A Semantic Category Matching Approach to Ontology Alignment

Tadashi Hoshiai[1], Yasuo Yamane[1], Daisuke Nakamura[2], and Hiroshi Tsuda[1]

[1] Fujitsu Laboratories Ltd., I. T. Media Laboratories,
211-8588 Kawasaki-shi, Kanagawa, Japan
{hoshiai, yamane.yasuo, htsuda}@jp.fujitsu.com
[2] Kyoto University, Graduate School of Informatics,
606-8501 Sakyo-ku, Kyoto, Japan
daisuke@lab7.kuis.kyoto-u.ac.jp

Abstract. We applied our semantic category matching (SCM) approach to the EON ontology alignment contest problems. Our approach found pairs of semantically corresponding categories from two different classification hierarchies such as Yahoo, based on natural language processing, similarity searching of huge vector spaces, and structural consistency analysis. The EON Contest's random name problems (#201, #202) could not be solved using conventional character string resemblance techniques. However, when we applied SCM to these problems, the results showed that SCM had improved the accuracy as compared to the conventional method (F-measure: 0.021=>0.949, 0.021=>0.580). Moreover, SCM exceeded the accuracy average in all problem areas by over 10 % as compared to conventional methods.

1 Semantic Category Matching

1.1 Outline

We applied semantic category matching (SCM) technology to the ontology alignment problem. Our method found pairs of semantically corresponding categories between two different classification systems. In the integration and interoperation of classification systems, this kind of technology is important. However, there are problems that cannot be solved using only the character string resemblance method because of the difference in the category names, category granularity, and the classification hierarchy formation principles.

Related works of SCM technology are the enhanced Bayes classification method by Agrawal [1] and the Identity test method by Ichise [2]. Agrawal's work is a content-oriented statistical approach, as much as ours. However, his work does not look at the entire hierarchical structure, and so therefore it is not suitable for large hierarchy classification systems. Ichise's work is not content-oriented approach but an extension-oriented approach, based on URL identification in web directories. However, we think that content-oriented approach is necessary for semantic analysis of text information. Furthermore, their works did not treat structural consistencies between the results of the category correspondence and hierarchical structures. Since these points are important for large system services and semantic approaches, we incorporated them into our method.

Semantic category matching is based on a statistical approach that takes sample documents from each category and hierarchical structure description data, and outputs all category pairs that semantically correspond with the two classification systems. Ontology alignment is a problem designed to find couples of corresponding classes.

While the purposes of SCM and ontology alignment are different, the problem structures of both are similar to each other, from the perspective of alignment between the hierarchical structures. Therefore, we applied our new SCM technology to problems that could not be solved by usual methods.

1.2 Elemental technology

We used the following elemental techniques, which we will explain sequentially, in SCM. An outline diagram of SCM is shown in Figure 1.

1) Hierarchical version of keyword extraction,
2) Similarity search category similarities, based on oblique coordinates and,
3) Structural consistency analysis.

1.2.1 Generating category feature vectors by hierarchical keyword extraction

A keyword extraction technique statistically analyzes documents classified by category. It finds keywords, which are words that occur frequently in the documents in a specific category, but exclude common frequently occurring words that appear regularly in other categories but that have a weak relationship to the category.

In keyword extraction technology, the following premises are given:

- High statistical correlation between word occurrences in the document and their classification categories.
- Sufficient classified documents to do a statistical analysis.
- Highly correlated nouns within a category.
- Subcategory word occurrence characteristics are succeeded by super categories along the classification hierarchy.

Under the above premises, the keywords are extracted automatically, based on the statistical correlation between the document's topic category and the word occurrences. In this case, we can select criteria measures that highly evaluate only words that have a high correlation to a specific category. In our research, we used Kullback-Leibler's information as follows:

$$P(w|C)\cdot\log\frac{P(w|C)}{Q(w)}$$

w: word, C: category, P(w|C): word occurrence probability in category C
Q(w): average word occurrence probability of every brother-category of C

The keywords are extracted according to the following procedures.

input document => morphological analysis => remove unnecessary words
=> count words => total word occurrences in each category
=> inherit subcategory's statistical characteristics to super-category
=> select higher-ranking words in each category

Finally, the words whose value of the criteria measure is higher (for instance, the higher 30 words) in each category are selected as keywords. Then, the category feature vectors, based on the word occurrence characteristics of each category, were output. Moreover, because the subcategory feature is weighted and inherited to a super category, the neighborhood of the classification's hierarchical structure was reflected in the features of the keywords, and the distance in the vector space.

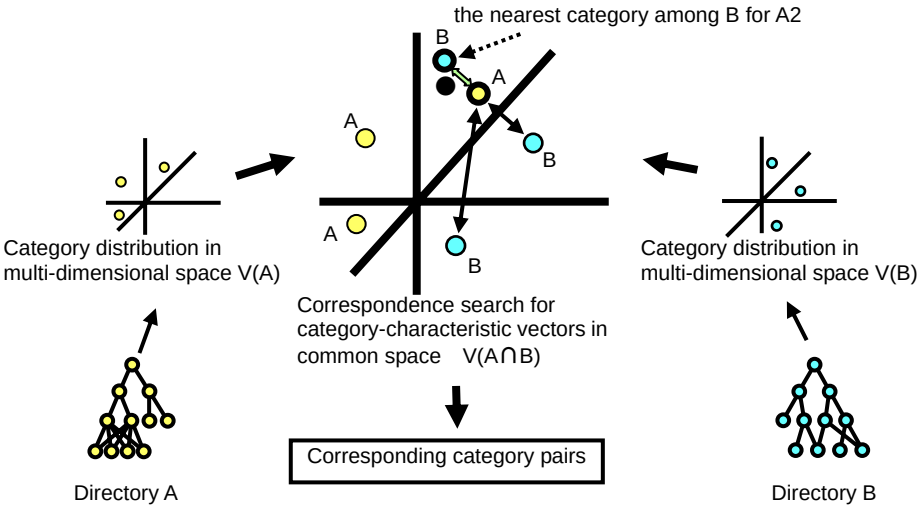

Figure 1. Outline of semantic category matching

1.2.2 Similarity category search based on oblique coordinates

Because each vector space (V(A) and V(B) in Fig. 1) formed by the two different classification systems generally has different coordinate systems (each axis coordinate corresponds to a keyword), we created and used a common feature vector space ($V(A \cap B)$ in Fig. 1) with vocabulary common to both systems, to compare the category vectors.

By using an oblique coordinate system [3] that reflects the correlationship between the keywords, the similarities in the word meanings can be woven into the coordinate system. As a result, the vector space becomes semantically more natural. For example, consider these keywords “sports”, “Olympics”, and “weather”. Because the first two words are more similar, the corresponding coordinate axes are set more closely than the last one. For each category in one classification system, the nearest neighboring categories of the other systems are output according to the order of their angle distances.

1.2.3 Structural consistency analysis

Structural consistency analysis focuses on whether the category couples correspond between the two classification systems and is formed naturally within their hierarchical structures, or not. After category couples are mutually and independently selected by a similar category search as the candidates of the results data of the entire system, it is necessary to decide which category couples are natural structural correspondences.

Figure 2 shows the naturalness of the correspondence between two hierarchical structures. When we call one of the category couples the reference couple (a solid-line arrow in Fig. 2), we can evaluate whether the neighboring couples (dotted-lines arrows in Fig. 2) can be placed well (or badly) in the hierarchical structure on both sides, by comparing them to the reference couple. Here, assuming that the categories of a reference couple are ‘a’ and ‘b’ in Fig.2, 'neighbor couple' indicates the category couple, whose category is near as link distance to another category. The link distance indicates the number of subcategory links that can be joined to ‘a’ or ‘b’.

The neighbor couple that contains subcategory ‘a1’ and ‘b1’ in Figure 2, is consistent with the reference couple (‘a’ and ‘b’) with respect to their hierarchical structures, because the category ‘a1’ is a subcategory of category ‘a’ and category ‘b1’ is subcategory of the category ‘b’. Conversely, the neighboring couple that contains subcategory ‘a2’ and ‘b2’ in Figure 2, the couple is not consistent with the reference couple with respect to their hierarchical structures, because category ‘a2’ is a super category of category ‘a’ and the category ‘b2’ is subcategory of the category ‘b’. If the degree of this consistency is provided according to a suitable measure, the structural consistency of the reference couple is obtained as the average consistency of all the neighboring couples. Finally, the structural consistency of the entire SCM is obtained as the structural consistency average of all the reference couples.

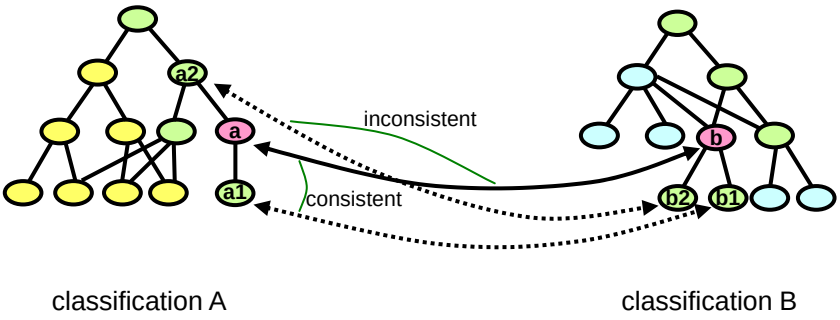

Figure 2. Structural consistency analysis

1.3 Adaptation for Semantic Category Matching

1.3.1 An SCM approach to the ontology alignment problem

Because both techniques have the same common structure from the point of view of correspondence between two hierarchical structures, we thought that we could apply SCM technique to ontology alignment (OA), even though the purpose of SCM was originally different from the purpose of OA.

Difference between ontology alignment and category matching:

Because the description unit for OA is class (or instance), and the description unit for an SCM is the category (object domain) of document topics, the granularity of OA is smaller than SCM. Furthermore, in OA, properties for both object's attributes and relations between objects, and also, the restricted condition of property can be described. On the other hand, we cannot describe any predefined logical relationships between any of the parts of a document in SCM, but XML tag's roles. Thus, the information described in OA is more detailed than the information described in SCM.

The idea of application of SCM to OA:

The class-instance relationship is common to both techniques: therefore if we interpret 'class' in ontology as 'category' in SCM, and interpret 'instance' as 'sample document', and the ontological description information is converted into a category name, the document ID, the category hierarchy relationship, and tag structure of XML documents in SCM, we can extract suitable keywords from the text of suitably selected tags in XML documents.

1.3.2 Outline of application of SCM to OA

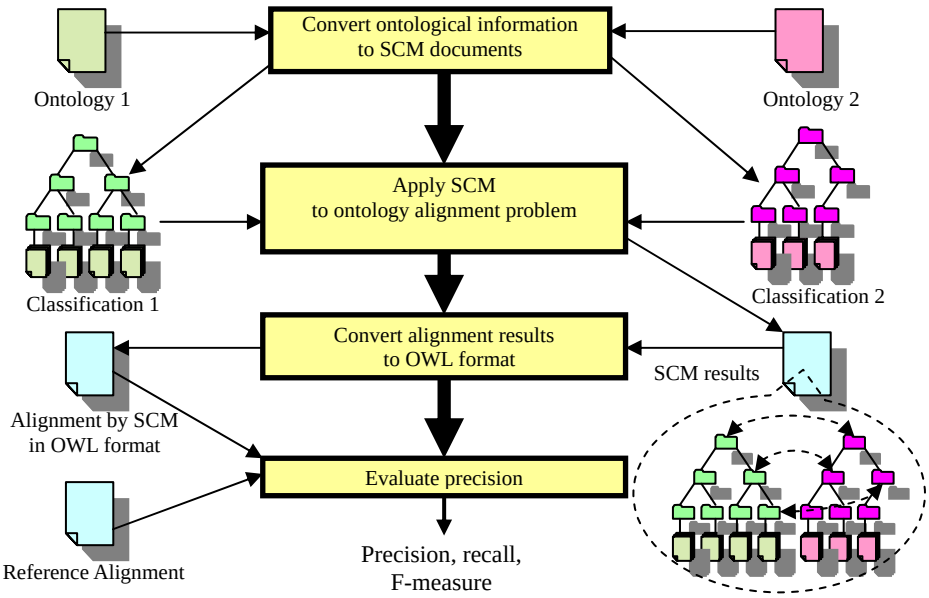

Figure 3. Application flow of SCM to OA

An outline of SCM application to the ontology alignment problem is as follows: (see Figure 3)

- The descriptions of two ontology sets (reference and target ontology) are converted into XML document sets in two classification systems and two category hierarchies in SCM format. (see Figure 4)
- SCM is applied to 2 sets of converted data to resolve the OA problem.

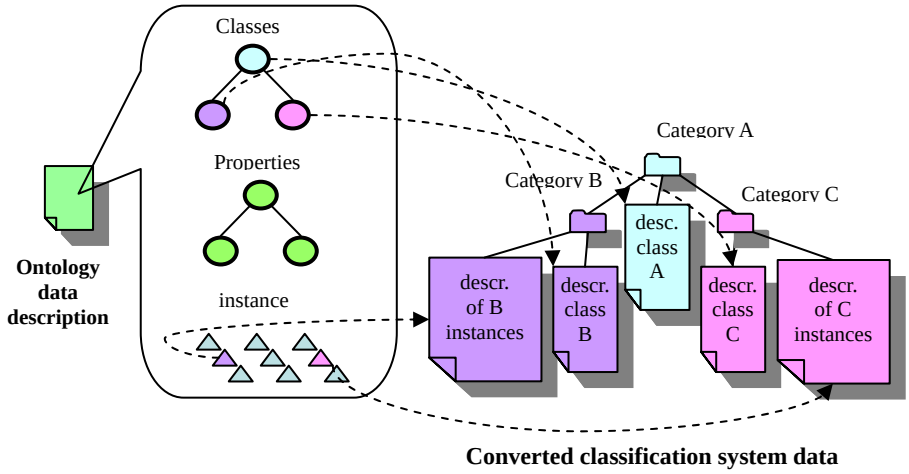

Figure 4. Conversion from ontology description to SCM input data

- SCM outputs a set of category-pairs that indicate the alignment between the two classification hierarchies. The results data are described in Ontolingua langage.
- The SCM results data are converted into the OWL alignment form of the EON contest.
- Finally, the alignment result accuracies (F-measure etc.) are calculated by ontoalign, the ontology alignment evaluation tool prepared by EON's promotion division.

2 Results

In first experiment, we applied SCM (version 1) to first version of contest test data, as much as string-based alignment method included in the ontoalign evaluation tool. Because these test data included bugs, we had to modify these data for enabling execution of programs, and so results data seem to be under a little influence of these modification. The accuracy data (F-measure) of results for applying SCM to each problem are listed in Table 1, along with the results of standard string-based alignment method.

Table 1. F-measures results of SCM and string-based methods in first experiment

test no.	101	102	103	104	201	202	204	205	206	221
String-based	.938	NaN	.948	.948	.021	.021	.753	.344	.423	.948
SCM v1	.990	NaN	.970	.980	.870	.500	.829	.579	.687	.909
Difference	.052	0	.022	.032	.849	.479	.076	.235	.264	-.039

test no.	222	223	224	225	228	230	301	302	303	304
String-based	.897	.897	.938	.948	.917	.854	.593	.411	.510	.804
SCM v1	.924	.916	.957	.978	.899	.890	.729	.468	.400	.820
Difference	.027	.019	.019	.030	-.018	.036	.136	.057	-.110	.016

Note: 'NaN' (not the answer) indicates 'division by zero' (the alignment number is zero) for calculating the F-measure. These are proper results, because test #102 has no proper alignment of classes or properties, therefore we can replace 'NaN' with 1.0 (that is, the results of the alignment are proper.).

After final revised version of test data was disclosed, we applied the little revised SCM (version 1.1) to the final version of test data in the second experiment. The results are listed in Table 2.

Table 2. F-measures results of SCM in second experiment

test no.	101	102	103	104	201	202	204	205	206	221
SCM v1.1	.995	NaN	.995	.995	.949	.580	.933	.699	.584	.925

test no.	222	223	224	225	228	230	301	302	303	304
SCM v1.1	.955	.908	.995	.995	.941	.953	.755	.468	.505	.886

For comparison between each method, the polygonal line graphs on results data of both experiments are shown in Figure 5.

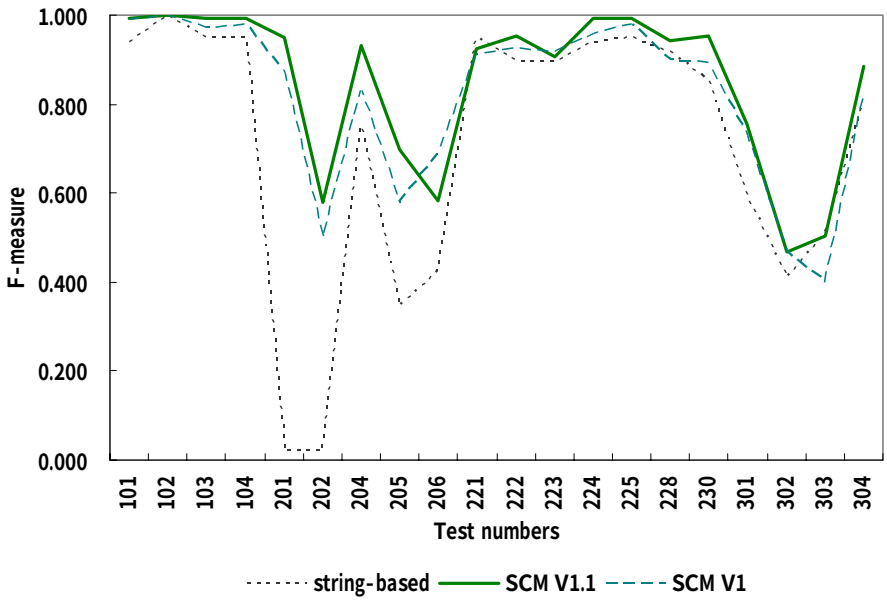

Figure 5. F-measure results of SCM and string-based method

2.1 Concept test (test no.: #101, #102, #103, and #104)

Test number 101 is a comparison test of the same ontology, and number 102 is a comparison test of quite different ontological domains (bibliography and food). The ontological structures of numbers 103 and 104 are close to that of number 101.

In test #102, results of both methods are exactly matched, that is zero alignment.

As the string-based method is based on the agreement/disagreement of the name character strings, whether or not the class names in the reference ontology are the

same as the class names in the target ontology, this method is suitable for this kind of tests. In the tests #101, #103, and #104, alignment result accuracies from the string-based methods attached to the ‘ontoalign’ evaluation tool were close to 100%.

Moreover, the SCM alignment results of other tests were generally superior to the results of the string-based methods.

By the way, the revised SCM (v1.1) results (0.995) included the alignment couple between ‘language’ property in reference ontology and ‘language’ property in target ontology. Though we think that this result is proper and results become 1.0, but the final version of reference alignment file (refalign.rdf) does not include this alignment couple.

2.2 Name diversity test (#201, #202, #204, #205, and #206)

The problems we focused on in this paper are those that involve naming diversity. These cannot be solved by string-based methods. The other hand, SCM is a content-oriented approach, and can solve these problems using content similarity between the semantically same classes in different ontologies. That is, even if there is a disagreement in the class names between both ontologies, when the description data of the classes and the instances belonged to their classes were statistically and structurally similar, we could obtain an ontological alignment. Semantic similarity of properties can be discussed as well as semantic similarity of classes.

In random name tests (#201, #202), there was no similarity in the name character strings between reference ontology and target ontology, and so the string-based method results were almost 0%. In contrast to this, the results of SCM was 87% and SCM v1.1 improved to 94.9% in the test #201 where comment sentences were available, and was 50.0% and improved to 58.0% in the test #202 where no comment sentences were available.

In test #204, #205, and #206, SCM v1.1 exceeded the string-based method by over 10%.

2.3 Hierarchy variation test (#221, #222, and #223)

In the no hierarchy test (#221), the SCM results fell below the string-based method results by a few percentage points. Conversely, in the flattened hierarchy test (#222) and the expanded hierarchy test (#223), the SCM results exceeded the string-based method by a few percent.

In SCM, because the hierarchical relationship of ‘subClassOf' is reflected in the calculation of the category feature vectors of both a super class and a subclass, the feature vectors are distributed close to each other in the vector space, even if the names in these classifications have no common character strings. In test number 221, we think the accuracy fell because of lost information in this hierarchical relationship.

2.4 Other systematic tests (#224, #225, #228, and #230)

In the no instances (#224), no restrictions (#225), no properties (#228), and flattened

entities (#230) tests, the SCM v1.1 results exceeded all of results of the string-based method by several percent.

2.5 Real ontology test (#301, #302, #303, and #304)

In BibTex/MIT test (#301), the SCM results exceeded the string-based method results by 10% or more. In the BibTex/UMBC test (#302) and BibTex/INRIA test (#304), the SCM results exceeded our expectations by several percent. In the BibTex/Karlsruhe test (#303) the SCM results fell by only 0.5%.

3 General comments

3.1 Results (strength and weaknesses)

Strength:
When there are semantically similar classes between both ontologies, even if the name of the class in one kind of ontology is different from that in another, SCM can find correspondences of these classes in both ontologies.

Weakness:
When there is little common vocabulary between the ontologies, there is a possibility that the system cannot identify the semantically similar category vectors in the feature vector space. (For example, the case there is no similarities in the instance description data.)

3.2 Improving the proposed system

Stemming:
Because we don’t process English stemming now, SCM cannot absorb inflection variations of English words (-s, -es, -ing, -ed, -er, -est, etc.). It is true that this changes the original word into a different one, thus decreasing the accuracy rate, but this influence is reduced by effects of correlation between semantically similar words in oblique coordinate vector space. Consequently we will use the stemming function in our system in the future. In our experiments, we performed Japanese morphological analysis and stemming.

3.3 New measures proposed

Path-weighted accuracy (P-measure):
Currently, we obtain an incorrect answer (accuracy 0) if the intended class is not described in results data. If there is correspondence between two classes that are semantically unrelated to the intended class, and the correspondence between two classes that are closely related to the intended class, it is clear that the latter

performance will be better than the former. Therefore, if we use the number of links between two classes of ‘subClassOf’ and define the semantic closeness between classes r (0 =< r =< 1) as accuracy, then the overall accuracy can be calculated as the average of all of accuracies of the correct answers).

4 Raw results

4.1 Links to the set of provided alignments

Currently, our company does not permit public access to URLs containing the alignment results data files.

5 Conclusions

We showed that there were large improvements in the accuracy during experiments when our category matching technology was applied to difficult ontological alignment problems, such as naming diversity. The EON contest’s random name problems (#201, #202) were difficult to solve using conventional techniques, based on character string resemblance. However, when we applied our category matching method, the SCM accuracy results showed some improvement over conventional methods (F-measure: 0.021=>0.949, 0.021=>0.580). Moreover, in all tests, the accuracy average surpassed that obtained in conventional tests by over 10 % on average.

In the future, I want to work on other ontology alignment problems and improve the accuracy of category matching technology much more.

Acknowledgements
I would like to thank Tomoya Iwakura for his suggestions and support on our program development.

References

1) Agrawal, R. and Srikant, R.: On Integrating Catalogs, in Proceedings of the Tenth International World Wide Web Conference (WWW-10), (2001) 603-612.
2) Ichise, R., Takeda, H. and Hon'iden, S.: Learning on the adjustment rules between hierarchical knowledge, Journal of JSAI, vol.17, no.3-F (2002) 230-238.
3) Yamane, Y., Hoshiai, T., Tsuda, H., Katayama, K., Ohta, M., Ishikawa, H.: Multi-Vector Feature Space Based on Pseudo-Euclidean Space and Oblique Basis for Similarity Searches of Images, in Proceedings of the First International Workshop on Computer Vision meets Databases (CVDB 2004), (2004) 27-34.
4) Hoshiai, T., Yamane, Y., and Tsuda H.: Category-level Retrieval among Heterogeneous Information Sources based on Category Matching, in Proceedings of the 6th SANKEN (ISIR) International Symposium, Osaka (2003) 73-76.