Document Classification using Implication Rules

László Kovács, Tibor Répási

Department of Information Technology, University of Miskolc

Abstract: The document categorization is a very importatnt part of the current information retrieval systems for the WEB. The main problem of the text categorization is to find appropriate methods to manage the large amount of different words In this paper we describe a proposal for a theasurus-based document classification. The graph edges are generated from the implication rules calculated witha frequent itemset uncovering algorithm.

1 Introduction

The largest information source is nowadays the WEB. There are several million documents available on the Internet. It would be very useful to have an information system that can manage the whole content of the Internet. One of the big problems in implementing such a WEB information system is the heterogeneity of the WEB documents.

Since the beginnings, numerous kinds of information query systems have been developed. However, only two kinds are in use in large-scale search engines. Firstly, the simplest kind is an indexing system, which contains indexes about the documents, created by a human expert team. These experts are responsible to its special field, and they're creating a – mostly hierarchical – concept structure that is the base of the document classification process. Since the concept structure is given the expert are reading the incoming documents and creating an index which links every document to one or more concepts. From time to time it may necessary to update the concept structure because of misdistributions of the documents on the concepts. Concepts with a very high number of documents linked have to be specialized to more concepts, and other concepts with a very low number of documents linked should be generalized to one concept. However, this method has its drawbacks in the – necessary high educated – human staff and its low performance. A great advantage of this method is the very high quality of the document classification.

Another kind of information system is more automatic, uses less human resources. Mostly, these systems are using techniques based on keyword lists. A keyword list contains keywords and a relevance value for each keyword. These structures may be used as vectors are generating a vector space, where one can define a metrics. Searches are based on the distances between the documents vectors and the keyword vector given by the query. This technique may precise enough, but the terrible amount of meta-data needed makes is necessary to limit the dimension of the vector space. One technique is to select some significant documents or creating the keyword/relevance vector of the most significant imaginary documents. These significant vectors can be the generators of a space which dimension is much lower than before. The individually documents are now identified by the nearest significant element.

To provide quality we have to be able to measure quality. At the first approach this is not a trivial task. Where is it possible to relate a numerical value to the quality of an answer given to a question? We have to clear the requirements to a search engine. Usually we use to audit two aspects:

?? What is ratio of really needed documents in the result to the wanted documents? In another phrasing, how many documents in the result set fit the needs of the user, and how many documents are undesirable. To be able to answer this question we need feedback from the user.

?? What is the number of documents fitting the needs of the user but not in the result set? This is a much more difficult question, because the there is no relation between the query and the document to be wanting but not in the result set. To solve this problem the best way is a longer user interaction while the user gives instructions to correct the result set.

In most of the different types of search engines, the result of a query is not a single document but a set of documents. These documents are close to the query document and are similar to each other in some way too.

2 Object Categorization Methods

In the categorization methods, the documents are assigned to a set of similar documents. There are two main types of categorization: if there exist pre-defined categories the documents should be assigned to, the categorization is called as classification. In this case, there exists a training set containing some documents with the corresponding class membership value. On the other hand, if no pre-defined classes are known, the grouping of the documents is called clustering. The categorization is based in both cases on the attribute values of the documents.

A cluster is a collection of objects that are similar to one other within the cluster and are dissimilar to the objects in other clusters. The text clustering is the problem of automatically grouping of free text documents. The groups are usually described by a set of keywords or phrases that describes the common content of the documents in the group. In the literature, there are several methods for the clustering process. The best known methods are the partitioning, the hierarchical and the density-based methods.

In the case of partitioning methods, the algorithm constructs *K* partitions or clusters of objects. Each object must belong to only one cluster. The algorithm starts with an initial clustering. It then uses an iterative relocation technique to improve the partitioning quality. The goodness of clustering is measured by using a distance function defined on a pair of objects. The cluster is usually represented either by the mean value of the objects (k-means algorithm) or by the object located near the centre (k-medoids algorithm). In the initial phase, K objects are selected as cluster centres. For each of the remaining objects, an object is assigned to the cluster with the minimal distance. After inserting a new element into the cluster, the mean value of the cluster is recomputed. The termination criteria is usually given by a threshold for the total distance value.

In the case of hierarchical methods, the clusters are generated by hierarchical decomposition of the object set. The bottom-up approach starts with each object forming a separate cluster. In the next steps, the clusters close to each other are merged into a single cluster in a successive way until a termination condition holds. Taking a top-down approach, the initial cluster includes all of the objects. In successive iterations a cluster is split up into smaller clusters.

A density-based clustering method grows regions with sufficient high density into clusters and can discover clusters of arbitrary shape. Initially it checks every object by counting the number of neighbourhood objects within a given radius. If the number of neighbourhood objects exceeds a given threshold the objects is insert into the list of cluster core objects. If a core objects lies within the neighbourhood of some other cluster the two clusters are merged into a common new cluster.

In the case of classification, a corresponding mapping function should be defined between the documents and the class memberships. Every document is described by a set of attribute values. Let $\mathbf{x} = (\mathbf{x}_1,\ldots,\mathbf{x}_n) \in \mathbf{R}^n$ denote the *d*-dimensional attribute vector. Every attribute vector is associated with a class c_j, where the total number of class is *m*. Thus, a classifier can be regarded as a function

$$g(\mathbf{x}) : \mathbf{R}^n \rightarrow \{c_1,\ldots,c_m\}$$

The optimal classification function is aimed at minimizing the misclassification risk. The *R* risk can be measured by an appropriate cost value. The risk value depends on the probability of the different classes and on the misclassification cost of the classes.

$$R(g(\mathbf{x}) \mid \mathbf{x}) = ?_{cj}\, b(g(\mathbf{x})?\ \ c_j)P(c_j \mid \mathbf{x})$$

where $P(c_j \mid \mathbf{x})$ denotes the conditional probability of c_j for the pattern vector **x** and $b(c_i?\ \ c_j)$ denotes the cost value of deciding in favor of c_i instead of the correct class c_j. The *b* cost function has usually the following simplified form:

$$b(c_i?\ \ c_j) = \quad 0, \text{ if } c_i = c_j \text{ and}$$

$$1, \text{ if } c_i \ ?\ c_j.$$

Using this kind of b function, the misclassification error value can be given by

$$R(g(\mathbf{x}) \mid \mathbf{x}) = ?_{\ g(x)?cj} P(c_j \mid \mathbf{x})$$

The optimal classification function minimizes the $R(g(\mathbf{x}) \mid \mathbf{x})$ value. As

$$?_{cj}\, P(c_j \mid \mathbf{x}) = 1$$

thus if

$$P(g(\mathbf{x}) \mid \mathbf{x}) \ ?\ \ max$$

then the

$$R(g(\mathbf{x}) \mid \mathbf{x})$$

has a minimal value. The decision rule which minimizes the average risk is the Bayes rule which assigns the **x** pattern vector to the class that has the greatest probability for **x**.

The Bayes classifier that minimizes the misclassification error is defined by

$$qB(\mathbf{x}) = \text{argmax}\ q_j(\mathbf{x})$$

where q_j is the a posteriori probability of class *j* at pattern **x**:

$$q_j(\mathbf{x}) = P(c_j \mid \mathbf{x})$$

The misclassification cost is equal to

$$R(g(\mathbf{x}) \mid \mathbf{x}) = 1 - qB(\mathbf{x})$$

The lower is the $qB(\mathbf{x})$ value the greater is this cost. The greatest cost is yielded if every class has the same probability for the pattern vector **x**:

$$q_j(\mathbf{x}) = 1/m$$

in this case

$$R(g(\mathbf{x}) \mid \mathbf{x}) = 1 - 1/m$$

The lowest misclassification value is equal to 0. This occurs if only one class has a nonzero probability for the pattern vector, i.e. $qB(\mathbf{x}) = 1$.

3 Mining Association Rules

The association rule is an implication rule among the objects stored in a transaction pool [1][6]. In the practice, the association rules can be used to detect hidden correlation among different objects. These objects may be for example, the items in the supermarket and the association means that the items are bought together during a trip to the super-market. The association rules have been shown to be useful for applications such, diagnosis, decision support, telecommunication systems and so on.

One of the key requirements for the accepted association rules is the high level support of the items, the rules are always based on itemsets with high frequency. The usual process for generation of the frequent itemsets is a very time-consuming process as the number of possible itemsets is an exponential function of the number of items.

Let us denote the set of all items in the database by $I = \{i_1,i_2,\ldots,i_n\}$. An itemset X is a non-empty subset of I. An itemset with k items is called a k-itemset. The database consists of a set of transactions, where a transaction is given by a Y itemset. An association rule is given by an implication of the form

$$A => B$$

where

$$A ? I, B ? I, A ? B = 0$$

The rule $A => B$ can be characterized by a support and by a confidence value where

$$\text{support}(A=>B) = P(A ? B)$$

$$\text{confidence } (A => B) = P(B|A)$$

An itemset is called a frequent itemset if its support is no less than a *min_sup* value. A rule is accepted if

$$\text{support}(A=>B) ? \text{min\_sup}$$

$$\text{confidence } (A => B) ? \text{min\_conf}$$

are met.

The best known method for the generation of the frequent itemsets is the apriori-algorithm. The algorithm generates first the frequent itemsets containing only one item. In the next phase, the frequent item-pairs are generated based on the result set of the previous step. In the following phases, the frequent k-itemsets are always generated from the frequent *(k-1)*-itemsets. The generation algorithm is based on the consideration that every *(k-1)*-subset of any frequent k item set must be a frequent *(k-1)*-itemset. The main drawback of this algorithm is that it consists

of three very time-consuming phases. According to the importance of this algorithm, there are several proposals in the literature to reduce the total cost.

4 Document categorization

A main property of the document classification is that both the document content and the class labels are words or word sequences. Due to this fact, the domain and value set are the same for the classification function. This allows the application of a special kind of classification method. A widely used approach for classification of documents is the usage of some kind of thesaurus. The thesaurus can be regarded as a generalization graph of terms or concepts. The generalization graph is a directed graph. A node in the graph may have several upper and lower neighbours. The elements of ancestor set are the generalization of the actual element.

The classification methods based on the thesaurus, usually starts with selection of the relevant words. The main problem of text categorization is the high number of attributes used to describe the text documents. In the word-based representations, each feature or dimension corresponds to a single word found in the document set. As a document set may contain several thousand of words, these results in a very high and impracticable dimensionality. To reduce the document space dimensionality some word reduction methods are applied in the pre-processing phase. The most common method to reduce the number of different words is to eliminate the words with low information value. These words are called stop-words. The stop words are collected into a dictionary or a list. Another way for reduction is based on the statistical properties of the words: the infrequent and the frequent words are filtered out from the original text. This reduction is based on the assumption that the words with low frequency are not a characteristic word for the document set. The weight of a word is measured by its frequency. The stemming algorithm is also used in text clustering to remove some pre or suffixes from the words in order to determine the common root of the words.

Regarding the problem of high dimensionality, different feature selection methods are available to reduce the size of the feature space. The best known methods for dimension reduction are among others the followings: stepwise elimination algorithm, decision tree based algorithms, principal component analysis or wavelet transformation.

In our proposal, the space reduction is based on eliminating the non-relevant words from the applied thesaurus. This kind of relevance is not the same parameter as it is used in the generation of the document description vectors. The relevance value measures how relevant a word is related to a topic class. For example, if every document containing a word w1 is assigned to class c_1, then w_1

is a very relevant word in the classification process. In our assumption a document may be assigned to several topic classes. On the other hand, if a word occurs in every kind of documents, independent from the topic class, the word is not relevant from the viewpoint of the classification. The goal of the procedure is to select the most relevant words for the classification process.

In our approach the relevance value can be expressed by examination of the association rules among the words[2]. A document is given namely be a pair

$$\{B,C\}$$

where B is the content of the document and C is the set of class description words. If in every documents where

$$w_1 \in B$$

the

$$c_1 \in C$$

is also met, then

- w_1 is a relevant word in classification of c_1

and

- $P(c_1 \mid w_1)$ conditional probability is equal to one.

Of course, the larger is the number of documents containing the word w_1, the greater is the reliability of the uncovered rule.

Similar to the terms used in discovering of the association rules, the conditional probability is called confidence and the number of corresponding documents is called support. However, it can be shown, that the association rules are exactly the partial implications. A partial implication rule

$$X \Rightarrow_p Y$$

is a triple where W and C are sets of attributes and p is the precision where

$$p = P(Y \mid X).$$

Association rules correspond to partial implication meeting the support and confidence constraints.

Based on this considerations, the thesaurus used for classification purposes contain only those implication edges and only those vertexes that are contained in the selected implications. The implications having a minimal support and confidence are selected for the classification purposes. The size of the generated thesaurus depends on the selected threshold values for the support and confidence parameters.

5 Building the thesaurus using the discovered implication rules

In the proposed method, the thesaurus used for the classification purpose will be generated from the set of implications discovered in the training set. Every document in the training set can be given as a pair d (B, C) where B denotes the set of words contained in the document, and C is the set of class description word. Both sets are subsets of the universal set of words. This property implies that the same word may be contained in both parts of the documents. This allows the generation of a multi-level thesaurus.

In our assumption, the C part contains the generalized words that describe the content of B. Given a d document, this document can support only the implication rules of the form

$$w_b \; ? \quad w_c$$

where w_b is a member of B and w_c id a member of C in the d document. The reason of this restriction is our assumption about the meaning of the B and C sets of words. As the generated implication graph will be used to classify the documents, the graph should contain only implications

$$W_b \; ? \quad W_c$$

where does not exist $W_{b'}$ with

$$W_{b'} \; ? \quad W_b$$

and

$$W_{b'} \; ? \quad W_c$$

This means that the left side of the implication rule should contain minimal number of words. At this stage of the research work, the W_c set is restricted to sets with only one element.

The first step of discovering the implication rules is to determine the frequent itemsets. Based on a modified version of the apriori algorithm, the initial step is to uncover the frequent pairs of the form (w_b, w_c). Every document containing both words in the corresponding section is an instance of the pair. The number of instances can be managed in a matrix performing a single scanning of the input document set. If the number of instances is larger than a given threshold, the pair is treated as a frequent pair.

This number of occurrences is called as support for the pair. The next step is to calculate the confidence value for every frequent pairs. The confidence is equal to

$$p = | (w_b, w_c| \; / \; |w_b|$$

If the confidence value is larger than a given threshold P_l, an implication rule is found. The parameter of the implication rule is the P_l confidence value. If the (w_b,w_c) pair denotes an implication rule, the pair is not extended with new elements, otherwise the base-words part is extended with other words to find further implication rules. According to the aproiri-princip, if $(w_1\ w_2 \mid w_c)$ is frequent then both $(w_1 \mid w_c)$ and $(w_2 \mid w_c)$ are frequent. According to this consideration, the candidates for the net level are generated from the frequent elements of the current level. After testing the support and confidence values, the candidate is either taken as implication rule or it will be extended with new elements for a next level testing. If the support value of the candidate sinks below the threshold, the recursive processing will be stopped. It is reasonable to use an other stopping criteria that doesn't permit candidates with very high number of words. Another parameter of the algorithm related to the computational complexity is the confidence value. The higher is the P_l confidence value the smaller number implication rule can be found.

After generating the set of implication rules, an implication graph can be built from the single rules. The generated graph can be considered as thesauri containing some of the basic words and class description words. This thesaurus can be used for document classification purposes. The input for the classification is a new document. As a result a set of class expressions is assigned to this document. The assignment is performed in the following steps:

- collecting the words contained in the document
- locating the generalization of the words using the thesauri
- generating the most significant description list of words

The words in the thesauri can be labelled to denote whether the word can be used as class or topic descriptor or not. If the words are labelled, the result should contain only words with topic label. Another benefit of this approach is that the thesauri used for classification is open for extension with any other thesauri. So an existing thesaurus may be used as a starting thesaurus during the training phase. At the end of the training, this initial thesaurus will be expanded with the class description words

The proposed algorithm contains an parameter for weighting of the basic words too. The inclusion of this parameter is based on the consideration that the words of a shorter definition or description are usually more important or more relevant than the words of a longer description. To eliminate the not important words, first a relevance value is assigned to the words. This relevance value is calculated on the usual way.

The words of low relevance are removed from the document. In the next phase, the relevance values of the words are recalculated based on the count of words in the document. The frequency value of a word-set is then weighted by the

relevance value of the words. Thus, the frequency of an expression may be any real number.

$$f = ?_{doc}(\text{avg}_{word}(rel))$$

The frequency value is equal to the sum of average relevance values for the expression. The summation is performed over the whole set of documents. This modification results in emphasizing of the significant words in the implication rules too.

6 Description of the document formats on the WEB

In our test system, the documents from the Reuters Corpus Vol 1 were used. The documents of the corpus are stored in XML format.

The XML is the dominant standard for representing and exchanging data on the Internet. The XML language is a subset of the SGML standard. The XML tags may have any arbitrary name. These names are used to describe the meaning of the corresponding document part and not to determine the presentation format. The tags of the document are treated as the structure or schema of the document. The tags may be nested in each other.

As the XML documents are self-describing, it is possible to interpret the XML documents in itself without any other schema description. As the XML documents are used to describe the objects and the relations in the modelled world, the schema of the XML documents should correspond to the relations of the real world. According to this requirement, it is desirable to restrict of the XML structure to a predefined schema. This part of the XML document is given by a DTD or by a XMLSchema schema language.

The Reuter Corpus Volume 1 is an archive of more then 800000 new stories. All of the stories are fully annotated using category or class codes for topic, region and industry sector. The stories cover a range of content typical of an English language newswire and can vary from a few hundred to several thousand words in length. The number of used topic codes is about 100, the number of industry codes is about 380. The number of region codes is about 370.

The main benefits of the Reuters Corpus can be summarized in the following points:

- standard set of documents
- large amount of training documents
- standard XML format
- accurately classified documents

On the other hand, there are some disadvantageous properties in the document set:

- the topic code words are different from the basic words
- no thesaurus is given for the basic words.

The extension of the training set with the missing functionalities is planed to perform in the next phase of the research project.

Acknowledgement

This research is supported by FKFP 180/2001.

References

[1] R. Agrawal – R. Srikant: Mining sequential patterns, *International Conference on Data Engineering* , March 1995

[2] M. J. Zaki and M. Ogihara: Theoretical Foundations of Association Rules, *SIGMOD98* Issues in Data Mining and Knowledge Discovery (DMKD'98), USA, 1998

[3] S. Dumais and H. Chen: Hierarchical classfification of WEB content, *Proc. Of 23rd ACM Int. Conf. on Research and Development in Information Retrieval*, Athens, 2000

[4] J. Yoon and V. Raghavan and V. Chakilam: Bitmap Indexing-based Clustering and Retrieval of XML documents, *citeseer.nj.nec.com*, 2001

[5] Y. Yang: An Evaluation of Statistical Approaches to Text Categorization, *Information retrieval*, Kluwer Academic Publishers , 1999

[6] J. Han – M. Kamber: *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 1999

[7] K. Hu and Y. Lu and C Shi: Incremental Discovering Association Rules: A Concept Lattice Approach, *Proceedings of PAKDD99*, Beijing, 1999, 109-113

[8] L. Kovacs – P Baranyi: Document Clustering Using Concept Set Representation, *INES02*, Opatija, Croatia, 2002

[9] D. Lewis - R. Schaipe - J. Callan – R. Papka: Training Algorithms for linear text classifiers, Proceedings of SIGIR'96, 1996