Looking for Similarity among Ontological Structures

Marcirio Silveira Chaves, Vera Lúcia Strube de Lima

Programa de Pós-Graduação em Ciência da Computação
Faculdade de Informática
Pontifícia Universidade Católica do Rio Grande do Sul
Av. Ipiranga, 6681
90619-900 Porto Alegre, RS
{mchaves, vera}@inf.pucrs.br

Abstract

The automatic mapping among Ontological Structures (OSs) has been a continuous concern as a task of integration and reuse of knowledge. Besides, this mapping can support the task of expansion and combination of OSs. However, the manual execution of such task is quite tedious and slow, so it is important to automate, at least partially, the mapping process. This paper describes an ongoing work that employs the similarity measure called String Matching (SM) proposed in (Maedche & Staab, 2002) to compare terms in distinct hierarchies. We apply SM to Portuguese language OSs aiming to finding lexically similar terms. We still present some experiments using the SM measure as well as a stemmer, trying to improve the preliminary results produced by SM.

1. Introduction

Nowadays, studies that focus the mapping among Ontological Structures (OSs) still include a considerable amount of manual work. The more recent proposals (Doan et al., 2002; Noy & Musen, 2001) are described as semi-automatic because they still lack techniques allowing the full automation of this process.

Noy and Musen (2001) assert that the manual work of mapping, merging or aligning OSs is accomplished, most of the cases, by hand. This manual mapping is slow (Uschold, 2001), tedious and susceptible to mistakes (Doan et al., 2002; Noy & Musen, 1999). Besides, this process is difficult to repeat and it is not practical.

In this work, OS is taken as a set of pre-defined terms explicitly connected by semantic relations, in a format readable by humans and machines. This notion includes collections of vocabularies and concepts.

The task of mapping one OS to another reflects a continuous interest on the reuse of available OSs. Ding and Foo (2002) mention that the mapping helps the task of expansion and combination of OSs. For example, on the context of information retrieval, as similar terms are found among OSs, a system can browse through combined OSs. This kind of approach could help improving user queries results.

For Prasad, Peng and Finin (2002) mapping OS_A to OS_B consists of a process where, for each concept in OS_A a correspondent concept with similar semantic has to be found in OS_B. If there is no correspondence in OS_B, the concept is not mapped. To help users or systems find similar concepts between OSs, similarity measures are used.

1.1. Similarity Measures

Similarity between conceptual models is difficult to measure and, to establish an adequate measure of similarity is a quite subjective task (Maedche et al., 2002).

Similarity measures are used in applications such as word sense disambiguation, summarization and text annotation, information retrieval and extraction, and automatic indexing, among others (Budanitsky & Hirst, 2000). Several similarity measures are found in the literature, each one of them applied to a specific situation.

The semantic similarity measures in (Resnik, 1995; Lin, 1998; Jiang & Conrath, 1997), for example, are based on the content of information of each term. This content is defined as the number of occurrences of a term, or any child term, in the same hierarchy in a corpus.

In the present work we do not use corpus but apply the similarity measures to terms belonging to hierarchies of OSs. We work with lexical similarity without concerning about the position of the term in the hierarchy.

We search for the similarity among Portuguese OSs using similarity measures among terms, namely String Match, at the lexical level. We also use a stemmer to improve the results produced by this measure. Some experiments and preliminary results are showed.

This paper is further organized as follows. In section 2, related works are presented. Preliminary experiments are described in section 3. Finally, in section 4 we give an outlook on some future works.

2. Related Works

2.1. Anchor-Prompt

Noy and Musen (2001) developed the algorithm Anchor-Prompt that works on a set of anchor-combinations[1] previously identified (by hand or automatically). The OSs used belong to the library of DAML program[2].

The algorithm receives the anchor-terms that constitute a path in a hierarchy of concepts or terms. After the length of this path is known, a rate is attributed to the similarity between each two terms in the same position on the path. For example, let A and D be anchor-terms in OS_A and OS_B. In OS_A composed by the terms A-B-C-D the length of path from node A to node D is 3; in OS_B composed by the terms A-M-N-D the length of the path from node A to node D is 3. In this case, the similarity between B and M

[1] Pair of related terms.

[2] DARPA Agent Markup Languages - http://www.daml.org/ontologies

and C and N will be higher because these terms are in the same relative positions on the path from A to D.

In spite of providing consistent mappings, the approach based on anchors has a strong limitation for OSs with different depths, that is, as an OS is deep (with several levels in the hierarchy) and the other OS is flat (with a few levels in the hierarchy). In this case, Noy and Musen assert that the algorithm does not fit.

The OSs used in our work have distinct depths in most of the cases, so the approach of anchor-terms is not suitable.

2.2. String Matching

Maedche and Staab (2002) present a two layer approach, lexical and conceptual, to measure the similarity between terms of different OSs. At the lexical level, Maedche and Staab considered the Edit Distance (ED) formulated by Levenshtein (1966). This measure considers the minimum number of modifications should occur to change a string into another using a dynamic programming algorithm. For example, ED(*computador, computadores*) is 2, because two operations of insertion transform the original string *computador* into *computadores*. The contribution of Maedche and Staab consists of the String Matching (SM) measure given by:

$$SM(T_i, T_j) := \max\left(0, \frac{\min(|T_i|, |T_j|) - ED(T_i, T_j)}{\min(|T_i|, |T_j|)}\right) \in [0,1].$$

The SM measure calculates the similarity between two terms (T_i,T_j). The length of the shortest term is represented by min(|T_i|,|T_j|). For example, to obtain the similarity between the terms (*computador, computadores*) the minimum length is 10 and the value of ED(T_i,T_j) is 2. Thus, the resulting value is 0,8.

The shortest length is considered in the numerator as well as in the denominator of this formula allowing pondering the number of changes appearing in the term with shortest length. In the previous example the value 0,8 corresponds to the similarity between the terms (*computador, computadores*). The SM measure always returns a value of similarity between 0 and 1, where one stands for perfect match and zero indicates a bad match. Maedche and Staab used German language OSs, specifically tourism domain, in their experiments.

3. Experiments with Portuguese Language

We apply the SM measure to Portuguese language OSs. These OSs come from two distinct sources, the first from São Paulo University[3] (OS_1) and the second from the Brazilian Senate[4] (OS_2).

The terms appearing in these OSs can be associated with one of two groups: one word terms and multiword terms.

When calculating the similarity by using the SM measure it is important to establish a threshold in the detection of similar terms. In our experiments were adopt the value 0,75 as a threshold, that is, terms that present values equal or above 0,75 are considered similar, otherwise they are not.

3.1. SM applied to One Word Terms

We first applied the SM measure to terms composed by only one word. Table 1 presents some results for the preliminary tests with Portuguese:

EO_1	EO_2	SM
profissão	procissão	0,89
denúncia	renúncia	0,88
asfalto	assalto	0,86
geoprocessamento	teleprocessamento	0,81

Table 1: Examples of terms considered similar by SM measure.

Despite SM measure has produced good results with one word terms, we can observe in Table 1 unlike terms with values above 0,75.

An alternative solution to this problem is the use of a stemmer. We used a stemmer that was specifically developed for Portuguese language (Orengo & Huyck, 2001) which presented good results when compared to Porter algorithm in (Orengo & Huyck, 2001) and when compared to another algorithm developed also specifically to Portuguese language in (Chaves, 2003).

Some results obtained with the application of this stemmer are shown in Table 2. Column "SM" shows the results to the terms in the first and second columns, while column "SMStem" presents values resulting from the application of the SM to the strings in the two last columns. These strings own a stronger semantic weight, what allows a more reliable result produced by SM and, consequently, by SMStem.

Despite the good results presented in Table 2, we still observe inconsistent values after the application of the stemmer as depicted in Table 3, where SM as well as SMStem present bad results with dissimilar terms.

The extract in Table 3 presents terms with similarity higher than 0,75 for measures SM and SMStem. This indicates that only the use of a stemmer is not enough to solve the similarity problem at the lexical level. In the next section we consider the treatment to multiword terms.

3.2. SM applied to Multiword Terms

For these experiments, ontologies were first preprocessed in order to eliminate blanks. This preprocessing has also been used for other experiments in the literature (Noy & Musen, 2001; Maedche & Staab, 2002).

In the same way that for one word terms, SM generates inconsistent results, some of which can be seen in Table 4.

Terms can be considered similar if the SM threshold is equal or above to 0,75, as stated in section 3.1, but the terms depicted in Table 4 have low semantic similarity in a human point of view.

So, to improve results like those in Table 4, we calculate the similarity between multiword terms regarding each word individually by means the string returned by the stemmer.

[3] Additional information available in *http://www.usp.br/sibi*

[4] Additional information available in *http://webthes.senado.gov.br/thes*

EO_1	EO_2	SM	SMStem	EO_1	EO_2
acampamento	acabamento	0,89	0,50	acamp	acab
antiguidade	ambiguidade	0.82	0.67	antigu	ambigu
antologia	oncologia	0.78	0.71	antolog	oncolog
funcionalismo	racionalismo	0.75	0.50	funcion	racion

Table 2: Examples of terms considered similar by SM and considered unlike by SMStem.

EO_1	EO_2	SM	SMStem	EO_1	EO_2
tumulos	tumultos	0.86	0.80	tumul	tumult
aceite	azeite	0.83	0.80	aceit	azeit
linho	vinho	0.80	0.75	linh	vinh
metrologia	nefrologia	0.80	0.75	metrolog	nefrolog
trova	tropa	0.80	0.75	trov	trop

Table 3: Examples of terms considered similar by SM and SMStem.

EO_1	EO_2	SM
aguasSubterraneas	ruasSubterraneas	0.88
comportamentoPolitico	comportamentoColetivo	0.86
direitoPrevidenciario	direitoPenitenciario	0.85
africaDoSul	americaDoSul	0.82
contratoColetivoDeTrabalho	convencaoColetivaDeTrabalho	0.77

Table 4: Examples of multiword terms considered similar by SM.

This approach is similar to the one used with one word terms. We apply the stemmer to each word in the term. So, our algorithm process the SM measure for each pair of stems returned. Finally, it returns the minor value found as result of similarity between the multiword terms. For example, SMStem(analiseDoSonho, analiseDoSolo) is changed in SM(analis, analis), SM(do, do) and SM(sonh, sol), (1, 1, 0,33), respectively. So, SMStem(analiseDoSonho, analiseDoSolo) is 0,33. According to SM, the similarity between these terms is 0,84. Considering the threshold 0,75, SMStem points that these terms are not similar, although they could be considered similar if using SM. More results are shown in Table 5.

EO_1	EO_2	SM	SMStem	EO_1	EO_2
pescaIntensiva	pescaExtensiva	0.78	0.67	pescIntens	pescExtens
ecologiaFlorestal	economiaFlorestal	0.75	0.67	ecologFlorest	economFlorest
biologiaDoSolo	ecologiaDoSolo	0.75	0.67	biologDoSol	ecologDoSol
plantasMarinhas	plantasDaninhas	0.75	0.33	plantMar	plantDan

Table 5: Examples of terms considered similar by SM and considered unlike by SMStem.

Table 5 presents cases where the application of the stemmer improves the results produced by SM. In these cases, similar terms detected by SM are considered unlike by SMStem measure. The reader may notice that these terms are really dissimilar and should not to be related between OSs.

Despite of the improvement with the stemmer, in some cases SMStem measure presented results quite near to SM according to Table 6, which shows terms with low semantic similarity. However, SM as well as SMStem present values allowing these terms to be considered similar. As for the one word terms, we also found inconsistent results produced by SMStem measure to multiword terms.

Maedche and Staab (2002) assert that SM helps detecting similar lexically similar strings in German. However, regarding the preliminary results, we notice that the SM measure is insufficient to detect similarity of terms in Portuguese. The stemmer algorithm seems to improve the preliminary results, however we still keep some inconsistent examples.

In some cases the stemmer has even introduced some errors, that is, common mistakes like overstemming[5] and understemming[6].

We hope an additional penalty can be set, associated with the changes in the resulting string, that is, changes in the root indicate a higher probability that the words are not similar.

[5] It occurs when the string removed was not a suffix, but part of the stem. For example, *gramática* is reduced to *gramá* and not *gramát*.

6 It occurs when the suffix is not removed. For example, *sistemático* is reduced to *sistemátic* and not to *sistemát*.

EO_1	EO_2	SM	SMStem	EO_1	EO_2
veiculosEspeciais	veiculosEspaciais	0.89	0.80	veiculEspec	veiculEspac
acionistaMinoritario	acionistaMajoritario	0,82	0,75	acionMinorita	acionMajorita
turismoDeImportacao	turismoDeExportacao	0.80	0.78	turDeImportaca	turDeExportaca
soloAcido	soloArido	0.80	0.85	solAcid	solArid
sociologiaDoRadio	semiologiaDoRadio	0.80	0.75	sociologDoRadi	semiologDoRadi

Table 6: Examples of terms considered similar by SM and SMStem.

4. Final Remarks and Future Work

In this paper we present an ongoing work that investigates alternatives to detect similar terms in Portuguese language ontologies. We believe that similarity of strings is not completely treated yet, and it can be useful to detect similarities as an initial step in a task of integration of OSs. This integration allows the reuse of information that reflects a concern of research on the semantic web approach.

We apply the SM measure to Portuguese language ontologies and present some preliminary results. It was possible to confirm that this measure alone is not enough to detect similarities. Besides, the use of a stemmer as a complement to SM presents also inconsistent results.

We are conscious that it is necessary to undertake a deeper evaluation in our experiments, once the measures and the stemmer used for this moment do not present completely reliable results.

As a future work we intend to apply a weight to changes accomplished on the root of words and use some heuristics to get more consistent results. Besides, we can use other measures of similarity and compare the results.

Acknowledgements

Marcirio Silveira Chaves is supported by the research center HP-CPAD (Centro de Processamento de Alto Desempenho HP Brasil-PUCRS). We would like to thank Viviane Orengo that kindly provided us the stemmer used here.

5. References

Budanitsky, A. & Hirst, G., 2000. Semantic Distance in WordNet: An Experimental, Application-oriented Evaluation of Five Measures. *Workshop on WordNet and Other Lexical Resources*, in the North American Chapter of the Association for Computational Linguistics (NAACL-2000, Pittsburgh, PA.

Chaves, M. S., 2003. Um estudo e apreciação sobre algoritmos de stemming para a língua portuguesa. *IX Jornadas Iberoamericanas de Informática*. Cartagena de Indias - Colômbia, 11-15 agosto de 2003. (CD-ROM)

Ding, Y. & Foo, S., 2002. Ontology Research and Development Part 2 - A Review of Ontology Mapping and Evolving. *Journal of Information Science*, 28(5): (pp. 375–388).

Doan, A., Madhavan, J., Domingos, P., & Halevy, A., 2002. Learning to Map between Ontologies on the SemanticWeb. In *Proceedings of the World Wide Web Conf. (WWW- 2002)*, Honolulu, Hawaii, USA.

Jiang, J. J. & Conrath, D. W., 1997. Semantic similarity based on corpus statistics and lexical taxonomy. Proceedings of the International Conference on Research in Computational Linguistics (ROCLING), Taiwan.

Levenshtein, I. V., 1966. Binary codes capable of correcting deletions, insertions and reversals. *Cybernetics and Control Theory*, 10(8):(pp. 707–710).

Lin, D., 1998. An information-theoretic definition of similarity. Proceedings of the 15th International Conference on Machine Learning. Morgan Kaufmann, San Francisco, CA, (pp. 296-304).

Maedche, A. & Staab, S., 2002. Measuring Similarity between Ontologies. In *Proceedings of the European Conference on Knowledge Acquisition and Management - EKAW-2002*. Madrid, Spain, October 1-4, (pp. 251–263).

Maedche, A., Motik, B., Silva, N. & Volz, R., 2002. MAFRA - A Mapping Framework for Distributed Ontologies. *Proceedings of the 13th European Conference on Knowledge Engineering and Knowledge Management EKAW*, Madrid, Spain.

Noy, N. F. & Musen, M., 1999. SMART: Automated Support for Ontology Merging and Alignment. In *Twelfth Banff Workshop on Knowledge Acquisition, Modeling, and Management* - Banff, Alberta, Canada.

Noy, N. F. & Musen, M. A., 2001. Anchor-PROMPT: Using Non-Local Context for Semantic Matching. In *Proceedings of the Workshop on Ontologies and Information Sharing at the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-2001)*, Seattle, WA.

Orengo, V. M. & Huyck, C., 2001. A *Stemming* Algorithm for Portuguese Language, In: *Eigth Symposium on String Processing and Information Retrieval (SPIRE 2001)*, Chile. (pp. 186-193).

Prasad, S., Peng, Y. & Finin, T., 2002. Using Explicit Information to Map Between Two Ontologies. In *Proceedings of the 1st International Joint Conference on Autonomous Agents and Multi-Agent Systems - Workshop on Ontologies in Agent Systems (OAS)* - Bologna, Italy. 15-19 July.

Resnik, P., 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *Proceedings of the XI International Joint Conferences on Artificial Intelligence (IJCAI)*. (pp. 448-453).

Uschold, M., 2001. Where is the Semantics in the Semantic Web? In *Workshop on Ontologies in Agent Systems*, Montreal, Canada.