

A Study on Automatic Ontology Mapping of Categorical Information

Naijun Zhou

Department of Geography, Land Information and Computer Graphic Facility

University of Wisconsin – Madison

nzhou@wisc.edu

http://www.lic.wisc.edu/DG_Project/DGhomepage.html

Abstract

Semantic heterogeneity of information is a major barrier of information and system interoperability. Defining ontology of data and mapping ontologies among heterogeneous information repositories is one approach to achieve interoperability. This paper focuses on the ontology mapping of categorical information, which usually have a tree structure with categories and subcategories. Subcategories can be considered as the definition of their upper level categories. Methods of automatic mapping of categorical information using Naïve Bayes classifier are discussed, and improved algorithms for categorical ontologies mapping are proposed and compared to a standard word-by-word matching algorithm.

1 Background

Geographic information is being collected and provided by various information producers. Sharing and retrieving geographic information over the World Wide Web and thereby achieving GIS interoperability is highly desired. However, semantic heterogeneity, which occurs when information from multiple sources is defined and represented differently, is a major barrier of the GISystem and geographic information interoperability (Sheth, 1999). A very simple example is, *light industry* and *small manufacturing* could refer to the same land use types in different land use type classification systems, but machine-based Web search engines cannot recognize this similarity and will treat them as different types.

The *semantic web* is proposed to add machine-processable information to web-based data in order to realize interoperability (Berners-Lee, 2001). At the core of research on the semantic web, *ontology* has been studied to provide semantic information to assist communication among heterogeneous information repositories (Fonseca 2001). Ontology mapping, by identifying how different ontologies are mapped and related, is a mechanism of the communication among ontologies. Of the methods of ontology mapping, there are manual ontology mapping (i.e., defining the mappings by hand or assisted by computer) (Cruz et al, 2002) and automatic ontology mapping (i.e., producing the mappings by intelligent machine agents) (Wiesman et al, 2001). Automatically mapping ontologies is more desirable and practical in some cases considering the high heterogeneity and huge volume of information the Web contains.

Land use information, as one of the major sources of geographic information on the Web, is highly heterogeneous in syntax, structure and semantics (Wiegand et al, 2002). The heterogeneities arise because land use data are produced and provided by a variety of agencies having different definitions, standards and applications of the data. Solving the problem of semantic heterogeneity, for example, the categorical land use types in various land use classification systems, is difficult but very useful for information sharing on the Web.

2 Ontology mapping using Naïve Bayes classifier

2.1 Ontology mapping for categorical information

Although automatic ontology mapping is essential in solving the problem of semantic interoperability, it is still poorly understood. Many mapping techniques are based on word matching, using a dictionary (for

example, WordNet), or applying logical reasoning techniques. However, the mapping of categorical information is unique and requires special consideration. Categorical data usually have a tree structure, of which the upper level category includes subcategories. Subcategories can be considered as the definition of their upper level categories. A category and its subcategories together make an ontology of this category, thus the mapping of categories can be transformed into a problem of ontology mapping of related subcategories.

This paper aims to identify the categorical information mapping based on the meaning of the information. Land use classification systems are studied as an example. It is observed that, in land use classification systems, subcategories actually define the meaning of the category. For example, in North America Industrial Classification System (NAICS), subcategories of *Animal food manufacturing*, *Grain and oilseed milling*, etc, describe the meaning of category *Food Manufacturing*. Therefore, a classification system can be considered a set of ontologies of land use types, with subcategories as the definition of the higher level category. In this way, ontology mapping of land use types is the mapping of meanings.

This research applies the document classification technique to the categorical information mapping. The mapping of land use types can be transformed into a task of classifying a type in one land use classification system into candidate types in another system, and assigning one or several candidate mappings with a mapping degree respectively. Naïve Bayes classifier, as one of the most successful classification techniques (Lewis, 1998) in machine learning and information retrieval, is studied and enhanced for automatic categorical information mapping.

2.2 Naïve Bayes classifier

The Naïve Bayes classifier is based on a probability theorem of Bayes' rule (Mitchell, 1997). The Bayes' rule lets us use causal knowledge to make diagnostic inferences. Consider an example of air flight delay, it may be difficult to obtain the probability of snowing given a fact that an air flight has been delayed, i.e. $\Pr(\text{snow} | \text{flight delay})$. But we could already know well the probability of flight delay if there is a snow, i.e. $\Pr(\text{flight delay} | \text{snow})$. With Bayes' rule, we can calculate the $\Pr(\text{snow} | \text{flight delay})$ using the $\Pr(\text{flight delay} | \text{snow})$:

$$\Pr(\text{snow} | \text{flight delay}) = \Pr(\text{flight delay} | \text{snow}) * \Pr(\text{snow}) / \Pr(\text{flight delay}).$$

Bayes' rule has been successfully applied in document classification, namely Naïve Bayes classifier. To classify a source document into one of the candidate categories, we may instead only need to know how each candidate category is close to the source document. With predefined document categories c_i , $i = 1, \dots, n$, for a source document d with n words (w_1, w_2, \dots, w_n) , the document d can be classified to a candidate category c_j with a probability of

$$\Pr(c_j | d) = \frac{\Pr(c_j) \Pr(d | c_j)}{\Pr(d)} \propto \frac{\Pr(c_j) \prod_{i=1, n} \Pr(w_i | c_j)}{\Pr(d)}$$

w_i is a word in the source document d , and c_j is a candidate category to be assigned from the source document, $\Pr(w_i | c_j)$ is the probability that the category c_j contains the word w_i . Assuming each type of document (d) to be classified will appear with a same probability, the classification with the highest matching probability is the mapping result:

$$\arg \max \Pr(c_j) \prod_{i=1, n} \Pr(w_i | c_j)$$

2.3 A standard algorithm of the $\Pr(w_i|c_j)$: Laplace Estimator (LE)

It is noted that the algorithm to calculate the mapping probability, $\Pr(w_i|c_j)$, determines the efficiency and accuracy of the Naïve Bayes classifier. Laplace Estimate (LE) is a standard word-by-word matching algorithm to calculate $\Pr(w_i|c_j)$:

$$\Pr(w_i | c_j) = \frac{N(w_i, c_j) + 1}{N(c_j) + T}$$

$N(w_i, c_j)$ is the number of times a word (w_i) appears in a category (c_j), $N(c_j)$ is the total number of words (duplicated words are counted as well) in category c_j , and T is the total unique word number in all candidate categories c .

2.4 Improved algorithms of the $\Pr(w_i|c_j)$ for categorical information

The word-by-word matching algorithm is a frequency-based research assuming the more frequently a word appears the more important the word is. However, this is problematic in classifying categorical information, for example, land use types, where words are used as the definition of a category and usually appear only once. It is observed that the frequent words, which are called stop words and sometimes need to be ignored, are probably the least important in defining a concept.

To fully consider the uniqueness of categorical information, this paper develops two improved algorithms of $\Pr(w_i|c_j)$, i.e., Category Counting Estimate (CCE) and Category Counting Estimate with Semantics (CCE-S). CCE is based on an assumption that, if a source document (d) is similar to a candidate category (c_j), then the number of word matching between d and the category c_j will be higher than the matching with other categories. In this way, the existence of a word, instead of the frequency of a word, is incorporated into the algorithm to calculate the matching probability. Let a source category to be mapped to a candidate category (c_j) has a word w_i , the CCE algorithm is

$$\Pr(w_i | c_j) = \frac{N(w_i, c_j)}{N(c_j)}$$

$N(c_j)$ is the total number of words (duplicated words are not considered) in a candidate category c_j . $N(w_i, c_j) = N(w_i) * N'(w_i, c_j)$, where $N(w_i)$ is the number of a word (w_i) appears in the source category, and $N'(w_i, c_j)$ is the number of a word (w_i) appears in a candidate category (c_j). This algorithm is expected to perform better than LE because it fully considers the uniqueness of categorical information by calculating an individual word's relationship with a whole category instead of only with the matched words.

LE and CCE are algorithms of word matching, not considering any semantic aspect of the categorical information. Using the semantic component of information has been considered a promising solution for ontology mapping. With semantic definition, relationships among words can be extended from exact matching to synonym, hypernym, hyponym, overlap, etc. This paper develops two semantic relationships of words, synonym and superset, to retrieve additional information of synonym and superset words. An example of synonym words is *music* and *audio*, and an example of superset words is *dental* and *medical*. Additionally, because the synonym is actually an exact matching while the superset is a partial matching, this paper weights the matching by assigning a weight of 1 to synonym words, and a weight of 0.5 to superset words when calculating the matching probability.

2.5 Result

The algorithms of CCE and CCE-S improve ontology mapping for categorical information by considering the features of categorical data. The algorithms perform better than the standard LE algorithm in terms of

accuracy and completeness. On conclusion is that, for well-defined land use types, the algorithms of LE, CCE and CCE-S provide similar mapping results. However, when the land use types are incompletely defined and the definition of a type is scattered in different categories, CCE and CCE-S outperform the LE by providing more accurate and complete mapping results. CCE-S particularly has the potential in improving ontology mapping of categorical data because it consults additional semantic relationships of the words. The Naïve Bayes classifier also provides a probability to each mapping as the degree of belief of the mapping. This belief degree is especially useful in cases that one ontology is mapped to multiple ontologies but with different degree of confidence.

3 Conclusion

This paper studies the categorical information mapping with an example of land use types. This approach uses sub-categories as the definition of a category in order to implement the ontology mapping of meaning other than just of words. Improved algorithms for the Naïve Bayes classification technique are developed to improve the accuracy and completeness of the map.

Although this research studies the mapping of land use types, it has the potential to be generalized in all domains of geospatial information because geospatial data are usually defined or described by metadata. Semantic metadata can be interpreted and mapped between different data sources in order to realize the interoperability of heterogeneous geospatial data over the Web.

Acknowledgements

This work was supported in part by the National Science Foundation, Grant No. 091489. Discussions with Steve Ventura, Mark Harrower and Nancy Wiegand are highly appreciated.

Reference

- Berners-Lee, Tim (2001). The semantic Web. *Scientific American*, 284(5), 35-35.
- Cruz, Isabel, Afsheen Rajendran, William Sunna, and Nancy Wiegand (2002). Handling semantic heterogeneities using declarative agreements. *Proceedings of ACM GIS*, November 2002.
- Fonseca, Frederico (2001). *Ontology-driven geographic information systems*. Ph.D. thesis, University of Maine.
- Lewis, David D. (1998). Naïve (Bayes) at forty: The independence assumption in information retrieval. *Proceedings of ECML-98, 10th European Conference on Machine Learning*, 1998.
- Mitchell, T. (1997). *Machine Learning*. McGraw Hill.
- Sheth, Amit P. (1999). Changing focus on interoperability in information systems: From system, syntax, structure to semantics. In Goodchild, M., Max Egenhofer, Robin Fegeas and Cliff Kottman (eds.), *Interoperating Geographic Information Systems*. Kluwer Academic Publishers, MA, 5-30.
- Wiegand, Nancy, E.D. Patterson, Naijun Zhou, Steve Ventura, Isabel F. Cruz (2002). Querying heterogeneous land use data: problems and potential. *Proceedings of National Conference on Digital Government Research*, 115-122, 2002.
- Wiesman, Floris, N. Roos and P. Vogt (2001). Automatic ontology mapping for agent communication. MERIT-Infonomics Research Memorandum series. <http://www.infonomics.nl>.