Remote Sens. **2015**, *7*, 5805-5827; doi:10.3390/rs70505805

OPEN ACCESS

remote sensing

ISSN 2072-4292

www.mdpi.com/journal/remotesensing

Article

Active Collection of Land Cover Sample Data from Geo-Tagged Web Texts

Dongyang Hou [1,2], Jun Chen [2,*], Hao Wu [2], Songnian Li [1,3], Fei Chen [2,4] and Weiwei Zhang [2]

[1] School of Environment Science and Spatial Informatics, China University of Mining and Technology, Xuzhou 221116, China; E-Mails: houdongyang1986@cumt.edu.cn (D.H.); snli@ryerson.ca (S.L.)

[2] National Geomatics Center of China, 28 Lianhuachi West Road, Beijing 100830, China; E-Mails: wuhao@nsdi.gov.cn (H.W.); cfei0618@163.com (F.C.); zhangweiwei@nsdi.gov.cn (W.Z.)

[3] Department of Civil Engineering, Ryerson University, 350 Victoria Street, Toronto, ON M5B 2K3, Canada

[4] School of Geosciences and Info-Physics, Central South University, Changsha 410083, China

* Author to whom correspondence should be addressed; E-Mail: chenjun@nsdi.gov.cn; Tel.:+86-10-6388-1088.

Academic Editors: Yoshio Inoue, Chandra Giri and Prasad S. Thenkabail

Received: 5 January 2015 / Accepted: 29 April 2015 / Published: 7 May 2015

Abstract: Sample data plays an important role in land cover (LC) map validation. Traditionally, they are collected through field survey or image interpretation, either of which is costly, labor-intensive and time-consuming. In recent years, massive geo-tagged texts are emerging on the web and they contain valuable information for LC map validation. However, this kind of special textual data has seldom been analyzed and used for supporting LC map validation. This paper examines the potential of geo-tagged web texts as a new cost-free sample data source to assist LC map validation and proposes an active data collection approach. The proposed approach uses a customized deep web crawler to search for geo-tagged web texts based on land cover-related keywords and string-based rules matching. A data transformation based on buffer analysis is then performed to convert the collected web texts into LC sample data. Using three provinces and three municipalities directly under the Central Government in China as study areas, geo-tagged web texts were collected to validate artificial surface class of China's 30-meter global land cover datasets (GlobeLand30-2010). A total of 6283 geo-tagged web texts were collected at a speed of 0.58 texts per second. The collected texts about built-up areas were transformed into sample data. User's accuracy of 82.2% was achieved, which is close to that derived from formal expert validation. The preliminary results show that geo-tagged web texts are valuable ancillary data for LC map validation and the proposed approach can improve the efficiency of sample data collection.

Keywords: sample data; land cover; validation; deep web crawler; geo-tagged web texts

1. Introduction

Land cover (LC) map validation requires reliable sample data, which describes the land cover type, spatial coverage, geographic locations, collecting time as well as other factors [1–4]. In general, sample data is collected through field surveys, existing land cover maps, or interpretation of remotely sensed images [5–11]. These traditional methods are costly, labor-intensive and time-consuming tasks [2,6,8,12] and are becoming even more difficult when sample data needs to be collected over a large area or the entire earth [5,13,14]. Recently, web-based methods have been developed to enable people from anywhere of the world to provide sampling information about their familiar regions through the web, such as uploading geo-tagged photos or labeling the interpretation results on images for land cover map validation [5,15–18]. These methods helped reduce the costs and operation time of LC sample data collection to some extent. However, their success depends largely on the willingness and qualification of volunteers, *i.e.*, whether they have time, experience and good faith motivation to complete this kind of non-paid work [16]. Improving the efficiency and quality of web-based sample data collection is becoming an important research topic in the area of LC map validation.

There are rich and massive geo-tagged resources available on the web, which are provided and published by real estate enterprises, tourist agencies and other professional organizations, as well as volunteers [19]. With the development of geo-tagging technologies, more and more geo-tagged resources will become available online such as geo-tagged photos, geo-tagged videos and geo-tagged web texts [19]. They often contain valuable information about land cover type, geographic locations, time, and other factors, which are essential for LC map validation [2,3,19]. Over the last few years, geo-tagged photos have been used to assist in LC map validation [2,3]; however, no automatic approach has been found to collect them. Utilizing web data mining methods (e.g., web crawler), active collection of sample data, *i.e.*, automatically collecting geo-tagged web resources to assist in LC map validation, may be realized [20–24].

Developing such an active data collection approach requires extensive investigations to answer some fundamental questions. Land cover map validation has specific requirements regarding sample data contents [1] and the current web data mining methods may not satisfy these requirements. How to define land cover keywords and formalize extraction rules for automatic gathering of geo-tagged web resources and extracting of land cover type, geographic location, time and other factors become important issues to handle first. Further, the collected raw geo-tagged text data from different sources and in different formats usually specify the occupied area instead of spatial coverage directly, which is one of important factors in LC map validation. How to calculate spatial coverage using the extracted occupied area and geographic locations becomes essential for transforming the collected raw data into candidate sample data for LC map validation.

Among all types of geo-tagged resources over the web, more and more geo-tagged texts are emerging on the web [2,25]. Similar to the definition of geo-tagged photos in Wikipedia, geo-tagged web texts represent text data associated with geographic locations [26]. These data may describe land cover types (e.g., built-up areas and wetland parks), geographic locations and building times, which are valuable information for LC map validation. This paper attempts to examine the potential of geo-tagged web texts as a new cost-free sample data source to help validate land cover maps and proposes an active data collection approach. Firstly, a deep web crawler, which aims to collect web texts from deep web in information retrieval domain [27,28], is customized for automatic collecting of geo-tagged web texts. It uses a maximum-frequency method to define land cover keywords and the XML Path Language (XPath) to formalize extraction rules. Secondly, GIS-based buffer analysis with occupied areas and geographic locations derived from crawling results is used to calculate spatial coverage through a corresponding data transformation process.

The remainder of this paper is organized as follows. Section 2 reviews related work about geo-tagged web resources, their applications, data collection and web-based sample data collection for LC map validation. Section 3 outlines the active collection of LC sample data from deep web, including the key technologies of deep web crawling and the data transformation process using geo-tagged web text about built-up areas. Preliminary results are presented in Section 4, followed by some discussions in Section 5. Section 6 concludes the paper.

2. Related Works

2.1. Geo-Tagged Web Resources and Their Applications

In the era of Web 2.0, the emergence of geo-tagged web resources, such as geo-tagged photos, geo-tagged video and geo-tagged web texts, has opened up a new world of possibilities for geographically-related research and applications [19,29,30]. A massive amount of geo-tagged web resources is collaboratively authored and community-contributed, and mostly generated by social media sites or networks [19]. Geographic locations of geo-tagged web resources are generally generated via hardware-based methods or software-based methods [19,31]. Hardware-based methods mainly employ built-in Global Positioning System (GPS) receivers in mobile devices such as smart phones and digital cameras to automatically tag geographic locations [19,20,31–33]. For example, Twitter allows users to add location information to their tweets based on their smart phones' geographic locations or web browsers [20]. The software-based methods usually use geo-tagging software with map interface (*i.e.*, Google map or Bing map) to record geographic locations manually [19,32]. For instance, Flickr supplies a map interface on which users can "drag" their photos to the locations where the photos were taken from [32]. In addition, triangulation with Wi-Fi signals, cellular radio and other sensor networks also have been used to generate geographic locations [19,34].

The massive amount of geo-tagged web resources is usually considered as a good source of knowledge discovery and has been extensively studied in many domains, such as gazetteers construction, travel recommendations, emergency management and sociolinguistic associations *etc*. [20,21,30,35–39]. For instances, Gao *et al.* [35] extracted crowd-sourced place names from big geo-tagged web resources using Hadoop. Majid *et al.* [21] and Vu *et al.* [40] identified semantically meaningful tourist locations from geo-tagged photos for tourist travel recommendations. Goodchild and Glennon [37] stated that the location of the fire, evacuation orders, the locations of emergency shelters, and much other useful information can be extracted from various geo-tagged web resources. Chae *et al.* [38] analyzed spatiotemporal distribution of tweets to identify public behavior patterns and response planning during natural disasters. Li *et al.* [20] and Eisenstein *et al.* [39] identified spatial, temporal and socioeconomic patterns from Twitter and Flickr data to discover sociolinguistic associations.

In addition, some initiatives have been taken in land use or land cover to explore the use of geo-tagged web resources for supporting land use and land cover classification and validation [2,3,22,41]. For example, Leung and Newsam [22] performed land-use classification using visual and textual features extracted from geo-tagged photos. With the help of unsupervised learning method and spectral clustering, Vanessa and Enrique [41] used geo-tagged tweets to automatically determine land use in urban areas by clustering geographical regions. Foody and Boyd [3] and Estima and Painho [2] extracted cover type, spatial coverage, geographic locations and collecting time from geo-tagged photos to validate LC map, which will be further discussed in Section 2.3. As discussed above, only geo-tagged photos have been used as a LC sample data source. However, geo-tagged web texts have not yet been utilized to assist in LC map validation.

2.2. Geo-Tagged Web Resources Collection

New methods and tools are needed to collect data from an immense amount of geo-tagged web resources because manual collection of this kind of data is both labor-intensive and time-consuming. Currently, there are two ways to automatically collect geo-tagged web resources: using Application Programming Interfaces (API) and using web crawling tools [42].

APIs, offered by some popular web content providers or social network sites such as Flickr and Twitter [20–23,43], allow users' programs to automatically obtain geo-tagged web resources and can be readily integrated into users' programs [21]. For instance, Li *et al.* [20] use Twitter's and Flickr's public APIs to collect tweets and photos from 21 January 2011 to 7 March 2011 within the bounding box of the contiguous United States for discovering sociolinguistic associations. But the public API has some limitations imposed on rights, calling times, special policies, and so on [24,42,44]. For example, Twitter restricts the number of tweets per API request and the limit varies based on the API used, *i.e.*, REST API, Search API or Streaming API. Furthermore, not every geo-tagged website such as real estate websites provides a public API.

To overcome limitations of public APIs, web crawlers become popular tools for the collection of geo-tagged web resources. A web crawler takes a series of seed Uniform Resource Locators (URLs) as its input, parses the seed web pages to obtain their contents and outgoing URLs, and determines what URLs to visit next based on certain criteria [45,46]. Web pages that are pointed by the outgoing URLs continue to be parsed and their contents satisfying certain relevance criteria are stored in a local repository. The above processes will continue until a desired number of web pages are obtained or local resources (such as storage) are exhausted [45,46]. Web crawling tools can be categorized into surface web crawler and deep web crawler according to the types of web pages found [27,47]. A surface web crawler is usually used by general purpose search engines to collect surface web pages, such as static web pages [47]. It parses web pages through a simple parsing engine of HTML. Therefore, it has no ability to deal with JavaScript, dynamic pages, *etc.*

In contrast, a deep web crawler aims to obtain data stored in JavaScript codes, forms, databases and other types of dynamic pages [27,28]. It parses web pages by JavaScript parsing engine or form analyzers together with HTML parsing engine [27,48]. Most geo-tagged websites are dynamic and part of their contents is stored in JavaScript codes, forms and databases [23,24]. Therefore, deep web crawling is usually used for collecting geo-tagged web resources. For example, Gao *et al.* [24] proposed a simulated browser-based crawler to solve the problem of parsing JavaScript codes for collecting Sina_Micro-blog and Tecent_Microblog, which are both geo-tagged websites. The crawler could only collect web texts, user ID, URL and other information, but not geographic locations. Our extensive review indicates that current deep web crawlers are not able to support requirements of LC sample data collection. A new deep web crawler, therefore, is needed to automatically collect geo-tagged web resources for LC map validation.

2.3. Web-Based Sample Data Collection for Land Cover Map Validation

With the advent of Web 2.0 technologies, some efforts have been made on developing in methods for collecting LC sample data over the web [5,15–17,49,50]. These efforts focus on two fronts: new cost-free sample data sources and collaborative volunteer-based collection methods. New data sources, such as volunteered geo-tagged photos posted on the web [2,3], have been explored for LC data validation. For instances, Foody and Boyd [3] used geo-tagged photos as sample data to validate the Globcover map's representation of tropical forest in West Africa. These geo-tagged photos were collected from a web-based collaborative project and interpreted by four volunteers. Estima and Painho [2] conducted a preliminary analysis of the adequacy of Flickr photos as a cost-free sample data source for validating the land use and land cover databases production. Comparing with field survey and desktop-based visual interpretation methods, these methods have lower costs because of the cost-free geo-tagged photos; however, time still needs to be spent on manual, laborious process of interpreting geo-tagged photos.

The collaborative volunteer-based methods require volunteers with local knowledge from any region of the world to upload their own geo-tagged photos or to collaboratively interpret images about their familiar regions over the web to support LC map validation [5,15–18]. The developments of three projects and/or systems are worth mentioning here. The first one is a Degree Confluence Project aiming at collecting geo-tagged photos from all the degree confluences by volunteers. It demonstrates the usefulness of geo-tagged photos for validating land cover maps [3,5], and has similar advantages and disadvantages discussed in the previous paragraph. The second one is the Virtual Interpretation of Earth Web-Interface Tool (VIEW-IT) [15]. It is a collaborative browser-based tool for sample data collection through interpreting high resolution satellite images in a crowd-sourcing manner [15]. The last one is Geo-Wiki, which is a collaborative volunteer-based tool to improve global LC map validation accuracy [16,17]. Geo-Wiki allows volunteers to upload or interpret various images, and it also can load pre-existing geo-tagged photos in Panoramio (a geolocation-oriented photo sharing website) and the Degrees Confluence Project web site [16]. Comparing with field survey and desktop-based visual interpretation methods, the collaborative volunteer-based methods have lower costs and shorter operation times, because volunteers usually make free contributions and their collaborative operation can help accelerate the interpretation process, especially for interpreting their familiar regions [15,16]. However, the collaborative volunteer-based methods also require efforts on interpreting images and may expect delays depending on how quickly volunteers make their contributions. In other words, this kind of method is "passive" in the sense that collection of geo-tagged web resources required for assisting timely LC map validation may not be possible. It is for these reasons that the active collection of sample data becomes very important.

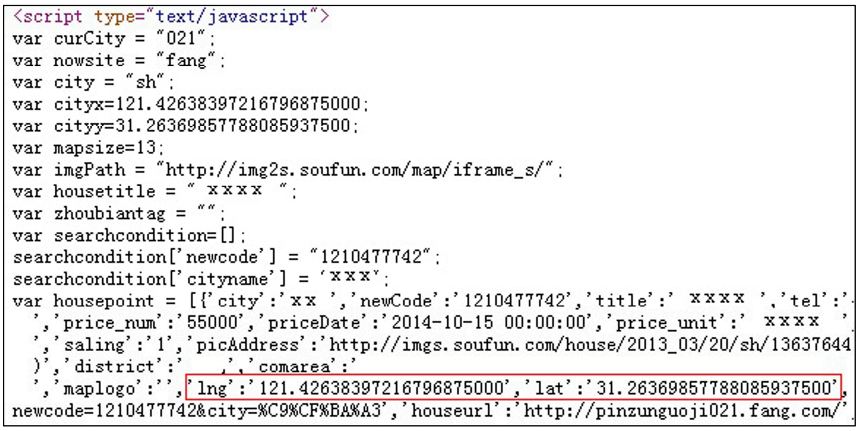
3. Methodology

Previous discussions and analyses conclude that geo-tagged web texts have not already been used to assist in LC map validation. To bridge this gap, this paper takes geo-tagged web texts about built-up areas for example to demonstrate that geo-tagged web texts have potential to become another new cost-free sample data source and can be actively collected to assist in LC map validation. The following presents frameworks of active collection of geo-tagged texts and web crawlers as well as how to transform geo-tagged web texts into sample data.

3.1. Conceptual Framework of Active Collection from Geo-Tagged Web Texts

According to the aforementioned analysis, geo-tagged web texts have not already been used to assist land cover map validation. To reach the goal, this paper takes geo-tagged web texts about built-up areas, for example, to demonstrate that geo-tagged web texts have the potential to become another new cost-free sample data source and can be actively collected to assist land cover map validation.

There are a huge amount of geo-tagged web texts about built-up areas in numerous real estate websites in China [51]. Although they are scattered everywhere on the web with different formats and provided by different organizations or people, they still share some common contents and structural characteristics. In terms of content, all geo-tagged web texts about built-up areas contain building names, building times, occupied areas, transport, geographic locations, and other descriptions. For simplicity, all of the above content except for the geographic location are collectively named as topic information in this study. Topic information is expressed in the form of Hyper Text Markup Language (HTML), while the corresponding geographic location is represented as map annotation and actual latitudes/longitudes that are hidden in JavaScript codes, as shown in the red box of Figure 1. In computer science, data stored in JavaScript codes, databases, forms, dynamic pages, and other types of objects are a part of deep web [27,28]. In other words, geographic locations of geo-tagged web texts belong to the deep web, which is the reason why this study selected deep web crawler to actively collect geo-tagged web texts about built-up areas.



```
<script type="text/javascript">
var curCity = "021";
var nowsite = "fang";
var city = "sh";
var cityx=121.42638397216796875000;
var cityy=31.26369857788085937500;
var mapsize=13;
var imgPath = "http://img2s.soufun.com/map/iframe_s/";
var housetitle = " xxxx  ";
var zhoubiantag = "";
var searchcondition=[];
searchcondition['newcode'] = "1210477742";
searchcondition['cityname'] = 'xxx';
var housepoint = [{'city':'xx ','newCode':'1210477742','title':' xxxx ','tel':'
  ','price_num':'55000','priceDate':'2014-10-15 00:00:00','price_unit':' xxxx  '
  ','saling':'1','picAddress':'http://imgs.soufun.com/house/2013_03/20/sh/13637644
  )','district':'    ','comarea':'
  ','maplogo':'','lng':'121.42638397216796875000','lat':'31.26369857788085937500',
newcode=1210477742&city=%C9%CF%BA%A3','houseurl':'http://pinzunguoji021.fang.com/'
```

Figure 1. Example JavaScript codes with a built-up area from the SouFun website.

One of the most important characteristics of geo-tagged web texts about built-up areas is their direct or indirect accordance with the content of sample data for LC map validation. As mentioned in Section 1, any data that mainly contains land cover type, spatial coverage, geographic location and collecting time may serve as candidate sample data. In geo-tagged web texts about built-up areas, land cover types can be automatically identified through keyword matching instead of image interpretation. For instance, built-up areas belong to artificial surfaces [9]. In addition, the occupied area of a built-up area can be used to determine whether the built-up area is uniform with minimum spatial coverage and its building time corresponds to collecting time. In general, one sample data should have several geographic locations, such as north, south, east and west locations in the Degree Confluence Project [5,16], whereas only one geographic location is labeled for built-up areas. Fortunately, GIS buffer analysis may help handle the difference in the numbers of geographic locations [52], where the occupied area and geographic location of the built-up areas are used to deduce the radius and center of the buffer. As such, geo-tagged web texts about built-up areas can serve as a candidate sample data source.

According to the characteristics of geo-tagged web texts, the objective of this paper is to actively collect geo-tagged web texts as a new cost-free sample data source to assist LC map validation. Figure 2 summarizes the conceptual framework of the active sample data collection, of which the deep web crawler and the data transformation process are two main modules of the proposed approach in the red box. Firstly, geo-tagged web texts about built-up areas are actively collected via a proposed deep web crawler. Referring to sample data requirements, the proposed deep web crawler should define topical keywords and formalize extraction rules to automatically collect and extract contents of geo-tagged web texts. Secondly, a data transformation process using GIS-based buffer analysis is performed to bridge the gap between geo-tagged web texts and candidate sample data, because the collected geo-tagged web texts may not directly meet sample data requirements. Finally, the transformed candidate sample data is used for LC map validation. The advantages of the proposed approach are twofold: (1) more efficient data collection due to the nature accordance of geo-tagged text contents with the content of sample data, without performing image interpretation; and (2) more active and automatic sample data collection using a deep web crawler.

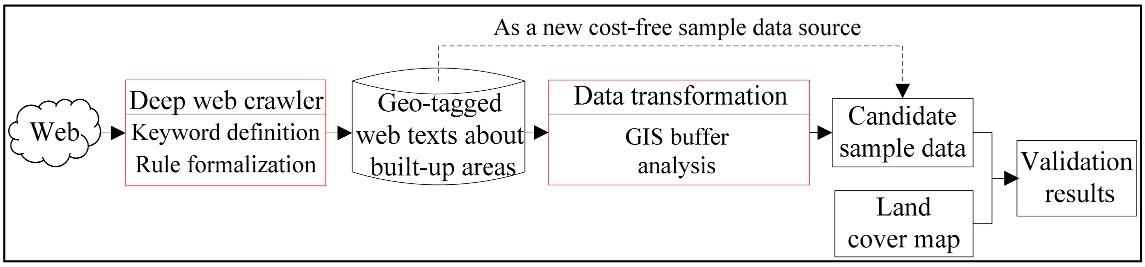


Figure 2. The conceptual framework of active sample data collection.

3.2. Deep Web Crawler for Geo-Tagged Web Texts

As mentioned in Section 3.1, geo-tagged web texts about built-up areas contain topic information and geographic location information. To obtain these data, the study proposes a deep web crawler to collect topic information and geographic locations. Figure 3 illustrates the crawling process realized by the proposed deep web crawler.

Similar to surface web crawlers, the proposed deep web crawler is given some seed URLs as initial values, and is followed by fetching web pages of seed URLs to extract topical URLs based on rule matching. These topical URLs, whose corresponding web pages may contain topical information, are stored in a topical URL queue. Next, two interconnected cycles run with topical URLs. The first one employs a HTML parsing engine together with some rules to extract topic information. The extracted topic information is stored in a topic database. In addition, topical URLs are also extracted in this cycle. Another cycle aims to extract geographic locations of geo-tagged web texts. Unlike the first cycle, it firstly utilizes JavaScript parsing engine and keywords of geographic locations to parse web pages of topical URLs. If geographic locations are obtained from JavaScript codes in the web pages, geographic locations are stored in the geographic database. Otherwise, map URL, whose corresponding web page contains geographic locations, is extracted from the page using keyword matching. Afterwards, geographic locations will be extracted by using JavaScript parsing engine and the keywords of the geographic locations. The two cycles continue running until the topical URL queue becomes empty or other conditions are met. Finally, the topic database and geographic database are combined into a database of geo-tagged web texts about built-up areas through the same topical URL.

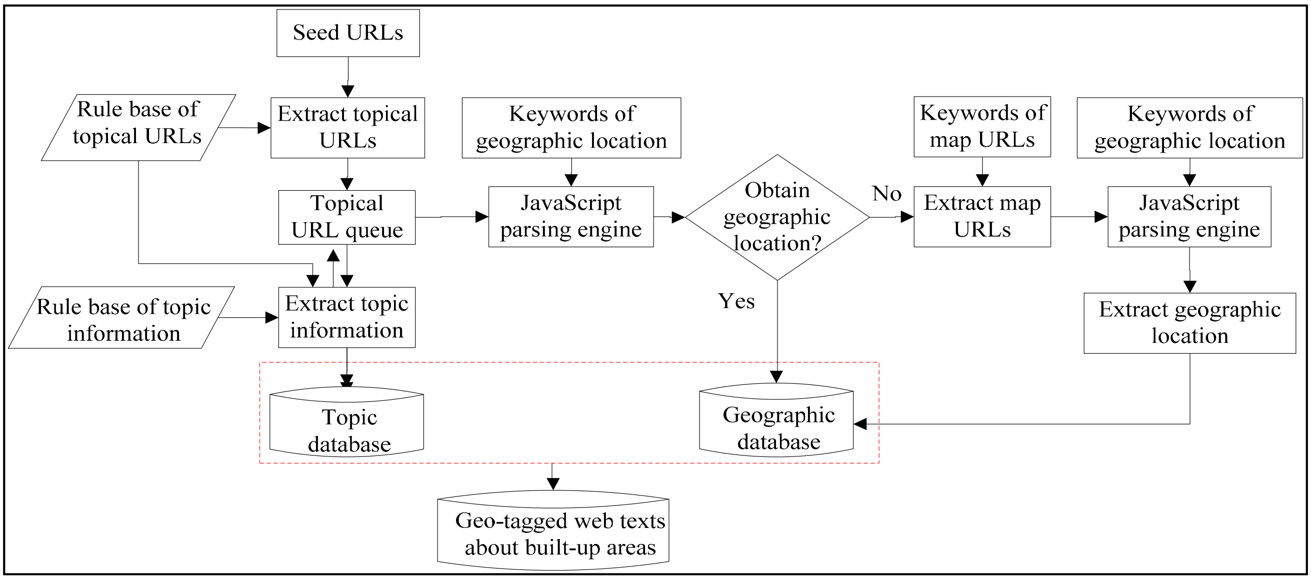

Seed URLs
Rule base of topical URLs
Extract topical URLs
Keywords of geographic location
Keywords of map URLs
Keywords of geographic location
Topical URL queue
JavaScript parsing engine
Obtain geographic location?
No
Extract map URLs
JavaScript parsing engine
Rule base of topic information
Extract topic information
Yes
Extract geographic location
Topic database
Geographic database
Geo-tagged web texts about built-up areas

Figure 3. The framework of the proposed deep web crawler.

Two rule bases are utilized to extract topical URLs and topic information, as shown in Figure 3. The two rule bases are constructed based on HTML structures of websites. They are formalized with XML Path Language (XPath). Occupied area, for instance, is defined in the structure of "<div class='lineheight'><strong> occupied area:</strong>XX square meters …</div>" in the website of SouFun, and the rule used to extract occupied area is "//div[@class='besic_inform']//td//strong|//div [@class='lineheight']//strong" based on XPath grammars.

In addition, two kinds of keywords are utilized to extract geographic locations and map URLs. The first kind is the keywords of geographic locations and they are extracted by analyzing their context words in JavaScript codes commonly used in real estate websites. The extracted keywords of geographic locations include "x", "y", "lat", "lng", "initLat", "initLng", "coordx", "coordy", "longitude" and "latitude". Keywords "x", "lng", "initLng", "coordx" and "longitude" correspond to longitude of geographic location; whereas keywords "y", "lat", "initLat", "coordy" and "latitude" represent latitude of geographic location. This paper exploits a maximum-frequency method, in which keywords are selected through descending order of term frequency in candidate words [46], to obtain the second kind of keywords of maps URLs. By doing so, the candidate words are extracted from map URLs and their anchor text from most real estate websites. In this paper, only the word "map" is selected as the keyword of map URLs due to its maximum frequency.

3.3. Data Transformation Process Using GIS Buffer Analysis

Geo-tagged web texts about built-up areas have some differences with other types of sample data sources, such as geo-tagged photos, aerial photographs and satellite images. One difference is that geo-tagged web texts about built-up areas do not rely on image interpretation. Another is the fact that there is only one geographic location labeled. In other words, the collected geo-tagged web texts cannot be directly used as sample data, and, thus, a data transformation process based on buffer analysis is developed to transform the collected raw texts into sample data (see Figure 4). The data transformation process mainly consists of the following steps:

(1) Geo-tagged web texts about built-up areas are collected by the proposed deep web crawler described in Section 3.1.
(2) The collected geo-tagged web texts are filtered to remove undesired texts according to some spatial and temporal restrictions of LC maps. For example, if geo-tagged web texts are not uniform with minimum spatial coverage and temporal coverage of LC maps, they should be discarded.
(3) GIS buffer analysis is exploited to obtain reference zones, which in turn transform the rest of geo-tagged web texts to candidate sample data, as discussed in Section 3.1.
(4) The pixel comparison is conducted to validate the accuracy of the LC maps. Details of pixel comparison are introduced in the following sections.

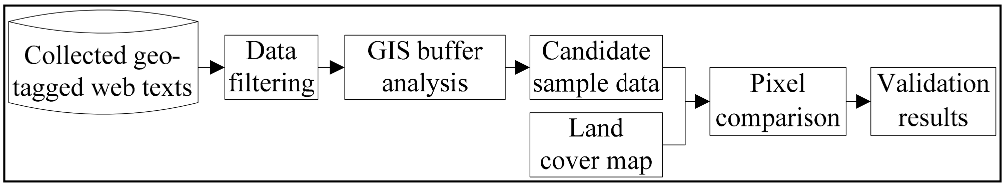


Figure 4. Data transformation process using GIS buffer analysis.

It is assumed that all built-up areas are squares in order to simplify GIS buffer analysis. Based on this assumption, five square buffer zones are designed, as shown in Figure 5. In these buffer zones, the geographic location of any geo-tagged web text has five possible locations, including (1) center; (2) upper left; (3) bottom right; (4) bottom left; and (5) upper right. In addition, the occupied area, which refers to the horizontal projection area of the land possessed or used by the built-up area and is collected from the web, can be used to deduce buffer radius by Equation (1). The unit of buffer radius is meter because the unit of the occupied area is usually square meter. However, the unit of pixel in land cover map often is longitude/latitude. Hence, a unit conversion is necessary to convert the meter to longitude/latitude in advance. Based on the geometric principles of squares, if one point location and the difference value between center points and other four directional points are known, the rest four points will be automatically obtained by a simple adding or subtracting. In this study, the differences in longitude and latitude between center point and other four directional points can be deduced from Equations (2) and (3), respectively.

$$r = \frac{\sqrt{area}}{2} \tag{1}$$

$$\Delta \ln g = \frac{r \times 180}{\pi \times R} \tag{2}$$

$$\Delta lat = \frac{r \times 180}{\pi \times R \times Cos(lat)} \tag{3}$$

where r represents buffer radius, *area* denotes occupied area. Variables Δln*g* and Δ*lat* represent the difference in longitude and latitude, respectively. π is the circular constant. Variable *R* represents earth radius which is set to 6,400,000 m [53,54]. Variable *lat* represents the latitude of the known points.

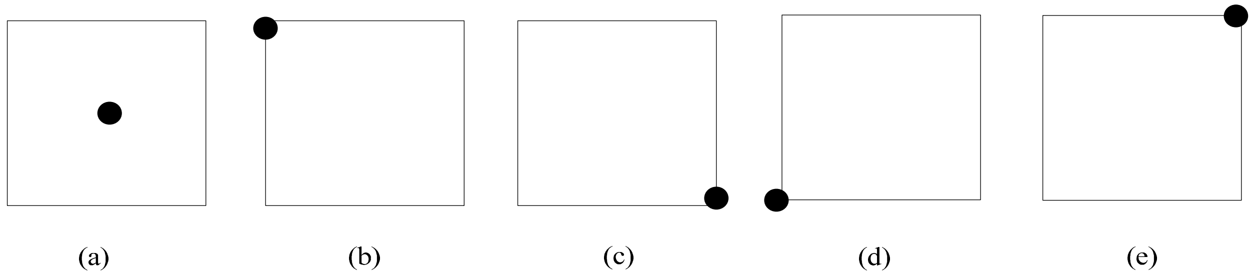


Figure 5. The five buffer zones. (**a**) Center; (**b**) Upper left; (**c**) Bottom right; (**d**) Bottom left; (**e**) Upper right.

As for pixel comparison, the pixel rate of artificial surfaces, which represents the fraction of pixels about artificial surfaces in all pixels in a reference zone, is first calculated by Equation (4). Five pixel rates can be obtained as every geo-tagged web text has five buffer zones. This study selects the maximum value of these five pixel rates as its final value. After that, the land cover types of geo-tagged web texts in land cover map are identified. The identification process is as follows: if the final pixel rate is greater than or equal to a given threshold, the geo-tagged web text is correctly identified as artificial surfaces; otherwise, the geo-tagged web text deems to be misidentified. Finally, user's accuracy, which refers to the probability that a sample point is correctly labeled as a certain land cover class in the map, is described as Equation (5) in this study [4].

$$pr = \frac{N_{as}}{N_{rz}} \tag{4}$$

where *pr* is the pixel rate of artificial surfaces, N_{as} and N_{rz} represent the number of pixel about artificial surfaces and the total number of pixel in the reference zone, respectively.

$$u = \frac{CR}{TC} \times 100\% \tag{5}$$

where variables of *u*, *CR*, and *TC* denote user's accuracy, the number of geo-tagged web texts correctly identified and the total number of geo-tagged web texts, respectively.

4. Experiments and Analysis

4.1. Experimental Data and Study Area

Six experiments were carried out on the land cover map of GlobeLand30-2010 to evaluate the performance of the proposed approach. The test data sets are available from website http://www.globeland30.org [9]. GlobeLand30 is the products of China's global land cover mapping project with 30-m resolution [9,55]. It provides global coverage for the years 2000 and 2010 with an overall accuracy of 80% or better [9,55]. In addition, GlobeLand30 adopts a classification scheme consisting of 10 first-level classes, including water bodies, wetland, artificial surfaces, cultivated land, permanent snow/ice, forest, shrubland, grassland, bareland and tundra [9]. The experiments only focus on artificial surfaces in the baseline year 2010 about GlobeLand30. The preliminary user's accuracy of artificial surfaces in the baseline year 2010 is 86.7%, validated by third-party experts [9]. Their sample data was collected by intensive field survey and interpretation of high-resolution images [9].

In this study, geo-tagged web texts were collected only from the website of SouFun Holdings Ltd. (Shijiazhuang, China), which operates one of the leading real estate web portals in China. The study area covers 37 cities located in Liaoning, Beijing, Tianjin, Shandong, Shanghai and Shanxi in China. For simplicity, these 37 cities were subdivided into six areas according to the administrative division of China in the experiments.

4.2. Results of Deep Web Crawler

The prototype deep web crawler was developed based on the Microsoft NET framework 3.5 with C Sharp programming language. In the prototype, the Microsoft ScriptControl was adopted to serve as JavaScript parsing engine. The prototype was then deployed on a desktop computer that has an Intel Pentium 4 CPU, 3.20 GHZ and 1 GB RAM with 6 MB bandwidth Internet connection. Table 1 reports 37 seed URLs used in the first experiments. The corresponding collection results and efficiency are given in Table 2.

Table 1. Seed Uniform Resource Locators (URLs) of the six experiments.

Study Areas	Seed URLs
Liaoning	http://newhouse.sy.fang.com/house/baojia.htm
	http://newhouse.fushun.fang.com/house/baojia.htm
	http://newhouse.yk.fang.com/house/baojia.htm
	http://newhouse.benxi.fang.com/house/baojia.htm
	http://newhouse.chaoyang.fang.com/house/baojia.htm
	http://newhouse.panjin.fang.com/house/baojia.htm
	http://newhouse.dandong.fang.com/house/baojia.htm
	http://newhouse.fuxin.fang.com/house/baojia.htm
	http://newhouse.huludao.fang.com/house/baojia.htm
	http://newhouse.liaoyang.fang.com/house/baojia.htm
	http://newhouse.tieling.fang.com/house/baojia.htm
	http://newhouse.anshan.fang.com/house/baojia.htm

Table 1. *Cont.*

Study Areas	Seed URLs
Beijing	http://newhouse.fang.com/house/baojia.htm
Tianjin	http://newhouse.tj.fang.com/house/baojia.htm
Shandong	http://newhouse.jn.fang.com/house/baojia.htm http://newhouse.binzhou.fang.com/house/baojia.htm http://newhouse.dz.fang.com/house/baojia.htm http://newhouse.dy.fang.com/house/baojia.htm http://newhouse.heze.fang.com/house/baojia.htm http://newhouse.jining.fang.com/house/baojia.htm http://newhouse.lc.fang.com/house/baojia.htm http://newhouse.linyi.fang.com/house/baojia.htm http://newhouse.qd.fang.com/house/baojia.htm http://newhouse.rz.fang.com/house/baojia.htm http://newhouse.taian.fang.com/house/baojia.htm http://newhouse.weihai.fang.com/house/baojia.htm http://newhouse.wf.fang.com/house/baojia.htm http://newhouse.yt.fang.com/house/baojia.htm http://newhouse.zaozhuang.fang.com/house/baojia.htm http://newhouse.zb.fang.com/house/baojia.htm
Shanghai	http://newhouse.sh.fang.com/house/baojia.htm
Shanxi	http://newhouse.baoji.fang.com/house/baojia.htm http://newhouse.hanzhong.fang.com/house/baojia.htm http://newhouse.weinan.fang.com/house/baojia.htm http://newhouse.xian.fang.com/house/baojia.htm http://newhouse.xianyang.fang.com/house/baojia.htm http://newhouse.sxyulin.fang.com/house/baojia.htm

Table 2. Collection results and efficiency of the proposed deep web crawler.

	Liaoning	Beijing	Tianjin	Shandong	Shanghai	Shanxi	Total Number
Collected Number	1264	485	478	2573	673	810	6283
Collecting Time(s)	2022	630	717	4640	1346	1528	10,883
Speed	0.63	0.77	0.67	0.55	0.5	0.53	0.58

As shown in Table 2, the number of geo-tagged web texts collected was 6283 in the six experiments, which expended 10,883 s with one thread. The speed of the proposed approach ranges from 0.5 to 0.67 (geo-tagged web texts per second), with an average value of 0.58. This indicates that the change in collecting speed is small; and thus, the efficiency of the proposed approach is robust and stable to collect geo-tagged web texts.

Previous studies show that in general it costs a team at least several hours to collect sample data with traditional approaches [6,15]. In contrast, the above experiments demonstrate that the proposed approach only requires one person and costs only thousands of seconds; therefore, it reduces the costs of labor and time comparing to those of other methods, which in turn improves the economic and computational efficiency.

In addition, an extra experiment was carried out to prove its ability to search multiple different websites. Two different websites of real estate companies are used as shown in Table 3. Unlike the six experiments above, geo-tagged web texts in the six areas were collected in one experiment, which may make the deep web crawler access the same website frequently. Therefore, sixteen free proxy IP addresses with four threads were used in turn in the experiment to avoid having our IP (Internet Protocol) address banned. Table 4 reports the corresponding collection results and efficiency. In the experiment, geo-tagged texts whose occupied area is empty were ignored, when the collected number was counted.

Table 3. Seed URLs from two different websites.

Study Areas	Seed URLs
Liaoning	http://db.house.qq.com/index.php?mod=search&city=dl#LXNob3d0eXBlXzE http://db.house.qq.com/index.php?mod=search&city=sy#LXNob3d0eXBlXzE http://db.house.qq.com/index.php?mod=search&city=fushun#LXNob3d0eXBlXzE http://db.house.qq.com/index.php?mod=search&city=jinzhou#LXNob3d0eXBlXzE http://db.house.qq.com/index.php?mod=search&city=yingkou#LXNob3d0eXBlXzE http://data.house.sina.com.cn/dl/search/ http://data.house.sina.com.cn/sy/search/
Beijing	http://db.house.qq.com/index.php?mod=search&city=bj#LXNob3d0eXBlXzE http://data.house.sina.com.cn/bj/search/
Tianjin	http://db.house.qq.com/index.php?mod=search&city=tianjin#LXNob3d0eXBlXzE http://data.house.sina.com.cn/tj/search/
Shandong	http://db.house.qq.com/index.php?mod=search&city=weihai#LXNob3d0eXBlXzE http://db.house.qq.com/index.php?mod=search&city=yantai#LXNob3d0eXBlXzE http://db.house.qq.com/index.php?mod=search&city=jn#LXNob3d0eXBlXzE http://db.house.qq.com/index.php?mod=search&city=qd#LXNob3d0eXBlXzE http://db.house.qq.com/index.php?mod=search&city=linyi#LXNob3d0eXBlXzE http://data.house.sina.com.cn/sd/search/ http://data.house.sina.com.cn/qd/search/
Shanghai	http://db.house.qq.com/index.php?mod=search&city=sh#LXNob3d0eXBlXzE http://data.house.sina.com.cn/sh/search/
Shanxi	http://db.house.qq.com/index.php?mod=search&city=xian#LXNob3d0eXBlXzE http://data.house.sina.com.cn/sx/search/ http://data.house.sina.com.cn/xy/search/

As shown in Table 4, a total of 10,362 geo-tagged web texts were collected in 7002 s with four threads. Its average speed of 0.37 is lower than that of the above experiments, because the experiment increases the sleep time of threads, which also aims to avoid accessing the same website frequently. Therefore, this also indicates that the efficiency of the proposed approach is robust and stable. Further, the experiment demonstrates that the proposed deep web crawler has the ability to collect geo-tagged web texts from different websites of real estate companies. This is because the keywords and extraction rules used in the proposed deep web crawler are extracted and formed by analyzing various real estate websites.

Table 4. Collection results and efficiency of the extra experiment.

	Liaoning	**Beijing**	**Tianjin**	**Shandong**	**Shanghai**	**Shanxi**	**Total Number**
Collected Number	1577	4034	1202	2514	2024	1711	10,362
Collecting Time(s)	7002						
Speed	10362 / (7002×4) = 0.37						

4.3. Assisting Validation and Results

To evaluate the potential of geo-tagged web texts as a sample data source, the collected geo-tagged web texts were used to validate artificial surfaces of GlobeLand30 in the baseline year 2010. The experiments were also carried out in the above six areas according to the designed data transformation process.

Data filtering was performed to remove the incorrect geo-tagged web texts. In the experiments, since the minimum map unit of artificial surfaces in GlobeLand30 is 8 × 8 pixels ((*i.e.*, 240 m × 240 m) block of 30 m, those geo-tagged web texts whose occupied areas are greater than or equal to 57,600 m^2 (*i.e.*, 240 m × 240 m) were reserved. The final numbers of reserved geo-tagged web texts between 2010 and 2015 are presented in Table 5.

Table 5. Number of reserved geo-tagged web texts between 2010 and 2015.

Building Time	**Liaoning**	**Beijing**	**Tianjin**	**Shandong**	**Shanghai**	**Shanxi**
2010	**139**	**74**	**67**	**282**	**184**	**80**
2011	168	48	71	402	91	76
2012	90	36	51	178	48	85
2013	58	25	51	204	46	54
2014	11	10	9	25	13	17
2015	0	0	0	0	1	1

On the other hand, as the temporal coverage of GlobeLand30 used in this study was the baseline year 2010, the collected geo-tagged web texts were accordingly filtered again, which means that only those whose building time was the year 2010 could be used to validate artificial surfaces. The numbers of geo-tagged web texts finally used are shown in bold in Table 5. Since these geo-tagged web texts are tagged with map datum WGS_84, the Web Map Service of GlobeLand30 with ESPG_4326 was therefore requested. Figure 6 presents the geo-tagged web texts in the six study areas on the GlobeLand30, and Figure 7 shows enlarged view of the geo-tagged web texts in Shanghai area. In Figure 6, the red arrows indicate geographic location of corresponding textual label and the black points represent geo-tagged web texts.

Once the qualified geo-tagged web texts were obtained, GIS buffer analysis and pixel comparison were carried out subsequently. Figure 8 shows the pixel rate of artificial surfaces in reference zones and their numbers in the six study areas. As can be seen clearly, the number of geo-tagged web texts, whose rate value falls in interval (0.9, 1], is the most frequent in all six study areas, especially in Shandong. Although the numbers of geo-tagged web texts do not vary much between other intervals, a threshold line may still be drawn between interval (0.5, 1] and interval (0, 0.5] where a significant change can be observed between the two sides of the line (see the red line in Figure 8).

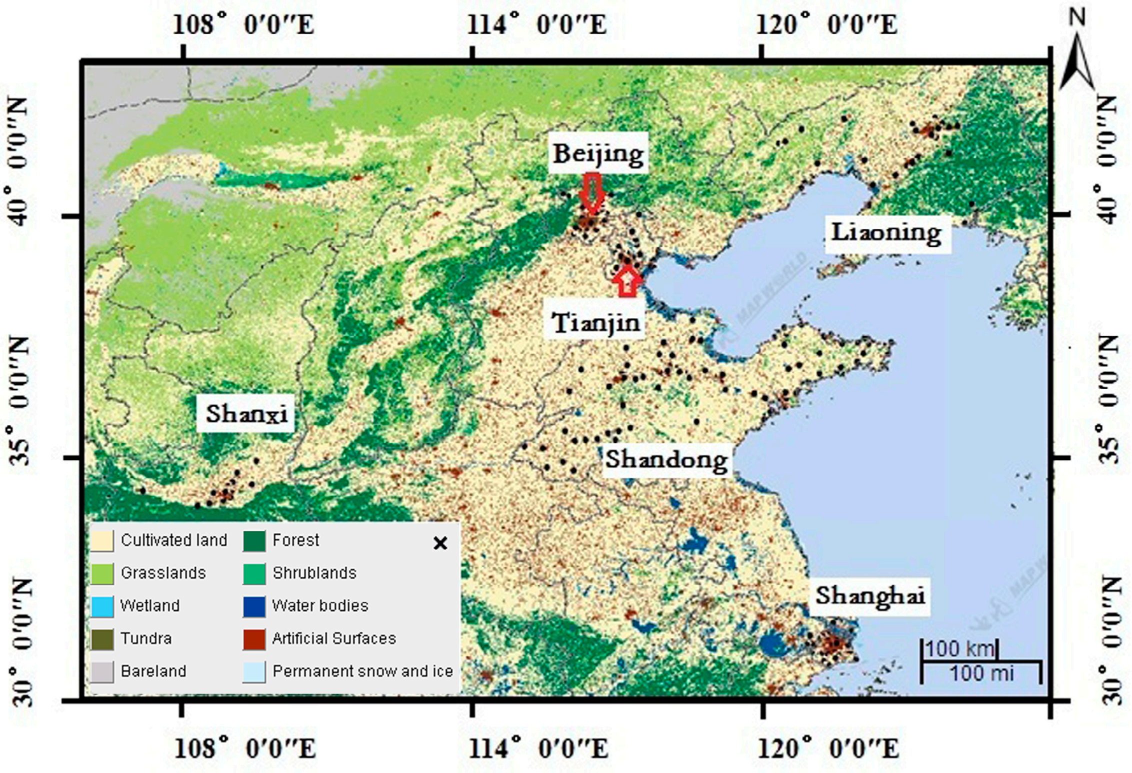


Figure 6. Distribution of geo-tagged web texts in the six study areas on GlobeLand30-2010.

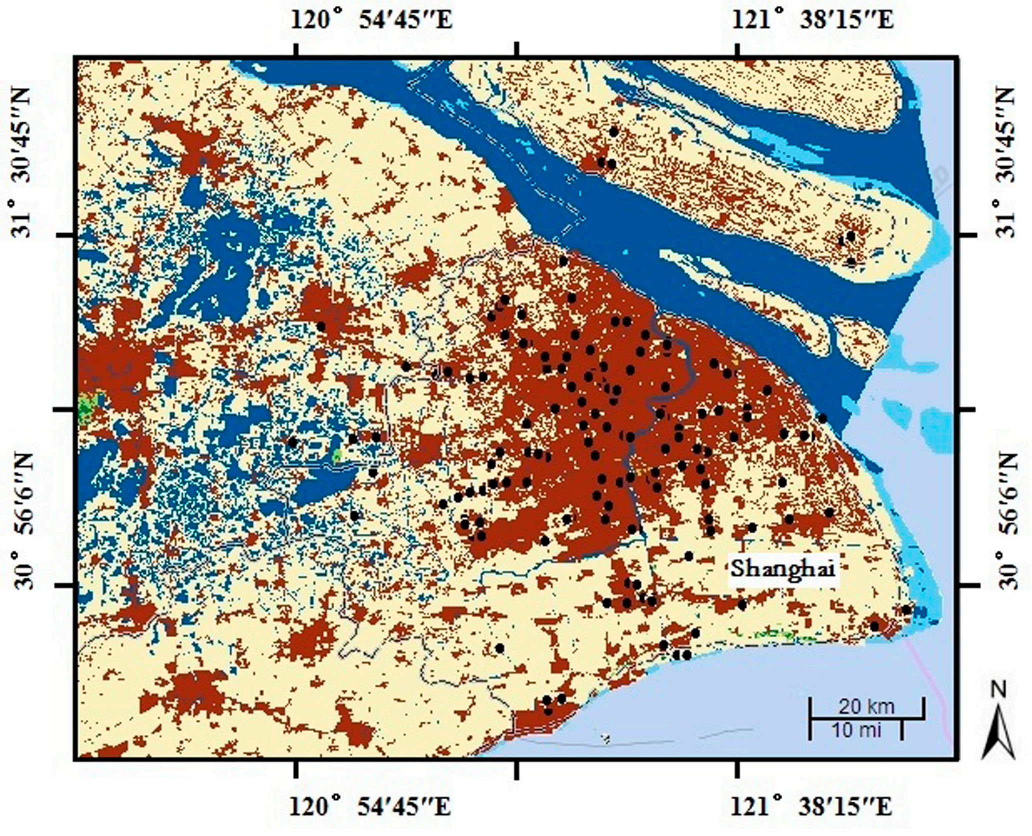


Figure 7. Distribution of geo-tagged web texts in Shanghai on GlobeLand30-2010.

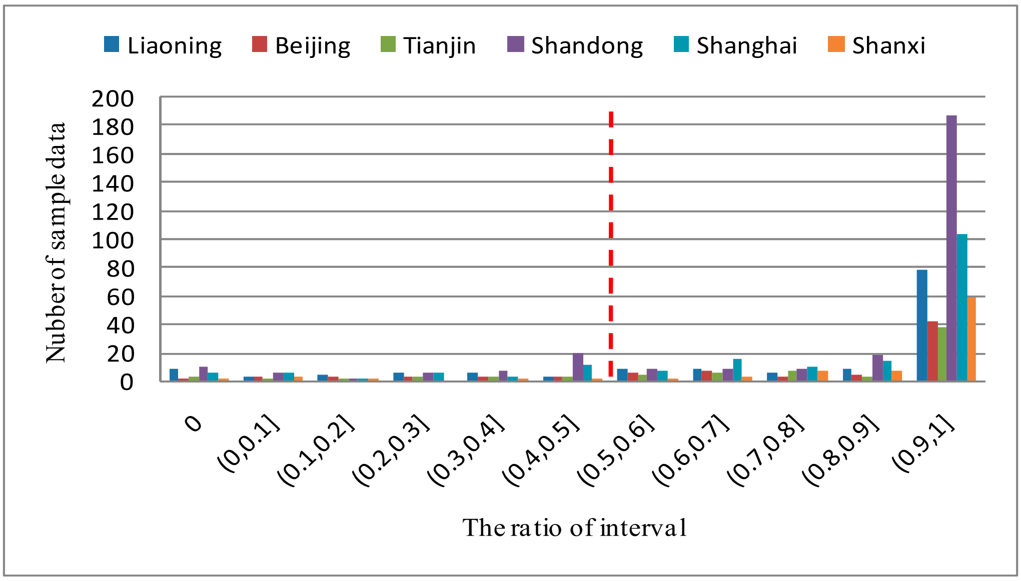


Figure 8. The pixel rate of artificial surfaces in reference zone.

If the pixel rate of artificial surfaces in one reference zone is greater than or equal to 0.5, the reference zone is most likely to be artificial surfaces. This is because GlobeLand30 has 10 classes and based on common sense, built-up areas are often decked by plants and waters. For this reason, the 0.5 pixel rate was set as the threshold value to identify whether or not the geo-tagged web text belongs to artificial surfaces in GlobeLand30. In other words, if the rate is greater than or equal to 0.5, it means that the geo-tagged web text is correctly classified; vice versa. After repeating this computation process, the number of *CR* (*i.e.*, the number of geo-tagged web texts correctly classified) and the user's accuracy of GlobeLand30 in the year 2010 were obtained, as shown in Table 6.

Table 6. User's accuracy of artificial surfaces in GlobeLand30-2010.

	Liaoning	Beijing	Tianjin	Shandong	Shanghai	Shanxi	Total Number
CR	108	60	55	232	151	73	679
TC	139	74	67	282	184	80	826
User's accuracy	77.70%	81.08%	82.09%	82.27%	82.07%	91.25%	82.20%

It can be seen from Table 6 that in the six study areas the user's accuracy calculated by correctly identified geo-tagged web texts ranges from 77.70% to 91.25%. The preliminary user's accuracy of artificial surfaces validated by third-party experts [9] in the baseline year 2010 is 86.7%, while the average value of the proposed approach is 82.2%, which is only 4.5% lower than the former one. The small gap indicates that the proposed approach is comparable to the expert validation approach and, thus, the collected geo-tagged web texts have a potential to serve as a new cost-free sample data source to assist land cover map validation.

5. Discussions

Our experiments confirm that geo-tagged web texts about built-up areas are indeed valuable ancillary data for land cover map validation and that the active approach based on deep web crawler is efficient, robust and stable to collect land cover sample data.

Certainly, our findings on using geo-tagged web texts to validate land cover map is similarly to those of several web-based approaches and tools using geo-tagged photos to support land cover map validation [5,15–17,49,50]. For example, Foody and Boyd [3] have confirmed the potential value of geo-tagged photos in the Degrees of Confluence project as the useful and spatially extensive data to aid land cover map validation. However, we used geo-tagged web texts about built-up areas rather than geo-tagged photos as sample data to support land cover map validation, while the filed survey data, existing land cover maps or various images are used in the traditional approaches [5–11]. One of the challenges facing web-based approaches/tools is to attract experienced volunteers who are willing to spend time uploading geo-tagged photos or labeling their interpretation results [16]. To meet this challege, we developed an active approach based on a deep web crawler to automatically collect geo-tagged web texts in other websites as sample data, while most of candidate sample data obtained from filed survey data, existing land cover maps or various images rely on manual operation [2,6,8,12,16,17].

Although the proposed approach helps automate the task of land cover map validation, the traditional approaches of sample data collection are still valid and necessary. Complementary to each other, they can be used to create a good balance between efficiency and accuracy required for land cover sample data collection. For instance, geo-tagged web texts collected by the proposed approach may be considered as ancillary knowledge to help traditional approaches interpret images, while some types of sample data that cannot be crawled on the web can still be collected by traditional approaches.

Despite the advantages and benefits of the proposed approach, there is still a lot of work that needs to be done to improve the active collection approach of land cover sample data from geo-tagged web texts, which are shown as follows:

(1) Usually, land cover map validation should be done through a rigorous sampling scheme. However, this study just aims to demonstrate that geo-tagged web texts have the potential to become another new cost-free sample data source, rather than to perform rigorous validation. Therefore, sampling schemes are ignored to keep the study focused and simple. After the sample data about other land cover types is collected from geo-tagged web texts, a rigorous sampling scheme suitable for online validation will be integrated into the proposed approach in the future.

(2) Only geo-tagged web texts of built-up areas were collected using the proposed approach to validate artificial surfaces. However, an extensive verification with other land cover types would provide more valuable insights into the proposed approach. Web texts about cultivated land, shrubland, and tundra are difficult to find for the time being, and their occupied areas are seldom offered by websites. Fortunately, a lot of texts about wetland, water bodies, permanent snow/ice, forest, grassland, and bareland are geo-tagged on the web (*i.e.*, Wikipedia and Baidu Encyclopedia [56,57]) and their occupied areas are also offered in detail. For instance, geographic location, occupied area and other descriptions about the wetland area of the Pantanal can be collected from the Wikipedia website [58]. However, the content and structural characteristics of other land cover types are different from artificial surfaces class and their shapes are irregular, which makes the data transformation more difficult. A unified description model for representing the land cover types of wetland, water bodies, permanent snow/ice, forest, grassland, and bareland is under development now. Further, some new rules

for extracting the required factors from multiple different web sources are to be summarized. In addition, a new data transformation process will need to be designed for irregular shapes based on minimum bounding box or convex hull in the future. As part of the long-term research objectives of the team [59], when this work is completed, the proposed approach can be used to collect other land cover types from different websites by simply replacing the modules of keywords definition, extraction rules and data transformation.

(3) The quality of geo-tagged web texts directly affects land cover map validation results. Currently, geo-tagged web texts about built-up areas used in this study are from commercial companies, which may be credible to some extent. Except for this data source, there are also a large number of geo-tagged web texts contributed by individual volunteers with no quality assurance. It is difficult to assess the quality of these voluntarily contributed contents quantitatively, but the increasing emotional comments on the content that somehow indicate "approval" or "disapproval" may help at least measure the quality to certain extent. Future work will include sentiment analysis of comments to assess their quality for more accurate validation.

6. Conclusions

Geo-tagged web texts have great potential in providing valuable geo-location information. As massive geo-tagged texts become available on the web, a natural question is how we can utilize this vast amount of textural information to extract necessary sample data for land cover map validation. It is for this reason that the study reported in this paper investigated an active sample data collection approach using deep web crawling techniques to collect geo-tagged web texts to support land cover map validation. This active collection approach consists of two steps. Firstly, a deep web crawler was customized in terms of land cover-related keywords and string-based rule matching. The land cover-related keywords were defined by the maximum-frequency method and the extraction rules were formalized with the XPath Language. Secondly, a buffer analysis-based data transformation processing was conducted to convert the collected raw web texts into land cover sample data according to some spatial and temporal restrictions.

Experiments were conducted to collect geo-tagged web texts from the website of SouFun, the largest real estate broker company in China, in 37 cities of three provinces and three municipalities directly under the Central Government in China to validate the artificial surface class of GlobeLand30-2010. Thirty-seven URLs from the website were selected and used as the seed URLs of the customized deep web crawler. A total of 6283 raw geo-tagged web texts were collected at an average speed of 0.58 texts per second. Among them, 826 raw geo-tagged web texts being consistent with GlobeLand30-2010 in temporal and spatial dimensions were converted into sample data and were used alone to validate artificial surface class of GlobeLand30-2010. The average user's accuracy of 82.2% was achieved. It is close to the user's accuracy of 86.7% derived in a formal validation by third-party experts. The experimental results show that geo-tagged web texts can serve as very useful ancillary data to assist in land cover map validation and the proposed approach can improve sample data collection efficiency.

Acknowledgments

This work has been funded by the Ministry of Science and Technology of China (Project #2012BAK12B01) and the National Science Foundation of China (Project #41231172 and #41301412). The authors would like to thank the journal editor and reviewers for their constructive comments, which helped improve the paper and Zelang Miao from the Hong Kong Polytechnic University for his assistance in paper editing.

Author Contributions

Dongyang Hou developed the concept, designed the active collection method, implemented the method, performed the major part of experiments, and drafted the manuscript. Jun Chen and Hao Wu made substantial contributions to conceptual design, methodological development and preparation of the manuscript. Songnian Li participated in project discussions and made important contributions to the revision and final editing of the manuscript. Fei Chen and Weiwei Zhang participated in the discussion of the idea and performed part of the experiments.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Stehman, S.V. Sampling designs for accuracy assessment of land cover. *Int. J. Remote Sens.* **2009**, *30*, 5243–5272.
2. Estima, J.; Painho, M. Flickr geotagged and publicly available photos: Preliminary study of its adequacy for helping quality control of corine land cover. In *Computational Science and Its Applications—ICCSA 2013*; Murgante, B., Misra, S., Carlini, M., Torre, C., Nguyen, H.-Q., Taniar, D., Apduhan, B., Gervasi, O., Eds.; Springer: Heidelberg, Germany, 2013; Volume 7974, pp. 205–220.
3. Foody, G.M.; Boyd, D.S. Using volunteered data in land cover map validation: Mapping west African forests. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.***2013**, *6*, 1305–1312.
4. Olofsson, P.; Foody, G.M.; Herold, M.; Stehman, S.V.; Woodcock, C.E.; Wulder, M.A. Good practices for estimating area and assessing accuracy of land change. *Remote Sens. Environ.* **2014**, *148*, 42–57.
5. Iwao, K.; Nishida, K.; Kinoshita, T.; Yamagata, Y. Validating land cover maps with Degree Confluence Project information. *Geophy. Res. Lett.* **2006**, *33*, L23404.
6. Tsendbazar, N.E.; de Bruin, S.; Herold, M. Assessing global land cover reference datasets for different user communities. *ISPRS J. Photogramm. Remote Sens.* **2015**, *103*, 93–114.
7. Comber, A.; See, L.; Fritz, S.; van der Velde, M.; Perger, C.; Foody, G. Using control data to determine the reliability of volunteered geographic information about land cover. *Int. J. Appl. Earth Obs. Geoinf.* **2013**, *23*, 37–48.

8. Zhao, Y.; Gong, P.; Yu, L.; Hu, L.; Li, X.; Li, C.; Zhang, H.; Zheng, Y.; Wang, J.; Zhao, Y.; *et al*. Towards a common validation sample set for global land-cover mapping. *Int. J. Remote Sens.* **2014**, *35*, 4795–4814.
9. Chen, J.; Chen, J.; Liao, A.; Cao, X.; Chen, L.; Chen, X.; He, C.; Han, G.; Peng, S.; Lu, M.; *et al*. Global land cover mapping at 30 m resolution: A POK-based operational approach. *ISPRS J. Photogramm. Remote Sens.* **2015**, *103*, 7–27.
10. Manakos, I.; Chatzopoulos-Vouzoglanis, K.; Petrou, Z.I.; Filchev, L.; Apostolakis, A. Globalland30 mapping capacity of land surface water in Thessaly, Greece. *Land* **2014**, *4*, 1–18.
11. Tong, X.; Zhang, X.; Shan, J.; Xie, H.; Liu, M. Attraction-repulsion model-based subpixel mapping of multi-/hyperspectral imagery. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 2799–2814.
12. Stehman, S.V.; Olofsson, P.; Woodcock, C.E.; Herold, M.; Friedl, M.A. A global land-cover validation data set, II: Augmenting a stratified sampling design to estimate accuracy by region and land-cover class. *Int. J. Remote Sens.***2012**, *33*, 6975–6993.
13. Bastin, L.; Buchanan, G.; Beresford, A.; Pekel, J.F.; Dubois, G. Open-source mapping and services for web-based land-cover validation. *Ecol. Inform.* **2013**, *14*, 9–16.
14. Wulder, M.A.; White, J.C.; Magnussen, S.; McDonald, S. Validation of a large area land cover product using purpose-acquired airborne video. *Remote Sens. Environ.* **2007**, *106*, 480–491.
15. Clark, M.L.; Aide, T.M. Virtual Interpretation of Earth Web-Interface Tool (VIEW-IT) for collecting land-use/land-cover reference data. *Remote Sens.* **2011**, *3*, 601–620.
16. Fritz, S.; McCallum, I.; Schill, C.; Perger, C.; Grillmayer, R.; Achard, F.; Kraxner, F.; Obersteiner, M. Geo-Wiki.Org: The use of crowdsourcing to improve global land cover. *Remote Sens.* **2009**, *1*, 345–354.
17. Fritz, S.; McCallum, I.; Schill, C.; Perger, C.; See, L.; Schepaschenko, D.; van der Velde, M.; Kraxner, F.; Obersteiner, M. Geo-Wiki: An online platform for improving global land cover. *Environ. Model. Softw.* **2012**, *31*, 110–123.
18. Han, G.; Chen, J.; He, C.; Li, S.; Wu, H.; Liao, A.; Peng, S. A web-based system for supporting global land cover data production. *ISPRS J. Photogramm. Remote Sens.* **2015**, *103*, 66–80.
19. Zheng, Y.-T.; Zha, Z.-J.; Chua, T.-S. Research and applications on georeferenced multimedia: A survey. *Multimed. Tools Appl.* **2011**, *51*, 77–98.
20. Li, L.; Goodchild, M.F.; Xu, B. Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartogr. Geogr. Inf. Sci.* **2013**, *40*, 61–77.
21. Majid, A.; Chen, L.; Chen, G.; Mirza, H.T.; Hussain, I.; Woodward, J. A context-aware personalized travel recommendation system based on geotagged social media data mining. *Int. J. Geogr. Inf. Sci.* **2013**, *27*, 662–684.
22. Leung, D.; Newsam, S. Exploring geotagged images for land-use classification. In Proceedings of the ACM Multimedia 2012 Workshop on Geotagging and Its Applications in Multimedia, Nara, Japan, 29 October–2 November 2012.
23. Lu, G.; Liu, S.; Lü, K. MBCrawler: A software architecture for micro-blog crawler. In Proceedings of the 2012 International Conference on Information Technology and Software Engineering, Beijing, China, 8–10 December 2012.

24. Gao, K.; Zhou, E.-L.; Grover, S. Applied methods and techniques for modeling and control on micro-blog data crawler. In *Applied Methods and Techniques for Mechatronic Systems*; Springer: Berlin, Germany, 2014; Volume 452, pp. 171–188.
25. Schmidt, V.; Binner, J. A semi-automated display for geotagged text. In *City Evacuations: An Interdisciplinary Approach*; Preston, J., Binner, J.M., Branicki, L., Galla, T., Jones, N., King, J., Kolokitha, M., Smyrnakis, M., Eds.; Springer: Heidelberg, Germany, 2015; pp. 107–116.
26. Wikipedia. Geotagged Photograph. Available online: http://en.wikipedia.org/wiki/Geotagged_photograph (accessed on 27 October 2014).
27. Manvi, M.; Dixit, A.; Bhatia, K.K. Design of an ontology based adaptive crawler for hidden web. In Proceedings of the 2013 International Conference on Communication Systems and Network Technologies (CSNT), Gwalior, India, 6–8 April 2013.
28. Piccinini, H.; Casanova, M.; Leme, L.P.; Furtado, A. Publishing deep web geographic data. *Geoinformatica* **2014**, *18*, 769–792.
29. Luo, J.; Joshi, D.; Yu, J.; Gallagher, A. Geotagging in multimedia and computer vision—A survey. *Multimed. Tools Appl.* **2011**, *51*, 187–211.
30. Cao, L.; Friedland, G.; Larson, M. GeoMM'12: ACM international workshop on geotagging and its applications in multimedia. In Proceedings of the 20th ACM International Conference on Multimedia, Nara, Japan, 29 October–2 November 2012.
31. Senaratne, H.; Bröring, A.; Schreck, T. Using reverse viewshed analysis to assess the location correctness of visually generated VGI. *Trans. GIS* **2013**, *17*, 369–386.
32. Ahern, S.; Naaman, M.; Nair, R.; Yang, J.H.-I. World explorer: Visualizing aggregate data from unstructured text in geo-referenced collections. In Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, Vancouver, BC, Canada, 18–23 June 2007.
33. Fujisaka, T.; Lee, R.; Sumiya, K. Discovery of user behavior patterns from geo-tagged micro-blogs. In Proceedings of the 4th International Conference on Uniquitous Information Management and Communication, Suwon, Korea, 14–15 January 2010.
34. Crampton, J.W.; Graham, M.; Poorthuis, A.; Shelton, T.; Stephens, M.; Wilson, M.W.; Zook, M. Beyond the geotag: Situating "big data"and leveraging the potential of the geoweb. *Cartogr. Geogr. Inf. Sci.* **2013**, *40*, 130–139.
35. Gao, S.; Li, L.; Li, W.; Janowicz, K.; Zhang, Y. Constructing gazetteers from volunteered big geo-data based on Hadoop. *Environ. Urban Sys.* **2014**, doi:10.1016/j.compenvurbsys.2014.02.004.
36. Popescu, A.; Grefenstette, G. Deducing trip related information from flickr. In Proceedings of the 18th International Conference on World Wide Web, Madrid, Spain, 20–24 April 2009; pp. 1183–1184.
37. Goodchild, M.F.; Glennon, J.A. Crowdsourcing geographic information for disaster response: A research frontier. *Int. J. Digit. Earth* **2010**, *3*, 231–241.
38. Chae, J.; Thom, D.; Jang, Y.; Kim, S.; Ertl, T.; Ebert, D.S. Public behavior response analysis in disaster events utilizing visual analytics of microblog data. *Comput. Graph.* **2014**, *38*, 51–60.
39. Eisenstein, J.; Smith, N.A.; Xing, E.P. Discovering sociolinguistic associations with structured sparsity. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, OR, USA, 19–24 June 2011.

40. Vu, H.Q.; Li, G.; Law, R.; Ye, B.H. Exploring the travel behaviors of inbound tourists to Hong Kong using geotagged photos. *Tour. Manag.* **2015**, *46*, 222–232.
41. Frias-Martinez, V.; Frias-Martinez, E. Spectral clustering for sensing urban land use using Twitter activity. *Eng. Appli. Artif. Intell.* **2014**, *35*, 237–245.
42. Boanjak, M.; Oliveira, E.; Martins, J.; Mendes Rodrigues, E.; Sarmento, L. TwitterEcho: A distributed focused crawler to support open research with twitter data. In Proceedings of the 21st International Conference Companion on World Wide Web, Lyon, France, 16–20 April 2012; pp. 1233–1240.
43. Tsou, M.-H.; Yang, J.-A.; Lusher, D.; Han, S.; Spitzberg, B.; Gawron, J.M.; Gupta, D.; An, L. Mapping social activities and concepts with social media (Twitter) and web search engines (Yahoo and Bing): A case study in 2012 US Presidential Election. *Cartogr. Geogr. Inf. Sci.* **2013**, *40*, 337–348.
44. Lee, R.; Wakamiya, S.; Sumiya, K. Discovery of unusual regional social activities using geo-tagged microblogs. *World Wide Web* **2011**, *14*, 321–349.
45. Batsakis, S.; Petrakis, E.G.; Milios, E. Improving the performance of focused web crawlers. *Data Knowl. Eng.***2009**, *68*, 1001–1013.
46. Hou, D.; Wu, H.; Chen, J.; Li, R. A focused crawler for borderlands situation information with geographical properties of place names. *Sustainability* **2014**, *6*, 6529–6552.
47. Raghavan, S.; Garcia-Molina, H. Crawling the hidden web. In Proceeding of the 27th International Conference on Very Large Data Bases (VLDB 2001), Rome, Italy, 11–14 September 2001.
48. Zeng, W.; Li, M. Survey on the research of deep web crawler. *Comput. Syst. Appl.* **2008**, *17*, 122–126.
49. Foster, A.; Dunham, I.M. Volunteered geographic information, urban forests, & environmental justice. *Comput. Environ. Urban Syst.* **2014,** doi:10.1016/j.compenvurbsys.2014.08.001.
50. Weiss, M.; Baret, F.; Block, T.; Koetz, B.; Burini, A.; Scholze, B.; Lecharpentier, P.; Brockmann, C.; Fernandes, R.; Plummer, S.; *et al*. On Line Validation Exercise (OLIVE): A web based service for the validation of medium resolution land products. Application to FAPAR products. *Remote Sens.* **2014**, *6*, 4190–4216.
51. Yan, Z.; Li, Q.; Dong, Y.; Ding, Y. An ontology-based integration of Web query interfaces for house search. In Proceedings of the International Conference on Information and Automation, Changsha, China, 20–23 June 2008.
52. Li, X.; Zhang, L.; Liang, C. A GIS-based buffer gradient analysis on spatiotemporal dynamics of urban expansion in Shanghai and its major satellite cities. *Procedia Environ. Sci.* **2010**, *2*, 1139–1156.
53. Oliveira, D. Ionosphere-magnetosphere coupling and field-aligned currents. *Revi. Bras. Ensino Físi.* **2014**, *36*, 1–5.
54. Arcavi, A. The role of visual representations in the learning of mathematics. *Educ. Stud. Math.* **2003**, *52*, 215–241.
55. Chen, J.; Chen, J.; Liao, A.; Cao, X.; Chen, L.; Chen, X. Concepts and key techniques for 30 m global land cover mapping. *Acta Geo. Et Carto. Sin.* **2014**, *43*, 551–557.
56. Wikipedia Encyclopedia. Available online: http://www.wikipedia.org/ (accessed on 30 January 2015).
57. Baidu Encyclopedia. Available online: http://baike.baidu.com/ (accessed on 30 January 2015).
58. Pantanal-Wikipedia, the Free Encyclopedia. Available online: http://en.wikipedia.org/wiki/Pantanal (accessed on 30 January 2015).

59. Chen, J.; Wu H.; Li S.; Chen F.; Han G. Services oriented dynamic computing for land cover big data. *J. Geo. Sci. Tec.* **2013**, *30*, 551–557.