Associating Terms with Text Categories

Osmar R. Zaïane
Department of Computing Science
University of Alberta
Edmonton, AB, Canada
zaiane@cs.ualberta.ca

Maria-Luiza Antonie
Department of Computing Science
University of Alberta
Edmonton, AB, Canada
luiza@cs.ualberta.ca

ABSTRACT

Discriminating between text articles and automatically classifying documents is an essential task for many applications. With the prevalence of digital documents and the wide use of e-mail and web documents, text categorization is regaining interest and is becoming a central problem in digital text collections. There have been many approaches to solve this problem, mainly from the machine learning community. This paper proposes a new fast method for building a text classifier using association rule mining by discovering associations between terms and topical categories of documents.

Keywords: Text Categorization, Text Mining, Association Rules, Classification.

1. INTRODUCTION

Automatic categorization of text has been a relevant research issue since the inception of digital documents. Extensive research has been done in this field from as early as the sixties [6], with applications ranging from automatic indexing for information retrieval to information filtering and word sense disambiguation. Most techniques adopted for text categorization are statistical-based or machine-learning-based such as regression models, Bayesian networks, decision trees, neural networks, support vector machines, etc. [5] presents an excellent survey on classifiers for text documents. Nowadays, text categorization becomes fundamental given the large number of on-line documents that have to be sorted and grouped. For example large companies could use text classifiers for in-coming e-mail triage and

memo categorization. Recently, the categorization of Web pages and Web sites has brought up considerable interest such as the study of hubs and authorities [3]. Text classifiers can be used to classify web pages, in-coming emails, memos, news and any other text collection. Building a text classifier usually necessitates a training set consisting of a collection of text documents already associated with topical categories. Once a classifier is built with the training set, a test set, consisting of documents with known categories, is classified and the found class labels compared to the existing categories to determine the effectiveness of the classifier.

In this work we have investigated the use of association rules mining in building a text classifier. To the best of our knowledge there is no reported work exploiting association rules for the categorization of text. The proposed solution in this paper differs from previous methods in that it consists of discovering associations between words and categories and builds a classifier based on the association rules extracted from the training set. Our approach has proven efficient, fast and generating a classifier that can be updated incrementally.

2. BUILDING A FAST TEXT CLASSIFIER

2.1 Association Rule Mining

Association rule mining has been extensively investigated in the data mining literature. Many efficient algorithms have been proposed, the most popular being apriori [1] and FP-Tree growth [2]. Association rule mining typically aims at discovering associations between items in a transactional database. Given a set of transactions $D = \{T_1,..,T_n\}$ and a set of items $I=\{i_1,..,i_m\}$ such that any transaction T in D is a set of items in I, an association rule is an implication $A \rightarrow B$ where the antecedent A and the consequent B are subsets of a transaction T in D, and A and B have no common items. For the association rule to be acceptable, the conditional probability of B given A has to be higher that a threshold called minimum

confidence. Association rules mining is normally a two-step process, wherein the first step frequent item-sets are discovered (i.e. item-sets whose support is no less than a minimum support) and in the second step association rules are derived from the frequent item-sets.

2.2 Text Categorization and Association Rules

In our approach, we model text documents as transactions where items are words or phrases from the document. After pre-processing a text document, by eliminating stopwords (i.e. terms that are too frequent in the data collection and insignificant) and stemming (i.e. transforming words in their canonical form), documents are represented by sets of cleansed words $d_i=\{t_1,..,t_n\}$ as well as the category to which they belong C_j. Thus, each document in the training set belonging to a category C, and with n terms would be represented by a transaction $\{t_1,..,t_n, C\}$. We used the *apriori* algorithm [1] for mining frequent item-sets in transactional databases in order to find frequent sets of words in the documents of the training set. Given the frequent sets of words and the topical category assigned to the transaction from which they were extracted, association rules are deduced with constraints on the antecedent and consequent of the rules such that the antecedent always contains words while the consequent is exclusively a topical category. In other words, we are only interested in rules of the form W➔C where W is a set of words and C is a topical category.

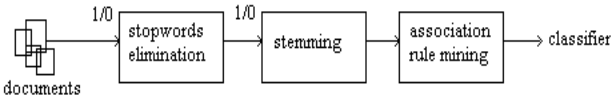

Figure 1. Association-rule-based Text Categorization.

Preceding the discovery of association rules from the documents are two optional tasks for cleansing the document terms: stopwording and stemming. Both tasks are either on or off depending upon user preferences (See Figure 1). While stopwording considerably reduces the search space and thus increases

the speed of building the document classifier, stemming does not add significant enhancements. However, for some domain applications stemming could be relevant. Contrary to other methods that select keywords for text categorization, our approach considers every word from the document that is not in the stopword list. This allows us to handle any text regardless of the domain application.

We have considered two different approaches for association rules mining from the text corpus given as a training set. In the first approach, every individual topical category is considered separately and different rules are extracted from documents of each category independently. In the second approach, the documents of all categories are considered together. In the first case, the classifier obtained is a set of rules for each category. This approach is simple, but does not take into account terms of other categories for discrimination. In the second case the classifier is a set of rules for the whole data set regardless of the category. Intuitively, this has the advantage of taking into account frequent terms of different categories simultaneously for discrimination.

3. EXPERIMENTS

We used the Reuters-21578 text collection [4] as a benchmark. This text collection is commonly used for text categorization experiments by the information retrieval community. The collection, consisting of a set of classified news articles, is arranged into a set of 22 compilations. To test the effectiveness of our algorithm, we used the first 21 compilations for the training phase, and the remainder for the testing phase. We also selected some documents from the 21 compilations of the training set for testing. From the Reuters collection we used the TOPICS set category, which contains 135 economic subject categories. Some categories in the data collection are larger than others. When all the categories are put together, the constructed database has about 10,000 transactions for training and about 500 documents for testing. In the database each transaction is represented by the words that are contained in a document.

Table 1. Preliminary results categorizing Reuters-21578

	Training Time	**Categorization time per document**	**Average precision**
Association Rules by Category	10 minutes 45	0.023 sec	0.775
Association Rules for all 135 Categories	14 minutes	0.019 sec	0.610

Table 1 illustrates the effectiveness of our text categorization algorithm when using all 135 categories from the benchmark collection. The results depicted in the table are in the case when stopwords are eliminated but no stemming is applied. We tested two classifiers: one obtained by generating association rules for each of the 135 categories separately, and one obtained by training the classifier on all categories together generating approximately 947 distinct association rules over the whole collection of articles. These preliminary tests show that training the classifier on separate categories is more effective and the precision is in the order of 77.5% given a support threshold of 30% and confidence threshold of 70%. Note that this precision is very difficult to directly compare to other classifiers since most classifiers applied to the same Reuters-21578 collection elect to use only the 10 most populated categories instead of all 135, a practice that makes the categorization task easier [5]. The categorization time is given for the average classification of one document. It is also important to note that the training time for the classifier is relatively small for 10,000 documents (less than 15 minutes on a Pentium III 800Mhz dual processor machine). This allows the reconstruction of the classifier for any new collection without significant time constraints. It is also possible with our approach to incrementally add new rules to the classifier when new documents are available for training.

The use of association rule mining can lead to the construction of very effective and flexible classifiers for text categorization of large digital text corpora. Moreover, these classifiers can be built efficiently in a reasonably short time. We intend to test our classifiers with other large text collections and perform more efficiency tests to examine the impact and worthiness of stopwording and stemming with the different association-rule-based classifiers. Moreover, in our implementation, we considered a global support threshold for frequent item-set filtering. We are in the process of adding relative and partial support concepts to take into account intra and inter-category relationships especially given the fact that documents are not evenly distributed in the given categories. Furthermore, some terms may appear commonly in some categories, but are not present in others. We are considering absence of words in documents as a discriminator for text categories as well.

4. REFERENCES

[1] Agrawal, R., Srikant, Fast Algorithm for Mining Association Rules, Proc. VLDB Conf., 487-499, Santiago, Chile, 1994

[2] Han, J., Pei, J., Yin, Y., Mining Frequent Patterns without Candidate Generation, Proc. ACM-SIGMOD, Dallas, 2000

[3] Kleinberg, J.M., Authoritative sources in a hyperlinked environment, ACM-SIAM Discrete Algorithms, 1998

[4] The Reuters-21578 Text Categorization Test collection, http://www.research.att.com/~lewis/reuters21578.html

[5] Sebastiani, F., Machine learning in automated text categorization, Tech. Rep. IEI-B4-31-1999, Consiglio Nazionale delle Ricerche, Pisa, Italy, 1999

[6] Sparck Jones, K., Willet, P., Readings in Information Retrieval, Morgan Kaufmann Publishers, 1997