Altering Document Term Vectors for Classification - Ontologies as Expectations of Co-occurrence

Meenakshi Nagarajan, Amit Sheth
Kno.e.sis, College of Eng. & Computer Science
Wright State University
Dayton, OH, USA
{nagarajan.5, amit.sheth}@wright.edu

Marcos Aguilera, Kimberly Keeton,
Arif Merchant, Mustafa Uysal
HP Labs
Palo Alto, CA, USA
{first.last}@hp.com

ABSTRACT

In this paper we extend the state-of-the-art in utilizing background knowledge for supervised classification by exploiting the semantic relationships between terms explicated in Ontologies. Preliminary evaluations indicate that the new approach generally improves precision and recall, more so for hard to classify cases and reveals patterns indicating the usefulness of such background knowledge.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Abstracting methods, Dictionaries, Indexing methods, Linguistic processing, Thesauruses

General Terms

Design, Experimentation

Keywords

Supervised Document Classification, Background domain knowledge, Vector Space Models, Ranking semantic relationships

1. INTRODUCTION

The Web has many services that make its vast amount of information more usable, like search engines, shopping bots etc. Many such services use *classification* [5] to organize documents into a set of predefined classes or categories. A classifier uses training data and their correct categories (as provided by an accurate source, like a human). Classifiers then infer significant patterns that allow them to classify new content based on this training data. Albeit successful in eliminating a lot of subsequent human involvement, classifiers are limited by the information inherent to the training data. Recognizing this drawback, prior research has proposed to augment training data with external information to help classifiers learn categories better (like dictionaries [16], or subclass/super class relationships between terms [13]). These approaches have been successful, but are limited to specific forms of outside information in the kind of relationships between terms that they exploit. In this work, we propose to use a more general framework to leverage background knowledge in classification: domain Ontologies. Such knowledge can be incorporated into many classification schemes. Here, we focus on a particular scheme based on Vector Space Models (VSMs) [15], which represent documents and categories as a vector of their most important terms. Similarity measures between document and category vectors are used to determine how well a document fits in a category.

A drawback of VSM is that it treats a document as a bag of words and ignores the dependence between terms. Techniques have been developed to normalize term vectors using the term ordering [6] and statistical/synonym/hierarchical dependencies. However, terms in a document may not co-occur frequently or be related in any of the above ways; they may be related by named relationships, like "responsible for". We extend the intuition of exploiting relationships between terms, by using named semantic relationships in addition to the aforementioned relationships for normalizing term vectors. *The contribution of this work is a new method to alter a document's basic TFIDF*[14] *weighted term vector by using additional domain knowledge from an Ontology to improve existing classifiers.*

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

2. ALTERING TERM VECTORS

The core of our approach lies in altering document term vectors in three steps as shown below:

1. The syntactic term vector V_{syn}: This basic TFIDF weighted vector consists of words and phrases in the document ordered by their relative importance. We use Lucene [4] to create the syntactic term vector and normalize it using WordNet[12] to account for synonyms.

2. The semantic term vector V_{sem}: This vector consists of terms that are in the document (V_{syn}) and are also instances in the Ontology, weighted by the TFIDF scores as in V_{syn}. We also disambiguate cases where multiple matches for a term are found in the Ontology [1]. This step guarantees a 'meaningful' reduction of the vector and establishes a semantic grounding of the terms in the document that overlap with instances in the Ontology. We assume a relatively complete domain model although we recognize that a minimal overlap between document terms and Ontology instances may result in a sparse vector.

3. The enhanced semantic term vector $V_{enh\text{-}sem}$: For every term T_i in V_{sem}, we use instantiations of that term in the Ontology to obtain the most relevant terms (T_rs) connected to $T_{i.}$; thereby meaningfully extending V_{sem} to include terms that are not explicitly mentioned in the document or corroborate the ones already present in the document. This involves two critical steps:

Ranking Semantic Relationships: We quantify weights of relationships in the Ontology to consider only the relevant terms connected by the most important relationships. Our past work in ranking semantic relationships (SemRank and others [3, 9]) use heuristics, semantic and information theoretic techniques to determine the rank of semantic relationships in an Ontology. The system uses ranks assigned by SemRank and additional human input to establish numerical scores on schema level relationships (Terror Agent → operates in → Place ← based in ← Terror Organization: 0.9.). These weights along with the weight functions determine what related terms affect the term vector and by how much.

Weight Functions: The second step is defining a weight function that alters the weights of old and new terms in the term vector so as to reflect their relative importance in the document. Our weight functions employ the strength of relationships between terms in the Ontology in addition to their statistical co-occurrence strengths. In extending V_{sem}, we either add new terms or alter weights of existing ones. The two cases to consider are: **Case1:** When T_i (a term in the document) is related to a T_r (in the Ontology) and T_r does not already exist in the document *(new term added to the vector)*: The new weights for the terms T_i and T_r are calculated using: ***T_r ' = TFIDF (T_r) + Σ(all related T_i s)[TFIDF (T_i) * (R $_{TiTr}$)] T_i ' = weight of T_i*** where R $_{TiTr}$ is the normalized strength of the relationship in the Ontology between T_i and T_r. *Not changing the weight of T_i is in line with our intuition that the weights of terms are affected only by terms that are* ***in*** *the document.* **Case2:** When T_i (a term in the document) is related to a T_r (in the Ontology) and T_r is already present in the document. *(Corroborating textual co-occurrence)*: The new weights for the terms T_i and T_r are calculated using: ***T_r ' = TFIDF (T_r) + Σ(all related T_i s) [TFIDF (T_i) * (R $_{TiTr}$) + Co-Occurrence $_{TiTr}$] T_i ' = TFIDF (T_i) + Σ(all related T_rs) [TFIDF (T_r) * (R $_{TiTr}$) + Co-Occurrence $_{TiTr}$]*** where co-occurrence $_{TiTr}$ is the co-occurrence strength between the two terms quantified using the relative position of the terms in a document (generated using Lucene).

3. EXAMPLE, SYSTEM ARCHITECTURE

For a document about the 'Abu Sayyaf' terrorist group, we show excerpts of the three vectors and changes in their contents and term weights. Italicized terms are in the document but not in Ontology; bolded terms are new terms from the Ontology related strongly to terms in the document and added to the term vector.

V_{syn} : < Abu Sayyaf .004137, Libya .00357, *Christian* .00286, …..>
V_{sem} : < Abu Sayyaf .004137, Libya .00357, Manila .002, … >
$V_{enh-sem}$: < Abu Sayyaf .255734, **Al Harakat Al Islamiyya** .255734, Libya .02739, **Iraq** .023, Manila .011, **Basilan** .01866, …>

The Semantic Document Classifier system that constructs the three vectors and uses the centroid-based classification algorithm [10] to evaluate the technique is presented here [7].

4. EVALUATION and RESULTS

Our dataset for evaluation is in the national security domain. The training and testing documents were obtained from sources listed in [7] while the categories and the Ontology were created by domain experts for a prior intelligence analytics application [2]. The classification was performed on the entire category set (60 categories), but we chose a small subset of the classification (8 categories) to analyze and explain the results clearly. Given the subjective notions of ranking semantic relationships, we performed a close **human intensive evaluation** by picking a subset (3 random samples of 15 documents each) of the classified documents from each category and testing for precision and recall metrics.

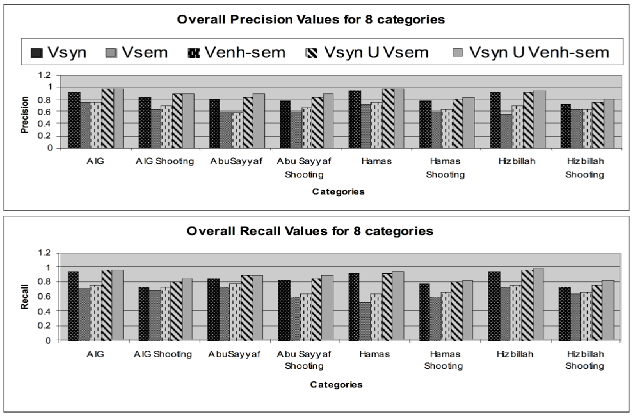

Figure 1 Overall Recall and Precision values for 8 sample categories

RESULTS: Evaluations (Figure 1 and others in [7]) indicate that the use of a domain Ontology along with the document contents i.e. V_{syn} U $V_{enh-sem}$ (union of the vectors with the higher weight of the term used if it occurs in both vectors), contributed to a marginally higher precision and recall. In most cases this vector combination generated a higher confidence in the classification (using the cosine dot product similarity of two vectors, the confidence is 0 if the two vectors are orthogonal and closer to 1 if they are similar). Maximum benefit of bringing such domain knowledge to bear was in classifying hard to classify documents. For example, precisely classifying documents related to shooting and bombing incidents are hard because of the overlap in several common buzz words. Use of an Ontology strengthened terms that related to the incident than the general buzz words like 'attack', 'gunmen' etc. While the overlap between document terms and Ontology instances in our evaluations was substantial, classification patterns also suggest the need for a rich domain model for effective deployment of such techniques.

5. DISCUSSION and CONCLUSIONS

The intuition behind this investigative work was to use a combination of statistical and domain information to alter document term vectors by amplifying weights of discriminative terms. Although preliminary, the results strongly indicate that there is a clear value in using semantic relationships between terms in documents to affect classification. Among others, some of the immediate investigations include, using different techniques to assign weights to Ontology relationships and measuring their effect on the classification; weighting semantic relationships present or implied in the document higher than those that are not; using negative training examples for the classification; testing this approach with other classifier algorithms and evaluating the approach on larger benchmark datasets subject to the availability of an Ontology in the domain.

6. REFERENCES

[1] Aleman-Meza B. et. al, Semantic Analytics on Social Networks: Experiences in Addressing the Problem of Conflict of Interest Detection. WWW 2006.

[2] Aleman-Meza B. et al., An Ontological Approach to the Document Access Problem of Insider Threat. *IEEE International Conference on Intelligence and Security Informatics*, 2005.

[3] Kemafor A. et. al. SemRank: Ranking Complex Relationship Search Results on the Semantic Web. WWW 2005.

[4]http://lucene.apache.org/java/docs/index.html Apache Lucene

[5] Baeza-Yates R., B. Ribeiro-Neto, Modern Information Retrieval. 1999 *Addison-Wesley*.

[6] Cavnar W.B., J.M. Trenkle, N-Gram-Based Text Categorization. 1994 *In Proceedings of the SDAIR*.

[7] Semantic Document Classification http://lsdis.cs.uga.edu/semdis/DocumentClassification.html

[9] Halaschek C. et. al. Discovering and Ranking Semantic Associations over a Large RDF Metabase. VLDB 2004

[10] Han, E. and Karypis, G., Centroid-Based Document Classification: Analysis Experimental Results *Principles of Data Mining and Knowledge Discovery*, 2000

[12] Miller George A, WordNet: A Lexical Database for English. 1995 *Communications of the ACM*, *38* (11). 39-41.

[13] Mladenic D. and M. Grobelnik., Feature selection for classification based on text hierarchy. *Automated Learning and Discovery* 1998.

[14] Salton G. and C. Buckley, Term Weighting Approaches in Automatic Text Retrieval. 1987 *Technical Report*

[15] Salton G. et al., A Vector Space Model for Automatic Indexing. 1975 C*ommunications of the ACM, vol. 18, nr. 11, pages 613–620.*

[16] Scott S. and S. Matwin., Text Classification Using WordNet Hypernyms. *Use of WordNet in Natural Language Processing Systems*, 1998.