Corpus+WordNet Thesaurus Generation for Ontology Enriching

Fernando M.B.M. Castilho[1], Roger L. Granada[1], Breno Meneghetti[1], Leonardo Carvalho[1], Renata Vieira[1]

[1]Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)

Ipiranga Av., 6681. FACIN. CEP 90169-900. Porto Alegre, Brazil.

E-mail: {fernando.castilho, roger.granada, breno.meneghetti, leonardo.carvalho}@acad.pucrs.br, renata.vieira@pucrs.br

Abstract

This paper presents a model to enrich an ontology with a thesaurus based on a domain corpus and WordNet. The model is applied to the data privacy domain and the initial domain resources comprise a data privacy ontology, a corpus of privacy laws, regulations and guidelines for projects. Based on these resources, a thesaurus is automatically generated. The thesaurus seeds are composed by the ontology concepts. For these seeds similar terms are extracted from the corpus using known thesaurus generation methods. A filtering process searches for semantic relations between seeds and similar terms within WordNet. As a result, these semantic relations are used to expand the ontology with relations between them and related terms in the corpus. The resulting resource is a hierarchical structure that can help on the ontology investigation and maintenance. The results allow the investigation of the domain knowledge with the support of semantic relations not present on the original ontology.

Keywords: ontology; enrichment; thesaurus.

1. Introduction

Accountability on private information exchange between companies and governments requires update management practices regarding laws and guidelines that will affect control mechanisms in place to detect privacy threats. This scenario is aggravated by the fact that origin and destination can adopt different rules and information privacy requirements.

The management of organizations' laws and practices may require support from semantic structures such as ontologies, which are often incomplete, considering the language actually in use in relevant textual documents, glossaries and other documents. Specific features present in domain texts must be considered and may require great effort for their specification. This scenario makes desirable the automation for the process of ontology enrichment based on corpus, in a way that the language explicitly contained in documents becomes part of the ontology or is mapped to it. A thesaurus, generated from privacy documents mediates the mapping between ontology and corpus.

This paper presents the ontology expansion using a thesaurus automatically generated, plus WordNet consulting to verify the matching between ontology concepts and the corpus.

2. Related work

Thesauri have supported the discovery of domain knowledge. Xing *et al.* (2009) argue that the structure and hierarchy of terms in the vocabulary of thesaurus can reduce the difficulty in building ontologies. The standardization of terms in thesauri provides clarity, completeness and consistency, and their hierarchy can guide ontology division according to domain knowledge to create new hierarchies. Zhuhadar *et al.* (2010) developed an ontology of the distance learning domain to support a multilingual course/lecture retrieval system. In this work, when a query is submitted, the ontology provides the recovery of classes and subclasses semantically related to the query, as well as the query translation to documents retrieval in other language. A thesaurus extends the recovery capabilities, showing the relationships between domain concepts, plus tips to help distinguish multilingual concepts and relationships between ontology entities. Bawakid & Oussalah (2010) developed a system to categorize documents, which uses the synonyms of WordNet expanding the terms in documents and having a better categorization. This work also calculates the degree of similarity of a document with several topics, through semantic categorization of texts. Kwak & Yong (2010) explore the identification of semantic similarity between ontology concepts and properties based on the semantic relations of hypernymy, hyponymy, holonymy and meronymy, as presented in WordNet. The resulting measures define a set of terms called Super Word Set to guide the searching for concept and property matches. WordNet is applied to provide lexical matching between ontology elements, in order to deal with problems related to polyssemy and synonym in ontology matching.

Ontologies for privacy management have been addressed in works such as (Solove, 2006), which defines a taxonomy that supports an ontology on the identification of privacy violations involving private information. Improving privacy accountability in the exchange of medical information between European countries is defended in the work of Rahmouni *et al.* (2009). An ontology is intended to bridge the gap between data protection laws and operational controls for this protection, as basis to a semantic application to support medical decisions concerning patients' sensitive information handling, and the creation of mandatory privacy policies. Hu *et al*. (2008) point out that privacy policies are partly expressed by ontology rules, whose integration will require an effective control for their representation and fulfillment. Schäfer (2006) argues that

the automatic identification of ontological elements in text will be increasingly required, justified by the support to ontology maintenance, and by the matching between declared and undeclared rules.

3. Data privacy ontologies and corpus

In this work we use as starting point a set of ontologies on the domain of Data Privacy Regulation and Management, which were manually built based on the study of domain documents, and considering an existing system of inspection of privacy accountability compliance (Pearson *et al.*, 2009). The ontology modelling considered the following sources:

- A database of questions to the assessment of privacy risks (Pearson *et al.*, 2009).
- A description of risks involving the exchange of personal information (Bridi, 2010).
- Relevant privacy legislation terminology to assist the identification of laws and regulations involved in the exchange of information (Bruckschen *et al.*, 2010).

Our corpus[1] consists of 100 documents of privacy laws and acts, and project guidelines.

4. Thesaurus for ontology enrichment

According to Grefenstette (1993), a domain-specific thesaurus suggests alternative terms, useful to describe a domain concept. Moreover, a thesaurus can be associated to an ontology to expand and enrich its concepts, thus facilitating the domain understanding (Tomassen, 2011). This approach can also support ontology evolution, suggesting candidate terms to concepts and instances, automatically discovered in the domain.

In this work we propose to enrich an ontology by associating new terms with its concepts. Terms found in the corpus through thesaurus generation techniques are associated as similar to seeds in thesaurus, representing ontology concepts. Thus similar terms on corpus are associated to ontology concepts through the thesaurus.

Before applying the methods to the thesaurus automatic construction the corpus needs to be parsed. This process occurs due the fact that the seeds of the thesaurus, as the related terms, must be composed by nouns or noun-phrases. To parse the corpus we used the Stanford Parser[2] , obtaining XML files containing the annotated corpus.

Thesaurus generation was based on three following known methods:

- Grefenstette (1994) describes the automatic creation of a thesaurus from a corpus, using syntactic contexts to calculate the similarity between words. A syntactic context is any set of words that establish a syntactic relation with another word in the corpus. In our work the extracted syntactic contexts contain adjectives, nouns, subjects, and direct and indirect objects that change nouns or noun-phrases. The definition of semantically similar terms is accomplished through associating syntactic contexts with seeds, using the weighted Jaccard similarity measure.
- Kaji *et al.* (2000) describes the automatic thesaurus construction using a bag of words approach. This approach uses statistical methods that consist of term extraction, co-occurrence data extraction and correlation analysis. The term extraction process uses stop words to filter terms that should not be part of the thesaurus. As in this work we were searching terms semantically related to nouns, a stop word is considered any term which is not a noun or a noun-phrase. In the co-occurrence data extraction, pairs of semantically or contextually associated terms are collected using pairs of terms occurring within a window. In our work we used a window containing 20 terms. The last step is classifying the related terms using Mutual Information as correlation analysis.
- Yang & Powers (2008) describes the automatic generation using Latent Semantic Analysis (LSA) for finding non trivial semantic relations between terms. The steps look like Grefenstette's (1994), but before computing the similarity between terms, the LSA technique is applied. This technique uses matrix decomposition to find semantically related terms.

These methods, whose results are presented in Table 1, were applied over the privacy corpus. Our seeds for the thesaurus entries are ontology concepts (when the exact terms are found in the corpus). We then considered up to the 100 most similar terms to these seeds, according to the methods mentioned above.

Method	Similar terms
Grefenstette	19438
Kaji *et al.*	20004
Yang and Powers	22117

Table 1: Results of thesaurus generation methods.

Thesaurus generation methods were comparatively evaluated by data privacy experts, in other experiment (Castilho et al, 2011). In the present work we combine the results of the methods to enhance recall, and then consulted WordNet to enhance precision.

5. WordNet filtering

Works such as (Yang & Powers, 2005) and (Jean-Mary *et al*., 2009) establish the WordNet based similarity between terms from corpus, and between ontological elements, to verify and evaluate semantic relationships between terms and between ontology concepts.

In our work the thesaurus plays the role of a bridge between ontology concepts (seeds) and corpus (similar terms). The automatic thesaurus generation considered 248 seeds. The combination of the generation methods resulted in 54.549 similar terms for unique similar pairs. Each pair was consulted in WordNet to verify the existence of a semantic relation, resulting in 869 related terms for 104 ontology concepts. Concepts like *sensitive*

[1] http://www.cpca.pucrs.br/VisualizationTool/Resource/Corpus.html

[2] http://nlp.stanford.edu/software/lex-parser.shtml

data and *personal information* are not represented in WordNet. For this reason the filtered thesaurus presented not more than 104 seeds, representing ontology concepts. The filtering procedures to discover semantic relations between ontology concepts and terms from corpus, represented respectively by seeds and similar terms in thesaurus involve tasks for the extraction of relations using Java classes in WordNet and the creation of a new thesaurus reflecting the discovered semantic relations. The filtering process is presented in Figure 1.

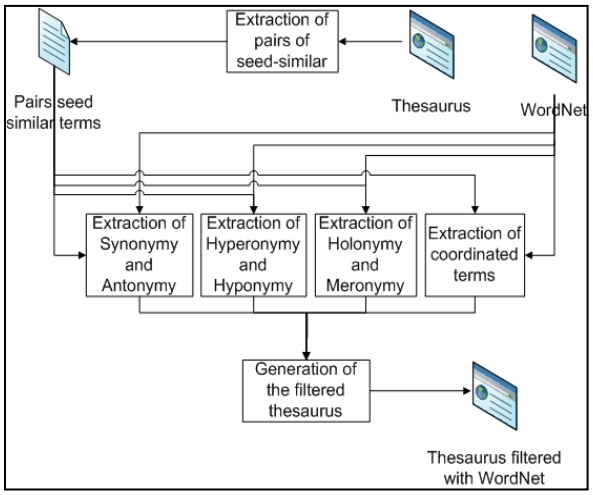

Figure 1: WordNet Filtering Process

Starting from the thesaurus automatically constructed from the corpus the pairs seed/similar terms are used to build a list to the extraction of semantic relations in WordNet. For each term in a pair its senses are retrieved from all the WordNet synsets where it occurs. Accordingly, the retrieved synsets compose the basis to verify the semantic relations involving the pairs of terms and their semantic relations with all the terms in each synset. The investigation of semantic relations is defined by the following:

- Synonymy and Antonymy –. Synonyms are identified by synsets in WordNet. An example of synonymy is found between the concept *agency* and the term *authority*. The relation of antonymy is exemplified by the relation found between the seed *employee* and the term *employer*, both similar to the concept *customer* in the thesaurus, However, the filtering process allows to understand that both have opposite meanings.

- Hypernymy and Hyponymy – The identification of such relations allow the identification of similar terms that establish a conceptual hierarchy within the ontology. Hyponymy defines a "is-a" relation, or a relation of specialization between concepts. Hypernymy in turn defines a relation of generalization. The ontology concept *company* is related to the similar term *service* in the thesaurus. The filtering process exposes this relation as *service* being a hyponym of *company*. Such ontology enrichment is derived from the similarity between *company* and the term *organization* first found in the thesaurus, and semantically enriched by the identification of *organization* as hypernymy of *company* after the filtering process.

- Holonymy and Meronymy – Two terms define a relation of holonymy when, for example, the seed denotes the whole and the similar term denotes the part. In other words, this relation allows the identification of similar terms that compose or are composed by the ontology concepts. Meronymy is the opposite relation, which denotes the part. An example of such relations can be found between the ontology concept *country* and the similar terms *department* and *province*, *country* being the holonym and *department* and *province*, the meronyms.

- Coordinated terms – This relation is established when two terms are not directly related, but instead are related to a common hypernym. Discovering terms related in such a way enriches the ontology with relationships between concepts and terms from the corpus that were not noticed in the previous direct relations. An example is, for the seed *activity*, the relation by hypernymy with the term *human action*. Through this relation the concept is coordinated to the terms *assessment* and *interference*. This kind of indirect relation enriches the ontology assigning to the concept new meaning related to human actions that are aligned with the privacy domain.

Each pair of terms is investigated to find a semantic relation in WordNet, in the following order: synonymy, antonymy, hypernymy, hyponymy, holonymy and meronymy. The first semantic relation found for a pair of terms is assigned to it and the investigation of the remaining relations in the sequence is not performed for this pair of terms. When all pairs of terms are assigned to a semantic relation the construction of the filtered thesaurus is carried out by updating the original thesaurus with relations between seeds and similar terms enriched by terms directly related to them by semantic relations or coordinated by relations with common hypernyms, in WordNet.
Figure 2 shows the filtering results to the seed *collection* and similar terms. An amount of 354 occurrences of the term were found on corpus and 19 terms were directly related to it as synonym, hypernymy or hyponym, and indirectly, as terms coordinated by common hypernyms.

```
Seed: collection(354 terms)(19 Related)
-Synonym
        collecting
-Hypernym
        publication
        request
-Hyponym
        category
        combination
        compilation
        data
        information
        set
-Coordinate
        application          :: Hypernym : request
        applications         :: Hypernym : request
        collect              :: Hypernym : petition
        collects             :: Hypernym : petition
        order                :: Hypernym : request
        reference            :: Hypernym : publication
        source               :: Hypernym : publication
        sources              :: Hypernym : publication
        systems              :: Hypernym : group
        volume               :: Hypernym : publication
```

Figure 2: Results of the filtering process for the seed "collection"

The results of the filtering process, with the distribution of semantic relation types over the generated thesaurus are

presented in Table 2 for synonymy, antonymy, hypernymy, hyponymy, holonymy, meronymy and coordinated terms respectively.

Syn.	Ant.	Hyper.	Hypo.	Hol.	Mer.	Coord. terms
113	3	76	78	0	3	596

Table 2: Semantic relations filtered with WordNet

This method generates corpus centered semantic relations, an especially useful resource for domain experts in need of corpus search/study. However, 144 ontology concepts such as *sensitive data* and *personal information* do not have an entry in WordNet, and remain without any relations determined by such resource, requiring deeper semantic processing. However the thesaurus automatic generation methods found for the seed *access* similar terms such as *cardholder_data*. The problem is that among good candidates for similar terms these methods also return terms to be discarded such as *international* for the seed *city*, or *conclusions* for the seed *hidden_pii*. These and other examples can be seen in a sample of the generated thesaurus[3] , based on our implementation of the methods.

Other examples of related terms not found in WordNet are *implicit consent* and *explicit consent*, similar to the concept *user consent*. The term *prior consent* in turn, similar to the seed *access* was not related to *user consent*, which is defined as "an action performed by the user to give consent to the use his/her data", an important definition in the domain.

The resulting domain corpus [4] and WordNet based thesaurus[5] along with other related resources, such as ontologies are available at the project Web site.

6. Concluding remarks

Our work presented the extraction of semantic relations between ontology concepts and a corpus in the privacy domain. Corpus based thesaurus generation techniques were applied to ontology concepts. Concepts and similar terms were submitted to WordNet consulting aiming at the specification of their semantic relations. As a result of this work we also make available the privacy corpus and the resulting corpus related thesaurus.

The WordNet based filtering of thesaurus generation is proposed as a support for domain ontology enrichment. In this way we can align ontology concepts and terms on the corpus, identifying synonyms, hypernyms, hyponyms, holonyms, meronyms and coordinate terms. Terms with such semantic relations are linked to the ontology concepts and may be subject to investigation for ontology enriching or updating. We are aware that WordNet is not a complete resource for the privacy domain investigation and in the future we will deal with specificities such as domain more specific concepts.

Investigation efforts must be carried out to solve problems related to sense ambiguity in concepts and terms related to the ontology concepts, as well as in composed terms like *personal information*. An example is the term *act* that carries ambiguity due to its different senses such as defined in WordNet as "an action performed" and "a legal document codifying the result of deliberations of a committee or society or legislative body". The second sense is of special interest for our domain, while the first is not.

We also plan to perform in the future an evaluation of the resulting thesaurus as a source of ontology enrichment as well as an extrinsic evaluation of the resulting thesaurus for document indexing.

For future works we plan to work on verifying the consistency of ontology concepts to support ontology development tasks using the results of such a filtering process. As an example we can investigate whether concepts found as antonyms have been defined as disjoint classes in the ontology, among other resulting relations.

[3] http://www.cpca.pucrs.br/VisualizationTool/Resource/Thesaurus.php

[4] http://www.cpca.pucrs.br/VisualizationTool/.

[5] http://www.cpca.pucrs.br/VisualizationTool/Resource/Thesaurus-wn-Filtering

7. Acknowledgements

This paper was achieved in cooperation with Hewlett-Packard Brasil Ltda. using incentives of Brazilian Informatics Law (Law nº 8.2.48 of 1991). This work is also partially supported by CAPES, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brazil.

8. References

Bawakid, A., Oussalah, M. (2010), A Semantic-Based Text Classification System. In 9th International Conference on Cybernetic Intelligent Systems. Reading, UK: IEEE Computer Society, pp. 1--6.

Bridi, P.M. (2009). Utilizando tecnologias da web semântica na inferência de riscos de privacidade. Final graduation work, PUCRS, 62 p.

Bruckschen, M.; Northfleet, C.; Silva, D. M.; Bridi, P.; Granada, R.; Vieira, R.; Rao, P.; Sander, T. (2010). Named Entity recognition in the legal domain for ontology population. In Proceedings of the Seventh International Conference on Language Resources and Evaluation. Malta: European Language Resources Association, pp. 16--21.

Castilho, F.M.B.M., Granada, R.L., Vieira, R., Sander, T., Rao, P. (2011). Ontology Enrichment Based on the Mapping of Knowledge Resources for Data Privacy Management. In Proceedings of the 4th ONTOBRAS. Gramado, Brazil: UFRGS, pp. 85--96.

Grefenstette, G. (1993). Automatic thesaurus generation from raw text using knowledge-poor techniques. In Making sense of Words. Ninth Annual Conference of the UW Centre for the New OED and text Research.

Grefenstette, G (1994). Explorations in automatic thesaurus discovery. Norwell, MA: Kluwer Academic Publishers.

Hu, Y., Guo, H., and Lin, A. G. (2008). Semantic Enforcement of Privacy Protection Policies via the Combination of Ontologies and Rules. In IEEE International Conference on Sensor Networks, Ubiquitous and Trustworthy Computing. IEEE

Computer Society, pp. 400--407.
Jean-Mary, Y.R., Shironoshita, E.P., Kabuka, M.R. (2009). Ontology Matching with Semantic Verification. Web Semantics: Science, Services and Agents on the World Wide Web, 7, pp. 235--251.
Kaji, H., Morimoto, Y., Aizono, T., Yamasaki, N. (2000). Corpus dependent association thesauri for information retrieval. In Proceedings of the 18th International Conference on Computational Linguistics. Saarbrücken, Germany: Morgan Kaufmann, pp. 404--410.
Kwak, J., Yong, H. (2010). Ontology Matching Based on Hypernym, Hyponym, Holonym, and Meronym Sets in WordNet. International Journal of Web & Semantic Technology 1(2), pp.1--14.
Pearson, S., Rao, P., Sander, T., Parry, A., Paull, A., Patruni, S., Dandamudi-Ratnakar, V., Sharma, P. (2009). Scalable, Accountable Privacy Management for Large Organizations. In Workshops Proceedings of the 12th IEEE International Enterprise Distributed Object Computing Conference. Auckland, NZ: IEEE, pp.168--175.
Rahmouni, H.B., Solomonides, T. Mont, M.C.Shiu, S. (2009). Privacy compliance on European healthgrid domains: An ontology-based approach. In Proceedings of the Twenty-Second IEEE International Symposium on Computer-Based Medical Systems. Albuquerque, NM: IEEE, pp. 1--8.
Schäfer, U. (2006). OntoNERdIE-Mapping and Linking Ontologies to Named Entity Recognition and Information Extraction Resources. In Proceedings of the 5th International Conference on Language Resources and Evaluation, pp. 1756--1761.
Solove, D. J. (2006). A Taxonomy of Privacy. University of Pennsylvania Law Review, 154(3), p. 477.
Tomassen, S. L. (2011). Conceptual Ontology Enrichment for Web Information Retrieval. Doctoral Theses at Norwegian University of Science and Technology, Trondheim.
Xing, X., Li, R., Liu, K (2009). Building Ontology Base on Thesaurus. In Proceedings of the 2nd International Conference on BioMedical Engineering and Informatics. Tianjin, China: IEEE, pp. 1--4.
Yang D., Powers, D.M.W. (2005). Measuring semantic similarity in the taxonomy of WordNet. Computer Science, 38, pp. 315--322.
Yang, D., Powers, D.M.W. (2008) Automatic thesaurus construction. Computer Science, 74, pp. 147--156.
Zhuhadar, L., Nasraoui, O., Wiatt, R., Romero, E. (2010). Multi-language Ontology-Based Search Engine. In The Third International Conference on Advances in Computer-Human Interactions. Saint Maarten, Netherlands, Antilles: IEEE Computer Society, pp. 13--18.