AN ONTOLOGY-BASED CLUSTER ANALYSIS FRAMEWORK

Paweł Lula
Department of Computational Systems
Cracow Uniwersity of Economics
ul. Rakowicka 27, 31-510 Kraków, Poland
+48 12 293 5265
pawel.lula@post.pl

Grażyna Paliwoda-Pękosz
Department of Computer Science
Cracow Uniwersity of Economics
ul. Rakowicka 27, 31-510 Kraków, Poland
+48 12 293 5265
paliwodg@uek.krakow.pl

ABSTRACT
The main objectives of this paper are to discuss the various aspects of similarity calculations between objects and sets of objects in ontology-based environments and to propose a framework for cluster analysis in such an environment. The framework is based on the ontology specification of two core components: description of categories and description of objects. Similarity between objects is defined as an amalgamation function of taxonomy, relationship and attribute similarity. The different measures to calculate similarity that can be used in framework implementations are presented. The ontology-based data representation and the framework of cluster analysis can be useful in the area of Business Intelligence, e.g. clustering similar companies that profiles are described by ontology-based data.

Categories and Subject Descriptors
H.2.8 [Database Applications]: Data mining.

General Terms
Algorithms, Measurement

Keywords
Cluster analysis, ontology, similarity measures, data mining

1. INTRODUCTION
Cluster analysis is one of the most commonly used methods of data analysis. Usually, classical data analysis methods operate on flat data sets in which rows represent objects and columns correspond to variables (objects characteristics); all objects are homogeneous and the importance of each variable is identical. However, when using such data representation it is difficult to capture additional relationships between objects. It may be useful to enrich classical cluster analysis methods by using objects that are described using an ontology.

An ontology-based approach allows the analyst to represent the complex structure of objects, to implement the knowledge about hierarchical structure of categories as well as to show and use the information about relationships between categories and individual objects. However, it has to be noted that the ontology-based methods are more demanding than classical data mining methods in the following ways: performing calculations on complex objects is more challenging from a theoretical and numerical point of view and insufficient theoretical background can result in calculations that may be partly subjective.

2. GENERAL FRAMEWORK

2.1 Ontology definition
We assume that the ontology is a structure that consists of:

- *Categories description* (list of categories, the definition of each category, data types for all attributes, a category hierarchy schema and definitions of relationships between categories);
- *Objects description* (a category, attributes values, descriptions of relationships with other objects).

2.2 Framework scheme
The goal of cluster analysis is the division of a set of objects into homogeneous clusters. In the paper we will concentrate on the application of agglomerative hierarchical clustering [2] which in ontology-based environment will take the following steps: (1) calculation of distance (or similarity) matrix between every pair of objects using ontology-specific methods of calculation the distance (or similarity) between objects, (2) every object constitutes a separate cluster, (3) merging of the two closest clusters, (4) modification of the distance matrix – merged clusters are treated as the one object. Here the methods of counting similarity between an object and a cluster as well as methods of counting similarity between clusters in ontology-based environment are needed, (5) if the objects have not been divided yet into desired number of clusters then we move to the step 3.

We assumed that a common ontology is used for descriptions of all compared objects (a problem of ontology matching is out of the scope of this paper) and that the similarity measure is a real value normalized to the range [0; 1].

3. SIMILARITY BETWEEN OBJECTS

3.1 General scheme

The approach presented in this paper is the generalization of the methodology proposed by Maedche and Zacharias in [3]. The similarity between objects is an aggregation function (f_{agr}):

$$\mathrm{sim}(I_i, I_j) = f_{agr}\left(TS(I_i, I_j), RS(I_i, I_j), AS(I_i, I_j)\right)$$

where TS- taxonomy similarity, RS - relationship similarity and AS - attribute similarity.

3.2 Evaluation of taxonomy similarity

There are several approaches to calculating similarity (or dissimilarity) between classes on the hierarchy schema, e.g. similarity measures based on the path distance between classes in the tree (e.g. Wu and Palmer measure [1]), the upward cotopic similarity - the application of the Jaccard similarity to the superclasses of two categories [3], measures based on information theory (e.g. Resnik and Lin [5]).

3.3 Evaluation of relationship similarity

The idea of the relationship similarity is very simple: similar objects should have relationships with objects that are similar to each other. When we compare two objects O_1 and O_2 we should indicate all objects that have relationships with object O_1 and all objects that have relationships with O_2, calculate taxonomy similarity and/or attribute similarity between these two sets of objects and finally aggregate calculated similarities.

3.4 Evaluation of attribute similarity

The way in which the similarity between attributes' values is calculated depends on the data type of object's attributes. To compare numbers we can use the relative difference between numbers [3]. Jaccard similarity can be used to compare intervals as well as sets. We assume that nominal values similarity will be counted using the simple rule: if the nominal values are equal then the similarity measure is 1, otherwise 0. In order to compare strings measures based on the edit distance can be used, Jaro and Jaro-Winkler measures and measures based on a lexical similarity (e.g. measures based on WordNet: Leacock-Chodorow, Resnik, Lin) [1]]. Texts can be compared using a vector representation of texts. The similarity between texts can be calculated using distance measures between vectors [4]. Finally, in order to compare sequences of values the methods for string comparison can be used, e.g. edit distance.

An open issue is how to combine partial attribute similarity measures to calculate global attribute similarity between objects. From the theoretical point of view we can take into account arithmetic average, geometric average, harmonic average and quadratic average, to name a few. In practice, the weighted average is most commonly used. However, the problem arises in what weights should be assigned to attributes. Moreover, the way of dealing with incompatibly number of attributes has to be solved. If all attributes are obligatory the problem of aggregation similarity measures is not difficult. The greater incompatibility in the number of attributes the more complicated the calculations.

3.5 Aggregation formula

When taxonomy, relationship and attribute similarity measures are evaluated, it is necessary to combine them into one measure. The weighted average proposed in [3] seems to be a good idea. The aggregation formula and weights' values are essential for similarity assessment. The definition of aggregation function and the estimation of its parameters may be proposed by an expert or can be approximated on the basis of a learning set (for example by the neural network technique).

4. SIMILARITY BETWEEN SETS OF OBJECTS

When calculating similarity between two sets of objects we can consider different ways of incorporating similarities between objects that belong to these sets, e.g. single link, completely link, average link, centroid [2]. The choice of the proper method depends on the characteristic of the research problem.

5. IMPLEMENTATION

Currently, the implementation of the framework is being prepared. The main assumptions of this implementation are the following: the ontology description is prepared in the OWL language, the set of objects is also defined in the OWL, all algorithms are implemented in Java, the project implementation uses other Java packages related to ontology approach (e.g. Jena, SimPack). The results of calculations may be presented during the workshop.

6. CONCLUSIONS AND FUTURE WORK

The ontology-based clustering framework proposed in this paper is the generalization of the framework formulated by Maedche and Zacharias [3]. It gives the possibility of incorporating various kinds of measures to determine similarity between objects and set of objects in the three dimensions: taxonomic, relationship and attribute. Though the basis of ontology-based cluster analysis has been done, much work remains. First, the implementation of the framework in Java environment should be completed. Second, extensive tests should be conducted in order to evaluate different kind of measures and amalgamation functions in various application domains.

7. REFERENCES

[1] Euzenat, J. and Shvaiko, P. 2007. Ontology Matching. Springer-Verlag. Berlin Heidelberg.

[2] Han, J. and Kamber, M. 2006. Data Mining: Concepts and Techniques (2nd edition). Morgan Kaufmann.

[3] Maedche, A. and Zacharias, V. 2002. Clustering Ontology-Based Metadata in the Semantic Web. Lecture Notes In Computer Science. Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery. Springer-Verlag London. Vol. 2431, 348-360.

[4] Manning, C. and Schütze, H. 1999. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA.

[5] Zhang, X., Jing, L., Hu X., Ng M. and Zhou, X.2007. A Comparative Study of Ontology Based Term Similarity Measures on PubMed Document Clustering. http://www.pages.drexel.edu/~xz38/pdf/209_Zhang_DASFAA07.pdf