# Web Taxonomy Fusion using Topic Maps-driven Ontological Concepts and Relationships

Ing-Xiang Chen[1] and Cheng-Zen Yang[2]

[1] Ericsson Taiwan Ltd., 11F, No.1 Yuandong Road,
Banqiao Dist., New Taipei City, Taiwan, 220, R.O.C.
ing-xiang.chen@ericsson.com
[2] Department of Computer Science and Engineering, Yuan Ze University,
135 Yuan-Tung Road, Chungli, Taiwan, 320, R.O.C.
czyang@syslab.cse.yzu.edu.tw

**Abstract.** Since most of the Web taxonomies and catalogs are organized in conceptual hierarchies, taxonomy fusion can be viewed as a specialized case of hierarchical ontology coalition in real-world applications. Hence, different kinds of semantic information can be further extracted to facilitate Web taxonomy fusion, such as intra-ontological concepts and inter-ontological relationships. This paper proposes approaches to effectively improve the accuracy of Web taxonomy fusion by using a taxonomy fusion model based on the ontological concepts and relationships of Topic Maps. Specifically, a novel fusion model based on inter-ontological mapping as well as intra-topic concept is presented to outperform a Naïve Bayes (NB) classifier and a Support Vector Machine (SVM) by 20% to 30% in $F_1$ measure over real-world Web taxonomies.

**Keywords:** Web taxonomy fusion, Topic Maps, Hierarchical ontology, Intra-ontological information, Inter-ontological information

## 1 Introduction

As more and more semi-structured digital contents are organized into taxonomy-based ontologies (i.e., hierarchical ontologies), information fusion on large-scale ontologies becomes an important issue in digital content management. Mergers and acquisitions among enterprises are practical examples in which the large amount of taxonomical semi-structured data of an enterprise is integrated into the categorized ontologies of another enterprise. For realizing the fusion work, the ultimate goal is to develop an integrated view with a single ontology or a small set of ontologies to which all partakers will conform as noted in [1]. However, a general fusion work on semi-structured ontologies faces several severe challenges. First, ontologies to be integrated may be created independently in reality, and thus exist in different formats or structures. Second, the semantical or conceptual diversity existing in different ontologies further complicates the ontology fusion problem.

Since the assertion of Topic Maps is to outline the real-world things into topics with names, and the ontological knowledge can be clearly described based

on the TAO spirit of Topic Maps without tagging, Topic Maps can convey the knowledge of resources through a view of virtual maps, in which the resource subjects and the relationships between them are distinctly depicted [2]. Topic Maps are thus suitable to facilitate semantic-level ontology integration and describe Semantic Web. In this paper, we propose a taxonomy fusion model based on the ontological concepts of Topic Maps, in which semantic concepts, hierarchical relations, and physical instances within the ontological knowledge are utilized to facilitate taxonomy integration. In advance, inter-ontological mapping relationships are studied to complement the loss of semantics in the hierarchical structure and enhance the semantic-level integration. An information fusion mechanism based on the Topic Maps-driven (TM-driven) ontological concepts is thus illustrated to facilitate Web information fusion with the scenario of taxonomy integration.

## 2 Related Work

In previous studies on Web taxonomy fusion, different kinds of implicit information embedded in the source taxonomy are explored to help information fusion. These implicit source features can be mainly categorized into four types. First, the co-occurrence relationships of source objects are studied to enhance a Naïve Bayes classifier based on the concept that if two documents are in the same source category, they are more likely to be in the same destination category [3]. Second, latent source-destination mappings are explored to improve the integration performance with an effective learning scheme in [4]. Third, a cluster shrinkage (CS) approach, in which the feature weights of all objects in a document category are shrunk toward the category centroid, is proposed [5]. Therefore, the cluster-binding relationships among all documents of a category are strengthened. Fourth, the parent-children information embedded in hierarchical taxonomies is intentionally extracted [6]. Based on the hierarchical characteristics, these approaches are extended to improve the integration performance.

Even though, the semantic information embedded in the source taxonomy has not been discussed in past studies. The semantic conceptual relationships existing in Web taxonomies are particularly ignored. This observation thus motivates us to study the effectiveness of intra/inter-ontological information for taxonomy fusion.

## 3 Topic Maps-driven Integration

We propose an integration model based on TM concepts to merge semantic resources from diverse sources. As described in the basic definition of TAO, semantic resources from different sources can be primitively transformed into TM concepts using topics, associations, and occurrences. In the following, both syntactic and semantic mappings are addressed to fulfill more comprehensive ontology matching.

### 3.1 Matching Semantic Resources

For syntactic mappings between TM concepts and general ontological knowledge, ontologies are directly mapped to TM based on the definition of TAO [2]. by individually mapping *concepts*, *relations*, and *instances* to *topics*, *associations*, and *occurrences*. For semantic mappings between different concept domains, the mappings between most similar semantic concepts need to be determined. Here, we define a knowledge ontology as a triple $O = (C, R, I)$, where $C$, $R$, and $I$ represent *concepts*, *relations*, and *instances*, respectively.

In our model, a similarity function $\delta$: $O_s \times O_d \rightarrow \mathbb{R}^s$ is used to express the similarity between two ontologies. Since each concept domain, $C \in O$, consists of its basic constitutes, namely, *instances*, the centroid of these instances can be used to represent the concept domain. We thus estimate $\delta(C_{s_i}, C_{d_j})$ by calculating the centroid similarity between $C_{s_i}$ and $C_{d_j}$ and counting the correct and incorrect mapping number of $I$ between each concept domain $C_{s_i} \in O_s$ and its corresponding concept domain $C_{d_j} \in O_d$. Here, $\delta(C_{s_i}, C_{d_j})$ is determined by the similarity between the centroids and the the number of correct mapping instances. Ontology matching between $O_s$ and $O_d$ can be therefore determined.

### 3.2 Integration Process

In the TM-driven ontology integration procedure, *intra*-TM and *inter*-TM information is employed to extract ontological knowledge and find the concept mappings.

**Integration with Intra-TM Concepts** To establish the internal semantic concepts, a weighting scheme is designed to control the impact of the semantic concepts of each hierarchical level. Equation 1 calculates the enhanced feature weight of each instance $I$, where $L_k$ is the relevant concept feature weight assigned as $1/2^k$ with a $k$-th level depth, $L_x$ denotes the hierarchical weight of the concept feature $x$, $f_{x,I}$ is the original weight of feature $x$, $f'_{x,I}$ represents the enhanced feature weight, and $\lambda$ is used to control the magnitude of relation. The weight of each hierarchical concept is exponentially decreased as $1/2^k$, since the higher-level concept domains shall have weaker semantic relationships to the concept domain where the instance is located. The weight $f_{x,I}$ is assigned by $TF_x / \sum TF_n$, where $TF_x$ is the term frequency of $x$, and $n$ denotes the number of the stemmed terms in each instance. The feature weight $L_k$ of each concept domain is exponentially decreased and accumulated based on the increased levels.

$$f'_{x,I} = \lambda \times \frac{L_x}{\sum_{k=0}^{n} L_k} + (1 - \lambda) \times f_{x,I} \tag{1}$$

**Integration with Inter-TM Concepts** In the TM-driven model, inter-Topic Maps information (inter-TM) is further extracted to enhance the semantic-level

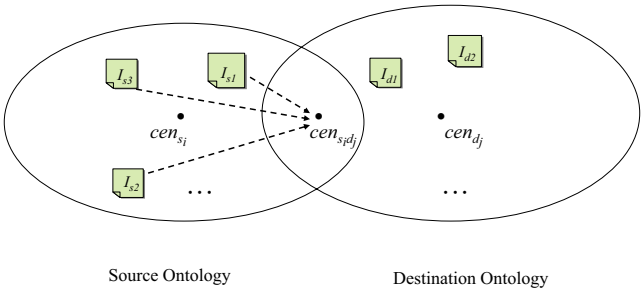


**Fig. 1.** The idea of common concept shrinkage (CCS).

ontology integration. The main idea is to augment the feature space of each source instance with the mapping destination concepts to help the corresponding integration process. In the augmentation process, our model first calculates the similarity between the source concepts and the destination concepts using a cosine similarity function to find the potential augmentation mappings. Each semantic concept is represented by its centroid which is calculated by averaging the feature vectors associated with the instances in the concept domain. To obtain the centroid of the set of instances in each concept, the instances are transformed into feature vectors, and each feature weights are averaged by the number of instances in the concept domain, namely, $cen = \frac{1}{|C|}\sum_{I \in C} I$, where $|C|$ represents the number of instances $I$ in the concept domain $C$.

For each source concept ($C_{s_i}$) and each destination concept ($C_{d_j}$), we obtain $cen_{s_i}$ and $cen_{d_j}$ to represent $C_{s_i}$ and $C_{d_j}$, respectively. Each destination concept $C_{d_j}$ will have a mapping relationship to a source concept $C_{s_i}$ that has the highest similarity to $C_{d_j}$. In order to obtain the common conceptual information of $C_{s_i}$ and $C_{d_j}$, the instances of $C_{s_i}$ are shrunk to $cen_{s_id_j}$ using the cluster shrinkage algorithm in [5]. The idea of common concept shrinkage (CCS) is illustrated in Figure 1, and its algorithm is depicted in Figure 2.

---

for each pair of mapped concepts $C_{s_i}$ and $C_{d_j}$ {
  compute its centroid: $cen_{s_id_j} = \frac{1}{|C_{s_i}|}\sum_{I_{s_i} \in C_{s_i}} I_{s_i}$;
  for each instance $I_{s_i} \in C_{s_i}$ {
    replace it with $I'_{s_i} = \alpha \cdot cen_{s_id_j} + (1-\alpha)I_{s_i}$,
      where $0 \leq \alpha \leq 1$;
  }
}

---

**Fig. 2.** The CCS algorithm for each pair of mapped concepts.

**Table 1.** The experimental categories and the numbers of documents.

| | Y Class | Y Test | G-Y | G Class |
|---|---|---|---|---|
| Autos | 25 | 416 | 1,090 | 14 |
| Movies | 27 | 1,311 | 5,181 | 27 |
| Outdoors | 26 | 188 | 2,391 | 23 |
| Photo | 22 | 201 | 612 | 9 |
| Software | 16 | 675 | 5,719 | 59 |
| Total | 116 | 2,791 | 14,993 | 132 |

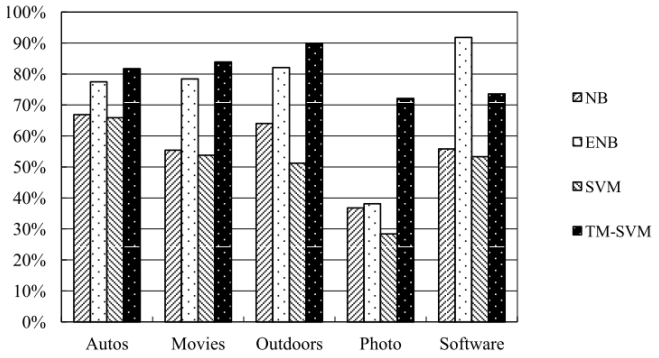


**Fig. 3.** Performance comparison of different models in Y→G integration.

## 4 Applications on Web Taxonomy Fusion

In the current Web environment, many information resources exist in hierarchical relationships or tree structures such as Web taxonomies and enterprise catalogs. When the nature of Web taxonomy is further considered, a Web taxonomy is essentially constituted of concepts (categories), relations (hierarchical structures), and instances (documents). Therefore, the proposed TM-driven integration model can be intuitively transformed to a hierarchical taxonomy integration scheme for Web taxonomy fusion.

In the experiments, five directories from Yahoo! (Y) and Google (G) were extracted to form two experimental taxonomies. Table 1 shows these directories and the number of the extracted documents after ignoring the documents that could not be retrieved. The documents appearing in only one category were used as the training data (|G-Y|), and the common documents (Y∩G) were used as the testing data (|Y Test|). In the experiments, we measured the integration performance with $F_1$ measure. We have conducted experiments on integrating Yahoo! to Google (Y→G) using the TM-driven model. The experimental results show that using SVM with both *intra*-TM and *inter*-TM conceptual semantic information (TM-SVM) can significantly and consistently improve the integration performance of SVM in most cateogries in Y→G fusion.

The experimental results show that TM-SVM can get the most improved integration performance using the TM-driven model with $\lambda$=0.1 and $\alpha$=0.25. To

demonstrate the performance of the TM-driven model, TM-SVM is further compared with the enhanced Naïve Bayes (ENB) model [3] over the five categories. According to the performance comparison in Figure 3, TM-SVM performs best on average and outperforms ENB in four out of the five categories. In "photo", TM-SVM has a significant boost because *inter*-TM can find correct mappings between subcategories with more true positive documents, such as "photographers", and "techniques_and_style", and reduce the number of false positives as well. In further analysis of ENB, ENB performs best only in "software" because over 80% documents belonging to "software" are distributed in three main subcategories, namely, "desktop_customization", "operating_systems", and "internet", in both source and destination taxonomies. In such a case, the performance improvement of the TM-driven model is hence not as significant as ENB.

## 5 Conclusion

Although different methodologies have been studied for Web information fusion, integrating and merging semantic resources is still a major challenge and research issue for Web information fusion because of the large-scale, dynamic, heterogeneous, and hyperlinked nature. In this paper, an integration mechanism based on the TM-driven framework is proposed for Web information fusion. Augmented with both the intra-category and inter-category features, the resources in the source category can be more precisely integrated into the correct destination category to advance Web information fusion. The experimental results show that *inter*-TM-SVM can achieve the best $F_1$ score in most cases in Y→G taxonomy fusion. The results also show that the ontological concepts and relationships extracted by the TM-driven integration model can help $F_1$ enhancements in a significant portion of all cases.

## References


1. Alasoud, A., Haarslev, V., Shiri, N.: A Hybrid Approach for Ontology Integration. In: Proceedings of the 31st VLDB Conference, Trondheim, Norway (2005)
2. Garshol, L.M., Moore, G.: Topic Maps—Data Model (2008)
3. Agrawal, R., Srikant, R.: On Integrating Catalogs. In: Proceedings of the 10th International Conference on World Wide Web (WWW 2001), Hong Kong (2001) 603–612
4. Sarawagi, S., Chakrabarti, S., Godbole, S.: Cross-training: Learning Probabilistic Mappings Between Topics. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (2003) 177–186
5. Zhang, D., Lee, W.S.: Web Taxonomy Integration Using Support Vector Machines. In: Proceedings of the 13th International Conference on World Wide Web (WWW 2004). (May 2004) 472–481
6. Wu, C.W., Tasi, T.H., Lee, C.W., Hsu, W.L.: Web Taxonomy Integration with Hierarchical Shrinkage Algorithm and Fine-grained Relations. Expert Systems with Applications **35**(4) (2008) 2123–2131