TEXT CLASSIFICATION IN ROMANIAN FOR AN INTELLIGENT NEWS PORTAL

Traian REBEDEA[1], Ştefan TRĂUŞAN-MATU[2]

Rezumat. *Articolul prezintă un modul de clasificare a textului pentru un portal de ştiri în limba română. Tehnicile de prelucrare statistică a limbajului natural sunt combinate cu scopul de a realiza funcţionarea complet autonomă a portalului. Fiecare ştire este colectată automat dintr-un mare număr de surse de ştiri folosind clasificarea web. De aceea, tehnicile de îmbogăţire a cunoştinţelor programului sunt folosite pentru clasificarea automată a fluxului ştirilor. Mai întâi, termenii sunt clasificaţi folosind un algoritm aglomerativ răsturnat; clasele rezultate corespund subiectelor principale. De aceea, sunt preluate mai multe informaţii despre fiecare dintre subiectele principale din mai multe surse de ştiri. Apoi, algoritmii de clasificare a textului sunt aplicaţi pentru etichetarea automată a fiecărei clase de ştiri (al căror număr este predeterminat). Au fost folosite peste o mie de ştiri, atât pentru antrenarea cât şi pentru evaluarea clasificatorilor. În articol este prezentată o comparare completă a rezultatelor obţinute prin fiecare metodă. Mai mult, sunt prezentate problemele specifice care apar datorită particularităţilor limbii române şi sunt discutate soluţiile găsite.*

Abstract. *The paper presents a text classification module for a news portal for the Romanian language. Statistical natural language processing techniques are combined in order to achieve a completely autonomous functionality of the portal. The news items are automatically collected from a large number of news sources using web syndication. Afterward, machine-learning techniques are used for achieving an automatic classification of the news stream. Firstly, the items are clustered using a bottom-up agglomerative algorithm and the resulting groups correspond to the main news topics. Thus, more information about each of the main topics is acquired from various news sources. Secondly, text classification algorithms are applied to automatically label each group of news items in a predetermined number of classes. More than a thousand news items were employed for both the training and the evaluation of the classifiers. The paper presents a complete comparison of the results obtained for each method. Moreover, specific problems that arose due to the particularities of the Romanian language are presented and the solutions found are discussed.*

Keywords: natural language processing, text clustering, classification, news portal, intelligent agent

[1] Drd. Eng., University "Politehnica" Bucharest, Department of Computer Science and Engineering.

[2] Prof. Dr. Eng., University "Politehnica" Bucharest, Department of Computer Science and Engineering; corresponding member of the Academy of Romanian Scientists (trausan@gmail.com).

1. Introduction

The latest studies show that the number of users of the World Wide Web has doubled in the last five years, exceeding today one billion [1]. Meanwhile, the number of web domains has tripled to almost 100 million in the same period, from which more than half are active [2]. Due to these conditions, both the demand and the production of information have continuously increased. Thus, the volume of online information has reached an impressive quantity, as shown by the reports of the biggest search engine players [3] on the market: over 20 billion web items are indexed for search. Even powerful search engines like Google, that beneficiates from the help of the PageRank algorithm developed by Brin and Page [4], provide answers that can contain a great quantity of useless information.

Another important problem on the Web is informational redundancy, because a large part of the information can be found in different sources, with slight or even no variations. However, this redundancy can sometimes be used for the benefit of the user, as computer programs can be built in order to exploit it. In the last decade, the most important news sources worldwide have adopted web syndication as a novel means of providing their news items to their readers, as well as to other web sites or applications that are willing to use them. XML-based formats were designed so that computer programs called feed readers or aggregators can easily use the syndicated content. Aggregators are useful as they automate the process of periodically collecting news feeds and presenting them to the user in an easy to follow manner. Still they do not solve the two essential problems: processing a large quantity of information and news redundancy.

The paper continues with a section introducing the idea of intelligent news processing and its main concepts. The third section covers the most important techniques used for text clustering and classification. The next section contains the description of a news portal for the Romanian language that uses intelligent processing. The paper ends with conclusions and references.

2. Intelligent News Processing

When dealing with large volumes of data, it is important to find a method to determine the importance of the processed data. This is also valid for news headlines, especially when dealing with news collected from different sources. In this case, manually assigning an importance to each piece of news can be a difficult and a time consuming effort. In an initial phase of the news feeds' processing, online web portals were developed using the number of views for each news headline as the main criteria for determining the importance of each piece of news. In the last period of time, a number of researchers as well as companies world-wide, have undertaken the task of intelligently processing news feeds. There are two main reasons for calling these methods intelligent.

Firstly, and the most relevant, this technique automatically determines the main headlines for some period of time (e.g. a day) and offers a classification of these subjects by taking into account the number of different news headlines that compose each subject. In this case, the process of determining the importance of a subject does not use the subjective opinion of the users, but the objective one of the news providers. Secondly, this methodology of processing news items uses various techniques, including natural language processing, information retrieval and machine learning.

The news taken from various feeds can be processed in two different manners: The first one is news clustering, that means determining the most important subjects. The second manner is news classification, which assigns each piece of news into one of a number of predetermined news categories. In addition, some of the papers in the field of data extraction from news [5, 6] propose a more detailed processing, emphasizing specific named entities recognition – such as persons, countries, organizations and companies' names – and using this knowledge for the grouping and classification of news. Certainly, there is a multitude of approaches in the field of intelligent news processing but, however, most of them are still in the research stage. Labels (very significant key-word or group of key-words) assigning to every group of news resulted from the grouping is an interesting proposal. Thus, the user may follow more easily the list of subjects connected to a certain label. The central issue of this approach is the quality of the labelling, as no algorithms have been discovered which are able to provide satisfying results [6]. A new method is offered by the NewsJunkie system [7], which proposes the determination of information novelty for a piece of news. Thus, in the case of a flow of news on a common topic, recounted over a longer period of time, the system should find the truly new information and should filter the articles recounting events that have already been presented. A similar idea is proposed by Ranking a Stream of News [8].

3. Text Clustering and Classification Techniques

This section contains an overview of the clustering and classification methods and how they are used in natural language processing. Any piece of news can be seen as an array, where each element represents the frequency of a term in the text associated with the news, the headline and the description fetched using web syndication. In this manner, each piece of news is defined as a vector in a very large vector space. Each vector has as dimension the number of different terms in the bag of texts that are processed. Alternatives to this representation would be to use Boolean (0/1) variables for each dimension or the TF-IDF weight [9, p. 56-57]. Both clustering and classification algorithms need a measure to compute the similarity between two items, in this case text documents. The characteristics space used to represent documents has as many dimensions as the number of distinct terms in all the documents. Although any metric distance can be converted into a similarity measure, this technique does not offer good results for text, due to the curse of dimensionality [10]. Other measures are used for computing similarity between text documents [11] and one of the most used is the cosine of the angle formed by the two vectors that correspond to the documents that are compared.

Clustering algorithms can be divided using a wide range of criteria [11]. A first classification is considering the process of forming the groups, either top-down (divisive) or bottom-up (agglomerative). Top-down processing starts with all the data considered in a single group, which is then refined into subgroups. Agglomerative algorithms consider each element as a separate group, which are then merged to form larger ones. In relation to the result of the clustering process, the algorithms can be hierarchical or flat. Hierarchical clustering is implemented using greedy algorithms that construct the resulting tree either top-down or bottom-up. Divisive techniques always choose the least coherent cluster to be split at each step while agglomerative clustering unifies the clusters that are most similar. These algorithms may use three distinct methods to compute the similarity between two groups in order to find the most similar ones. The single link clustering computes the similarity of two groups as the similarity between the most similar items in each group. This technique corresponds to an algorithm of finding the minimum spanning tree for the set of points that define the items that need to be grouped. The complete link method uses the similarity between the two least similar elements from each group and produces better results than the single link as it uses the global quality of a group, not the local one. Single link clustering produces elongated clusters but it has a better time complexity than the complete link method. Average-link clustering has the performances of complete link and the complexity of single link and it defines the similarity of the groups as the average similarity between all the elements in each group.

Text documents classification [9, pp. 124-169] is one of the applications of classifiers in natural language processing and it is used for a wide number of purposes varying from news and e-mail categorization to automatic classification of large text documents. The most relevant aspects of text classification is that text is unstructured and the number of features is very high, greater than a thousand, unlike database mining where features ale usually less than a hundred. Nearest neighbour (NN) classifiers [9, pp. 133-138] are based on the idea that similar documents should be in the same class. The training phase is very simple and it consists of an indexing of the training data for each category, therefore it is very fast. For classifying a new item, the most similar *k* indexed documents are determined. The item is assigned to the class that has the most documents from the ones selected above. An improved variant of the this classifier uses a score for each class, defined as the sum of all the documents from that class that are in the most similar *k* ones. The document will then be assigned to the category that has the greatest score. For improving the accuracy even more, offsets for each class can be used – these are added to the score. Training the classifier to find the best values for *k* and the offsets for each category can transform it into a powerful tool, that offers “an accuracy that is comparable to that of the best text classifiers” [9, p. 135].

4. Intelligent News Classification in Romanian

Combining the facilities offered by technologies like web syndication with the advantages provided by text mining techniques allows the creation of a news portal, which is able to function with a minimum of human intervention. It is intended to offer a viable alternative for traditional news portals based on its advantages like the autonomy towards an administrator as well as the methodology used to present the news based on the importance of the headline.

4.1. Text Processing and Clustering

The news feeds need to be processed before applying the clustering and classification techniques. Because the Romanian language uses diacritical marks (special characters) that are not used by all the news sources, these need to be eliminated, regardless the encoding scheme used by the provider. HTML tags and entities and stop words are also eliminated from the text. The resulting text is tokenized and each term is stemmed using a special Romanian stemmer created especially in order to reduce the number of words. Implementing a suffix elimination algorithm for the Romanian language is extremely difficult because the rules for declination are very complicated and they affect the inner structure of the words, not only the trailing part. For example, the rules for eliminating the plural of the nouns are very numerous and affect words that are not in the plural form. The same applies for some important groups of verbs. Each suffix stripping rule defined by the algorithm can bring disadvantages, therefore our system uses only a small set of solid rules. These rules reduce the number of terms with 20-25%, depending on the number of processed news articles.

After the initial processing phase presented above, the characteristic vector for each piece of news is extracted. Because the features’ space is very large and an item contains only a few terms, the vectors are very sparse, therefore a set ordered by these terms is a useful representation.

The clustering algorithm uses a hierarchical bottom-up strategy, with hard assignment and average-link used for computing inter-cluster similarity. It does not generate the entire tree, as the grouping process ends when two clusters similar enough can not be found. At each step, two clusters are merged if their similarity is greater than a threshold. Actually, the algorithm uses two different thresholds, a higher value for creating a first set of clusters that contain very similar items, and a lower one. The upper threshold avoids merging news items that have a slightly similar subject, because it will ensure that the clusters formed in this first phase cover only one subject. In order to increase the processing time while using the memory efficiently, a buffer is used for memorizing the similarities between the clusters.

4.2. The Classification of the News Clusters

Determining the hottest subjects improves the quality of the information that is presented to the users, but some of them are only willing to read the news concerning a particular topic of interest. In order to achieve this demand, a classification of the news clusters has been implemented. The categorization of news clusters has the main advantage that a cluster holds more features' information and, therefore, is more probable to be correctly classified than a single piece of news. The categories that have been used are Romania, Politics, Economy, Culture, International, Sports, High-Tech and High Life. These classes have been chosen in order to overlap as little as possible and to have training data provided by news sources via specialized feeds for each category. A number of different nearest neighbour classifiers were compared using the cross-validation technique. The results are presented in the next section. Three variations of NN were used: the first one is simple k-NN, while the second one considers k-NN classifiers with a score for each category computed by summing the similarities of each of the *k* documents that are in the same class.

The third method is not exactly a NN as it uses a slightly different approach. Instead of determining the similarity between the document that needs to be classified and its nearest neighbours, this algorithm computes the similarity between the document and the centroid of each category, choosing the one that it is most similar to. For this reason, the method is called nearest centre (NC) or centre-based NN and it works very well when the objects from a category are evenly distributed around its centroid. Unlike the classical NN classifiers, this method needs a simple training phase that computes the features' vector of each category by summing the vectors of each element part of that category from the training set. This operation has linear complexity on the dimension of the training data, but as the vector of a class becomes bigger, the time needed to sum two vectors increases considerably. Losing some time for training comes with saving time on the classification phase, because the complexity drops from *O(dimension of the training set)* to *O(number of classes)*. Using cross-validation, the training set was divided into two subsets: two thirds of the training sets were used for the training phase and the last third was used for the classification. The number of elements in each subset, for each category, can be found in *table 1*. The number of articles is not very high and unequally distributed, as there is a difference in size between the category with the least articles in the training set and the one with the most articles. These have affected the training process and, therefore, the performance of the classifiers. The presented parameters will improve when the training corpus will grow. Using the data mentioned above, the following classifiers were tested: one of the nearest centre type (NC), one of the NN type and two of the k-NN type (for *k = 3* and *k = 5*). The tested parameters were: the accuracy of the classifier (defined as the number of correctly classified articles over the total number of news), duration of training and duration of classification. In addition, for each classifier the confusion matrix was generated, however, they will not be presented in this paper.

Table 1

Dimensions of the training and validation data sets used for evaluating the classifiers

Category	Dimension of the training data set	Dimension of the validation data set
Romania	640	325
Politics	313	157
Economy	182	92
Culture	88	45
International	481	241
Sport	316	158
High-Tech	76	38
High Life	84	43
Total	2180	1099

The best accuracy is offered by the *nearest centre* algorithm *(NC)* which outclasses with a small margin the NN and 3-NN algorithms (2%-6%), using the sum of the similarity of the nearest neighbours for each category. Although the NC algorithm needs a more time-consuming training phase, it compensates this with its classification speed – being a few times faster than the NN algorithms. The NC classifier, using cosine similarity, was the best one to employ both from the point of view of accuracy, as well as of duration of classification. In *table 2*, the confusion matrix for this classifier is presented – the precision and recall for each category, as well as globally. Finally, the accuracy of the classifier and the F1 parameter, the harmonic mean of precision and recall, are presented: Average recall = 0.59, Average precision = 0.62, Accuracy = 0.64, F1 = 0.61.

Table 2

Results of the validation phase for the Nearest Centre classifier, using cosines similarity.

	Rom	Pol	Eco	Cul	Int	Spo	Teh	HL	*Tot*	Prec
Rom	177	28	32	11	37	12	10	18	*325*	0.54
Pol	15	117	10	0	12	2	1	0	*157*	0.75
Eco	16	2	57	2	10	1	2	1	*92*	0.63
Cul	11	1	0	21	3	3	2	3	*45*	0.48
Int	37	12	16	4	162	3	3	4	*241*	0.67
Spo	5	3	7	4	7	128	1	3	*158*	0.81
The	1	0	9	1	0	0	27	0	*38*	0.71
HL	6	0	5	10	4	0	2	16	*43*	0.37
Tot	*268*	*163*	*136*	*53*	*235*	*149*	*48*	*45*		
Rec	0.66	0.71	0.42	0.40	0.69	0.86	0.56	0.36		

Conclusions

· The paper presents an alternative to classical news portals, designed to solve the problems of large amounts of news and of information redundancy, by using the latter as an advantage.

· Moreover, the portal uses web syndication and natural language processing techniques in order to achieve a human independent functionality that may be called intelligent as it automatically determines the importance of the news headlines and their category.

· Web feeds offer a simple solution for fetching news from a large number of sources and clustering can be used to exploit similar news and grouping them into a single topic.

· The user is then presented with only the news topics, ordered by the number of pieces of news that cover each topic. Moreover, one has the possibility to choose their favourite news source with an article on that topic.

· Automatic classification of the news clusters has an important advantage over the classification of each piece of news because the features' of the cluster are more consistent than the ones of a single news item. This is of critical importance in this application because syndication offers only a short brief of each piece of news.

REFERENCES

[1] Internet World Stats: *World Internet Usage and Population Statistics,* online at http://www.internetworldstats.com/stats.htm, 2006.

[2] Netcraft: May 2006 Web Server Survey, 2006.

[3] Yahoo! Search Blog: *Our Blog is Growing Up And So Has Our Index*, online at http://www.ysearchblog.com/archives/000172.html, 2005.

[4] Page, L., Brin, S., Motwani, R. and Winograd, T., *The pagerank citation ranking: Bringing order to the web*. Technical Report, Stanford University, 1998.

[5] Ueda, Y., Oka, M., Yamashita, A., *Evaluation of the Document, Categorization in Fixed-point Observatory,* Proceedings of NTCIR-5 Workshop Meeting, Tokyo, 2005.

[6] Toda, H., Kataoka, R., *A Clustering Method for News Articles Retrieval System*, Special interest tracks and posters of the 14th international conference on World Wide Web, Chiba, 2005.

[7] Gabrilovich, E., Dumais, S., Horvitz, E., *Newsjunkie: Providing Personalized Newsfeeds via Analysis of Information Novelty.* Proceedings of the Thirteenth International World Wide Web Conference, 2004.

[8] del Corso, G., Gulli, A., Romani, F., *Ranking a Stream of News.* Proceedings of the 14th international conference on World Wide Web, Chiba, 2005, pp. 97–106.

[9] Chakrabarti, S., *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan Kaufmann Publishers, 2002.

[10] Ullman, J., *Data Mining* Lecture Notes, 2000.

[11] Strehl, A., *Relationship-based Clustering and Cluster Ensembles for High-dimensional Data Mining.* Doctoral dissertation, University of Texas at Austin, 2002.

[12] Manning, C., Schutze, H.: *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts, 2003.