Journal of Computers Vol. 35 No. 2, April 2024, pp. 285-301
doi: 10.53106/199115992024043502018

Rapid Production Method of Massive Thematic Maps Based on Geospatial Knowledge Extraction

Chuan Yin[1], Yanhui Wang[2,3*], Duoduo Yin[4], Wanzeng Liu[5], Hao Wu[6], Kexin Liu[1]

[1] School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing 102616, China
{yinchuan@bucea.edu.cn, 2108570021118}@bucea.edu.cn

[2] College of Resources Environment and Tourism, Capital Normal University, Beijing 100048, China
[3] School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing 102616, China
yanhuiwang@cnu.edu.cn

[4] Inner Mongolia water conservancy development center, Hohhot 010050, China
2190902114@cnu.edu.cn

[5] Information Service Department, National Geomatics Center of China, Beijing 100830, China
lwz@ngcc.cn

[6] Tianjin Geomatics Research Center, Tianjin, 300202, China
2200902181@cnu.edu.cn

Received 27 March 2024; Revised 21 April 2024; Accepted 23 April 2024

Abstract: Geospatial knowledge in massive academic papers can provide knowledge services such as location-based research hotspot analysis, spatio-temporal data aggregation, research results recommendation, etc. However, geospatial knowledge often exists implicitly in literature resources in unstructured form, which is difficult to be directly accessed and mined and utilized for rapid production of massive thematic maps. In this paper, we take the geospatial knowledge of the area studied as an example and introduce its extraction method in detail. An integrated feature template matching and random forest classification algorithm is proposed for accurately identifying research areas from the abstract texts of academic papers and producing thematic maps. Firstly, the precise recognition of geographical names is achieved step by step based on BiLSTM-CRF algorithm and improved heuristic disambiguation method; then, the area studied is extracted by the designed integrated feature recognition template of area studied using random forest classification algorithm, and a fast thematic map is designed for the knowledge of area studied, topic and literature. The experimental results show that the area studied recognition accuracy can reach 97%, the F-value is 96%, and the recall rate reaches 96%, achieving high accuracy and high efficiency of area studied extraction in text. Based on the geospatial knowledge, the thematic map can achieve the effect of fast map formation and accurate expression.

Keywords: area studied, BLFR model, BI-LSTM-CRF, improved heuristic disambiguation method, feature template, random forest

1 Introduction

Journal papers are one of the important carriers of knowledge of various disciplines in various fields, and they contain excellent research ideas, theories and achievements of scholars, which are the most cutting-edge, authoritative and accessible knowledge resources in various research fields, and contain many professional core knowledge including research problems, algorithmic models and other knowledge types [1]. More than 80% of all kinds of information involved in the field of mapping and geographic information are related to geospatial location, and geospatial knowledge is the main body for constructing geographic knowledge graph; therefore, geospatial knowledge is an important component for constructing geographic knowledge system oriented to the field of mapping and geographic information at present. According to different needs, geospatial knowledge in

* Corresponding Author

literature can be divided into types such as sampling area and area studied in which scientific research activities are located.

Area studied is a key kind of geospatial knowledge in journal papers, which refers to the area or geographical area selected by the authors of the papers for the scientific research theme in the process of scientific research and applicable to the research purposes such as data collection, experimental examination and result verification. It contains not only spatial location information, which corresponds to the coordinate interval on the map; but also semantic information, which indicates the geographical scope of the paper research. If the area studied in literature can be extracted and gathered, it is possible to provide users with knowledge services such as hotspot area discovery and location-based spatial data recommendation through statistical analysis and association rule mining. Therefore, as an important basic geographic knowledge, the extraction technology of area studied has shown important research value and broad application prospects [2]. Cartography integrates various types of geographic knowledge on a map to make it spatially presented, and the combination of knowledge and map outputs effective information in a more intuitive way, which facilitates information acquisition and instant analysis.

In the construction of digital cities, geospatial knowledge services are needed to provide geospatial knowledge on spatial patterns, distribution differences, spatial and temporal patterns, attribution mechanisms, etc. of terrain and features to support planning, management and decision-making studies [3]. The massive literature resources are one of the important carriers of geospatial knowledge. However, area studied knowledge based on massive literature is still relatively rare, and there exists the phenomenon that area studied knowledge is hard to find and cannot meet users' demand for geospatial knowledge services. The reason for this is that area studied mostly exists implicitly in unstructured form in various kinds of books, journal papers, dissertations, science and technology reports, patent descriptions and other literature resources, and the area studied knowledge contained in the literature carriers can neither be extracted and used systematically by automation nor managed by human in a very convenient way, which means that it is difficult to capture, share and reuse these knowledge, and restricts the This means that such knowledge is difficult to be captured, shared and reused, which restricts the full play of the role of such basic geographic data information. Therefore, there is an urgent need to break through the efficient and accurate extraction technology of area studied, and provide the necessary data foundation and technical support for the production, transformation, sharing and application of area studied knowledge. On the basis of knowledge extraction, the knowledge is exported in the form of maps, and the mapping method has a direct impact on the presentation, organization and use of knowledge, and the exploration of mapping expression will also be a work of this paper.

2 Related Work

2.1 Concept Definition

Area studied is a kind of geospatial knowledge key in the literature, which refers to the area or geographical range selected by the author of a thesis for the research topic in the process of scientific research and suitable for research purposes such as data collection, experimental examination and result verification. It not only contains spatial location information, corresponding to the coordinate interval on the map; it also contains semantic information, indicating the geographic scope of the paper's research, and implicitly contains the information of the place name. Presented in the abstract text as a place name, it is one or more of the extracted place names that express the semantics of the area studied.

Take the abstract of the article "Chemical Analysis of a Dust Event in Beijing", "In this paper, a chemical analysis was carried out by taking a dust event in Beijing in the spring of 2000 as an example. According to the backward trajectory analysis, this weather process originated in southern Mongolia, crossed central and western Inner Mongolia, raised dust and sand by high winds, and then entered Beijing from the northwest, filling the yellow soil..." as an example.

It contains four place names, "Beijing, Mongolia, Inner Mongolia, and Beijing". A place name is a textual identifier that people give to a physical or human geographic entity in a specific spatial location. The area studied is only "Beijing". The area studied is a place name in its own structure, but the area studied is a special type of place name with a unique semantic meaning, which is a subset of the set of place names.

In addition to defining the area studied and its task, the relationship between the area studied and the place name, and the issue of area studied features also need to be discussed.

1) The problem of place name disambiguation. For the place names identified from the text, semantic parsing is needed, i.e., to determine the unique geographical location of the link, to achieve the correspondence between the name and the location of the place name, and to complete the accurate recognition of the place name. For example, in the sentence, "..., and the effluent from Daxing sewage and chemical plants along the historical Liang Shui and North Canal are the main contributors to soil Hg contamination in the area studied...", there are three "Daxing" in the country, belonging to Beijing, Liaoning, and Inner Mongolia, respectively. Therefore, "Daxing" is an ambiguous place name.

2) The problem of feature template for area studied recognition. The features of the area studied are the key signs that distinguish the area studied from other place names in the text with special significance. The scientific and reasonable design of area studied feature template can be used to improve the accuracy of distinguishing area studied from non-area studied. The selection of area studied feature templates needs to be developed by considering the area studied contextual information and location, and combining the area studied itself with the features of the abstract text.

2.2 Related Research

Geospatial knowledge includes place name knowledge, geomorphic knowledge, and declarative knowledge [1]. Three major types of methods have been developed for place name recognition. It has been relatively mature, and the effect in recognition accuracy and efficiency can meet the requirements of the basic work of this paper [4], however, there is data dependence and lack of effective integration in geographic name disambiguation. And the research on the extraction of area studied knowledge as one of the important components of geospatial knowledge is still rare, while the extraction of geospatial knowledge in a broad sense needs to distinguish geospatial knowledge from other contents of the text on the basis of accurate recognition of geographical locations, including three parts of work: semantic disambiguation, geospatial knowledge extraction and mapping.

Place name disambiguation is the process of assigning unique geographical locations to place names. There are two main approaches for place name disambiguation: the data-driven approach and the heuristic rule approach [5]. The data-driven approach mainly contains statistical methods for co-occurrence of place names [6] and machine learning classification methods [7]. The data-driven approach is less used in the field of place name disambiguation because they lack sufficient size of training datasets and do not consider unregistered words. The heuristic rule approach first identifies all place names and their corresponding geographic locations from the text to form a set of candidate locations, and then designs a series of heuristic rules to select a unique geographic location from the set of candidate locations using a priori knowledge and text context. This method is consistent with the idea of understanding spatio-temporal semantics in texts, and has become the mainstream method in the field of place name disambiguation at home and abroad. The heuristic rule approach mainly uses maps [8], external resources [9], semantics [10], cognitive salience [11], a library of conceptual relationships of place names [12] and geographic relatedness [13] to construct rules that are used to inspire contextual semantics, so as to determine the unique linguistic meaning for the referred place names and thus perform place name disambiguation. Although constructing rules from different perspectives can eliminate the ambiguity of geographical names to a certain extent, it has the limitation that each rule is highly targeted and the focus of disambiguation is poorly correlated in scope. If the distribution characteristics of place names can be extracted from multiple sources of data and integrated with multiple methods for disambiguation, the disambiguation effect can be improved more comprehensively.

Area studied knowledge is one of the important components of geospatial knowledge. The methods of geospatial knowledge extraction can be divided into geographic markup-level recognition method, literature-level recognition method and event-level recognition method. Geographic markup-level recognition method infer geographic locations from mentioned place names [14]. Literature-level recognition method are used to infer a single geographic focus from literature. These two conventional geographic knowledge recognition methods have not yet adequately handled multiple geographic knowledge of events mentioned in literature, and at both resolutions, their resolution range is either at the place name level or at the literature level. But literature containing multiple events or an event containing multiple locations, both of which are common in the corpus. Event-level recognition methods are able to detect multiple precise location coordinates of events that may be described in the text. Dewandaru pointed out that the method starts with either place name recognition or geographic markup, followed by an event detection, often called event encoding step, and finally a place name detection step, usually as independent modules [15]. LDC proposed a new implementation of geographic event recognition based on ACE

model, which tightly combines event extraction and place name resolution. And place name resolution are usually handled separately [16]. The model decomposes events into triggers, related entities, parsed (fixed) locations and their semantic role parameters, especially numeric parameters. The model treats geographic markup and event extraction as sequence labeling tasks, so it uses LSTM-CRF neural network, the state-of-the-art sequence labeling model, as a statistical method. Profile, a geographic event detector, identifies and detects event locations and therefore has an event-level resolution range [17]. It uses an SVM classifier to distinguish between focus-located and non-focus-located entities. However, it is based on the rather strong assumption that there is only one main event in the literature and, therefore, only one geographic focus location for that event. This limitation prevents Profile from handling literature containing multiple events or an event containing multiple locations, both of which are common in the corpus, and another dataset confirms this situation as a common observation [18]. The Mordecai geographic event detector is probably the only one that explicitly defines the concept of an event and enforces the event (possibly several) with its location linked geographic event detectors [19]. It defines the geographic event paradigm, where literature can consist of multiple events, each of which can have multiple locations. However, although Mordecai represents a model of events and a way to geographically locate them, it does not model semantic actors and their parameters. Therefore, the ability to detect events depends only on the features used by the model, i.e., lexical (POS) labeling, pre-training, dependency on tags, and directed distance of words from tags [20]. It is effective on the political dataset trained by Mordecai, but may not be sufficient for a broader domain. Since Mordecai is not open source it cannot be compared for validation, but it does not use a place name disambiguation algorithm, which makes it vulnerable to place name ambiguity.

The interest of the field of geography in cyberspace mapping originated from the concern of space and distance [21-22]. Dodge was one of the early researchers who studied cyberspace mapping more systematically and consistently [23], especially his cyberspace atlas co-authored with Kitchen, which very comprehensively covers conventional cyber infrastructure maps and information space maps produced using abstract mapping methods. The objective and mathematical characteristics of cartography in the digital era [24], Zhilin Li proposed a "scale-driven, spatially-first" objective synthesis paradigm based on the laws of nature and a corresponding algorithm based on raster algebra [25]. P. Raposo evaluated this scale driven theory based on the "laws of nature" as "really suitable for map mapping" [26]. Jiang et al. proposed that a cyberspace map is a map that visualizes many aspects of cyberspace, such as the physical location of cyberspace and the status of traffic [27]. Yufen Chen noticed cyberspace mapping earlier and wrote an article about it [28]. Jun Gao focused on virtual reality technology for cyberspace mapping [29]. In recent years, Chun-Dong Gao et al [30] sorted out the current situation related to cyberspace mapping at home and abroad from the perspective of cyberspace geography. Tinghua Ai repeatedly proposed that cyberspace mapping is one of the emerging directions in the development of cartography [31]. Geospatial knowledge, as unstructured text data, is mainly displayed in the form of knowledge graphs, and the cartographic output lacks spatial representation, so this paper makes a new exploration on the cyberspace mapping of geospatial knowledge.

To address the above problems, this paper proposes a feature-template matching algorithm based on feature templates for accurate recognition of area studied and production of thematic maps from the abstract texts of academic papers. The main solution idea is to first extract all the place names in the text, disambiguate the ambiguous ones, determine the features of the area studied by comprehensive analysis, select a classifier with stronger generalization and better adaptation, and then construct a classifier to distinguish the area studied from the non-area studied by using the feature values of the place names as input, so as to extract the area studied of the text. The extracted area studied and other literature knowledge are used for thematic map design. This study aims to provide methodological support for the extraction of area studied of abstract texts and to provide a new perspective for research related to the extraction and use of geospatial knowledge, thus improving the convenience and utilization of geospatial knowledge applications.

3 Methods

Faced with the problem that area studied exist in unstructured form and are difficult to be extracted and utilized efficiently, the core of this paper is how to automatically extract area studied in literature with high accuracy. In this paper, we integrate a feature mining algorithm based on feature template matching and a classification algo-

rithm based on random forest. The model first preprocesses the text and realizes the extraction of place names in literature abstracts through common third-party tools, including linking the corresponding geographic locations for place names based on an improved heuristic disambiguation method to eliminate the ambiguous expressions of place names; then, on the basis of accurate recognition of place names, the feature templates of area studied are constructed to significantly distinguish their differences from other contents in the text, and the random forest classification algorithm is combined to extract. Then, we construct a knowledge graph and finally output a thematic map. This paper contains three parts: disambiguation, area studied recognition, and mapping. The model structure is shown in Fig. 1.

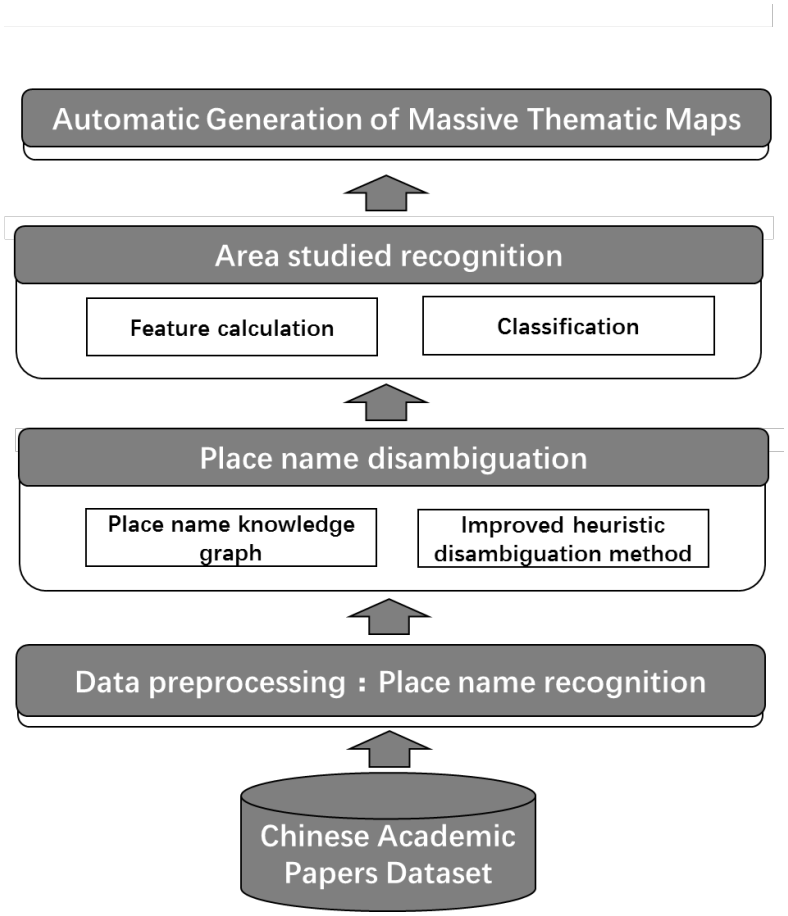

Fig. 1. Structure of the paper

3.1 Place Name Disambiguation

Although the place names have been extracted in advance by the tool, the ambiguity caused by the homonymy of place names can reduce the usefulness of the extraction results under the current situation of place name naming. In addition, the affiliation between the names is a factor to be considered in the construction of the area studied features. Therefore, this paper proposes an improved heuristic disambiguation method based on the place name knowledge graph. The entity type of the place name knowledge graph constructed in this paper is place name, and the place name disambiguation algorithm is shown in Fig. 2.

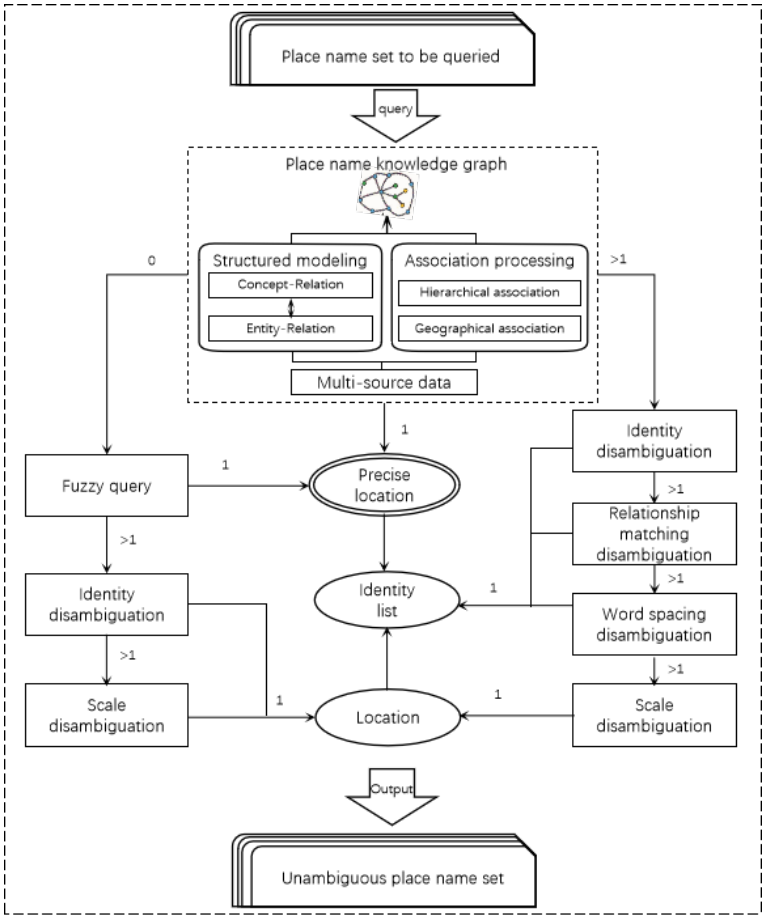

Fig. 2. Geographical name disambiguation

For the identified set of place names, matching retrieval is performed in the place name knowledge graph. In this paper, a 4-level place name knowledge graph is used, which is composed of 4 levels: provincial, city, county, township (town), and becomes the knowledge source for place name disambiguation. If a place name can be precisely located in the upper, lower and the same level by the initial query of the graph, there is no need to disambiguate it. Otherwise, there are two cases. (1) If the meaning cannot be uniquely determined, the name is considered as ambiguous. The disambiguation steps are as follows. Firstly, use the semantics to eliminate the irregularities caused by the simplified expressions. If the full name of a place name can be found in the query, the full name is used to disambiguate. If more than one full name of the place name is retrieved, the ambiguity is eliminated with the aid of the place names appearing in the abstract text. The place names disambiguated in the previous step are used as markers. Disambiguation is performed sequentially using the identified place name, the superior or subordinate place name of an ambiguous place name, and the closest neighboring place name within the threshold range. Assume the set of uniquely identified place names is U, and the set of ambiguous places is A. For each identifier place name $u_i \in U$, search for its superior and subordinate place names in the place name database, denoted as $H(u_i)$. During the disambiguation process, find the intersection of the place names in A with $H(u_i)$, formalized as:

$$C = A \bigcap (\bigcup_{u_i \in U} H(u_i)). \tag{1}$$

Then the number of place names successfully disambiguated this time, N is the number of elements in the intersection C:

$$N = |C|. \tag{2}$$

If the ambiguity has not been eliminated, the scale disambiguation is used, that is, the full name of the highest administrative level among multiple full names is chosen to eliminate the ambiguity. For each $a_i \in A$, the set of place names that can be linked to it in the standard place name database is $M(a_i) = \{m_{i1}, m_{i2}, \ldots, m_{ij}\}$, where each m_{ij} represents the j candidate matching place name for the i place name waiting for disambiguation. Each place name m_{ij} has an associated scale value $S(m_{ij})$, representing the administrative level of the place name. For each ambiguous place name a_i, select the place name m_{ij} with the largest scale as the result of disambiguation r_i:

$$r_i = \arg\max_{m_{ij} \in M(a_i)} S(m_{ij}). \tag{3}$$

(2) If the place name cannot be retrieved, it is judged as an ambiguous place name. The disambiguation steps are as follows. Fuzzy query in the place name database and retrieve all the full names containing it. If there is a unique result, the ambiguity will be eliminated. Otherwise, the ambiguity will be eliminated by identifying the place name. If there is still no unique location, the scale will be used for disambiguation. The algorithm flow is as follows.

The place names in the knowledge graph constitute the standard place name set P. For any paper $A_i \subseteq Doc$, traverse the set Li of place name it contains, and for each place name in Li, process as follows.

Step1: Firstly, $L_{i,j}$ is retrieved from P. For the unique place name of $L_{i,j} \subseteq$ P, $L_{i,j}$ is removed from Li and added to the unambiguous place name set Ti.

Step2: When $L_{i,j} \subsetneq$ P, sequentially using semantic, marking, and scale disambiguation methods until the geographic location is uniquely determined, it will be removed from Li and added to the disambiguation set Ti.

Step3: When $L_{i,j} \subseteq$ P and there are multiple matches, the recognition, relation matching, word spacing, and scale disambiguation methods are used sequentially until the geographic location is uniquely determined, then it is removed from Li and added to the unambiguous set Ti.

3.2 Area Studied Recognition

The feature template designed for the characteristics of the area studied will show a significant difference between the feature values on the area studied and the non-area studied, and through this difference the high-precision and high-efficiency distinction between the area studied and the non-area studied is achieved in the classifier. Therefore, it is necessary to construct the feature set in 3.3 for each word in the above extracted place name set, and classify the place names in the place name set into two categories according to the feature difference: area studied place names and non-area studied place names, to complete the extraction of area studied (Fig. 3).

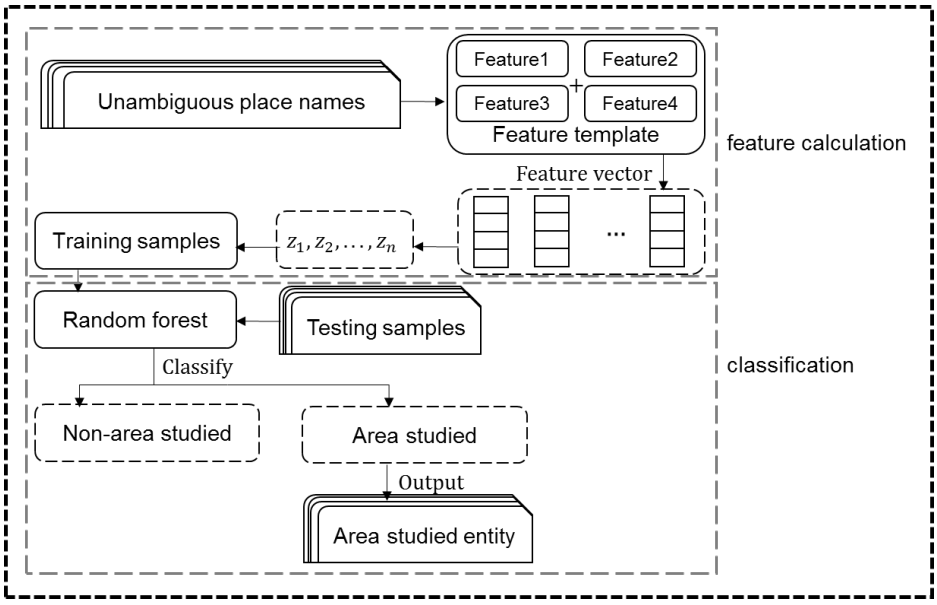

Fig. 3. Area studied recognition

Since this study targets the recognition of area studied in the abstract texts of academic papers, and there is no regularity in the composition of area studied. It is difficult to achieve satisfactory results if we rely only on the structure of the area studied itself. On the basis of accurate and meaningful recognition of place names, the area studied is extracted from the abstract text, from which some features related to the area studied need to be constructed to improve the final extraction effect. Therefore, it is crucial to select the appropriate features.

(1) Title association feature

The title of academic papers are generally concise and reflect the core content of the text, so the elements contained in the title, such as the research topic and the area studied, are bound to have different correlation degree to the core content. Then the sentences containing these elements in the text must also have a strong correlation with the core content, and the place name in the sentence with the greatest correlation is more likely to be the area studied in the paper. If the abstract contains a place name that appears in the title, especially the subtitle, the probability that the place name is the area studied will be greater than the probability that the place name is not the area studied. If the title does not include a place name, but a sentence in the abstract with high similarity to the title contains a place name, the probability that the place name is an area studied will also increase. Therefore, correlation degree between the sentence and the title is determined as one of the features to extract the area studied.

For anyone paper $A_i \subseteq Doc$, the set of titles is Ti, the set of abstract sentences is Si, and the set of sentences containing place names is a subset Si', that is $Si' = \{s | s \in Si \ \wedge s_{i,j}$ including place name$\}$. In this paper, we discriminate A_i with no place name in the title and discriminate the title similarity for each sentence in its Si', and write $c(s_{i,j}) \subseteq C$ for the title similar sentence and obtain the modified Ti′ .The title association feature of $l_{i,j}$ is noted as $\mathrm{p}(l_{i,j})$. The calculation formula is shown in Formula 4.

$$\begin{cases} \mathrm{p}(l_{i,j}) = 0, & l_{i,j} \in Si' \wedge l_{i,j} \notin Ti' \\ \mathrm{p}(l_{i,j}) = 1, & l_{i,j} \in Ti' \end{cases}. \tag{4}$$

The input of the feature algorithm is abstract, and the output is the title association feature values of all the place names in the abstract, by first discriminating papers without place name in the title and obtaining title similarity sentences $c(A_i)$, followed by $c(A_i)$ to replace the title Ti, corresponding to that paper to obtain the updated set of titles Ti′, final judging $l_{i,j}$ whether it appears in Ti′. If yes, then the feature value is 1, otherwise, the feature value is 0. In particular, for $l_{i,j}$ with affiliation that appears simultaneously in Ti′, the one with the smallest administrative division rank is taken to have an eigenvalue of 1, and the rest is 0.

(2) Location feature

In the abstract text of academic papers, it is customary to express the central idea of the article in the first two sentences of the paper or to make a summary of the last sentence. Therefore, if a place name appears in the first two sentences and the last sentence of the text, the probability that the place name is the area studied is greater than the probability that the place name appears in other sentences.

In this paper, we first define the distance for the sentence as follows. An article Ai is sequentially split into a set $Si = \{s_{i1}, s_{i2}, \ldots, s_{i,n}\}$ of sentences, which does not contain the titles ti. The distance of $s_{i,j}$ is its distance to the title, that is, its value is j. In this paper, the formula for calculating this feature is designed as showed in Formula 5.

$$\mathrm{d}(s_{i,j}) = \frac{j}{|s_i|}. \tag{5}$$

In the formula, j indicates sentence distance, $|S_i|$, is the set size of S_i, that is, the total number of sentences. The sentence distances are normalized by this calculation. For the first two terms s_{i1}, s_{i2}, and the last term $s_{i,n}$ of the set Si with n elements. That is, the location feature value of the place name appearing in sentences with j values of 1,2 and n is recorded as 1, otherwise, it is recorded as 0.

(3) Time association feature

Research time and research location often appear together in the abstract text of academic papers. The granularity of the research time varies from article to article, but the research time must contain different time words, such as a year. Therefore, sentences in the abstract that contain a time word and contain a place name have an in-

creased probability that the place name is the area studied. In the case of sentences containing time and containing multiple place name, the closer the place name is the time word the higher the probability that it is the area studied.

For anyone article $A_i \subseteq Doc$, the set of sentences is *Si*, from which a subset *Si′* of the set of sentences containing the place name and containing the time is extracted, that is $Si' = \{s|s \in Si \ \wedge s_{i,j}$ including place name and time$\}$. The formula for calculating the temporal association feature designed in this paper is shown in Formula 6.

$$\begin{cases} Y(l_{i,j}) = \min \left| \dfrac{L(l_{i,j}) - L(y_{i,p})}{|Si'|} \right|, \ l_{i,j} \in Si' \\ Y(l_{i,j}) = -1, \ l_{i,j} \in Si \wedge l_{i,j} \notin Si' \end{cases}. \tag{6}$$

In the formula, $L(l_{i,j})$ denotes the location of the j-th place name of the i-th article, $L(y_{i,p})$ denotes the location of the p-th time word of the i-th article, $|S_i'|$ denotes the length of the i-th abstract text, the time association feature of $l_{i,j}$ is denoted as $Y(l_{i,j})$.

(4) Trigger word feature

There are some fixed expressions in the sentence structure where the area studied is located. For example, the words "analyze", "study", "explore" and other words often appear before the area studied， and the words "located ", "for example", "for the scope of the investigation", "for the research object" and other words often appear after the area studied. These words are called trigger words, and the probability of a place name as an area studied increases when a trigger word appears in a sentence containing the place name.

For an article $A_i \subseteq Doc$, the set of sentences is *Si*, from which a subset *Si″* of the set of sentences containing the place name and containing the trigger words is extracted, that is $Si'' = \{s|s \in Si \ \wedge s_{i,j}$ including place name and trigger words$\}$. The formula for calculating the trigger word feature designed in this paper is shown in Formula 7.

$$\begin{cases} T(l_{i,j}) = 1, \ l_{i,j} \in Si'' \\ T(l_{i,j}) = 0, \ l_{i,j} \in Si \wedge l_{i,j} \notin Si'' \end{cases}. \tag{7}$$

In the formula, $l_{i,j}$ denotes the location of the *j*-th place name of the *i*-th article, T($l_{i,j}$) indicates the trigger word feature value of the place name.

Therefore, the templates developed for the area studied by making full use of the abstract text content and integrating the title association feature, location feature, time association feature, and trigger word feature is shown in Formula 8.

$$w_i[t_1, t_2, t_3, t_4] \rightarrow y_i. \tag{8}$$

Among them, the value of t_j (j = 1, 2, 3, 4) denotes the value of the j-th feature selected by the i-th place name entity, t_1, t_2, t_3, t_4 in order of title association feature, location, year association feature, and trigger word feature, y_i denotes the i-th place name entity and contains two values 1 and 0, indicating the area studied and non-area studied respectively.

3.3 Automatic Generation of Massive Thematic Maps

The massive literature contains two categories: academic journal papers and master's degree theses. In the literature, 14 disciplinary categories such as science, engineering, agriculture, medicine and literature are covered, and the major categories are subdivided into ponderous research directions. Among them, this paper focuses on academic journal papers in the field of surveying, mapping and geographic information related to this specialty to carry out research related to the area studied. The carrier of the area studied is the name of the place. In the actual research, the field of its research does not often appear directly in the paper, but the keywords express the topic of the paper's research and indirectly reflect the field information.

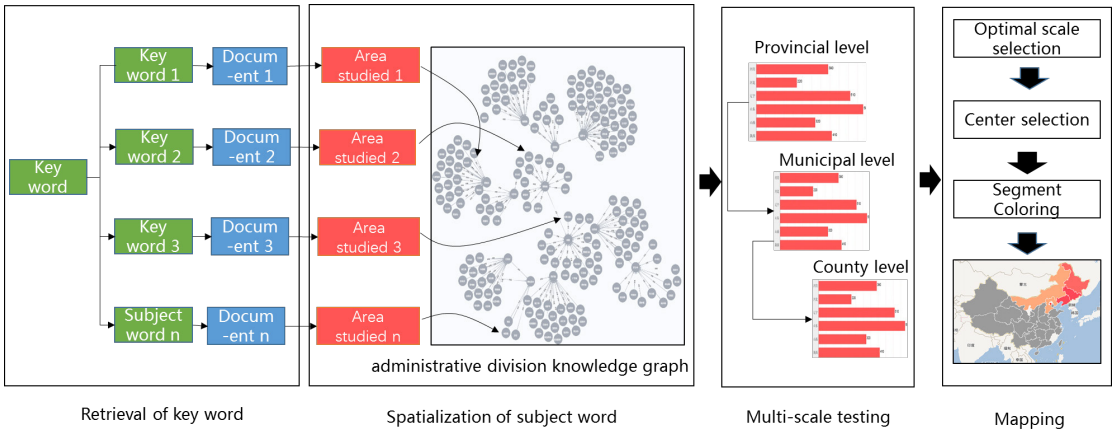

Fig. 4. Thematic map production process

The production process of literature knowledge thematic map is shown in Fig. 4. The topic of the study is first quickly determined by the keywords of the literature. The research topic is obtained by clustering multiple semantically similar keywords in the cluster center. Since the data stock of research literature is large and constantly updated in real time, the clustering method uses the leader-follower incremental clustering strategy [32]. The specific process is as follows: in the clustering process, the set of centers C of existing clusters is kept, C = {c1, c2..., c_j}, where c denotes the cluster center obtained by combining the keywords contained in its class clusters, and when a new keyword Ii is clustered, the spatial, temporal, and semantic similarity measures are used to calculate the spatial and temporal similarity between the new keyword Ii and the center c of each class cluster [33]. The calculation of spatial similarity is conducted using the cosine similarity method. Suppose I_i and c represent two information items, with their corresponding sets of geographic focal points being $G_{set}(I_i)$ and $G_{set}(c)$, respectively. By integrating their associated quantitative values of geographic focal points, we obtain their combined geographic location vectors $V(I_i)$ and $V(c)$. Then, the spatial similarity between I_i and c can be calculated to obtain $L\_sim(I_i, c)$.

$$L\_sim(I_i,c) = \frac{\overrightarrow{v(I_i)} \cdot \overrightarrow{v(c)}}{\left\|\overrightarrow{v(I_i)}\right\| \cdot \left\|\overrightarrow{v(c)}\right\|}. \tag{9}$$

Since all values in the vector are positive, the range of $L\_sim(I_i, c)$ is [0,1], with values closer to 1 indicating a higher degree of similarity. The method for calculating temporal similarity is quantified using a Gaussian function, with the following formula:

$$T\_sim(I_i,c) = e^{-\frac{(t_i - t_j)^2}{2\sigma^2}}. \tag{10}$$

In this formula, t_i and t_j represent the contents within the temporal context T of location-associated information I_i and c, respectively, and a is the standard deviation of the Gaussian function. The event's western similarity value ranges from [0,1], where the larger the difference in time, the closer the metric result approaches 0; the smaller the difference, the closer the metric result approaches 1. The method for calculating semantic similarity also uses the cosine similarity calculation method, with the following formula:

$$S\_sim(I_i,c) = \frac{\sum_{k=1}^{n} Vec(I_i)_k \cdot Vec(c)_k}{\sqrt{\sum_{k=1}^{n} (Vec(I_i)_k)^2} \cdot \sqrt{\sum_{k=1}^{n} (Vec(c)_k)^2}}. \tag{11}$$

In which *Vec* represents the word vector of the information item. The range of semantic similarity is [-1,1],

with larger values indicating higher similarity. If the spatio-temporal semantic similarity between the new information item I_i and the existing class cluster c, $Total\_sim(I_i, c)$ is greater than a certain threshold (0.7 in the study), then I_i is included in the class cluster c. The calculation formula is as in Formula 12.

$$Total\_sim(I_i, c) = L\_sim(I_i, c) \cdot T\_sim(I_i, c) \cdot S\_sim(I_i, c). \quad (12)$$

Where, $L\_sim(I_i, c)$, $T\_sim(I_i, c)$, and $S\_sim(I_i, c)$ denote spatial, temporal and semantic similarity, respectively. After the completion of spatio-temporal semantic similarity clustering for the keywords associated with the topic, the information that has coterminous relationship in spatial location, remains the same or similar in time, and describes the same or closely similar content semantically is integrated and consolidated. On this basis, scientific research literature is searched based on the keywords contained under the use of topics, and then indirectly associated with the research areas extracted from the literature, i.e., the relationship between topics and research areas is constructed through scientific research literature. The geographical names of the study area have been disambiguated and can be directly associated with the corresponding geographical entities in the geographical names database to complete the geographical mapping operation. The area studied processed according to the above ensure the accuracy of knowledge, but the huge amount of information is not conducive to knowledge mining and use, so the area studied are counted separately according to three levels: provincial, municipal, and county. That is, the area studied are counted by topic and by administrative scale. Then, the most suitable administrative division scale is selected for different themes to display the output in priority, and the location of the center point of the result map is also presented dynamically according to the themes. The whole map is color graded, and the results are displayed in a hierarchical color setting based on the frequency of area studied selection. Finally, the graphical decorations such as map name, output time, output unit and legend are drawn according to the knowledge of literature to complete the thematic mapping.

4 Experiment

4.1 Data Pre-processing

Considering that the existing publicly available paper datasets do not contain datasets with area studied annotation, they cannot meet the research requirements of this paper. Therefore, it is necessary to manually annotate the data and construct the dataset used for the experiment. In this paper, we use a dataset of Academic Papers Dataset of Surveying and Mapping Geographic Information provided by China Knowledge Centre for Engineering Sciences and Technology (CKCEST), and each data includes 2 fields of the title and abstract of academic papers. In terms of abstract length, these abstracts are mainly short texts. In this paper, 2241 data were randomly selected and stored in .xls format.

Data pre-processing can improve the efficiency and accuracy of model operation. In this paper, we extract the titles and abstracts of academic papers and pre-process them with word separation and manual annotation. Unlike text in Western flexion languages, text does not have annotations such as spaces between words to indicate word boundaries; therefore, word separation becomes the first task in text processing. In this paper, we use the BiLSTM-CRF to quickly split 2241 data into words and annotate the words, but the words are not annotated correctly, and the place name is misconceived and omitted. On this basis, it is necessary to manually annotate the place name and the area studied. The specific labeling method is: 2 people label the data at the same time, if the labels are consistent, then keep them, otherwise discuss to determine the labels. The label of the non-area studied is uniformly labeled with “ns”, the central sentence of the area studied and sentence is labeled with “st”, and the central sentence of the area studied and its sentence is not labeled with “sa”, so as to improve the accuracy of place name recognition. The data preprocessing and the algorithms used in the experiments are shown in Table 1.

Table 1. Algorithm models

Function	Algorithm model
Place name recognition	BiLSTM-CRF
Place name disambiguation	Improved heuristic disambiguation method
Area studied classification	Random Forest

4.2 Evaluation Index

The diversity of recognition methods and the multi-source nature of the input data make it a complex and challenging task to establish more uniform and appropriate evaluation metrics. In order to evaluate the performance of the area studied extraction model, we integrated the mainstream evaluation methods and related review literature and selected the following three metrics to evaluate the performance of the model: Recall, Precision, and F1 value. The formulas are as following:

$$\text{Recall}=\text{TA}/(\text{TA}+\text{FB}). \quad (11)$$

$$\text{Precision} =\text{TA}/(\text{TA}+\text{FA}). \quad (12)$$

$$\text{F1}= 2* \text{Precision} * \text{Recall} /(\text{Precision} + \text{Recall}). \quad (13)$$

Among them, TA indicates that the area studied is correctly identified, FA indicates that the non-area studied is identified as an area studied, FB indicates that the area studied is identified as a non-area studied, and TB indicates that the non-area studied is correctly identified.

4.3 Experimental Results

The experimental environment is CPU i9-9900K, graphics card NVIDA 2080Ti, memory 16G concurrent server, and training using the TensorFlow framework. The parameters of model were tuned in this framework, and we determined the most appropriate settings for each parameter after repeated testing. When the learning rate of the dataset of the local name recognition module is 0.001, the batch size is 64, the epoch is 30, and the number of hidden layers is set to 300, the model achieves the relatively best overall performance. In the random forest module, the model performs best when the most critical parameter, the number of trees in the tree structure, is set to 100.

Area Studied Feature Recognition. To verify the relevance of the four features selected by the constructed feature template to the area studied, we do the distribution statistics of the title association feature, location feature, time association feature, and trigger word feature on the area studied place name and non-area studied place name respectively, and their association degree is shown in Fig. 5.

Fig. 5 shows the distribution of the title association feature for the place name entities. Among all the place name entities with this feature, 91% of the area studied and 9% of non-area studied are associated with it, indicating high relevance of the area studied of the title association feature. Among the place name entities in a concentration of locations regulated by the location feature, 68% were area studied; while non-area studied accounted for the majority of place name entities without the location feature, reaching 65%. It indicates that the area studied has a positive correlation with the location feature. Among the place name entities with a correlation with time, 85% of the relative distance between the area studied and time words are less than 0.5, which reflects that research time and research location are closely related. Among the place name entities containing trigger words, 72% are area studied, indicating that area studied have a high positive correlation with trigger word features; while non-area studied of place name entities without trigger word features have comparable distribution with area studied and no significant negative correlation. Thus, it proves that the above four features have a good correlation with the area studied in the abstract text of academic papers.

Table 2 shows an example of a feature template.

Table 2. Example of a feature template

Entity number	w	y	Place name entities
1	[1, 1, -0.993684, 0]	1	Altai
2	[0, 1, -0.762557, 1]	0	Anhui Province
3	[1, 1, -0.858639, 1]	1	Beijing

Note. “w” and “y” are the labels indicating the current place name entity and the entity, respectively.

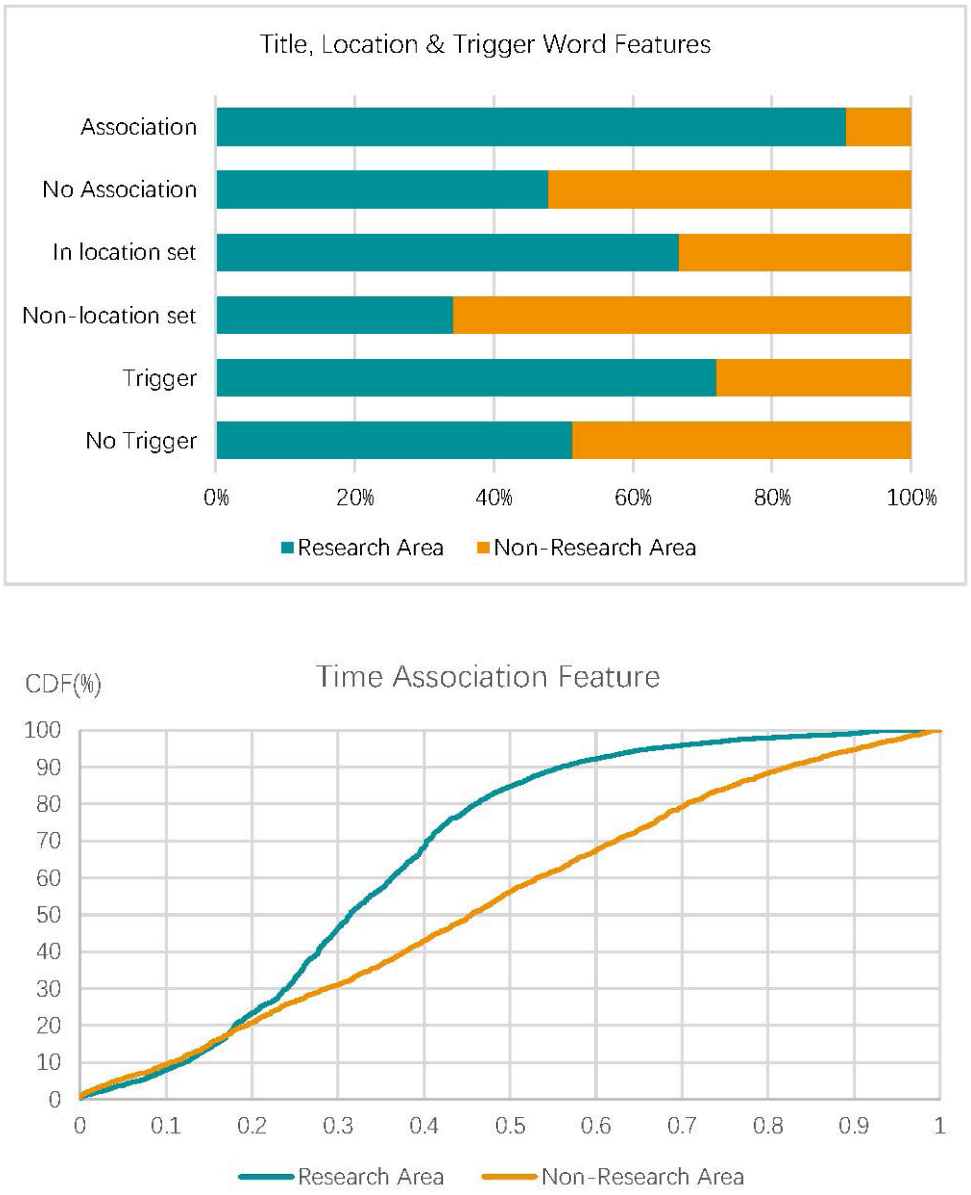

Fig. 5. Area studied feature

Area Studied Extraction. For the extracted place name, feature recognition is performed based on the feature template, and the extraction of the area studied is achieved using a binary classification method. Based on the features of 3.3, the set of feature values for each place name is constructed. The vector of feature values is input into the model, that is, if the place name is an area studied place name, its label is 1, otherwise, the label is 0.

To solve the binary classification problem of place name, this paper compares the test results of the eight most commonly used classification models, Gaussian Bayes, Bernoulli Bayes, logistic regression, support vector machine, gradient boosting decision tree, K-neighborhood classification, decision tree, and random forest, on the dataset, as shown in Table 3. With respect to the performance of the F1 metric, the F1 values of random forest and decision tree are 96% and 94%, respectively; K-neighborhood, gradient boosting decision tree, and support vector machine reach more than 70%, and Gaussian Bayes, Bernoulli Bayes, and logistic regression classifiers are between 60% and 70% effective. In terms of accuracy metrics, the random forest has the highest 97%, the decision tree is the second-highest reaching 93%, Bernoulli Bayes, K-neighborhood, and gradient boosting decision tree have accuracy above 75%, while Gaussian Bayes, logistic regression, and support vector machine have low-

er accuracy. In addition, Random Forest achieves the best recall of 96%. The best performance of random forest is seen from the performance of the eight models in the evaluation metrics, which verifies that random forest is more suitable for the task of area studied extraction in the abstract text of academic papers.

Table 3. Area studied extraction algorithm comparison

Algorithm	Precision (%)	Recall (%)	F1-score (%)
GaussianNB (NaiveBayes)	69	68	69
BernoulliNB	76	65	67
LogisticRegression	70	68	69
SVM	72	72	72
GradientBoostingClassifier	77	76	76
KNeighborsClassifier	79	77	78
DecisionTreeClassifier	93	96	94
RandomForestClassifier	**97**	**96**	**96**

Mapping Results. Spatial knowledge is difficult to describe spatio-temporal distribution and spatio-temporal behavior laws, and visual analysis based on map can quickly and accurately represent the spatial distribution characteristics of spatio-temporal information mapping elements. The interactive visual analysis view of spatial knowledge graph refers to the geospatial-temporal distribution view. The required spatial knowledge is retrieved and cartographically output to form a thematic map.

Fig. 6 shows the regional thematic map, the top three panels show the natural geographic partitions, and the geospatial-temporal distribution of Northeast, Northwest and North China from left to right, where different colors such as red, orange and yellow represent the frequency of occurrence in the paper, and the darker the color, the higher the frequency. In addition to the thematic map of physical geographic divisions, all knowledge of regions, features and place names involved in the literature can be presented in the form of thematic maps. The three maps below, from left to right, are the regional thematic maps of the Yellow River Basin, Beijing-Tianjin-Hebei, and the Yunnan-Guizhou Plateau, which were analyzed and studied or mentioned in the thesis research. The color indicates the extent of the regions, and the gray indicates the areas in the thesis that are relevant to these regions, and this relevance is either geographically relevant or research relevant.

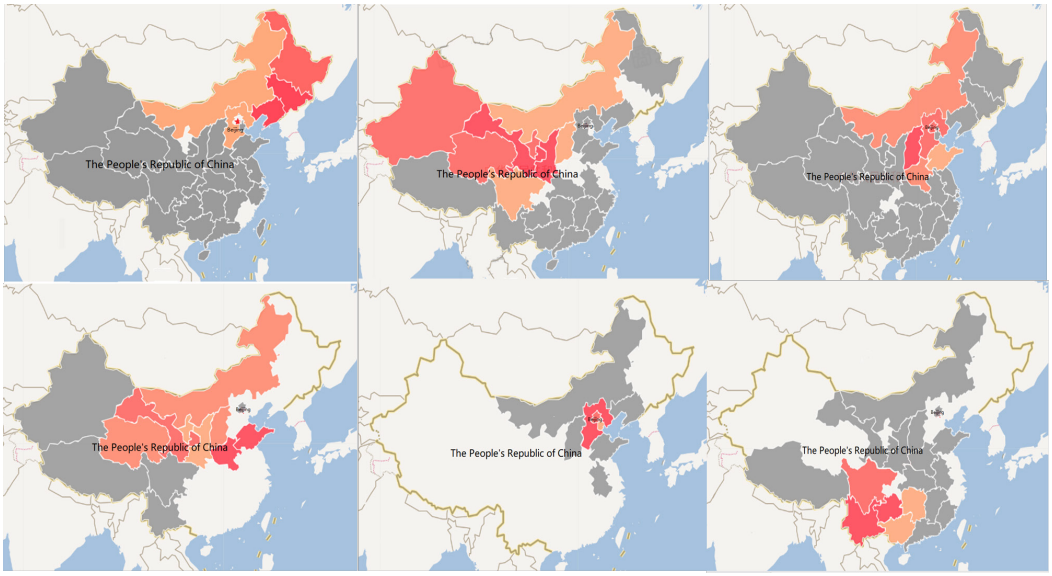

Fig. 6. Regional thematic map

The topics of the papers to be studied are extracted to form a thematic map, and Fig. 7 is color-coded in layers, with lighter to darker colors indicating a gradual increase in research heat. The red color indicates the hot spots of research. The thematic maps of haze, desertification, and soil erosion are shown from left to right. Under the theme of haze, Beijing is the most selected area studied among the published papers, which indicates that the seriousness of the haze problem in Beijing is worthy of attention and more studies have been conducted. And the map clearly shows that there are more studies on haze in the provinces surrounding Beijing southward, and all the eastern regions are involved. The attention of the research cold spots needs to be improved, and the studies of provinces with similar causes of haze formation are informative. This provides some ideas and references for researchers in the development of their papers and the selection of experimental areas.

Fig. 7. Thematic topic map

In addition to mining the subject matter of the dissertation and knowledge of the area studied to aid in providing reference for the dissertation research, knowledge of various types of knowledge in the literature, such as cultural, historical, and celebrity knowledge, is shown in Fig. 8 from left to right in the order of Youzhou, Zhongyuan, and Li Bai. Youzhou is taken from the ancient name of the place, and in the study of it, the geographical range is detailed, as well as the changes in the geographical range with historical changes, hence the two colors shown on the figure.

Fig. 8. Various thematic maps

5 Conclusion and Discussion

In this paper, we extracted the area studied from the abstract text of academic papers and used the literature knowledge to generate a thematic map. It consists of three modules: a feature calculation module based on feature templates, a binary classification module based on random forests, and a thematic map drawing based on spatial knowledge. Based on the feature template developed from the area studied features, the feature value of each disambiguated place name is calculated as the input value of the classification algorithm, and random forest is used as the classification algorithm to classify the place names into area studied and non-area studied, and the accuracy of extracting area studied reaches 97%, which demonstrates the excellent performance of the algorithm

for area studied extraction in the abstract text of academic papers. The thematic map design with the extracted area studied and other literature knowledge transforms the structured textual knowledge into geospatial display with intuitive visualization and rich thematic maps.

In this paper, we achieve good performance in extracting area studied on the abstract text of academic papers, but in the module of identifying geographical names, for the perspective of constraint normativity, only administrative divisional geographical names are identified without physical geographic entities, which makes the minimum scale of some area studied positioned to the provincial administrative units to which they belong, and also limits the accuracy of disambiguation of geographical names. Based on the research in this paper, the subsequent research expands the scope of place name recognition to include administrative division place names and natural geographic entity place names, refines and perfects the place names, and further narrows the scope of area studied positioning in order to extract the minimum scale area studied more accurately.

6 Acknowledgement

This research was supported by Scientific Research Projects of Beijing Municipal Education Commission—General Projects of Science and Technology Program (Surface Projects) (KM202110016003); The National Natural Science Foundation (NSFC) of China (Key Project #41930650).

References

[1] X.-M. Ma, Z.-G. Xuan, J.-N. Wu, A Topic Finding Method for Scientific and Technical Literature, in: Proc. 2010 International Conference on E-Product E-Service and E-Entertainment. IEEE, 2010.
[2] Y.-J. Hu, Geo-Text Data and Data-Driven Geospatial Semantics, Geography Compass 12(11)(2018) e12404.
[3] J. Chen, W.-Z. Liu, H. Wu, Z.-L. Li, Y. Zhao, L. Zhang, Basic issues and research agenda of geospatial knowledge service, Geomatics and Information Science of Wuhan University 44(1)(2019) 38-47.
[4] M. Marrero, J. Urbano, S. Sánchez-Cuadrado, J. Morato, J.M. Gómez-Berbís, Named entity recognition: Fallacies, challenges and opportunities, Computer Standards & Interfaces 35(5)(2013) 482-489.
[5] M. Gritta, M.-T. Pilehvar, N. Collier, A pragmatic guide to geoparsing evaluation, Language Resources and Evaluation 54(3)(2020) 683-712.
[6] W.-A. Gale, K.-W. Church, D. Yarowsky, One sense per discourse, in: Proc. HLT'91: Proceedings of the workshop on Speech and Natural Language, 1992.
[7] B. Martins, H. Manguinhas, J. Borbinha, W. Vaca, A geo-temporal information extraction service for processing descriptive metadata in digital libraries, E Perimetron 4(1)(2009) 25-37.
[8] E. Aldana-Bobadilla, A. Molina-Villegas, I. Lopez-Arevalo, S. Reyes-Palacios, V. Muñiz-Sanchez, J. Arreola-Trapala, Adaptive Geoparsing Method for Toponym Recognition and Resolution in Unstructured Text, Remote Sensing 12(18)(2020) 3041.
[9] D. Buscaldi, P. Rosso, A conceptual density-based approach for the disambiguation of toponyms, International Journal of Geographical Information Science 22(2008) 301-313.
[10] Y. Zhang, X. Wang, M. Chen, Y. Liu, A semantics-based method for extracting geographic scopes of texts, Chinese High technology letters 22(2)(2012) 165-170.
[11] X.-R. Tang, X.-H. Chen, X.-Y. Zhang, Research on Toponym resolution in Chinese text, Geomatics and Information Science of Wuhan University 35(8)(2010) 930-935.
[12] M. Karimzadeh, S. Pezanowski, A. MacEachren, J.O. Wallgrün, GeoTxt: A scalable geoparsing system for unstructured text geolocation, Transactions in GIS 23(1)(2019) 118-136.
[13] X.-G. Wang, R.-J. Zhang, Y. Zhang, Toponym resolution based on Geo-relevance and D-S theory, Acta Scientiarum Naturalium Universitatis Pekinensis 53(2)(2017) 344 -352.
[14] R.-S. Purves, P. Clough, C.-B. Jones, A. Arampatzis, B. Bucher, D. Finch, G. Fu, H. Joho, A.-K. Syed, S. Vaid, B. Yang, The design and implementation of SPIRIT: A spatially aware search engine for information retrieval on the Internet, International Journal of Geographical Information Science 21(7)(2007) 717-745.
[15] A. Dewandaru, D.-H. Widyantoro, S. Akbar, Event Geoparser with Pseudo-Location Entity Identification and Numerical Argument Extraction Implementation and Evaluation in Indonesian News Domain, ISPRS International Journal of Geo-Information 9(12)(2020) 712.
[16] Linguistic Data Consortium, ACE (Automatic Content Extraction) English Annotation Guidelines for Events V5.4.3 Linguistic Data Consortium. <https://www.ldc.upenn.edu/collaborations/past-projects/ace> , 2005 (accessed 08.08.2020).
[17] M.-B. Imani, S. Chandra, S. Ma, L. Khan, B. Thuraisingham, Focus location extraction from political news reports with

bias correction, in: Proc. 2017 IEEE International Conference on Big Data (Big Data), 2017.
[18] S.-J. Lee, H. Liu, M.-D. Ward, Lost in Space: Geolocation in Event Data, Political Science Research and Methods 7(4) (2019) 871-888.
[19] A. Halterman, Linking Events and Locations in Political Text, MIT Political Science Department Research Paper No. 2018-21, 2018. http://dx.doi.org/10.2139/ssrn.3267476
[20] A. Halterman, Geolocating Political Events in Text, in: Proc. The Third Workshop on Natural Language Processing and Computational Social Science, 2019.
[21] R. Dale, Book Reviews: Global financial integration: The end of geography. By Richard O'Brien. London: Pinter for Royal Institute of International Affairs, International Affairs 68(3)(1992) 531.
[22] J. Sempsey, Book Review: The death of distance: How the communications revolution will change our lives, Journal of the American Society for Information Science 49(11)(1998) 1041-1042.
[23] M. Dodge, The geographies of cyberspace: A research note, Communication Studies 12(4)(1998) 383-396.
[24] Z.-L. Li, B. Su, From phenomena to essence: envisioning the nature of digital map generalization, The Cartographic Journal 32(1)(1995) 45-47.
[25] Z.-L. Li, Algorithmic foundation of multi-scale spatial representation, CRC Press, London, 2007 (Chapter 1).
[26] P. Raposo, Scale-specific automated line simplification by vertex clustering on a hexagonal tessellation, Cartography and Geographic Information Science 40(5)(2013) 427-443.
[27] B. Jiang, F. Ormeling, Cybermap: The map for cyberspace, The Cartographic Journal 34(2)(1997) 111-116.
[28] B. Jiang, F. Ormeling, Mapping cyberspace: Visualizing, analysing and exploring virtual worlds, The Cartographic Journal 37(2)(2000) 117-122.
[29] J. Gao, Cartographic tetrahedron: explanation of cartography in the digital era, Acta Geodaetica et Cartographica Sinica 33(1)(2004) 6-11.
[30] C.-D. Gao, Q.-Q. Guo, D. Jiang, Z.-B. Wang, C.-L. Fang, M.-M. Hao, The theoretical basis and technical path of cyber-space geography, Acta Geographica Sinica 74(9)(2019) 1709-1722.
[31] T.-H. Ai, Development of cartography driven by big data, Journal of Geomatics 41(2)(2016) 1-7.
[32] R.-O. Duda, P.-E. Hart, D.-G. Stork, Pattern classification, John Wiley & Sons, New Jersey, 2006 (Chapter 2).
[33] H. Zhang, Research on Location-referenced Web Textual Information Extraction and Cartographic Visualization, [Doctoral dissertation] Wuhan: Wuhan University, 2019.