Information and Knowledge Management (2022)
Clausius Scientific Press, Canada

DOI: 10.23977/infkm.2022.030204
ISSN 2523-5893 Vol. 3 Num. 2

Geographic Information Linking Method Based on Machine Learning

Jun Shu

Wuhan University of Engineering Science, Wuhan, Hubei, 430200, China

Keywords: Machine Learning, Geographic Information, Geographic Features, Linking Methods

Abstract: With the continuous development of Internet geographic information(GI) system, various GI sources have sprung up. However, the representation and storage methods of GI of these GI data sources are different, resulting in significant differences in the accuracy and integrity of GI, which has caused many difficulties for GI integration. Therefore, this paper introduces the importance of geospatial relationship in GI data link, and discusses the GI link method based on machine learning(ML); The link results of support vector machine classification method and k-nearest neighbor classification method in GI data link method are discussed respectively. Both of them have achieved good experimental results, and the differences between them are analyzed in detail. At the same time, the GI data link method in this paper is compared with the geographic context related information link method using graph theory method. The experimental results show that the GI data link method proposed in this paper is obviously better than this method, which further proves the accuracy and effectiveness of the GI data link method proposed in this paper.

1. Introduction

The rapid development of information technology has gathered a large amount of network data. How to quickly and effectively obtain information from massive data is facing great challenges. Not only the characteristics of nodes but also the relationship between nodes should be considered. In location-based social networks, link prediction needs to consider the location characteristics of nodes in addition to the relationship characteristics between nodes. How to effectively use a large number of unlabeled sample data and obtain more information is also a difficult problem. Therefore, this paper studies and analyzes the GI link method based on ML.

Many scholars at home and abroad have studied and analyzed the method of GI link based on ML. Tempelmeier n proposes a new link discovery method osm2kg, which is used to predict the identity links between OSM nodes and geographical entities in a knowledge graph. The core of osm2kg method is a new potential and compact OSM node representation, which captures the similarity of semantic nodes in embedding. Osm2kg uses this potential representation to train the supervised link prediction model and uses the existing links between OSM and the knowledge map for training [1]. Kaya La uses GIS and RS technology. Landsat images are used for supervised classification using maximum likelihood technology. The classified image shows man-made surfaces, agricultural areas, forests, wetland, and water bodies according to the land cover legend of

Corine. ArcGIS (version 10.6) software is used to evaluate the accuracy before and after the correction process. The final overall kappa value is higher than 0.95 within 3 years [2].

In order to improve the flexibility and feasibility of GI data link method, the classification method in ML is used to realize the automation of GI data link. A general GI ontology is constructed to unify the labels of heterogeneous GI data sources, realize the mapping between heterogeneous GI data sources and general GI ontology, eliminate ambiguity, and realize the unification of geographic data representation. Then, analyze the characteristics of the extracted original GI data, extract the geographic feature data in the GI data, and realize the digitization of GI features through the similarity calculation method. At the same time, a multi-feature similarity calculation method combined with geospatial features is proposed. Finally, combining the classification method in ML with the proposed multi-feature similarity calculation method, a multi-feature fusion GI data link method is proposed to realize the automation of GI data link [3-4].

2. GI Linking Method of ML

2.1 Concept and Application of ML

ML, the learning process, includes adaptation to the environment, acquisition of knowledge, accumulation of experience, and self-improvement. How to improve the generalization ability of ML for different learning systems is not only the fundamental problem of ML research, but also the first of many research directions in the field of ML. Figure 1 clearly shows the whole process of ML.

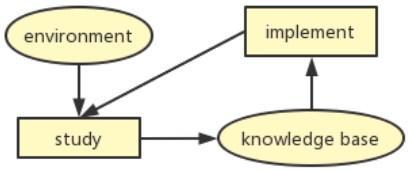

Figure 1: Basic process of ML

Data mining based on ML technology: because ML method is convenient for analyzing and modeling data and obtaining potential information in data, it has become an important method of data mining. Link prediction based on ML: ML technology has also been applied in the field of social network analysis. By mining social network data, we can build a dynamic network model and better design application services in online social networks [5]. Link prediction is a key problem in the field of social network analysis and has been widely used. For example, by predicting the relationship between users or between users and commodities in social networks, commodity recommendation can be carried out in the business field. In the framework of ML, the link prediction problem can be regarded as a simple class II classification problem. The prediction result 1 or 0 indicates that there is or does not exist a link. Using ML technology for link prediction is a relatively new research topic.

2.2 Semantic Mapping

With the rapid development of the semantic web, the GI data stored in the semantic web has increased significantly, but most of the GI on the Internet is still stored in structured databases. Different databases have different structures. How to query and obtain these structured data sources in the environment of the semantic web has become one of the important problems of information

integration.

In fact, the problem of geographic data interoperability between structured data sources and ontology models can be summarized as the problem of mapping between structured data sources and ontology patterns. In the current structured data source, the data source defines the structural relationship and integrity constraints of its attributes in the database, and there are many approximate correlation mappings between the attribute definition of a structured data source and ontology concepts: for example, the entity attributes in the structured data source schema can be mapped to the concept definition in Ontology [6]. Therefore, we can solve the problem of data interoperability between structured data sources and ontology patterns by using the mapping method between them. For the original structured data source, we can extract geospatial data through a mapping model on the basis of geographic ontology to form RDF data related to the semantic web.

In the research, karma is used as a tool to eliminate the heterogeneity between data sources, which can be well applied in geospatial data and services. We use karma to integrate the public text file of GI location attribute data. The location information and specific description information of each data of the data source exist in the JSON file, and the location data is stored in the file in Excel format. We use the original data obtained in the process of data acquisition and use karma call to execute complex geospatial reasoning services: extract the data in a certain area from the integrated data; the shortest distance between query data and other data; extract data integration results from information integration services in the required types [7-8]. Users can perform information integration, interactively invoke services, and visualize the results on the karma map.

In the research process, take the data as the need to standardize the expression of data in different formats and adjust it, and build a semantic model, and use ontology as the integration information source, levering class and attribute level, domain and scope information base and other ontology structures to integrate their data. We can also map the data source to multiple ontologies and map the attributes in the data source to the standard vocabulary defined by the ontology.

After the integrated acquisition of data, we use the karma tool to map the structured multi-data sources into the RDF ontology. In the karma visual interface, map the data sources and ontology patterns, and adjust their mapping relationship when necessary. Karma maps the attributes of the data source into the ontology model, that is, links the concept definition in the ontology and the attribute description in the data source, and makes the data source semantic [9].

2.3 GI Link based on Geospatial Features

Record link refers to identifying the same entity from different information sources, which can solve the heterogeneity and redundancy of data in different data sources. It can improve the quality and integrity of data and reduce the cost and cost of data collection. In the previous methods of record linking, it is mainly through defining semantic rules. However, by defining semantic rules, we ignore the unique spatial relationship features of geographic data, which play a essential role in links. In addition, in view of the different presentation languages used by different information sources, complex preprocessing work must be carried out before recording links. If we use the geospatial relationship of geographic data to record links, we can avoid the preprocessing of translations in different languages.

2.4 GI Link based on Classification and Geographic Features

Figure 2 is the workflow of the GI data link method proposed in this paper, which is mainly divided into four parts: preprocessing, data mapping, similarity calculation, and link. Preprocessing includes GI extraction, data formatting, data cleaning, etc., Data mapping includes the construction

of GI ontology and mapping of data from different GI data sources to realize data correction and eliminate data heterogeneity. Then, the similarity between geographic entities is calculated by multi-feature similarity method, and the feature weight is automatically determined by the classification method in ML to realize the automation of geographic data link [10].

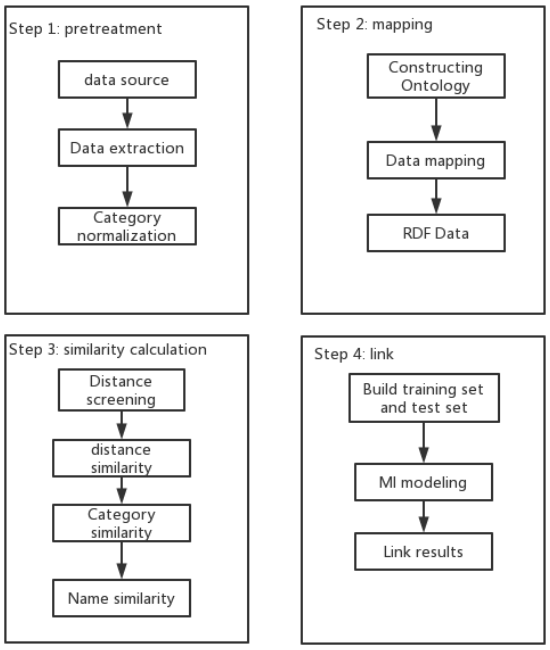

Figure 2: Workflow of GI data link method

In the process of GI data link, similarity calculation is an important basis to realize data link. Combined with entity geospatial features, a multi-feature calculation method integrating entity name, entity category, and geospatial topological relationships is proposed, and the GI data link method is realized combined with the classification algorithm in ML method.

GI linking Method Based on ML

The pseudo code of GI data link method based on classification and multi-features calculates the similarity of name, category and spatial relationship, respectively, and then obtains the final result of data link according to the calculated similarity value combined with the classification method of ML.

Firstly, by calculating the actual distance between GI entities and preliminarily screening through the distance, the first n linked entity pairs with the smallest distance increasing ranking are screened, which will significantly reduce the search space and calculation amount of data links [11]. Through previous experiments, when the distance between entities exceeds 150 meters, the similarity between them is very low, so the screening distance here is set to 150 meters. Next, judge the geospatial topological relationship between entities S1 and S2. When the spatial topological relationship between entities is "contained" or "contained", it is determined that the two entities represent the same entity, and the similarity of all features is 1.

When the names of entities are the same, we can think that the two entities represent the same entity, and the similarity of all features is 1. Because the representation and naming methods of different GI sources are different, the probability of identical entity names in a certain range is relatively low. After experimental observation and analysis, within a certain range, once the names of entities are identical, they all represent the same entity, indicating that the data link is successful.

Finally, different features are fused by the classification algorithm, and the final link result is

obtained by modeling the training data. In the field of spatial information, GI service is a service and application that follows the architecture and standards of a web services and provides GIS data analysis, visualization, and other functions under the network environment; broadly speaking, GI service refers to all services related to spatial information [12].

3. Evaluation Method of GI Data Link

Experimental evaluation refers to the effectiveness of the information integration system in integrating information resources and reflecting the ability of GI data links. The key of experimental evaluation is to determine the index of evaluation method quality and the standard of evaluation method quality. In order to more intuitively analyze the performance of the method to realize its function, this paper will evaluate the feasibility of this method from three angles: accuracy (precision), recall (recall), and F-measure, and compare it with the existing research methods to verify the feasibility and accuracy of this method. There are two final results of GI data link: positive case (match) and negative case (nomatch).

$$p_p = \frac{|TP|}{|Tp + FP|}, E_P = \frac{|TP|}{|TP + FN|}, P_P, E_P \in [0,1] \tag{1}$$

$$P_n = \frac{|TN|}{|TN + FN|}, E_n = \frac{|TN|}{|TN + FN|}, P_n, E_n \in [0,1] \tag{2}$$

Formula (1) is the accuracy and recall rate of positive cases, and formula (2) is the accuracy and recall rate of negative cases. TP represents the set of results that really matched successfully in the matching pairs returned during the data link process, TN is the set of results that really did not match successfully in the mismatch pairs returned during the data link process, FP is the set that misjudged the original negative case results as positive cases, and FN is the set that misjudged the original positive case results as negative cases.

In the process of data link, we pay more attention to the accuracy of link results than the recall rate of link results. In other words, it is necessary to reduce the value of FP. F-measure is the harmonic average of accuracy and recall rate, which can be considered by adjusting the weight.

4. Experimental Test and Analysis

4.1 Experimental Data

In the process of the experiment, the GI data of two urban areas in OpenStreetMap and WikiMapia, Google places, Los Angeles in the United States and London in the United Kingdom, are extracted as the experimental data. The specific ranges are [(-118.3801,33.9812), (-118.2630,34.0538)] and [(-0.1001,51.51), (-0.0835,51.5180)]. Table 1 describes the number of geographic entities extracted from different GI data sources for Los Angeles and London.

Table 1: Number of geographic entities extracted from different GI data sources

	OpenStreetMap	Wikimapia	Google Places
Los Angeles	1230	989	6386
London	1214	454	4490

Due to the difference in accuracy between GI data sources, there is a large gap in the number of geographic entities extracted from different GI data sources. It can be seen from Google place that it is the most necessary to link the GI from the geographic data source, and it is also the worst to link

the GI from the geographic data source.

Table 2 shows the preliminary link results between the two GI data sources and the actual manual verification link results. For example, 3004 / 793 indicates that there are 3004 groups of preliminary link results between two GI data sources OSM and WikiMapia in an area of Los Angeles, including 793 groups of correct link results verified manually, which is the basis for calculating the recall rate of data link results. 6000 pairs of link results are randomly selected from the actual manually verification link results as the training set to build the training model.

Table 2: Preliminary link results and actual link results between GI data sources

Places	OSM-WM	OSM-Google	WM-Google
Los Angeles	3004/793	5074/1116	3892/1050
London	5996/408	5313/3618	2270/1747

4.2 Experimental Results

In order to verify the accuracy of the GI data link method proposed in this paper, the accuracy (P), recall rate (R), and F-measure are used(β= 0.5) and evaluate the experimental results from multiple angles, as shown in Table 3.

Table 3: Comparison between this method and samal method

		paper's method(svm)			paper's method(knn)			Samal's method		
District	Source	P	R	F	P	R	F	P	R	F
Los Angeles	OSM-WM	0.976	0.812	0.936	0.971	0.818	0.936	0.327	0.896	0.375
	OSM-Google	0.954	0.806	0.925	0.941	0.837	0.918	0.275	0.810	0.317
	WM-Google	0.979	0.894	0.964	0.974	0.894	0.957	0.322	0.783	0.365
London	OSM-WM	0.938	0.817	0.912	0.916	0.846	0.901	0.090	0.799	0.109
	OSM-Google	0.889	0.917	0.902	0.870	0.944	0.884	0.685	0.914	0.721
	WM-Google	0.956	0.943	0.953	0.932	0.951	0.936	0.798	0.622	0.755

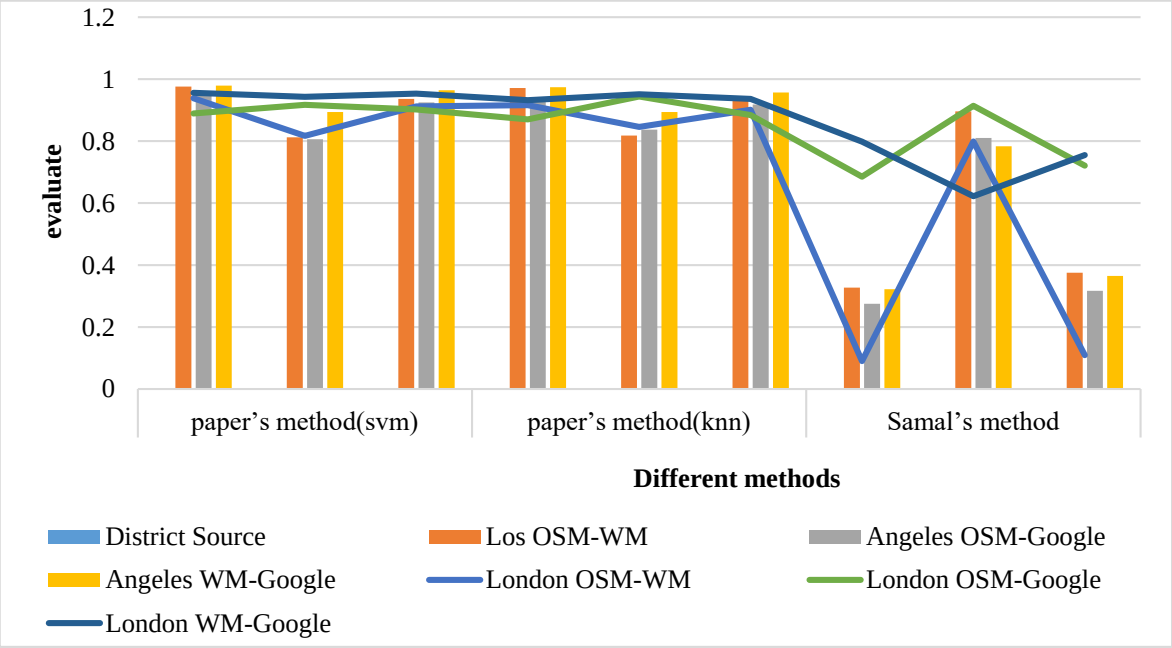

Figure 3: Data chart of comparison between this method and samal method

Figure 3 is the comparison result between the method in this paper and the method of using

graph theory to fuse geographic context related information and context free information for multisource feature matching. This method is implemented and applied to the data set of this paper. The results show that this method is significantly better than their method in accuracy, recall, and F value, which proves the feasibility and superiority of this method.

5. Conclusions

In a word, although the GI data link method proposed in this paper has good performance, there are still some deficiencies, which need further research and exploration. The use of ML methods needs to be further explored. At present, it is only the use of a single classification method. We can also consider combining a variety of classification methods or improving a certain classification method to improve the accuracy of GI data link results. This paper only carries out similarity matching between two GI data sources to realize GI data link. In the later work, the GI data link between multiple GI data sources can be considered; the research scope of this paper is mainly aimed at urban areas or more developed areas with large building density. The GI data link method in this paper has certain limitations, and the feasibility of GI data link in other areas needs to be further studied.

References

[1] Tempelmeier N, Demidova E. Linking Open Street Map with knowledge graphs — Link discovery for schema-agnostic volunteered GI[J]. Future Generation Computer Systems, 2021, 116(2):349-364.

[2] Kaya L A, E Kut Grgün. Land use and land cover change monitoring in Bandrma (Turkey) using remote sensing and GI systems [J]. Environmental Monitoring and Assessment, 2020, 192(7):1-18.

[3] Bayazidy-Hasanabad M, Vayghan S S, Ghasemkhani N, et al. Developing a volunteered GI-based system for rapidly estimating damage from natural disasters[J]. Arabian Journal of Geosciences, 2021, 14(17):1-13.

[4] Moreri K K. Using Kappa methodology to consider volunteered GI in official land administration systems in developing countries [J]. Spatial Information Research, 2020, 28(3):299-311.

[5] Saqr A M, Ibrahim M G, Fujii M, et al. Sustainable Development Goals (SDGs) Associated with Groundwater Over-Exploitation Vulnerability: GI System-Based Multi-criteria Decision Analysis[J]. Natural Resources Research, 2021, 30(6):4255-4276.

[6] Dehghani M H, Asghari F B, Mahvi A H, et al. Spatiotemporal variation of drying and salinity water basin on the quality of coastal aquifers using GI system[J]. Environmental Geology, 2019, 78(15):444.1-444.13.

[7] Alberto R T, Biagtan A R, Isip M F, et al. Hot spot area analysis of onion armyworm outbreak in Nueva Ecija using GI system[J]. Spatial Information Research, 2019, 27(6):673-680.

[8] Owolabi S T, K Madi, Kalumba A M, et al. A groundwater potential zone mapping approach for semi-arid environments using remote sensing (RS), GI system (GIS), and analytical hierarchical process (AHP) techniques: a case study of Buffalo catchment, Eastern Cape, South Africa[J]. Arabian Journal of Geosciences, 2020, 13(22):1-17.

[9] Sk M M, Ali S A, Ahmad A. Optimal Sanitary Landfill Site Selection for Solid Waste Disposal in Durgapur City Using GI System and Multi-criteria Evaluation Technique[J]. KN - Journal of Cartography and GI, 2020, 70(4):163-180.

[10] Peethambaran B, Anbalagan R, Shihabudheen K V, et al. Robustness evaluation of fuzzy expert system and extreme learning machine for GI system-based landslide susceptibility zonation: A case study from Indian Himalaya[J]. Environmental Geology, 2019, 78(6):231.1-231.20.

[11] Comber S, D Arribas-Bel. ML innovations in address matching: A practical comparison of word2vec and CRFs[J]. Transactions in GIS, 2019, 23(2):334-348.

[12] Borodinov A A. Development and research of algorithms for determining user preferred public transport stops in a GI system based on ML methods[J]. Computer Optics, 2020, 44(4):646-652.