Wikipedia の地理情報と GeoNames を組み合わせた地理データベースの構築

Construction of large geographical database by merging Wikipedia's Geo-entities and GeoNames

劉亦奇 [1]　吉岡真治 [1]

, Yiqi Liu [1], Masaharu Yoshioka[1]

[1] 北海道大学大学院情報科学研究科

[1]Graduate School of Information Science and Technology, Hokkaido University

Abstract: It is common to use on-line geographical databases for handling geographic information. For example, GeoNames is a free database for geographic information and Wikipedia is an encyclopedia that contains information of many geographic entities. However, since there is not enough linked information among these databases, it is not easy to combine the information stored in the both database. In this paper, we propose a method to find the corresponding entry between Wikipedia and GeoNames for constructing large Geographic database. In this approach, we made set of rules that represents the relationship between Wikipedia category information and feature class of GeoNames. We also propose a method to find out the corresponding Wikipedia's Geo-entities and GeoNames by using information of class correspondence, name, and country as constraints for matching.

1. Introduction

Nowadays, with increasing of geographical information systems, the geographical information processing is more and more important. In order to dealing with geographical information, it is necessary to use good geographical database. GeoNames[1] is one of the good and huge resources for that database. And there are several attempts to use this information for constructing large ontologies. For example YAGO2 (Yet Another Great Ontology 2) [1], which is one of the largest ontology for open use, expand their Wikipedia[2] based ontology by using GeoNames information. Their integration method is based on name matching and geographical coordinate matching. However, they miss many corresponding pairs among GeoNames entries and Wikipedia entries. We have also conducted an experiment to find corresponding pairs by using name matching and information about administrative are (e.g., country and state) [2]. We could find larger numbers of corresponding pair than YAGO2 cases. One of the reasons why we can achieve good coverage is that we use alternative names for matching. However, since we use alternative names, we found there are several errors due to this matching. For example, “Ueno” (上野) has entries for populated place and railway station. It is impossible to discriminate these entries by using geographical information, such as coordinates and administrative area.

Therefore, in this paper, we propose a method to use Wikipedia category information and feature class of GeoNames to solve this problem.

2. Integration of geographical information of Wikipedia and GeoNames

2.1 Integration method in YAGO2

YAGO (Yet Another Great Ontology) is one of the largest ontology constructed based on Wikipedia. In order to increase geographical information of YAGO,

[1] http://www.geonames.org/

[2] http://en.wikipedia.org/

they propose to use GeoNames as a resource to extract geological information. In this framework, they find 84,349 pairs of Wikipedia page and entry in GeoNames based on following procedures [1].

1. If the Wikipedia entity has the type yagoGeoEntity and shares its name with exactly one entity in GeoNames, we match them.
2. If the Wikipedia entity has the type yagoGeoEntity and shares its name with more than one entity in GeoNames, and we have coordinates for the Wikipedia entity, we match it to the geographically closest GeoNames entity (if its distance does not exceed 5km)

2.2 Integration method by using country and administrative code information

We have already proposed a method to make matching by using country and administrative code information. Followings are procedures to find pairs.

1. Selection of candidate Wikipedia pages Find Wikipedia page that belongs to the category of “Geography of *”, as a candidate geographical entry page. We also collect pages for subcategory (upto 3 steps). We use information that matches * for the information of area. Fr example, “Geography of Ohio” has parent category, “Geography of the United States,” we can identify the pages belong to “Geography of Ohio” are geographical entities that located in Ohio, the United States.
2. Matching between Wikipedia and GeoNames For each candidate page, we select candidate GeoNames entries that satisfy following condition.
   A) Name match: In Wikipedia, there are some conventions to identify information for disambiguation. Most of the case in geographical entity, they add area information after the name (e.g., “Columbus, Ohio”. We remove such information for name matching. We also use redirect information for finding out alternative description to the page. Name match is conducted by using all alternative names.
   B) Area, country match: Since all Wikipedia page has information about location, this information is used for matching entities between the page and GeoName entries.
3. Checking with pairs with multiple entries.
   A) When there are two or more Wikipedia pages exist for one GeoNames entry. Those pairs include disambiguated information
   B) When there are two or more GeoNames entries exist for one Wikipedia page, we check country and administrative code for those entries. If there are entries with different administrative code, those pairs include disambiguated information

By using this method, we can find pairs between 329,364 Wikipedia pages with 499,457 GeoNames entries. We evaluate the quality of matching results based on the information about Wikipedia and GeoNames relationship data supplied at DBPedia[3]. We confirm that we can achieve higher precision 99.5%, but lower recall 71.8%.

However, since GeoNames allows using same name for different classes, it includes inappropriate class matching. For example, “Ueno” (上野) has entries for populated place and railway station and the system extract relationship among Wikipedia page “Ueno, Tokyo” and those two GeoName entities.

3. Rules for representing Wikipedia Category and GeoNames feature class

Since GeoNames has multiple entries for same name that belongs to different feature classes, it is better to construct rules that represent the relationship among Wikipedia information and GeoNames. In this paper, we propose to construct these rules based on Wikipedia and GeoNames relationship data supplied at DBPedia. This data contains 86,547 pairs between Wikipedia page and GeoNames. However, due to modification of the Wikipedia page, we can only use 82,313 pairs.

3.1 Analysis between Wikipedia category and GeoNames feature class.

In order to understand the relationship among Wikipedia category and GeoNames feature class, we construct a list of Wikipedia categories for each GeoNames feature class. Table 1 shows a list of Wikipedia category for the Wikipedia page that corresponds to GeoNames feature class LK (Lake).

Table 1 List of Wikipedia categories for LK

Wikipedia category	Number of pages
LakesOfNewHampshire	41
LakesOfSweden	17

[3] http://wiki.dbpedia.org/Interlinking

LakesOfMichigan	13
CraterLakes	10
LakesOfWyoming	10
SalineLakes	10

Table2 Corresponding GeoNames feature class and Wikipedia category keyword

GeoNames feature class	Wikipedia category keyword
ADM1,ADM2,ADM3,ADM4,ADMD,PPL,PPLA,PPLA2,PPLA3,PPLA4,PPLC,PPLF,PPLG,PPLL,PPLQ,PPLR,PPLS,PPLW,PPLX,PPLX	Capitals,Cities,Regions,Suburbs,Borough,Towns,Departments,Governments,Populated_Places,Parishes,Provinces,States,District,Districts,Counties,Municipalities,Divisions,Quarters,Communities,Neighborhoods,Cantons,Villages,Communes,Prefectures
AIRQ,AIRB,AIRF,AIRP	Airfields,Airports
ISL,ISLS, ISLET,ISLF,ISLM,ISLT,ISLX	Island,Islands
HLL,HLLS,MT,MTS,VLC,PK,CONE	Mountains,Mountain,Volcanoes,Hills
MN, MNAU, MNC, MNCR, MNCU, MNFE, MNN, MNQ	Mines

During this rule construction process, we found following problems in GeoNames database and relationship pairs.

1. GeoNames has 656 feature classes and some of them is similar. For example, “AIRP'” means airport and “AIRF” 'means airfields. “MT” means mountain and “VLC” means volcano. However, it is little bit complicated for a user to select appropriate class to select, there are some errors in class selection (e.g., there are many volcanoes marked as “MT”). It is better to make some groups of class for reducing the mismatch problem due to this error.
2. There are many inappropriate pairs for Wikipedia page and GeoNames entry. For example, Curepipe (GeoNames id:934567) belongs to “RSTNQ” that means 'abandoned railroad station. However, corresponding Wikipedia page does not contain information about railway station. In addition, many pages are moved due to the disambiguation problems.

3.2 Rule construction

Based on this discussion, we construct rules that represent the relationship among groups of feature class in GeoNames and a set of keywords that are included in the Wikipedia category. In order to make group of GeoNames feature class, we check the similarity of using keywords for related to the GeoNames. For example, “MT” and “VLC” are also related to the keyword “Mountain” and “Volcano.” In addition, concept of “MT” and “VLC” are similar, we make the group that contains both classes.

Table 2 shows some examples of corresponding Geonames feature class and Wikipedia category keywords.

Since pair information in DBPedia is not large enough, we only made rules for 112 GeoNames feature class.

In order to evaluate the quality of these rules, we just use pair information in DBPedia. In this evaluation, we classify the data into three types.

- Matched: We can find at least one Wikipedia category for the given GeoNames feature class.
- Other: We can find at least one Wikipedia category associated with GeoNames feature class. But we cannot find appropriate GeoName feature class.
- NoInfo: There is no Wikipedia category associated with GeoNames feature class.

Table 3 shows the result of this classification process.

Table3: Classification results for internal data

Matched	Other	NoInfo
75,694	2,819	3,800

Precision and recall of a rule set based on this internal evaluation is 96.4% and 92.0%, respectively. However, NoInfo includes many entries with disambiguation pages (1056) that may not be a corresponding page for GeoNames.

In order to improve the quality of this rule, we plan to use template information for the next step.

4. Integration of Wikipedia and GeoNames

First, we construct candidate list for integration based on the method discussed in Section 2.2. We used database of Wikipedia dumped at 2011-09-01 and download GeoNames database at 2011-09-22.

Followings are list of modification from the previous procedure.

1. We construct candidate pairs of Wikipedia and GeoNames by using name matching for all

Wikipedia database. We found 8,863,723 pairs for 718,407 Wikipedia pages and 1,731,086 GeoNames entries.

2. In addition to “Geography of” category, we also extract country and administrative information by using Wikipedia category and disambiguation information. For example, from Wikipedia page, “Ueno, Tokyo,” we extract area related information “Tokyo” and check GeoNames database to find appropriate country and administrative code. In this case, we extract “Japan” and “Tokyo”. We can add country and area information to 545,472 Wikipedia pages.
3. We check pairs based on the country and administrative are information. We found 1,117,088 pairs for 467,015 Wikipedia pages and 819,348 GeoNames entries
4. For 113 1,117,088 pairs, we use rules for Wikipedia category and GeoNames feature class for classification. Table 4 shows the result of classification.

Table4: Classification results for extracted pairs

Matched	Other	NoInfo
845,777	161,462	109,849

Based on the analysis of internal evaluation, we cannot say all of matched data is true, but we can show the possibilities to find more related pairs than previous method.

In addition, we also plan to use geographical coordinate information to remove inappropriate pairs that can identify using this information.

5. Conclusions

In this paper, we propose a method to integrating geographical information of Wikipedia and GeoNames by using class matching rule between Wikipedia category and GeoNames feature class. During this rule construction process, we found several errors exist in GeoNames. Construction of correspondence rule may be helpful to find out this kind of problem. We also confirm that this rule may be helpful to ensure the quality of automatic generated pairs between Wikipedia page and GeoNames entry. However further analysis is needed for verification.

Acknowledgement

This research was partially supported by a Grant-in-Aid for Scientific Research (B) 21300029, from the Japan Society for the Promotion of Science.

References

[1] Johannes Hoffart, Fabian Suchanek, Klaus Berberich, and Gerhard Weikum: YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia, Research Report MPI-I-2010-5-007, Max-Planck-Institut fur Informatik, November 2010

[2] Hitoshi Takenaka and Masaharu Yoshioka: Extraction of geo-spatial relationships among geographical name by using Wikipedia, The 25th Annual Conference of the Japanese Society for Artificial Intelligence, CD-ROM 2J3-NFC2-2, 2011.