5.2. Efficiency Analysis*

The crawling speed of collected information, also known as crawling time, is often considered as a metric for evaluating the efficiency of focused crawler. Crawling speed is highly dependent on network bandwidth, capacity of the machine, crawling strategy, number of crawlers used, seed URLs, web resources of given topic, and many other factors [57]. In order to test the efficiency of our proposed focused crawler, the topic "North Korea Nuclear Issue" is used as an example. The basic parameters are set as follows: the number of threads is 10, the threshold is 0.5, the number of seed URLs is 10 and the two topics are both represented with five common keywords and five place names as shown in Section 5.1. Under the same conditions and topics, two independent experiments are made and the crawling time is recorded. The average crawling time of the two experiments is shown in Figure 7.

It can be seen in Figure 7 that our proposed focused crawler downloads 385 webpages in 540 s. It means that the crawling speed of our proposed focused crawler is 0.69 webpages per second. It is consistent with the polite access interval of one webpage per second [58]. Moreover, webpages about the borderlands situation, in some main web portals, increase by several hundred pages every day,

therefore, our proposed focused crawler can meet the daily demand for collecting borderlands situation information.

**Figure 7.** The average crawling time of the two experiments.




Additionally, it can also be seen in Figure 7 that the number of downloaded webpages scales linearly with the crawling time in seconds. Their slope values (crawling speed) between two adjacent points range from 0.5 to 0.9, with an average value 0.71, indicating that the changes in crawling speed is small. This means that, when collecting information, our proposed focused crawler is stable in efficiency.

### *5.3. Effectiveness Analysis*

#### 5.3.1. Effectiveness Metrics

Two most frequent and basic effectiveness metrics for focused crawler are precision and recall [13,59]. Precision represents the fraction of relevant webpages in crawled webpages [13] and the higher precision value implies that the focused crawler has a better ability to filter out irrelevant webpages. Recall is the fraction of crawled relevant webpages in the total relevant webpages [13] and the higher recall value means the focused crawler has better capacity to obtain relevant webpages.

According to the definition of precision and recall, they can be calculated with Equations (18) and (19).

$$p = \frac{CR}{TC} \tag{18}$$

$$r = \frac{CR}{TR} \tag{19}$$

where $p$ and $r$ denote precision and recall respectively, $CR$ represents the number of crawled relevant webpages. $TC$ represents the total number of crawled webpages and $TR$ denotes the total number of relevant webpages in the whole web. However, since the total number of relevant webpages for given topic $TR$ is unknown, the true recall is difficult to compute. Therefore, in Section 5.1, some test

webpages are constructed to compute precision and recall. That is to say, the total number *TR* of relevant webpages in the sample data is known in advance and it is 100.

Although precision and recall are not related to each other in theory, high precision is achieved almost always at the expense of recall and high recall is achieved at the expense of precision in practice [51]. Thus, a trade-off metric *F*-score, which is the harmonic mean of precision and recall, is adopted in this paper [51,59], as shown in Equation (20). 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 [51].

$$F = \frac{2 \times p \times r}{p + r} \times 100\% \tag{20}$$

where, *F* denotes the value of *F*-score and other variables are the same as above.

### 5.3.2. Results and Analysis

In this experiment, the traditional best-first focused crawler, whose major parts are shown in Equations (1), (2) and (10), is also implemented for comparison. Figure 8 shows the results of the relevance through the traditional best-first focused crawler. Figures 9 and 10 represent the relevance of common keywords and the relevance of place names through our proposed focused crawler, respectively. In these three figures, axis x represents the number of webpages, where the number of webpages 1 to 100 implies irrelevant webpages and number 101 to 200 represent relevant webpages, axis y denotes the relevance value, and the dashed line is the dividing line between actual irrelevant and relevant webpages. In addition, based on several extra experiments, we set 0.65 as the threshold value in the traditional best-first focused crawler because the traditional best-first focused crawler can obtain the best results in this threshold value, and we set 0.5 as the threshold value in our proposed focused crawler because of the same reason as the traditional best-first focused crawler.

**Figure 8.** Relevance in the traditional best-first focused crawler.

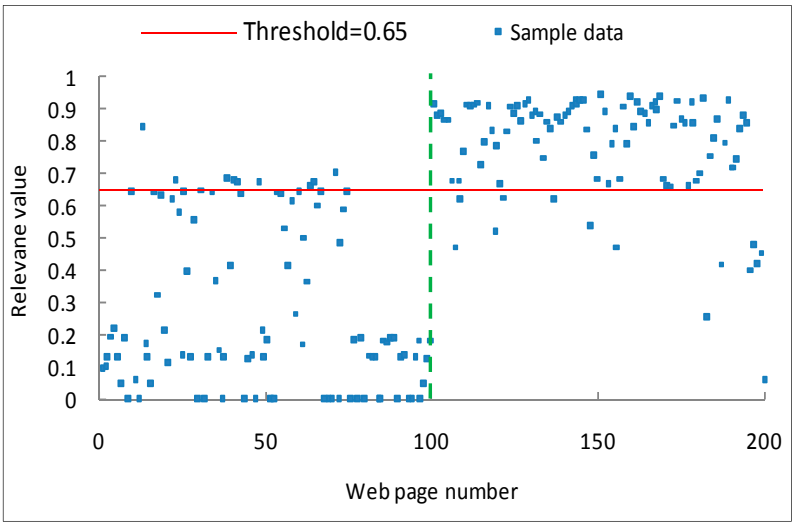


**Figure 9.** Relevance of common keywords.

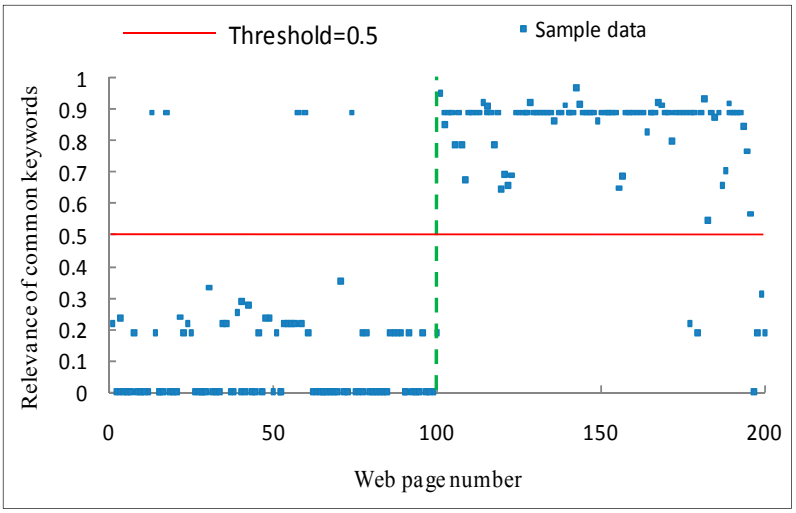
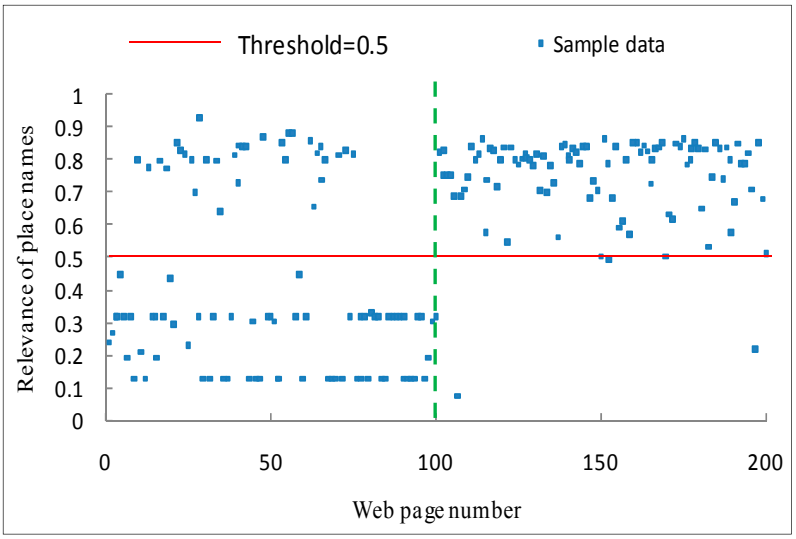



**Figure 10.** Relevance of place names.




It can be found in Figure 8 that the traditional best-first focused crawler can collect 95 webpages (webpages above the red line), which contains 86 actual relevant webpages. Therefore, the total number of crawled webpages *TC* is 95, and the number of crawled relevant webpages *CR* is 86. In Figure 9, there are only five webpages from number 1 to 100 (numbered 13, 18, 58, 60, and 74), of which relevance values of common keywords are greater than the threshold value 0.5. In Figure 10, only the place names relevance value of webpage 13 among the above five webpages are greater than the threshold 0.5. In addition, there are 92 webpages from number 101 to 200 of which relevance values of both common keywords and place names are greater than the threshold value 0.5. Thus, from Figures 9 and 10, it is concluded that our proposed focused crawler can collect 93 webpages, among which 92 webpages are actually relevant to the topic. That is to say, in our proposed focused crawler the total number of crawled webpages *TC* is 93 and the number of crawled relevant webpages *CR* is 92.

According to Equations (18)–(20), we can compute the precision, recall and F-score values of the traditional best-first focused crawler and our proposed focused crawler. Table 1 shows the results of the evaluation.

**Table 1.** The precision, recall and F-score values.

| | CR | TC | TR | Precision | Recall | F-score |
|---|---|---|---|---|---|---|
| Traditional Focused crawler | 86 | 95 | 100 | 90.53% | 86% | 88.21% |
| Proposed focused crawler | 92 | 93 | 100 | 98.9% | 92% | 95.3% |

It can be seen from Table 1 that the precision of the traditional best-first focused crawler is lower (8.37%) than the one of our proposed focused crawler, which means that our proposed focused crawler has a better ability to filter out irrelevant webpages than the traditional best-first focused crawler. Besides, it is also found that the recall of the traditional focused crawler is lower (6%) than the recall of our proposed focused crawler. This means that our proposed focused crawler also has a better ability to obtain relevant webpages than the traditional best-first focused crawler. In addition, it can be found that the *F*-score of the traditional best-first focused crawler is much lower (7.09%) than the *F*-score of our proposed focused crawler, which means our proposed focused crawler is more effective than the traditional best-first focused crawler.

## 6. Conclusions

Traditional focused crawlers have some defects in collecting borderlands situation information, because they just treat place name in the same way as other common keyword, which may reduce the effectiveness of focused crawlers. In order to solve this problem, a novel focused crawling method considering place names' spatial properties is proposed in this paper. Compared to traditional focused crawlers, this method represents given topic using a two-tuple-based method, in which topic is divided into place names and common keywords separately. Then, spatial relations of place names are introduced to calculate the relevance between given topic and webpages, and the calculation process is divided into two steps. The first step is to calculate the relevance using common keywords, just as the traditional focused crawlers do. The second step is to calculate the relevance between place names in given topic and in the webpages. Spatial relations of equal, contain, contained, overlap and disjoint assigned with different weights are utilized in the relevance calculation algorithm. Based on proposed methods, a focused crawler with geographical properties of place names is designed. Furthermore, an information collection prototype is implemented for borderlands situation information.

As shown in the experiments, the efficiency of the proposed focused crawler can meet the daily demand for collecting borderlands situation information and its *F*-Score value is increased by around 7% compared with traditional best-first focused crawler. It means that the proposed focused crawler is more effective than the traditional best-first focused crawler.

Even with its increased effectiveness, there are still some limitations for our proposed focused crawler. Firstly, this focused crawler must rely on a toponym ontology or gazetteer. The toponym ontology used in the paper is incomplete and there is no spatial extent for each place name. In future, we will supplement more place names and spatial relations into our toponym ontology. Furthermore, we will try our best to add spatial extent as a property, which enables our focused crawler to compute spatial relations in real-time. Secondly, our proposed focused crawler utilizes place names and their spatial relations, but it does not take temporal relations into account. Therefore, we will extend our

focused crawler with temporal element in the topic definition, relevance calculation and URL priority assignment. Thirdly, borderland situation information fetched by the proposed focused crawler is unstructured, which is not very suitable to monitor the tendencies of borderlands situation. Therefore, we will extract structured information in the form of event from the webpage repository and then cluster these events to obtain borderlands situation by association rule. On this basis, we will quantize the clustered events to obtain borderlands situation and monitor their tendencies.

## Acknowledgments

This study was funded by the National Science Foundation of China (Project #41301412), Ministry of Science and Technology of China (Project No. 2012BAK12B00) and the National Science Foundation of China (Project #41231172). All the authors gratefully thank the reviewers and editor.

## Author Contributions

Dongyang Hou played an important role in the development of the idea, implementing the focused crawler prototype, drafting and revising the manuscript. Hao Wu helped to develop the idea, and contributed to revise the abstract, related works and conclusion of the manuscript. Jun Chen also contributed to develop the idea and to revise the structure and introduction of the manuscript. Ran Li participated in the discussion of the idea. All authors read and approved the final manuscript.

## Conflicts of Interest

The authors declare no conflicts of interest.

## References

1. Chen, J.; Ge, Y.; Hua, Y.; Wang, F.; Yang, S.; Qu, B.; Li, R. Digital border-land: Conceptual framework and research agenda. *Bull. Surv. Mapp.* **2013**, *2*, 1–4.
2. Baumgartner, N.; Gottesheim, W.; Mitsch, S.; Retschitzegger, W.; Schwinger, W. BeAware!—Situation awareness, the ontology-driven way. *Data Knowl. Eng.* **2010**, *69*, 1181–1193.
3. Chen, J.; Ge, Y.; Cheng, Y.; Li, R.; Cao, Y. Borderlands modeling and understanding with GISs: Challenges and research agenda. *ISPRS Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci.* **2013**, *1*, 15–22.
4. Hu, H.; Ge, Y.; Hou, D. Using web crawler technology for geo-events analysis: A case study of the Huangyan Island incident. *Sustainability* **2014**, *6*, 1896–1912.
5. Chapman, M.S.; Ciravegna, P.F. Focused data mining for decision support in emergency response scenarios. *Management* **2006**, *4*, 6–14.
6. Menczer, F. Complementing search engines with online web mining agents. *Decis. Support Syst.* **2003**, *35*, 195–212.
7. Tsytsarau, M.; Palpanas, T. Survey on mining subjective data on the web. *Data Min. Knowl Discov.* **2012**, *24*, 478–514.
8. Zhong, Z.; Liu, Z. Ranking events based on event relation graph for a single document. *Inf. Technol. J.* **2010**, *9*, 174–178.

9. Almpanidis, G.; Kotropoulos, C.; Pitas, I. Combining text and link analysis for focused crawling—An application for vertical search engines. *Inf. Syst.* **2007**, *32*, 886–908.
10. Shi, Q.; Shi, Z.; Xiao, Y. VSEC: A Vertical Search Engine for E-commerce. In *Recent Progress in Data Engineering and Internet Technology*; Springer: Berlin, Germany, 2012; Volume 2, pp. 57–63.
11. Wilkas, L.R.; Villarruel, A. An introduction to search engines. *J. Soc. Pediatr. Nurs.* **2001**, *6*, 149–151.
12. Hsu, C.-C.; Wu, F. Topic-specific crawling on the Web with the measurements of the relevancy context graph. *Inf. Sys.* **2006**, *31*, 232–246.
13. Peng, T.; Liu, L. Focused crawling enhanced by CBP-SLC. *Knowl. Based Syst.* **2013**, *51*, 15–26.
14. Chakrabarti, S.; van den Berg, M.; Dom, B. Focused crawling: A new approach to topic-specific Web resource discovery. *Comput. Netw.* **1999**, *31*, 1623–1640.
15. Du, Y.; Pen, Q.; Gao, Z. A topic-specific crawling strategy based on semantics similarity. *Data Knowl. Eng.* **2013**, *88*, 75–93.
16. Derungs, C.; Purves, R.S. Measuring topographic similarity of toponyms. In Proceedings of the 15th AGILE International Conference on Geographic Information Science, Avignon, France, 24–27 April 2012.
17. Siemiński, A. Using WordNet to measure the similarity of link texts. In Proceedings of the First International Conference ICCCI, Wroclaw, Poland, 5–7 October 2009; Springer: Berlin, Germany, 2009; pp. 720–731.
18. Wu, H.; Liao, A.; He, C.; Hou, D. Topic-Relevance based crawler for geographic information web services. *Geogr. Geo Inf. Sci.* **2012**, *28*, 27–30.
19. Alam, M.H.; Ha, J.; Lee, S. Novel approaches to crawling important pages early. *Knowl. Inf. Syst.* **2012**, *33*, 707–734.
20. Catanese, S.A.; de Meo, P.; Ferrara, E.; Fiumara, G.; Provetti, A. Crawling facebook for social network analysis purposes. In Proceedings of the International Conference on Web Intelligence, Mining and Semantics, Sogndal, Norway, 25–27 May 2011; Association for Computing Machinery Press: New York, NY, USA, 2011.
21. Gjoka, M.; Kurant, M.; Butts, C.T.; Markopoulou, A. Practical recommendations on crawling online social networks. *IEEE. J. Sel. Area Commun.* **2011**, *29*, 1872–1892.
22. Batsakis, S.; Petrakis, E.G.; Milios, E. Improving the performance of focused web crawlers. *Data Knowl. Eng.* **2009**, *68*, 1001–1013.
23. Bedi, P.; Thukral, A.; Banati, H. Focused crawling of tagged web resources using ontology. *Comput. Electr. Eng.* **2013**, *39*, 613–628.
24. Liu, J.; Lu, Y. Survey on topic-focused web crawler. *Appl. Res. Comput.* **2007**, *24*, 26–29.
25. Hersovici, M.; Jacovi, M.; Maarek, Y.S.; Pelleg, D.; Shtalhaim, M.; Ur, S. The shark-search algorithm—An application: Tailored Web site mapping. *Comput. Netw. ISDN Syst.* **1998**, *30*, 317–326.
26. Pant, G.; Menczer, F. Topical crawling for business intelligence. In *Research and Advanced Technology for Digital Libraries*; Springer: Berlin, Germany, 2003; pp. 233–244.
27. Srinivasan, P.; Menczer, F.; Pant, G. A general evaluation framework for topical crawlers. *Inf. Retr.* **2005**, *8*, 417–447.

28. Ehrig, M.; Maedche, A. Ontology-focused crawling of Web documents. In Proceedings of the 2003 ACM Symposium on Applied Computing, Melbourne, FL, USA, 9–12 March 2003; Lamont, B., Ed.; Association for Computing Machinery Press: New York, NY, USA, 2003; pp. 1174–1178.
29. Ye, Y.; Ouyang, D. Semantic-Based focused crawling approach. *J. Softw.* **2011**, *22*, 2075–2088.
30. Liu, W.; Du, Y. An improved topic-specific crawling approach based on semantic similarity vector space model. *J. Comput. Inf. Syst.* **2012**, *8*, 8605–8612.
31. Yang, X.; Sui, A.; Tang, Z. Topical Crawler based on multi-level vector space model and optimized hyperlink chosen strategy. In Proceedings of the 9th IEEE International Conference on Cognitive Informatics (ICCI), Beijing, China, 7–9 July 2010; Sun, F., Wang, Y., Lu, J., Zhang, B., Kinsnor, W., Zadeh, L., Eds.; IEEE: Piscataway, NJ, USA, 2010; pp. 430–435.
32. Liu, Z.; Du, Y.; Zhao, Y. Focused crawler based on domain ontology and fca. *J. Inf. Comput. Sci.* **2011**, *8*, 1909–1917.
33. Vestavik, Ø. Geographic Information Retrieval: An Overview. Available online: http://wenku.baidu.com/link?url=Kirme_ZKvLyl7S41NPL5Jiq4rYFHf57Sf6Cq931F-voKdnlJ24Uz738gSIaQUKkDFdL_vlrG-mHZXPSvjigVcVMV4oaVOj9mOoAJyn3s6Rm (accessed on 10 May 2014).
34. Jones, C.B.; Purves, R.S. Geographical information retrieval. *Int. J. Geogr. Inf. Sci.* **2008**, *22*, 219–228.
35. Silva, M.J.; Martins, B.; Chaves, M.; Afonso, A.P.; Cardoso, N. Adding geographic scopes to web resources. *Comput. Environ. Urban Syst.* **2006**, *30*, 378–399.
36. Vasardani, M.; Winter, S.; Richter, K.-F. Locating place names from place descriptions. *Int. J. Geogr. Inf. Sci.* **2013**, *27*, 1–24.
37. Purves, R.S.; Clough, P.; Jones, C.B.; Arampatzis, A.; Bucher, B.; Finch, D.; Fu, G.; Joho, H.; Syed, A.K.; Vaid, S. The design and implementation of SPIRIT: A spatially aware search engine for information retrieval on the Internet. *Int. J. Geogr. Inf. Sci.* **2007**, *21*, 717–745.
38. Frontiera, P.; Larson, R.; Radke, J. A comparison of geometric approaches to assessing spatial similarity for GIR. *Int. J. Geogr. Inf. Sci.* **2008**, *22*, 337–360.
39. Khodaei, A.; Shahabi, C.; Li, C. SKIF-P: A point-based indexing and ranking of web documents for spatial-keyword search. *Geoinformatica* **2012**, *16*, 563–596.
40. Fu, G.; Jones, C.B.; Abdelmoty, A.I. Ontology-based spatial query expansion in information retrieval. In *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE*; Springer: Berlin, Germany, 2005; pp. 1466–1482.
41. Kozanidis, L.; Stamou, S. Automatic construction of a geo-referenced search engine index. Available online: http://www.dblab.upatras.gr/download/nlp/NLP-Group-Pubs/j09-IJWA_Geo-Referenced_Index.pdf (accessed on 10 May 2014).
42. Li, W.; Yang, C.; Yang, C. An active crawler for discovering geospatial web services and their distribution pattern—A case study of OGC Web Map Service. *Int. J. Geogr. Inf. Sci.* **2010**, *24*, 1127–1147.
43. Patil, S.; Bhattacharjee, S.; Ghosh, S.K. A spatial web crawler for discovering geo-servers and semantic referencing with spatial features. In *Distributed Computing and Internet Technology*; Springer: Berlin, Germany, 2014; pp. 68–78.

44. Ahlers, D.; Boll, S. Adaptive geospatially focused crawling. In Proceedings of the 18th ACM Conference on Information and Knowledge Management, Hong Kong, China, 2–6 November 2009; Association for Computing Machinery Press: New York, NY, USA, 2009; pp. 445–454.
45. Birkin, M.; Malleson, N. The spatial analysis of short-term population movements with social media data. Available online: http://www.geocomputation.org/2013/papers/28.pdf (accessed on 10 May 2014).
46. Gelernter, J.; Cao, D.; Carley, K.M. Extraction of spatio-temporal data for social networks. In *The Influence of Technology on Social Network Analysis and Mining*; Springer: Berlin, Germany, 2013; pp. 351–372.
47. Zhang, Y.; Gao, Y.; Xue, L.; Shen, S.; Chen, K. A common sense geographic knowledge base for GIR. *Sci. China Ser. E Technol. Sci.* **2008**, *51*, 26–37.
48. ChinaNews Net. North Korea Announced that it was Planning a Third Nuclear Test. Available online: http://news.163.com/13/0124/11/8LVU9J3J0001121M.html (accessed on 10 May 2014).
49. XinHua Net. The Iran Nuclear Issue: An Important Step in Bumpy Road. Available online: http://news.xinhuanet.com/2013-10/17/c_117761284.htm (accessed on 10 May 2014).
50. Chen, J.; Li, C.; Li, Z.; Gold, C. A voronoi-based 9-intersection model for spatial relations. *Int. J. Geogr. Inf. Sci.* **2001**, *15*, 201–220.
51. Liu, B. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, 2nd ed.; Springer-Verlag New York Incorporated: New York, NY, USA, 2010; pp. 217–218.
52. Eaglet. Pan Gu Segment. Available online: http://pangusegment.codeplex.com/ (accessed on 10 May 2014).
53. Stanford University Protégé. Available online: http://protege.stanford.edu/ (accessed on 10 May 2014).
54. Rob Vesse. DotNetRDF—Semantic Web, RDF and SPARQL Library for C#/.Net. Available online: http://www.dotnetrdf.org/default.asp (accessed on 10 May 2014).
55. Apache Software Foundation. Lucene.net. Available online: http://blogs.apache.org/lucenenet/ (accessed on 10 May 2014).
56. OpenLayers 3. Available online: http://www.openlayers.org/ (accessed on 10 May 2014).
57. Menczer, F.; Pant, G.; Srinivasan, P.; Ruiz, M.E. Evaluating topic-driven web crawlers. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, LA, USA, 9–12 September 2001; Association for Computing Machinery Press: New York, NY, USA, 2001; pp. 241–249.
58. Dill, S.; Kumar, R.; McCurley, K.S.; Rajagopalan, S.; Sivakumar, D.; Tomkins, A. Self-similarity in the web. *ACM Trans. Int. Technol.* **2002**, *2*, 205–223.
59. Manning, C.D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: New York, NY, USA, 2008; pp. 142–143.