ISBN 978-981-11-3671-9
Proceedings of 2017 the 7th International Workshop on Computer Science and Engineering (WCSE 2017)
Beijing, 25-27 June, 2017, pp. 654-658
doi: 10.18178/wcse.2017.06.114

An Automatic Approach to Build Geographical Knowledge Base from Geographical Data Sources for Earthquake Emergency Response

Qin-Yong Li, Jian-Gong Song, Jiang-Hua Lv [+] and Shi-Long Ma

Beihang University, Beijing, China

Abstract. Geographical knowledge base plays crucial role in earthquake emergency. However, how to automatically build a such knowledge base is still a challenge for practical use. In this research, we propose a novel method to extract geographical knowledge from geographical data sources for earthquake emergency. We first design a flexible hierarchical concept model, then we propose an extracting algorithm to extract knowledge from multiple data sources. To handle the place name disambiguation problem, we suggest a Hausdorff distance based method called MRBH-dist. We represent the extracted knowledge base in RDF format and use it via Protégé. This work has already applied successfully to National Earthquake Society Service Project (NESSP) in China.

Keywords: earthquake emergency management, knowledge base, ontology; place name disambiguation

1. Introduction

Strong earthquakes cause tremendous damages to human-beings life and property. Earthquake Emergency Response (EER) is a significant aspect of alleviating the disaster impacts. Many counties of the world has built knowledge-based system for EER, such as HAZUS [1] and USGS PAGER [2]. It is no doubt that the process of EER is a knowledge/data-intensive task and needs multi-discipline knowledge collaboration. Nevertheless, knowledge is the most crucial resources in the process of EER. Therefore, how to effectively and efficiently build a geographical knowledge base is an essential subject for earthquake emergency response.

In the last few years, with fast development in semantic web, knowledge base and knowledge-based systems have gained growing research interests. Geographical knowledge base is widely used and researched in many domains, such as disaster emergency response, transportation planning and so forth. However, manually building a geographical knowledge base is a difficult and time-consuming task and it is easy to make unobserved mistakes. Recently, researchers have suggested building geographical knowledge base or discovering geographical knowledge through Web resources, such as Web Pages, DBPedia and so on. Disadvantages of these methods have two aspects. On the one hand, they lack the support for ***domain-specific*** geographical knowledge, which is hosted by authorities or other organizations and hardly obtained on the Web. Domain-specific data is one kind of valuable resources for decision making or research and should not be overlooked. For example, active fault data plays a crucial role in estimating damage degree in EER. On the other hand, with the advent of the age of Big Data, many organizations run Spatial Data Infrastructure(SDI) for research and/or social service. Whereas data in SDI is very helpful to support decision when emergency situation occurs or to conduct a cross discipline research project. Therefore, extracting geographical knowledge from these data sources is also a crucial issue for knowledge discovery.

Unfortunately, to date, only few studies are conducted about extracting knowledge from these data sources. This study therefore set out to suggest a method to extract geographical knowledge from these data

[+] Corresponding author. Tel.: + 86 (010) 82316463; fax: +86 (010) 82316463.
E-mail address: jhlv@nlsde.buaa.edu.cn

sources. The proposed approach involves first designing a domain-specific concept model using Protégé for knowledge extracting and secondly utilising the instance and relation extraction algorithm to extract geographical knowledge. A critical issue in upper process, place name disambiguation (PND), is researched in order to enhance the correctness of the extracted knowledge base. We propose a Hausdorff distance based method called MBRH-dist.

The reminder of this work is organized as follows: Section 2 summaries previous work on building geographical knowledge base and place name disambiguation methods; Section 3 formally state definitions and research objectives of this work; We present the concept modelling, MBRH-dist based place name disambiguation algorithm in Section 4. Section 5 describes the application of proposed method in real world application: NESSP in China. Finally, Section 6 summaries the conclusions of this work as well as suggests future works.

2. Related Work

In this section, we highlight related work about extracting geographical knowledge from multiple data sources.Cao C. et.al present a practical method of extracting both names of geographical entities and their relations from the Web, which names is OMKast-Googling [3]. The method is composed of three major phases: 1) design a list of geographical entities, and 2) building a knowledge extractor, and 3) develop several methods for handling problems or errors in the generated knowledge base. Unfortunately, to extract domain-specific knowledge is overlooked by these work.

Ontologies play a key role in knowledge base and knowledge-based system. Fu G.H. et.al reports on how ontologies developed in the EU Semantic Web project SPIRIT are used to support retrieval of documents that are spatially relevant to users' queries [4]. The query expansion techniques presented in this work are based on both a domain and a geographical ontology. In [5], Fu et.al describes the geo-ontology designed for SPIRIT system and points out that similarity checking of datasets is an essential step in the process of knowledge extraction. The validity and effect of the different measures are studied by building a prototype geo-ontology utilising different datasets. Place name disambiguation [6~7] is an important task for improving the accuracy of geographical information retrieval.

3. Problem Formulation and Research Objective

In this section, we define some terms which will used in rest discussions and propose our research goal.

Definition 1. **Geographical data source.** A geographical data source is composed by a collection of Dataset: $GDS = \{dataset_1, dataset_2, \ldots, dataset_n\}$, the size of geographical data source, denoted as $|GDS|$.

Definition 2. Data set. Datasets in geographical data source are defined as a triple structure: $Dataset = (name_{ds}, GeoAttrSet_{ds}, FeatureSet_{ds})$, where: 1) $name_{ds}$ denotes the name of a dataset which identifies a dataset from the others; 2) $GeoAttrSet_{ds}$ represents a collection of geographical attributes, such as Coordinate Reference System (CRS), Unit of Measure (UoM) and so forth; 3) $FeatureSet_{ds} = \{feature_1^{ds}, feature_2^{ds}, \ldots, feature_n^{ds}\}$, is a range of geographical features.

Definition 3. Geographical feature. A geographical feature (feature for short) in dataset represents a geographical (or spatial) object and is defined as a quadri-tuple structure: $Feature = (id_f, AttrSet_f, type_f, footprint_f)$, where: 1) id_f is the identifier of a feature; 2) $AttrSet_f$ denotes a collection of key-value pair; 3) $type_f = \{\text{Point, Line, Polygon}\}$; 4) $footprint_f$ represents the footprint of a feature, which is a ordered range of coordinates. $footprint_f = \langle c_1, c_2, \ldots, c_k \rangle$, $\forall c_i \in footprint_f$, $c_i = \langle x_i, y_i \rangle$, which represents a coordinate on surface of the earth;

Definition 4. Ontology. An ontology is defined as a quadri-tuple structure: $\mathcal{O} = (C, R, \leq_C, \text{rel})$,where: 1) C: a set of concept; 2) R: a set of relations defined by C; 3) $\leq_C$: a concept hierarchy, which defines the is_a relationship among concepts; 4) *rel*: a function which specifics the relation R, rel: $R \rightarrow C_1 \times C_2$ or rel: $R(C_1, C_2)$;

For $\forall c_1, c_2 \in C$, $c_1 \leq_C c_2$, then c_1 is a *sub class* of c_2 and c_2 is called a *super class* of c_1; if $c_1 \leq_C c_2$ and there is no $c_3 \in C$ with $c_1 \leq_C c_3 \leq_C c_2$, then c_1 is a *direct sub class* of c_2, and c_2 is a *direct super class* of c_1. We use $\leq_{C_d}$ to denote this.

Definition 5. Domain and **Range**. For a given relation $r \in R$, we define its domain and range by $\text{dom}(r) = \pi_1(\text{rel}(r))$ and $\text{range}(r) = \pi_2(\text{rel}(r))$, $\forall r \in R, \text{dom}(r) \in C, \text{range}(r) \in C$.

Definition 6. Spatial Relation. $SR = TR \cup DR$, where TR denotes *topology relationship* and DR denotes *direction relationship*. $TR = \{\text{part\_of}, \text{contains}, \text{next\_to}, \text{int ersecs}\}$ and $DR = \{\text{north\_of}, \text{south\_of}, \text{east\_of}, \text{west\_of}\}$.

Definition 7. Knowledge base. A knowledge base is a structure: $KB = (C_{KB}, R_{KB}, I, U)$, where C_{KB} is a set of concepts; R_{KB} is a set of relations; I denotes a set of instances, $\forall i \in I$, i = (*id*, *name*, *concept*), where *id* is a URI and identifies a instance, *name* is the name of a instance and *concept* is the concept of a instance, *concept* $\in C_{KB}$; U is defined as $U : R_{KB} \to P(I \times I)$ called relation instances, or rule bas, where $P(I \times I)$ denotes the power set of $I \times I$.

Research Objective: The goal of this research is that given a collections of *GDS* = $\{g_1, g_2, \cdots, g_n\}$, and a user-designed ontology O, we want to conduct a mapping $\delta : GDS \times O \to KB$, where KB is a geographical knowledge base we extracted from *GDS* according to O, where KB is consistent, in other words, there is no inconsistent knowledge in KB.

4. Proposed Method

4.1. Concept Modelling

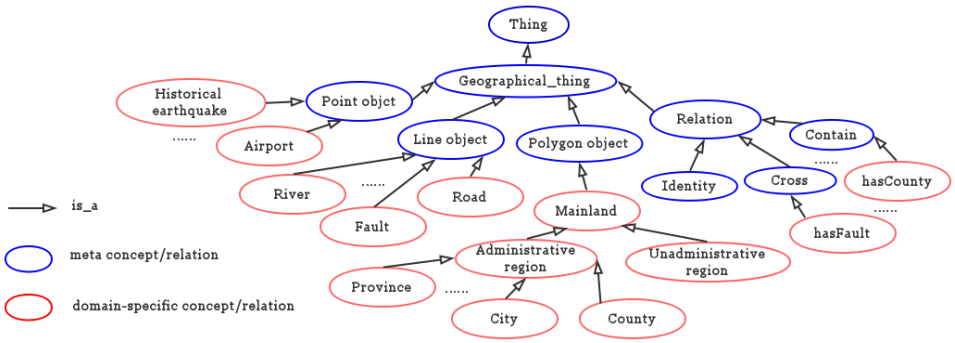

Fig. 1: Concept modelling of specific geographical data sources.

We design a hierarchy structure for our geographical knowledge extraction, as shown in Fig.1. In our designed ontology, concepts mainly divide into two groups: *meta-concept* and *domain-specific concept*, respectively. Meta Concepts, such as Point Object, Line Object and so forth, is used to describe domain-free concepts which could be used in any domain. Domain-Specific Concepts are used by practical situation. Every concept in our ontology is a sub-concept of main concept *Thing* using relation *is_a*. The advantage of designed ontology is its flexibility to represent geographical knowledge entities and relations between entities.

4.2. Extracting Algorithm and Place Name Disambiguation

The proposed geographical knowledge extracting algorithm is described in Algorithm I. The algorithm mainly divided into two parts: for each feature in dataset, we test its name whether in *KB* already. If the feature is not appeared in *KB*, then we add the feature to *KB* else we runs a place name test, called MBRH-dist, on the feature to figure out whether put the feature to *KB* or not.

Algorithm I Extracting new knowledge to *KB*
Input: ***f***: the feature in a data set which to be extracted
t: threshold for footprint similarity computing;

KB : knowledge base
Output: *KB'* : a new knowledge base which is filled by extracted knowledge

```
1: foreach inst ∈ I_KB do
2:     if inst.name = f.name then
3:         if inst.concept = f.concept then
4:             if MBR-HD(inst.fp, f.fp) ≥ t then
5:                 /*existed instance in KB */
6:                 continue;
7:             end
8:             else
9:                 /* identify that two place with same
                       place name, a new feature is extracted */
10:                I_KB ← f.name;
11:            end
12:        end
13:        else
14:            /* identify that two place with same
                   place name, a new feature is extracted */
15:            I_KB ← f.name;
16:        end
17:    end
18:    else
19:        if MBR-HD(inst.fp, f.fp) ≥ t then
20:            /* identify that a place with different place name */
21:            U_KB ← is_a(inst.name, f.name);
22:        end
23:        else
24:            I_KB ← f.name;
25:        end
26:    end
27: end
```

Algorithm I Extracting new knowledge to $\mathcal{KB}$.

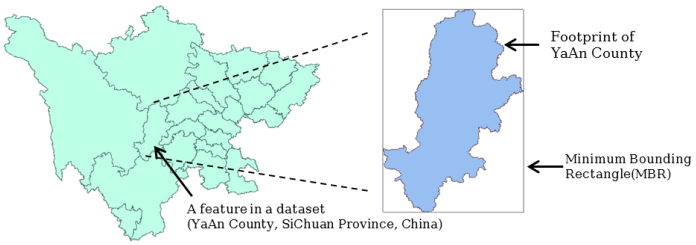

Fig. 2: Feature, its footprint and minimum bounding rectangle.

We solve the place name disambiguation problem using Hasudorff distance. Hausdorff distance (H-dist) is a well-known method to assess distance between tow point sets. Minimum Bounding Rectangle (see Fig. 2) Hausdorff Distance (MBRH-dist) based footprint similarity computing method. According to Hausdorff distance, we suggest a place name disambiguate method MBRH-dist as follows:

Assume that there are two point sets: $A=\{a_1,a_2,\cdots,a_m\}$ and $B=\{b_1,b_2,\cdots,b_n\}$. The Hausdorff distance between *A* and *B* can be represented as follows:

$$H(A,B)=\max\{h(A,B),h(B,A)\} \tag{1}$$

in equation (1), distance functions $h(A,B)$ and $h(B,A)$ are termed as *forward* and *backward* Hausdorff distances of A to B:

$$h(A,B)=\max_{a_i\in A}\{\min_{b_j\in B}\{d(a_i,b_j)\}\} \tag{2}$$

in equation (2), for simplicity, we will take $d(a_i,b_j)$ as the Euclidean Distance between a_i and b_j.

We define similarity between two features using Hasudorff distance as:

$$\text{MBR-HD}(F_1, F_2) = H(\text{MBR}(F_1.fp), \text{MBR}(F_2.fp)) \quad (3)$$

where F_i denotes feature, MBR(•) denotes the Minimum Bounding Rectangle of the feature.

5. Application in NESSP

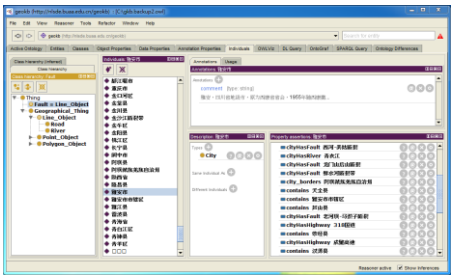

Fig. 3: Extracted geographical knowledge base represented by RDF using Protégé

YaAn county, which center longitude and latitude is (102.1E, 29.0N), is located at west of SiChuan province(West of China). It's square almost 15100 KM^2. YaAn county contains the city part, MingShan county, TianQuan county, ShiMian county and LuShan county. On the east of YaAn county is ChengDu and MeiShan, and on the sourth of YaAn county is LeShan. On the west of YaAn county is Tibetan Autonomous Prefecture of Garzê and ABA Qiang Tibetan Autonomous Prefecture. There is several active faults in YaAn county: XiHe-MeiGu fault, LongMengShan fault...

Fig. 4: Query geographical knowledge from extracted knowledge base about YaAn County(In Chinese).

This research is part of the *National Earthquake Social Service Project* (NESSP). The main motivation of NESSP is to construct the basic infrastructure and IT-Service for earthquake emergency response in China. Data for this study were collected from two sources which are National Geometrics Centre of China (NGCC), which is the data supplier for NESSP and http://www.diva-gis.org/. The extracted geographical knowledge base is represented as a RDF format using Protégé, which is a well-known ontology editor, as shown in Fig.3. We use this geographical knowledge base to query YaAn County information during the process of YaAn Earthquake, which was occurred in April 20, 2013, as shown in Fig.4. Application shows that our method is suitable for practice in real situation.

6. Conclusions and Future Work

In this work, we propose a domain-specific concept model for earthquake emergency response and a knowledge extracting algorithm and a Hausdorff distance based place name disambiguation method MBRH-dist. We show the application of proposed method in NESSP during the earthquake of YaAn.

7. Acknowledgements

Authors thank the anonymous reviewers for their insightful and constructive comments. This research work is co-supported by grants from Chinese National Science Foundation(NO41474013, NO.61300007); Special Fund for Central University Fundamental Research(NO.YWF-15-GJSYS-106, YWF-14-JSJXY-007), Self-Conducted Explore Research Program from State Key Laboratory for Software Developing Environment(NO.SKLSDE-2015ZX-09, SDLSDE-2014-ZX-06).

8. References

[1] http://www.fema.gov/hazus

[2] http://earthquake.usgs.gov/data/pager/

[3] Cao,C, Wang S., Jiang L.: A Practical Approach to Extracting Names of Geographical Entities and Their Relations from the Web. In: *Knowledge Science,Engineering and Management*, KSEM 2014. vol. 8793, pp. 210–221 (2014)

[4] Fu, G., Jones, C.B., Abdelmoty, A.I.: Ontology Based Spatial Query Expansion in Information Retrieval. *Lecture Notes in Computer Science* - ODBASE2005 3761,11466–1482 (2005)

[5] Fu, G., Jones, C., Abdelmoty, A.: Building a Geographical Ontology for Intelligent Spatial Search on the Web. In *Proceedings of IASTED International Conference on Databases and Applications* pp. 167–172 (2005)

[6] Hu Y, Janowicz K, Prasad S. Improving wikipedia-based place name disambiguation in short texts using structured data from DBpedia *The Workshop on Geographic Information Retrieval*. ACM, 2014:8.

[7] Nakatoh T, Yin C, Hirokawa S. Extraction and Disambiguation of Name of Place from Tourism Blogs *First ACIS International Symposium on Software and Network Engineering*. IEEE Computer Society, 2011:73-78.