Concept-Based Semantic Annotation, Indexing and Retrieval of Office-Like Document Units

Saša Nešić, Fabio Crestani, Mehdi Jazayeri
Faculty of Informatics, University of Lugano
Lugano, Switzerland
sasa.nesic@unisi.ch
fabio.crestani@unisi.ch
mehdi.jazayeri@unisi.ch

Dragan Gašević
School of Computing and Information System,
Athabasca University
Brookhaven National Lab
Athabasca, Canada
dgasevic@acm.org

ABSTRACT
The paper presents an ontology-driven approach to semantic annotation, indexing and retrieval of fine-grained units of document data. In this approach, the document units and the user query are both represented by weighted vectors of ontological concepts. To determine the relevance of the document units to given query, we measure the similarity between their concept vectors. Results of the preliminary evaluation conducted with a prototype implementation are promising. We present brief analysis of these results.

Categories and Subject Descriptors
H.3 [Information Storage and Retrieval]: Miscellaneous

1. INTRODUCTION
The main goal of ontology-driven information retrieval is to enhance search by making use of available semantic annotations and their underlining ontologies [1]. Accordingly, central to the ontology-driven information retrieval is the problem of having substantial amount of accurate semantic annotations. The approach that we present in this paper is based on the semantic document model (SDM) [2] that we developed to make office-like document units be uniquely identified, semantically annotated and linkable across document boundaries. In our semantic annotation model, we first lexically expand descriptions of ontological concepts in order to enhance syntactic matching. Next, we expand a set of syntactic matches with semantically related concepts (i.e., semantic matches) discovered by applying the concept exploration algorithm - CEA, which explores the annotation ontology starting from the syntactic matches and discovers relevant semantic matches. Moreover, we calculate the annotation weights for all discovered matches and form the annotation concept weight vector. Similarly to the annotation of document units, we represent the user queries by corresponding concept weight vectors. The retrieval model utilizes the generated concept vectors of the document units and the user query, and by measuring the similarity between them determines the search results.

2. SEMANTIC ANNOTATION MODEL
The semantic annotation of document units DUs refers to the process of discovering concepts from domain ontologies whose instances appear in the DUs and their linking to DUs via the annotation interface specified by SDM. The annotation is performed in three steps: i) the lexical expansion of concept descriptions, ii) the syntactic matching and iii) the semantic matching.

The objective of the lexical expansion is to expand concept descriptions in the annotation ontology with related terms from lexical dictionaries such as WordNet. In our approach we consider three dimensions of lexical relations: synonym, hyponym and hypernym. The results of the lexical expansion are the expanded sets of concept labels.

In the syntactic matching, we analyze the content of a DU and check if some of the concept labels, including those from the lexical expansion, appear in the DU. For the concepts whose labels appear in the DU, (i.e., the syntactic matches) we calculate their weights by taking into account the following: 1) the concept labels’ relevance factor (determined in the lexical expansion), 2) the labels’ frequency in the DU and 3) the inverse document unit frequency of the concept labels in a collection of all DUs annotated by the given annotation ontology. Let us consider an example document unit d that is being annotated. As a result of the syntactic matching we get the initial set of the annotation concepts \( \tilde{d} = [c_1, c_2, ..., c_r] \); (i.e., syntactic matches) for the document unit d and the corresponding concept weight vector \( \overrightarrow{W_C}(d) = [w_{c_1}, w_{c_2}, ..., w_{c_r}] \) composed of the calculated weights of the concepts.

The objective of the semantic matching is to extend the set of syntactic matches with semantically related concepts from the annotation ontology [3]. For this purpose we introduce the Concept Exploration Algorithm (CEA). The algorithm takes an input concept and traverses the ontology graph to discover semantically related concepts. In short, the algorithm calculates semantic distances between the input concept and the other concepts in the ontology, within a given path distance (i.e., number of hops in the ontology graph), and retrieves those concepts whose semantic distances from the input concept are less than a given semantic
distance matches constraint. By applying the algorithm to all syntactic matches for the document unit $d$, we discover the set of the document unit’s semantic matches and form the expanded concept vector $\vec{d} = [c_1, c_2, ..., c_i, c_{i+1}, ..., c_m]$ of $d$. For each of the semantic matches $c_{i,j}$ the algorithm calculates the semantic distance $SDist'(c_{i,j}, c_i)$ from the initial semantic match $c_i \in \vec{d}$. The weight $w_{c_{i,j}}$ of the semantic match $c_{i,j}$ for the document unit $d$ is then calculated by the following formula:

$$w_{c_{i,j}} = w_i \cdot \beta^{-SDist'(c_{i,j}, c_i)}; \quad \beta > 1$$

where $w_i$ is the weight of the syntactic match $c_i$ and $\beta$ is a generic coefficient, which optimal value depends on the domain of and should be experimentally determined.

3. INDEXING AND RETRIEVAL

Based on the concept vectors and the corresponding concept weight vectors, generated during the semantic annotation, we build an inverted concept index of the semantic documents repository. The index contains a list of concepts (i.e., concept identifiers) from the annotation ontology, each of which is assigned a list of document units it annotates. For each document unit in the concept’s list, the index also stores the weight of the concept for the document unit.

The first step in the semantic document search that we propose is ‘making sense of the user query’, that is, finding out the semantic meaning of the query. We model the semantic meaning of the query by means of a weighted concept vector composed of concepts from the domain ontology. In other words, we treat the user query (i.e., free text query) the same way as a document unit in the annotation process and as a result we got the semantic query represented by a query concept vector and a query concept weight vector. Having formed the semantic query, we first check the concept index to find all document units which match at least one concept from the query, and then calculate the similarity between the found document units and the query. The similarity between the query and the document units is calculated as the similarity between the query’s concept weight vector and the document unit’s concept weight vector. The search finishes by ranking the document units based on their similarity to the query and retrieving the ranked list of the document units.

4. PRELIMINARY EVALUATION

In order to evaluated our approach, we have developed a prototype consisting of two modules: the semantic document authoring module and the semantic document retrieval module. The authoring module transforms MS Office documents (i.e., Word and PowerPoint) into semantic documents and does the semantic annotation and indexing of $DU$s with a selected domain ontology. The retrieval module form the semantic query and execute it against the concept index of the semantic document collection.

We have designed the preliminary evaluation as a proof of concept. As the document set that we used was composed of 150 Wikipedia articles from the series List of mammals of Europe. As the annotation ontology we used Ontology of Mammals of the World (MAMO) which contains over 5,000 domain concepts. MAMO is an OWL ontology which conforms to the SKOS specification. We have transformed the evaluation document set into five collections semantic documents applying five different annotation/indexing options: $AO_1$ - simple syntactic matching, $AO_2$ syntactic matching with the lexically expanded concept descriptions, $AO_3$, $AO_4$ and $AO_5$ comprise all the features (i.e., lexical expansion, syntactic matching and semantic matching) of the proposed semantic annotation and indexing but differ in the value of the $SDC$ (semantic distance constraint) parameter of CEA. For each annotation option Table 1 shows the number of ontological concepts that have been used in the annotation and the total numbers of syntactic and semantic matches. The results indicate that the proposed annotation model (i.e., $AO_4$, $AO_5$ and $AO_5$) has enlarged the amount of semantic annotations.

<table>
<thead>
<tr>
<th>Annotation options</th>
<th>Number of concepts</th>
<th>Number of syn. matches</th>
<th>Number of sem. matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>$AO_1$</td>
<td>211</td>
<td>1524</td>
<td>-</td>
</tr>
<tr>
<td>$AO_2$</td>
<td>343</td>
<td>3182</td>
<td>-</td>
</tr>
<tr>
<td>$AO_3$</td>
<td>672</td>
<td>3182</td>
<td>6714</td>
</tr>
<tr>
<td>$AO_4$</td>
<td>795</td>
<td>3182</td>
<td>11102</td>
</tr>
<tr>
<td>$AO_5$</td>
<td>924</td>
<td>3182</td>
<td>23716</td>
</tr>
</tbody>
</table>

Table 1: Annotation data for each option ($AO_1$-$AO_5$)

To evaluate the performance of the proposed retrieval model we formed five queries related to the data of the evaluation document set and executed them against each of the five semantic document collections. The measured values of the precision and recall for each of the collections showed that the proposed semantic annotation model increases the performances of $DU$s retrieval in terms of both, precision and recall. Moreover, by comparing the values of the precision and recall of $AO_1$, $AO_4$, $AO_5$, which all include the semantic matching and differ only in the value of the semantic distance constraint ($SDC$), we realized that there is an optimal value for the concept semantic distance ($SDist'$). The semantic matches with the concept semantic distance higher than the optimal value reduce performances.

5. CONCLUSION

In this paper we present an ontology-driven approach to semantic annotation and indexing of office-like document units, which we developed in order to improve the retrieval of such document units. Results of the preliminary evaluation conducted on the chosen document-set and the annotation ontology have shown the improvements of retrieval performance compared to simple syntactic matching which is applied in most existing ontology-driven information retrieval approaches. In the future work we plan to perform more large-scale evaluation on document sets from different domains by using different annotation ontologies.

6. REFERENCES