Sense Cluster Based Categorization and Clustering of Abstracts

Davide Buscaldi[1], Paolo Rosso[1], Mikhail Alexandrov[2], and Alfons Juan Ciscar[1]

[1] Dpto. Sistemas Informáticos y Computación (DSIC),
Universidad Politécnica de Valencia, Spain
{dbuscaldi, prosso, ajuan}@dsic.upv.es
[2] Center for Computing Research,
National Polytechnic Institute, Mexico
dyner1950@mail.ru

Abstract. This paper focuses on the use of sense clusters for classification and clustering of very short texts such as conference abstracts. Common keyword-based techniques are effective for very short documents only when the data pertain to different domains. In the case of conference abstracts, all the documents are from a narrow domain (i.e., share a similar terminology), that increases the difficulty of the task. Sense clusters are extracted from abstracts, exploiting the WordNet relationships existing between words in the same text. Experiments were carried out both for the categorization task, using Bernoulli mixtures for binary data, and the clustering task, by means of Stein's MajorClust method.

1 Introduction

Typical approaches to document clustering and categorization in a given domain are to transform the textual documents into vector form, by using a list of index keywords. This kind of approaches has also been used for clustering etherogeneous short documents (e.g. documents containing 50-100 words) with good results. However, term-based approaches usually give unstable or imprecise results when applied to documents from one narrow domain.

Previous works on narrow-domain short document classification obtained good results by using supervised methods and set of keywords (*itemsets*) as index terms [3].

In this work, we exploited the linguistic information extracted from WordNet in order to extract key *concept clusters* from the documents, using the method proposed by Bo-Yeong Kang *et al.* [5], which is based on semantic relationships between the terms in the document. Concept clusters are used as index words.

Various methods have been tested for the categorization and clustering task, including Bernoulli mixture models, which have been investigated for text categorization in [4]. Text categorization procedures are based on either binary or integer-valued features. In our case, due to the low absolute frequency observable in short documents, we used only the information if an index term was or not in the abstract, thus obtaining a binary representation of each document.

A. Gelbukh (Ed.): CICLing 2006, LNCS 3878, pp. 547–550, 2006.

2 The MajorClust Clustering Method

We use the MajorClust method described by B.Stein [6], with the standard vector model for document representation. To evaluate the closeness between two documents, the well-known cosine measure is used, with some modifications for term weighting discussed in [1] in order to take into account the fact that abstracts usually introduce the reader to the possibilities of a suggested approach or method, while the full papers give its more or less detailed explanation.

The idea of the MajorClust method is very simple: it distributes objects to clusters in such a way that the similarity of an object to the assigned cluster exceeds its similarity to any other cluster. MajorClust method works as follows: first, every object is considered a separate cluster. Then the objects are joined to the nearest cluster. In the process of cluster construction, the objects can change their cluster in contrast, for instance, to the nearest neighbor method.

3 Bernoulli Mixture-Based Classifiers

A finite mixture model is a probability (density) function of the form:

$$p(\mathbf{x}) = \sum_{i=1}^{I} p(i)p(\mathbf{x}|i) \tag{1}$$

where I is the number of mixture components and, for each component i, $p(i)$ is its prior or coefficient and $p(i)$ is its component-conditional probability (density) function. It can be seen as a generative model that first selects the i-th component with probability $p(i)$ and then generates $\mathbf{x}$ in accordance with $p(\mathbf{x}|i)$.

A Bernoulli mixture model is a particular case of (1) in which each component i has a D-dimensional Bernoulli probability function governed by its own vector of parameters or *prototype* $p_i = (p_{i1}, \ldots, p_{iD})^t \in [0,1]^D$,

$$p(\mathbf{x}|i) = \prod_{d=1}^{D} p_{id}^{x_d}(1 - p_{id})^{1-x_d} \tag{2}$$

As with other types of mixtures, Bernoulli mixtures can be used as class-conditional models in supervised classification tasks. Let C denote the number of supervised classes. Assume that, for each supervised class c, we know its prior $p(c)$ and its class-conditional probability function $p(\mathbf{x}|c)$, which is a mixture of I_c Bernoulli components,

$$p(\mathbf{x}|c) = \sum_{i=1}^{I_c} p(i|c)\, p(\mathbf{x}|c,i). \tag{3}$$

Then, the optimal Bayes decision rule is to assign each pattern vector ($\mathbf{x}$) to a class $c^*(\mathbf{x})$ giving maximum a posteriori probability, or, equivalently,

$$c^*(\mathbf{x}) = \arg\max_c \left(\log p(c) + \sum_{i=1}^{I_c} p(i|c)\, p(\mathbf{x}|c,i) \right) \tag{4}$$

Maximum likelihood estimation of class-conditional mixture parameters (component coefficients and Bernoulli prototypes) can be reliably accomplished by the well-known EM algorithm [4].

4 Experiments and Results

The experiments have been conducted on the set of CiCling2002[1] conference abstracts, consisting in 48 abstracts related to computational linguistics thematics grouped into the following 4 categories: *linguistic*, *ambiguity*, *lexicon*, *text processing*. The intersection of vocabulary for the documents from the most different second and forth groups was about 70%. This implies that the selected domain is narrow.

The semantic indexing consists in extracting *concept clusters* from the documents. A concept cluster is composed by two or more document nouns that are connected by one or more of the following relations (R): identity, synonymy, hypernymy, meronymy. Except for identity, that is, word count, the others are defined in WordNet. Each cluster obtain a weight proportional to the number of nouns in the cluster and depending on the type of the relations connecting them, according to [5].

4.1 Categorization

Each document was represented as a bit vector, with 1 indicating the keyword (or key-sense cluster) was in the document and 0 elsewhere. The size of the vocabulary (d) was $d = 465$ when using full-text indexing and $d = 331$ for the semantic indexing technique. Due to the limited size of the corpus, the testing method was the leaving-one-out, using each document vector as test set and all the remaining documents as training.

Each average was computed from 4 runs, each one entailing a randomly initialised EM-based learning of a Bernoulli mixture per class. For simplicity, we did not try classifiers with class-conditional mixtures of different number of components.; i.e., an I-component classifier means that a mixture of $I_c = I$ Bernoulli components was trained for each abstract c. The average error obtained using the standard indexing and the Bernoulli mixtures classifier was 81.2%, whereas the average error obtained using sense clusters as indices was 48.8%. However, the use of mixtures does not seem to be useful since the errors do not change significantly with respect to the number of components in the mixture. The reason could be due to the small size of the corpus, that does not allow to estimate accurately the probabilities.

4.2 Clustering

The procedure of evaluating the clustering quality is called cluster validation. For testing cluster validity we used the index of expected density of clustering

[1] http://www.cicling.org/2002/

($\bar{\rho}$) defined in [6], and the F-measure in the form presented in [2] in order to evaluate the clusters usability (i.e., the correspondance between the results of automatic and human clustering).

Each abstract was represented with a feature vector, constituted by the weights of the index sense clusters (0 if the sense cluster is not present in the abstract). Results were compared with the indexing method based on words [1], according to well-known *tf* and *tf-idf* techniques.

The obtained F-measure was 0.44 with the sense cluster indexing and 0.64 with the standard *tf-idf*, whereas the obtained $\bar{\rho}$ was 0.08 and 0.56, respectively. Therefore, sense cluster indexing did not improve the results for the text clustering task as it did for the categorization one.

5 Conclusions

The use of semantic indexing seems to improve results in the case of categorization, although the small size of the corpus does not allow to appreciate the use of the multivariate Bernoulli mixture model. Semantic indexing did not allow to obtain better results for clustering. Further investigation will be done over the weights assigned to the WordNet relationships, and using larger collections like the Medline[2] one.

Acknowledgments

We would like to thank R2D2 CICYT (TIC2003-07158-C04-03) and ICT EU-India (ALA/95/23/2003/077-054) research projects for partially supporting this work.

References

1. Alexandrov, M., Gelbukh,A., Rosso,P.: An Approach to Clustering Abstracts. NLDB05, Alicante, Spain, 2005.
2. Eissen, S.,M., B. Stein: Analysis of Clustering Algorithms for Web-based Search. Practical Aspects of Knowledge Management, LNAI N 2569, Springer, 2002, pp.168178.
3. Hynek, J, Jezek,K., Rohlikm,O.: Short Document Categorization Itemsets Method. PKDD-2000, Springer, LNCS N 1910, 2000, 6 pp.
4. A. Juan and E. Vidal: On the use of Bernoulli mixture models for text classification. Pattern Recognition, 35(12):2705-2710, 2002.
5. Kang, B., Kim, H., Lee, S.: Performance Analysis of Semantic Indexing in Text Retrieval. CICLing 2004, Lecture Notes in Computer Science, Vol. 2945. Springer-Verlag, 2004
6. Stein, B., S. M. Eissen, F. Wissbrock: On Cluster Validity and the Information Need of Users. Proc. 3-rd IASTED Intern. Conf. on Artificial Intelligence and Applications (AIA'03), Acta Press, 2003, pp. 216221.

[2] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi