A Text Categorization Perspective for Ontology Mapping

Position paper

Xiaomeng Su
Dept. of Computer and Information Science
Norwegian University of Science and Technology
N-7491, Norway
xiaomeng@idi.ntnu.no

Abstract

This position paper addresses the problem of ontology mapping which is pervasive in context where semantic interoperability is needed. A preliminary solution is proposed using external information, i.e. documents assigned to the ontology to calculate similarities between concepts in two ontologies. Text categorization is used to automatic assign documents to the concepts in the ontology. Based on the similarities measure, a heuristic method is used to establish mapping assertions for the two ontologies.

1. Background & Problem

Lately, there has been much research related to the new generation web – semantic web. The hope is that the semantic web can alleviate some of the problems with the current web, and let computers process the interchanged data in a more intelligent way. In an open system like the Internet, which is a network of heterogeneous and distributed information systems (IS), mechanisms have to be developed in order to enable systems to share information and cooperate. This is commonly referred to as the problem of interoperability. The essential requirement for the semantic web is interoperability of IS. If machines want to take advantage of the web resources, they must be able to access and use them.

Ontology is a key factor for enabling interoperability in the semantic web [Bernees-Lee01]. An ontology is an explicit specification of a conceptualisation [Uschold96]. It includes an explicit description of the assumptions regarding both the domain structure and the terms used to describe the domain. Ontologies are central to the semantic web because they allow applications to agree on the terms that they use when communicating. Shared ontologies and ontology extension allow a certain degree of interoperability between IS in different organizations and domains. However there are often cases where there are multiple ways to model the same information and the problem of anomalies in interpreting similar models leads to a greater complexity of the semantic interoperability problem.

In an open environment, ontologies are developed and maintained independently of each other in a distributed environment. Therefore two systems may use different ontologies to represent their view of the domain. Differences in ontologies are referred to as ontology mismatch [Klein01]. The problem of ontology mismatch arises because a universe of discourse, UoD, can be specified in many different ways, using different modelling formalisms. In such a situation, interoperability between systems is based on the reconciliation of their heterogeneous views. How to tackle ontology mismatch is still a question under intensive research.

As pointed out in [Wache01], three basic architectures to cope with ontology mismatch can be identified: *single ontology approaches*, *multiple ontologies approaches* and *hybrid approaches.* An illustration of each of them is given in figure 1.

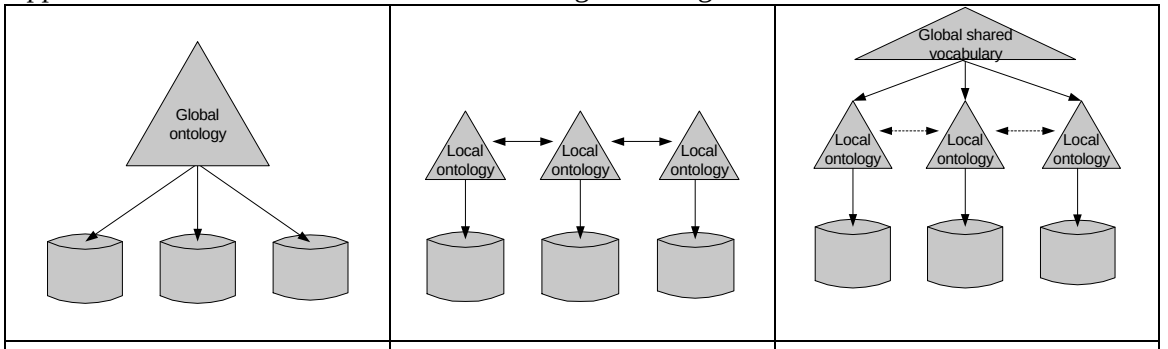

Figure 1a: Single ontology approach	*Figure 1b: Multiple ontologies approach*	*Figure 1c: Hybrid approach*

Figure 1 Architectures to cope with ontology mismatch.

In the Single ontology approach, a global ontology provides a *shared global ontology* to specify the semantics. All systems or information sources are related to the global ontology i.e. they are unified. The global ontology can be a combination of modularised sub ontologies.

In the Multiple ontologies approach, each information source has its own *local ontology*, which doesn't necessarily use the same vocabulary. Each ontology can be developed independently because there is a loose coupling between the ontologies. To achieve interoperability the ontologies must be brought together by mapping rules (links).

In the Hybrid approach, the basic features of the two other approaches are combined in order to overcome some of their disadvantages. Like the multiple ontologies approach, each source has its own *local ontology*. But the local ontologies are developed from a *global shared vocabulary* in order to make the alignment of ontologies easier. The shared vocabulary defines basic terms for the domain, which can be combined to describe more complex semantics in the local ontologies.

A single ontology approach, which is based on tight coupling, most often is too rigid and does not scale well in a large open environment. Adding a new source will most often lead to a new unification process [Wiederhold99]. In our opinion, a multiple or hybrid approach is more appropriate, allowing a degree of local autonomy to coexist with partial interoperability. In both of the latter cases, developing means to facilitate mapping between two ontologies is necessary.

A web portal scenario can be used to illustrate the ontology mapping problem. A Web portal is a web site that provides information content on a common topic, for example a specific city or a specific interest (like ski). A web portal allows individuals that are interested in the topic to receive news, find and talk to other interested people, build a community, and find links to web resources of common interest. Normally, web portals can define an ontology for the community. This ontology defines terminologies for describing content and serves as an index for content retrieval. One example of an ontology-based portal is The Open Directory Project [ODP], a large, comprehensive human-edited directory of the Web. Say, for example, that there are two web portals about topic sports. One of them uses ODP, while the other is based on a sub portion of Yahoo! Category. Users may want to share or exchange information between the portals. In that context, means that allow ontologies to map terms to their equivalents in other ontologies, must be developed.

2. Scope

The word ontology has been used to describe artefacts with different degrees of structure. These range from simple taxonomies (such as the Yahoo hierarchy), to metadata schemes (such as the Dublin Core), to logical theories. In our context, the scope and assumption of our work are the following:

1) An *ontology* is a set of elements connected by some structure. Among the structures, we single out hierarchical IS-A-relation and all the others we call them other relations. A *classification hierarchy* is a typical example organized by hierarchical IS-A-relation. Note that attribute (or slot) has not been taken into consideration at that stage.
2) The pair of ontologies, that are subject to be mapped, are homogenous and their elements have significant overlap.
3) There exist different ontology representational languages [Su02], we assume that it is possible to translate between different formats. In practice, a particular representation must be chosen for the input ontologies. Our approach is based on Referent Modelling Language (RML) [Soelvberg98].

The overall process of ontology mapping can then be defined as:

Given two ontologies A and B, mapping one ontology with another means that for each concept (node) in ontology A, try to find a corresponding concept (node), which has same or similar semantics, in ontology B and vice verse. To be more exact, we need to

a) define the semantic relationships that can exist between two related concepts.
b) develop algorithm, which can discover concepts that have similar semantic meaning.

Thus, the result of a mapping process is a set of mapping rules. Those mapping rules connect concepts in ontology A to concepts in ontology B.

Approaches from different communities have been proposed in the literature to deal with this problem. The intention of this work is to draw experiences from the related areas and base on those experiences, to formulate our own solution.

3. Related work

For dealing with semantic heterogeneity among distributed, autonomous information sources there exist approaches in the multi database and information systems area for years. In [Batini86], a variety of database schema integration methods were studied and the schema integration process can be divided into three major phases: schema comparison, schema conforming and schema merging. [Rahm01] claims that a fundamental operation in the manipulation of schema information is match, which takes two schemas as input and produces a mapping between elements of the two schemas that correspond semantically to each other. Schema matching approaches were classified into schema-level matchers and instance-level matchers. Schema-level matchers only consider schema information, including the usual properties of schema elements, such as name, description, data types, relationship types constraints and schema structure. As complementary methods, instance-level approaches can give important insight into the contents and meaning of schema elements. Instance-level approaches can be used to enhance schema-level matchers in that evaluating instances reveals

a precise characterization of the actual meaning of the schema elements. In general more attention should be given to the utilization of instance-level information to perform match [Rahm01]. The mapping returned by a match operation may be used as input to operations to merge schemas or mediate between schemas.

In the research area of knowledge engineering, a number of ontology integration methods and tools exist. Among them, Chimaera [McGuinness00] and PROMPT [Noy00]are the few which have working prototypes. Both tools support the merging of ontological terms i.e. class and attribute names from various sources. The processes start by running a matching algorithm on class names in the pair of ontologies to suggest the merging points. The matching algorithm either looks for an exact match in class names or for a match on prefixes, suffixes, and word root of class names. A user can then choose from these matching points, or proceed on his own initiative. PROMPT provides more automation in ontology merging than Chimaera does. For each merging operation, PROMPT suggests the user to perform a sequence of actions on copying the classes and their attributes, creating necessary subclasses and putting them in the right places in the hierarchy.

More recent work includes OntoMerger, an ontology translation service [OntoMerge]. The merge of two ontologies is obtained by taking the union of the axioms defining them, and then adds bridging axioms that relate the terms in one ontology to the terms in the other. XML namespaces are used to avoid name clashes. The service accepts a dataset as a DAML file in the source ontology, and will respond with the dataset represented in the target ontology also as a DAML file.

Text categorization, the assignment of free text documents to one or more predefined categories based on their contents, is an important component in many information management tasks. A number of statistical classification and machine learning techniques has been applied to text categorization [Aas99], including Rocchio, Naïve Bayes, Nearest neighbour, Support Vector Machine, voted classification and neural networks. More recently, there have emerged some preliminary studies trying to apply text categorization techniques into merging and mapping ontologies. [Lacher01] presents an approach using supervised classification (Rocchio) for ontology mapping. [Agrawal01] uses techniques well known from the area of data mining (association rules) for the task of catalogue integration.

[Stumme01] proposes a method called FCA-MERGE, based on the theory of formal concept analysis, for merging ontologies following a bottom up approach and the method is guided by application-specific instances of the given source ontologies that are to be merged.

4. Proposed Solutions.

We divide this part into three sub sections. The first two are in correspondence with the two sub-problems respectively, which we have outlined in previous sections. The third sub section suggests some of the uses of this approach in several domains.

4.1 Meta model for mapping

A general implementation of the mapping process compares each ontology A element with each ontology B element and determines a similarity metric per pair. Only the combinations with a similarity value above a certain threshold (or top- ranked lists) are considered as match candidates. Various mapping methods are distinguished with respect to using what information to compute the similarity value and how to compute it.

In order to discuss definitions of similarity and to support development of novel mapping approaches, we need to define a metamodel for mapping. In [Hakkarainen99], a notion of

correspondence assertion is introduced for that purpose. We will adopt that correspondence assertion metamodel as a base for discussing different types of mappings.

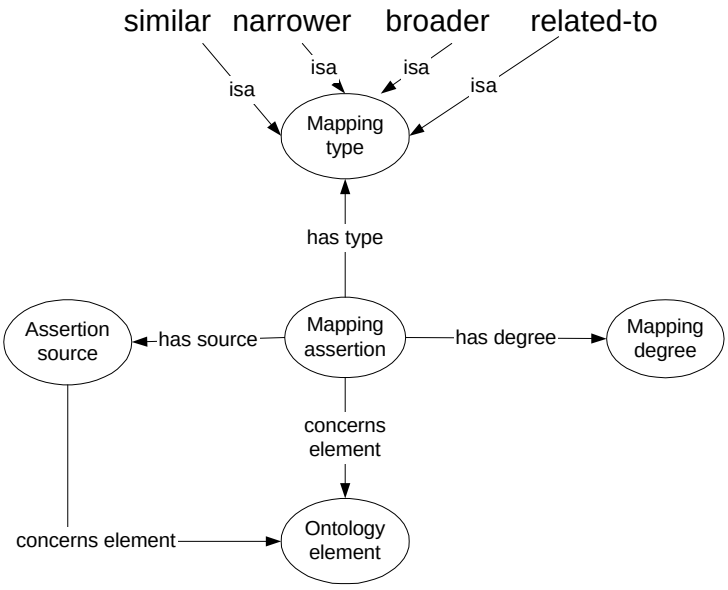

Figure 2 mapping assertion metamodel

The metamodel in Figure 2 has the following meaning: a mapping assertion is an objectification of the relationship between two ontology elements and supports further description of that relationship. A mapping assertion is uniquely assigned to two ontology elements. It has also a mapping degree in order to provide a way of ranking the outputs. A mapping type is also attached to a mapping assertion, which specifies how the pair of ontology elements is related. The intention of the assertion source is to provide an explanation why the particular assertion is chosen (linguistic derived, for instance). For a mapping process, mapping assertion is the core output of the process. How to generate the mapping assertions will be discussed next.

4.2 Discovering Mapping

In the context of database, a schema defines the *intension* of the database and the instances of data define *extensions*. The same can apply to ontology, where an ontology is a intentional description of a Universe of Discourse (UoD), and the set of instances, which conform to that ontology is the extension of the UoD.

If we think of an ontology as a taxonomy of a domain and each node in the taxonomy as a category which has documents assigned to it, the ontology then is the intension of the UoD and the sets of documents form the extension of the UoD. In our approach, the process of mapping ontologies can be supported by analysing the extension of concepts to derive corresponding intentional descriptions. In other terms, if we use the taxonomy of automatic schema matching, which is proposed in [Rahm01], our approach is in line with the so called instance-level approaches.

The intuition is that given two ontologies A and B, for each node a_i in A, we calculate a similarity measure $sim(a_i, b_j)$ where b_j belongs to B. Then the node with the highest similarity will be ranked on top. Information retrieval techniques are used when we calculate $sim(a_i, b_j)$.

Figure 3 depicts the general architecture of the suggested mapping process. The approach takes the two ontologies and a document set as input. Notice that documents are relevant to both ontologies.

The first step is to assign documents to concept nodes of the ontology using some text categorization techniques. The assigning of documents to concept nodes is necessary since there exist ontologies, where no external information is given in the format of reference with documents. However, if this situation is given, or in other words, we have in our possession two ontologies, similar to the one depicted in Figure 4, where documents have already been assigned to specific categories, we can skip the first step and use the two ontologies -- O_A' and O_B' directly as input for the Mapper. The Open Directory Project is an example, where documents have already been assigned to categories.

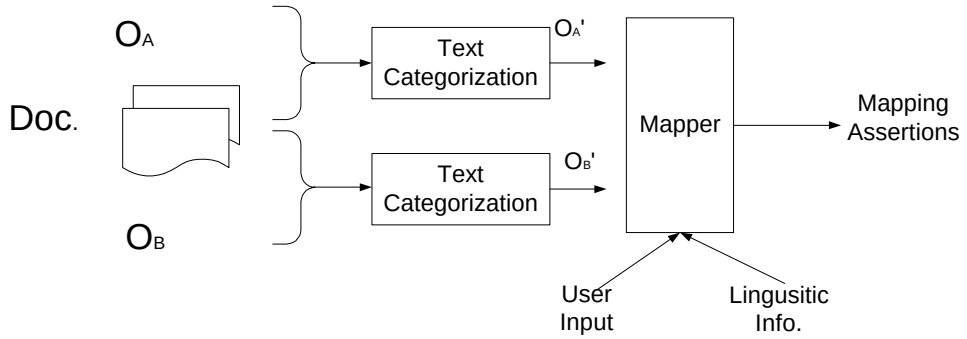

Figure 3 Architecture of Ontology Mapping

The second step takes the two intermediate ontologies as input and produces mapping assertions as the main output. The algorithm used in the Mapper is based on information retrieval techniques.

The intuition is that a feature vector for each node can be calculated based on the document assigned to it. Following a classic Rocchio algorithm[Aas99], the feature vector for node a_i is computed as the average vector over all document vectors that belong to node a_i. Following the same idea, the feature vector of any non-leaf node is computed as the centroid vector of all its sub nodes. By doing that hierarchical information can be taken into consideration to some extent.

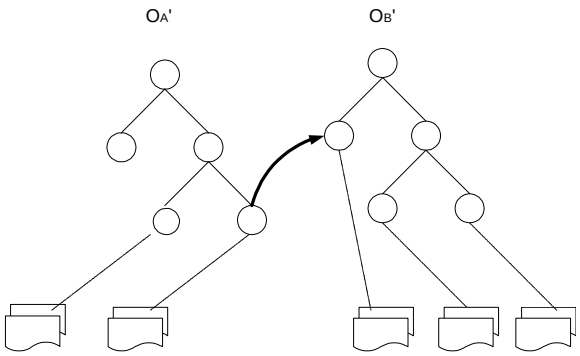

Figure 4 Ontology Mapping based on feature vectors for concepts.

With the feature vector at hand, we then measure the similarity of the two nodes by for instance the cosine measure of the two vectors or by using the Jaccard similarity measure.

As indicated in [Rahm01], using just one approach is unlikely to achieve as many good mapping candidates as one that combines several approaches. Therefore our approach should factor into other information as well. Among them are:

- Linguistic information. A matching algorithm of class names will be deployed to give a boost for nodes, which have the same or similar names (prefix, suffix, or word root) with the compared one. WorldNet can be used to provide synonym information.
- User specified information. User may supply further relationships about elements in the mapped ontologies. For example, user may explicitly define that concept a_i in ontology A is a broader concept of b_j in ontology B.

Therefore, the final approach would be a hybrid combination of several methods. The algorithm is semi-automatic since it produces a set of suggestions for possible correspondences, letting the user to be in control of accepting, rejecting or changing the assertions. Furthermore, the users will be able to specify mappings for elements for which the system was unable to find satisfactory match candidates.

4.3 applications

The approach can be used in different settings.

- Documents retrieval and publication between different web portals. Users may conform to their local ontologies through which the web portals are organized. It is desirable to have support for automated exchange of documents between the portals and still let the users keep their perspectives.
- Product catalog integration. In accordance with [Fensel01], different customers will make use of different classification schemas (UNSPSC, UCEC, and ECLass, to name a few). We need to define links between different classification schemas that relate the various concepts. Establishing such a connection helps to classify new products in other classification schemas.
- Service matching. Assuming there are some service description hierarchies (the MIT process handbook for instance) and that the provider and the requester are using different classification schemas. Imaging, some how, we can compute a feature vector for each node. Then the matching can be conducted by calculating the distance between the representative feature vectors.

5. Working schedules

Our way of working consists of the following phases.

1) Survey of ontology mapping methods and analysis of the ontology mapping process.
2) Survey of applicable parts of text categorization.
3) Development of an ontology mapping algorithm based on text categorization techniques.
4) Application of 3) in a case study
5) Analysis of empirical observations from 4) and evaluating its usage.

6. Useful resource.

McCallum, Andrew Kachites. "Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering." http://www-2.cs.cmu.edu/~mccallum/bow. 1996.

“Correspondence assertion metamodel.” In Sari’s phd thesis.

Reference:

[Aas99] K. Aas, and L. Eikvil: *Text Categorisation: A Survey*. Norwegian Computing Center, Oslo, 1999

[Agrawal01] R. Agrqwal and R. Srikant *On Integrating Catalogs*. Proceeding of the WWW-11, Hong Kong, 2001.

[Batini86] C.Batini and M. Lenzerini. *A comparative Analysis of methodologies for Database Schema Integration.* ACM Computer Surveys. 18(4). 1986

[Berners-Lee01] Berners-Lee, T., Hendler, J., Lassila, O. *The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities.* In Scientific American, Mai 2001. Online at: http://www.scientificamerican.com/2001/0501issue/0501berners-lee.html

[Fensel01] Fensel D., Ding Y., Omelayenko B., Schulten E., Botquin G., Brown M., Flett A.: *Procuct Data Integration for B2B E-Commerce*. IEEE Intelligent Systems 16, 2001.

[Hakkarainen99] S. Hakkarainen *Dynamic Aspects and Semantic Enrichment in Schema Comparison.* Phd. Thesis. Stockholm University, 1999

[Joachims98] Joachims, T. *Text categorization with support vector machines: learning with many relevant features.* Proceeding of European Conference on Machine Learning, 1998.

[Klein01] Klein, M. *Combining and relating ontologies: an analysis of problems and solutions.*In proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI-01), Workshop: Ontologies and Information Sharing, Seattle, USA, 2001.

[Lacher01] M.S. Lacher and G. Groh: *Facilitating the exchange of explicit knowledge through ontology mappings.* Proceeding of FLAIRS’2001, AAAI press, 2001.

[McGuinness00] McGuinness D., Fikes R., Rice J., and Wilder S. *:An environment for merging and testing large ontologies.* Proceedings of the 7th International Conference on Principles of Knowledge Representation and Reasoning. Colorado, USA.

[Noy00] Noy N. and Musen M. *PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment.* Proceedings of the AAAI-00 Conference. Austin, USA.

[ODP] http://dmoz.org/

[OntoMerge] http://cs-www.cs.yale.edu/homes/dvm/daml/ontology-translation.html

[Rahm01] Rahm E., Bernstein P.A. *A survey of approaches to automatic schema matching.* VLDB journal 10(4): 334-350, 2001

[Soelvberg98] Soelvberg A. *Data and what they refer to.* In Concept Modelling: Historical perspectives and future trends. In conjuction with 16th Int. Conf. On Concept Modelling. Los Angelse, USA, 1998.

[Stumme01] G. Stumme and A. Maedche *FCA-Merge: Bottom-up Merging of Ontologies.* Proceedings of the International Joint Conference on Artificial Intelligence IJCAI'01, Seattle, USA, 2001

[Su02] X. Su and L. Ilebrekke *A Comparative Study of Ontology Languages and Tools* In Proceeding of the 14th Conference on Advanced Information Systems Engineering (CAiSE'02), Toronto, Canada, May 2002.

[Uschold96] Mike Uschold and Michael Gruninger, *Ontologies: Principles, Methods and Applications.* Knowledge Enginerring Review 11(2), June 1996.

[Wache01] Wache, H., Vögele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., Hübner, S. *Ontology-Based Integration of Information -A Survey of Existing Approaches.* In proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI-01), Workshop: Ontologies and Information Sharing, Seattle, USA, 2001.

[Wiederhold99] Wiederhold, G., Mitra, P., Jannink, J., *Semi-automatic Integration of Knowledge Sources*. In proceedings the 2nd International Conference on Information Fusion (FUSION'99), California, USA, 1999.

[Yang99] Yang, Y and Liu, Y. *A re-examination of text categorization methods.* Proceedings of the 22nd Annual International ACM SIGIR Conference On Research and Development in Information Retrieval (SIGIR'99), 42-49 1999