Semantic Matching based on Enterprise Ontologies

Andreas Billig, Eva Blomqvist, and Feiyu Lin

Jönköping University, Jönköping, Sweden
{bill|blev|life}@jth.hj.se

Abstract. Semantic Web technologies have in recent years started to also find their way into the world of commercial enterprises. Enterprise ontologies can be used as a basis for determining the relevance of information with respect to the enterprise. The interests of individuals can be expressed by means of the enterprise ontology. The main contribution of our approach is the integration of point set distance measures with a modified semantic distance measure for pair-wise concept distance calculation. Our combined measure can be used to determine the intra-ontological distance between sub-ontologies.

1 Introduction

Our approach presented in this paper is a step towards reducing the information overload and assisting the human user in processing and evaluating the information entering a company from the surrounding world. An enterprise ontology represents, among other things, the business interests and processes of the company, and the information demand of individual users can be expressed in terms of this enterprise ontology. The core algorithm needed for evaluating how distant (or similar) two parts of the enterprise ontology are is the main focus of this paper, in order to support a relevance evaluation of incoming information.

The following section presents background and motivation. In Section 3 the problem is outlined and our semantic distance algorithm is described in detail. Section 4 focuses on a small example to illustrate the usefulness of the method. Finally, in Section 5 some conclusions are drawn and future work is outlined.

2 Background

This section describes the motivation and background of our approach, for example the notion of semantic distance. With respect to ontologies we adopt the classic definition from [1], describing an ontology as a formal explicit specification of a shared conceptualisation. In our research we do not, at this time, restrict ourselves to a specific ontology formalism, but assume the possibility to reduce the ontology to a semantic net-like structure (i.e. a labelled directed graph). Our research focuses mainly on domain and application ontologies within enterprises,

i.e. an ontology describing the necessary views and concepts, with the intention of structuring and retrieving information.

The term semantic distance is in this paper used as the inverse of semantic similarity. For semantic distance (or similarity) of ontologies many measures exist to measure distance of concept pairs within one ontology (see the survey in [3]). As described in [6] there are edge-based, information content-based, and feature-based approaches. We want to rely only on the ontology without incorporating documents for distance measurement, thus we do not consider information content approaches. Feature-based methods focus on property definitions. Our approach is based on semantic network-like representations where taxonomies play a more crucial role than property definitions. Still we incorporate user defined relations as well, but we focus on edge-based methods, as the one in [5].

Different aggregation schemes can then be applied to determine distance between sets of objects. Most of these assume a way to determine the distance of two individual objects, from the two sets. In [7] different distance measures are discussed and compared, as well as in [8]. There exist measures such as the simple Hausdorff metric, considering only the most extreme points in the set, and others such as the family of optimal mappings, also incorporating the cardinalities of the sets.

Document relevance in classical IR is commonly measured with respect to some query (a set of keywords), and the document is treated as a bag-of-words [9]. Some systems also use ontologies, as in [10] where RDF annotations are used to represent the meaning of documents. Also in [11] documents are ranked according to an ontology. However, the annotations for the documents are manually defined, in our case we need to avoid manual annotation. Therefore we choose to align the representation of the document to the ontology automatically, using simple string matching (see [2]).

3 The Semantic Matching Approach

This section describes our approach to semantic matching, calculating the semantic distance between two parts of an ontology.

3.1 General Framework

First of all, we rely on an enterprise ontology (see Figure 1). Secondly, we assume that for all users a profile, a set of concepts of the enterprise ontology, exists. As documents, the approach may deal with any kind of machine processable information that can be mapped to a list of terms. The matched part of the term list together with the enterprise ontology form the extended enterprise ontology (XO, as described later in Section 3.3).

To calculate the relevance of a document to a user's interests the following process steps are needed, (1) processing of the document to construct a corresponding term list, (2) aligning the term list to the enterprise ontology (constructing the XO), and (3) calculating the semantic distance.

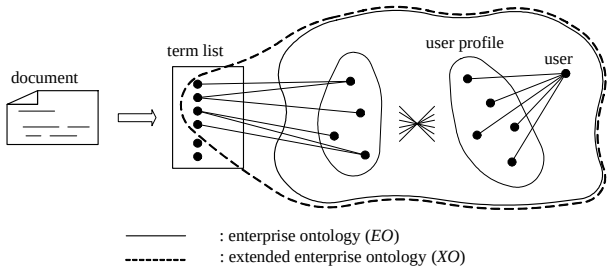

Fig. 1. Relevant information objects.

3.2 Preliminaries

Representing an ontology as a triple set is adequate for our purposes and aligned to RDF [12]. Because there is no need to consider literals here, we restrict ourselves to the set *Res* of (RDF-)resources (subsequently called *concepts*).

An *ontology* is a set of triples (s, p, o) over *Res* where s, p, o is called *subject*, *property*, and *object*, respectively. Let O be an ontology, then *root* denotes the taxonomic root, *Prop* denotes the set of properties occurring in O and $propCard(s, p) = \big|\{o \mid (s, p, o) \in O\}\big|$. An *(ontology) path of length n between concepts x and y* is a sequence of triples $(x, p_1, o_1), \ldots, (s_n, p_n, y)$ where $o_i = s_{i+1}$ for $i = 1, \ldots, n-1$. If triple t is within a path p we also write $t \in p$. *Path* resp. $Path(x, y)$ denote the *set of all paths* resp. the *set of paths between x and y*. A *triple weight* is a function from O to $\mathbb{R}$. A *path weight* is a function from *Path* to $\mathbb{R}$. For a triple weight f and path x the *canonical path weight* κ_f is defined by $\kappa_f(x) = \sum_{i=1,\ldots,n} f(x_i)$ where $x = x_1, \ldots, x_n$. The normalisation of the canonical path weight is defined by $\kappa_f^*(x) = \frac{\kappa_f(x)}{m}$ where $m = max_{y \in Path}\ \kappa_f(y)$. Note that κ_f^* is normalised if f is normalised. The *minimal path weight between x and y according to path weight f restricted by property set R* is defined as: $MinPathWeight(x, y, f, R) = min_{p \in P}\ f(p)$ where $P = \{p \in Path(x, y) \mid \forall (u, v, w) \in p\ .\ v \in R\}$.[1] A *property weight range* is a function $f : Prop \longrightarrow \mathbb{R} \times \mathbb{R}$ where $r_1 \leq r_2$ holds for every $(r_1, r_2) \in f(Prop)$.

3.3 Process Steps

The first step of the process is generating the term list, which is done through standard text processing methods (mainly tokenisation, stop-word removal and stemming, see [9]). The following steps of the process are described below.

Term List Alignment The term list is related to the enterprise ontology through string matching (see [2]). For this purposes we choose a threshold $\mathtt{t}$ for the string similarity level. Assuming that $L = TL(d)$ is the term list of

[1] If there is no property restriction, i.e. $R = Prop$, we write $MinPathWeight(x, y, f)$.

processing document d, the *string match of* L is defined by $SM(L) = \{x \in L \mid \exists y \in EO \;.\; \sigma(x,y) \geq \mathsf{t}\}$, where σ is any string matching method that deliverers values between 0 and 1. Then the XO is constructed with the matched terms of the term list as concepts and σ as an ontological relation. The result is the following ontology, where sm stands for the string match relationship:

$$XO = EO \cup \{(x, \mathsf{sm}, y) \mid x \in SM(L),\ y \in EO\}$$

Semantic Distance Next, is the calculation of the semantic distance between $SM(L)$ and the users' interest profiles. The interest profile of user u is a set of concepts, i.e. $UP(u) = \{r \mid (u, interestedIn, r) \in EO\}$.

Among the edge-based methods the one proposed by Sussna in [5] includes the facility to incorporate user defined relations as well as taxonomic relations, and additionally allows for weighting of relations. Our adaptation of this to the RDF environment is the triple weight ω:

$$\omega(x,p,y) = \frac{w(x,p) + w(y,p)}{2 * max(l(x), l(y))} \;,$$

$$w(x,p) = max_p - \frac{max_p - min_p}{propCard(x,p)} \text{ , and}$$

$$l(x) = length\big(MinPathWeight(x, root, f, \mathsf{rdfs{:}subClassOf})\big) \;,$$

where x, p, y are resources, (min_p, max_p) is a property weight range, and path weight f is equal to the path length. According to [5] the property weight range of rdf:label should be set to $(0,0)$ and the ranges of rdfs:subClassOf and rdf:type should be $(1,2)$. The property weight range of user defined relations can be set to the same range, or a value range can be determined experimentally.

To use this together with the string matching score in XO we define:

$$\Delta(x,p,y) = \begin{cases} \dfrac{\omega(x,p,y)}{max(\omega(EO))}, & \text{if } (x,p,y) \in EO \\ |1 - \sigma(x,y)|, & \text{otherwise,} \end{cases}$$

where σ is a string matching method. Note that Δ is normalised with the maximum value of ω for EO and by definition of σ. Furthermore string similarity σ is inverted to string distance because it has to be aligned to concept distance and concept set distance. Because we are interested in the distance between two arbitrary concepts instead of adjacent concepts we define $\Delta^*(x,y) = MinPathWeight(x, y, \kappa^*_\Delta)$ where x, y are concepts from XO and κ^*_Δ is the normalised canonical path weight.

To calculate the semantic distances of concept sets within ontologies we chose the *minimal linking* approach of [7] which embrace the whole sets (instead of acting locally, like the *Hausdorff* measure). This approach is based on the set of

linkings M^l between concept sets A and B. For $r \in M^l$ it holds $\forall a \in A\ \exists b \in B$ with $(a, b) \in r$ and vice versa. For the distance calculation those elements from M^l is chosen for which the sum of concept distances Δ^* is minimal:

$$dist^l(A, B) = min_{r \in M^l} \left(\sum_{(a,b) \in r} \frac{\Delta^*(a, b)}{|r|} \right)$$

Finally we introduce a measure for the document d and the user interests u:

$$\Delta^l(d, u) = dist^l(SM(TL(d)), UP(u))$$

4 Document Ranking

Our method could be useful in a range of applications. As a first case we apply it to ranking of incoming e-mails, with respect to user profiles expressed by the enterprise ontology. This case is part of a project called Media Information Logistics (MediaILog), which aims to develop and introduce semantic technologies to improve information supply within companies in the Swedish media industry.

4.1 Ranking Example

In this example the objective is to rank two incoming e-mails (documents) with respect to user profiles in the enterprise ontology. A small part of an enterprise ontology is used for understandability purposes (illustrated in Figure 2). The numbers associated with each relation are the distances between adjacent concepts (user defined relations also have the weight range $(1, 2)$). Two profiles are used, profile A and profile B in Figure 2 (in the real world case the language would be Swedish). The document texts can be viewed in Table 1.

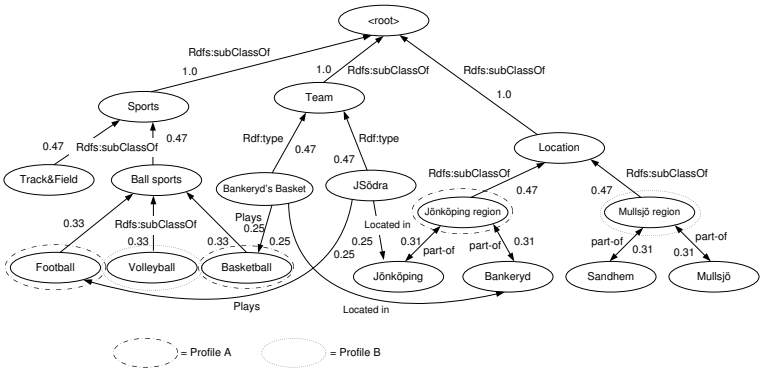

Fig. 2. The example ontology and the two profiles.

Table 1. Text content of the documents.

Doc#	Text
1	Hi! This weekend JSödra plays a training match against SandhemsIK. It starts on Saturday at 12.00. /Andreas Andersson, JSödra
2	Hi! Bankeryd's Basket invites you to a promotional evening with free snacks and a presentation of our plans for the future of our clubs in the region. The target group is newspapers and local politicians, please spread the word to anyone that might be interested! Regards, Arne Bokvist, Bankeryd's Basket

The term lists are extracted from the texts using standard text processing methods (tokenisation, stop-word removal and stemming, as in [9]). The term lists are matched against the concept labels using string matching (here the Jaro-Winkler measure from [2], with threshold 0.90). The result can be seen in Table 2. So far we do not involve the unmatched concepts in our calculation, but a penalty score could easily be incorporated. The matched terms are temporarily added to the EO to form the XO. An illustration of the XO:s can be seen in Figure 3.

Table 2. The matched terms in the term lists and their string matching scores.

Text	Term	Concept	Score	Text	Term	Concept	Score
1	JSödra	JSödra	1.0	2	Bankeryd's	Bankeryd's Basket	0.91
1	SandhemsIK	Sandhem	0.94	2	Basket	Basketball	0.92
2	Bankeryd's	Bankeryd	0.98				

Now, the distances between all concept pairs containing one concept from the extension (representing the text) and one concept in the current profile, can be computed. The values shown in Table 3 result from text 1 and profile A. Finally, these values are aggregated using the point set measure. The resulting ranks are shown in Table 4.

Table 3. Distances between concepts in XO representing profile A and text 1.

XO	Profile	Path	Dist.	XO	Profile	Path	Dist.
JSödra	Football	2	0.073	SandhemsIK	Football	7	0.62
JSödra	Basketball	4	0.27	SandhemsIK	Basketball	7	0.62
JSödra	Jönköping region	3	0.19	SandhemsIK	Jönköping region	4	0.38

The small differences between the ranking values are due to the very small ontology and due to that the texts are of relatively similar topics. Still it is possible to note benefits of our approach, especially compared to approaches using

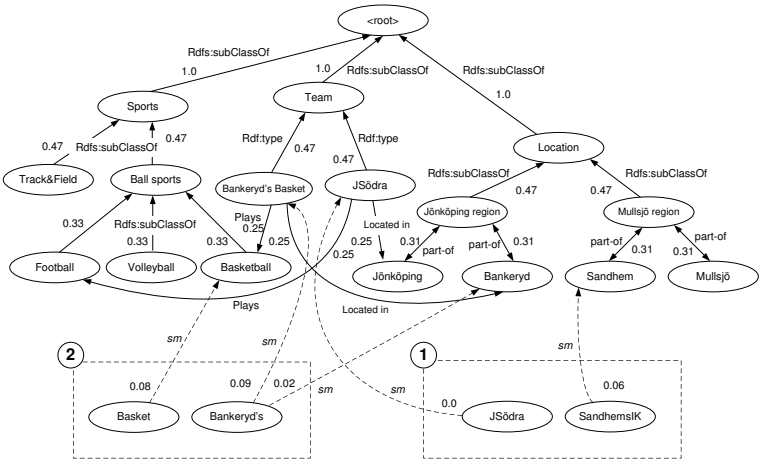

Fig. 3. The extended Enterprise Ontology.

keywords. Exact matching of keywords would not have been able to determine the relevance of either document, since none of the profile concepts are mentioned in the texts. Inexact string matching would work better but still with no way of determining the relevance of the first document, without using the ontology to discover that for example JSödra is a football team.

Table 4. Ranking of the texts with respects to profiles A and B.

	Document 1	Document 2
Profile A	0.76	0.86
Profile B	0.82	0.70

5 Conclusions

In this paper we have presented a general measure for computing the distance between two sets of concepts in the same ontology. The novelty of this approach is mainly that it modifies an existing point set distance measure and then combines it with a modified version of a commonly used distance measure between pairs of concepts in the same ontology. This is a general approach that can be useful in many application scenarios. This is also what we envision in the future of the MediaILog project, for example by assessing news items from news agencies or updates on websites. Intuitively, it is easy to see the basic usefulness of this approach, since it takes into account more background knowledge than a simple keyword-based approach, but still a thorough evaluation is also needed.

Future work concerning the distance measure intend to use a first application to measure and compare its performance to other systems and measures. There may be need for optimisation to reach a suitable time performance, although many parts can also be computed "off-line". We envision the option to use cut-off thresholds for the path computations to reduce the number of possible paths.

Acknowledgements

This work was mainly conducted within the MediaILog research project, financed by the Swedish foundation *Carl-Olof och Jenz Hamrins Stiftelse*. Special thanks to the three anonymous reviewers for valuable comments on how to improve this paper.

References

1. Gruber, T.: A translation approach to portable ontology specifications. In: Knowledge Acquisition. Volume 5. (1993) 199–220
2. Cohen, W., Ravikumar, P., Fienberg, S.: A Comparison of String Distance Metrics for Name-Matching Tasks. In: Proc. of IJCAI-03 Workshop on Information Integration on the Web (IIWeb-03), August 9-10, 2003, Acapulco, Mexico. (2003)
3. Blanchard, E., Harzallah, M., Briand, H., Kuntz, P.: A typology of ontology-based semantic measures. In: Proc. of the Open Interop Workshop on Enterprise Modelling and Ontologies for Interoperability. (2005)
4. Shvaiko, P., Euzenat, J.: A Survey of Schema-based Matching Approaches. Journal on Data Semantics **IV** (2005) 146–171 LNCS Springer-verlag.
5. Sussna, M.: Word sense disambiguation for free-text indexing using a massive semantic network. In: Proceedings of the second international conference on Information and Knowledge Management, ACM Press (1993)
6. Raftopoulou, P., Petrakis, E.: Semantic Similarity Measures: a Comparison Study. Technical report, Technical University of Crete. Department of Electronic and Computer Engineering (2005)
7. Eiter, T., Mannila, H.: Distance measures for point sets and their computation. Acta Informatica (1997)
8. Ramon, J., Bruynooghe, M.: A polynomial time computable metric between point sets. Acta Informatica **37**(10) (2001)
9. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison Wesley (1999)
10. Chirita, P.A., Ghita, S., Nejdl, W., Paiu, R.: Semantically enhanced searching and ranking on the desktop. In: Proc. of the ISWC Workshop on The Semantic Desktop Next Generation Personal Information Management and Collaboration Infrastructure, Galway, Ireland. (2005)
11. Castells, P., Fernndez, M., Vallet, D.: An adaptation of the vector-space model for ontology-based information retrieval. IEEE Transactions on Knowledge and Data Engineering **19** (2007) 261–272
12. Klyne, G., Carroll, J.J.: Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C Recommendation. W3C Consortium (2004)