Towards a Supervised Rocchio-based Semantic Classification of Web Pages

Shereen Albitar, Bernard Espinasse, Sébastien Fournier

Laboratory of systems and information science (LSIS)
Domaine Universitaire de St Jérôme Avenue de l'Escadrille Normandie-Niemen
13397 Marseille cedex 20

{shereen.albitar,bernard.espinasse,sebastien.fournier}@lsis.org

Abstract . In this work, we present and compare several methods for web page classification focusing particularly on supervised classification methods. Rocchio, the method we choose for its efficiency, is tested on the reference corpus "20newsGroups", adopting different similarity measures. Results illustrate some limitations mainly related to ignoring text semantics. In order to overcome these limitations, this work proposes to extend the original Rocchio through vector conceptualization replacing terms with related concepts and by adopting semantic similarity measures, all using semantic resources as ontologies, towards semantic Rocchio-based classification method.

Keywords: Web page Classification, Semantic classification, Information Retrieval, Rocchio, Similarity mesures, Conceptualization

1 Introduction

Nowadays, traditional keyword-based indexing techniques seem to be unable to respond efficiently to user queries through existing search engines. This is often related to the explosive increase in published information on the web, and also to the indexing techniques neglecting search context [1]. In order to save Internet users' time spent on checking traditional search engines' answers, it appears appropriate to apply classification techniques to these answers considering their contents. Web page classification is currently a challenging research topic, particularly in areas such as information retrieval, recommendation, personalization, user profiles etc.

Comparing the heterogeneous structure of web pages to plain text documents, web page classification can be considered as a particular case of text classification as many features can be extracted from different parts of a web page's HTML code (title, metadata, header, URL, …) [2, 3]. Despite these differences, the principles of plain text classification also apply to web page classification.

adfa, p. 1, 2011.

At the beginning, text classification was completely a manual task realized by experts. Then, it was automated by the use of rules that generally bind the occurrence of certain keywords or "features" in a document to its association to a specific class. However, rule implementation and maintenance demand a lot of time and effort from experts, in addition to their limited adaptability to their original context dynamics and for each new context [2].

Consequently, learning-based techniques appeared, introducing new methods for classification. Generally based on a supervised learning, these methods use training corpus to learn decision criteria in order to be able to discriminate relevant classes. These criteria are often crystallized in induced rules, or statistical estimations. Such supervised methods require training corpus preparation through manual tagging, that associates its documents to their relevant classes. Even if this preparation effort is significant, it is nevertheless much smaller than the effort required for rule implementation [4]. Most popular methods for text classification: Naïve Bayes Classifier (NB), Support Vector Machines (SVMs), Rocchio, and K Nearest Neighbor (KNN).

Despite its attractive simplicity, *NB* classifier [5], also called "The Binary Independence Model", has critical weaknesses. The unrealistic independence hypothesis of this model considers each feature independently for calculating class prototype during training phase. More effecient classification is possible using *SVMs* [6-8], nevertheless the number of features characterizing documents is crucial to learning complexity, so it is essential to this method to eliminate noisy and irrelevant features [4]. *KNN* [9] is also sensitive to noisy examples in training set in addition to its slow classification when using important corpus [4].

Compared to other methods for text classification, Rocchio (or centroïd-based classifier) has many advantages [10]. First, learned classification model summarizes the characteristics of each class through a centroïd vector, even if these characteristics are not all present simultaneously in all documents. This summarization is relatively absent in other classification methods except for NB that learns term-probability distribution functions summing up their occurrences in different classes. Another advantage is the use of similarity measure that compares a document to class centroïds taking into account summarization result as well as term occurrences in the document in order to classify it. NB uses learned probability distribution only to estimate the occurrence probability of each term independently to other terms in a class summarization or to document co-occurring terms.

Vector-based representation (binary or TF/IDF) used by all methods permits semantic enrichments. Conceptualization is the process of enriching these vectors by concepts related to their features that can be retrieved using a certain background knowledge base [11]. In addition, both KNN and Rocchio enable the use of knowledge bases in decision making through new semantic similarity functions [12].

As a conclusion, we consider Rocchio an adequate baseline classifier for text and web page classification. Its efficiency, simplicity, and extendibility with semantic resources in addition to other advantages lead us to choose Rocchio for the rest of this work. The main contribution of this paper is evaluating the performance of Rocchio applied to the corpus "20NewsGroups" and also studying the effects of corpus organization and distribution on its performance. Furthermore, we argue some semantic solutions in order to overcome its limitations.

In section 2, thorough experimentations are realized on the reference corpus "20NewsGroups" using several similarity measures. Analyzing statistical results, we can relate the majority of Rocchio method's limitations to omitting semantics throughout classification process. Section 3 presents an extension to Rocchio overcoming these limits, so document is represented using vector of concepts instead of terms, introducing semantic classification. Finally, we conclude with an assessment of our work, followed by different research perspectives.

2 Evaluating Rocchio Classifier Using Different Similarity Measures

In previous section, Rocchio classifier has been chosen for its efficiency and semantic extendibility. This section presents an experimental study of Rocchio-based classification of text documents referring to some implementation details. Using five similarity measures [13] (Cosine, Jaccard, Pearson, Averaged Kullback-Leibler Divergence, and Levenshtein distance) separately in experimentations enables us to evaluate Rocchio's performance independently to similarity calculation in decision making. Afterwards, results of different system configurations applied to three different variations of 20NewsGroups corpus are compared and analyzed. Finally, we discuss certain limitations in these results and relate some of them to the absence of semantic aspects in classification process.

2.1 Rocchio Implementation details

Rocchio or centroïd based classification [10] for text documents is widely used in Information Retrieval tasks, in particular for relevance feedback [14]. In Rocchio, VSM is adopted for document representation through applying four preprocessing steps (Tokenization, Stemming, StopWord Removal and Weighting). Multiple weighting schemes might be used to represent the corresponding importance of each term in a document [15] like idf, idf-prob, Odds Ratio, χ^2 etc. According to TF/IDF, the most popular schema, the score of a term tj in document di is estimated as follows:

$$\mathrm{w_{ij}} = \mathrm{tf_{ij}} * \log(\mathrm{N}/\mathrm{df_j}) \qquad (1)$$

tf_{ij} : Frequency of term tj in document di.

N: Number of documents.

df_j: Number of documents that contain term tj.

The result of applying vector space modeling to a text document is a weighted vector of features:

$$d_i = (w_{i1}, w_{i2}, w_{i3}, \dots, w_{in}) \qquad (2)$$

For centroïd-based classification, each class is represented by a vector positioned at the center of the sphere delimited by training documents related to this class. So a class's centroïd is the vector that summarizes features of training documents' vectors belonging to the considered class. During training and classification phases, documents are represented by vectors following the VSM as detailed earlier. Having n classes in the training corpus, n centroïd vectors {C1,C2,.....,Cn} are calculated through the training phase. In order to classify a new document x, first we use the TF/IDF weighting scheme to calculate the vector representing this document in the space. Then, resulting vector is compared to all centroïds of n candidate classes using a similarity measure. So the class of the document x is the one represented by the most similar centroïd.

$$\arg \underset{\mathrm{i=1,2,\ldots n}}{\mathrm{Max}} \left(SimFun(x, Ci\,)\right) \tag{3}$$

2.2 20NewsGroups corpus

20NewsGroups corpus [16] is a collection of 20,000 newsgroups documents almost evenly divided in twenty news classes according to their content topic assigned by authors. This collection is divided according to the percentages (60:40) into training corpus and test corpus respectively. Corpus organization is illustrated in table 1. Some classes cover similar topics for example (comp.sys.ibm.pc.hardware & comp.sys.mac.hardware), whereas others concern relatively different ones as (rec.autos & sci.crypt).

Table 1. Twenty actuality classes of 20NewsGroups corpus

Computer	comp	graphics		**Science**	sci	crypt	
	comp	os	ms-windows		sci	electronics	
	comp	sys	ibm		sci	med	
	comp	sys	mac		sci	space	
	comp	windows	x	**Politics**	talk	politics	misc
Sports	rec	autos			talk	politics	guns
	rec	motorcycles			talk	politics	mideast
	rec	sport	baseball	**Religion**	talk	religion	misc
	rec	sport	hockey		alt	atheism	
Forsale	mis	forsale			soc	religion	christian

2.3 Experimentations and results

In these experimentations, three variations of the 20NewsGroups corpus are used: (i) the original corpus, (ii) six chosen classes, and finally (iii) the original corpus reorganized in more general six classes as they are assembled in table 1. Training is realized on each of these variations and so class centroïds are calculated. As for test, on each variation of the corpus, five experimentations are executed applying five similarity measures. For most classification tasks, classifier's accuracy [17] exceeded 90%. Thus, we use F1-Measure [17] for performance comparison.

Experimentations on the original corpus

Our first test concerns the original corpus with twenty classes. As illustrated in Fig. 1**.** , similarity measures' performance varies according to the treated class (columns follow the same order of legends from the left to the right). For instance, the class (talk.religion.misc) is vast compared to other religious classes so it can draw in their documents (False negative) resulting in its relatively low value of F-measure. This observation is so called: general class issue. Classes related to computers seem to use similar vocabulary so the classifier cannot be able to distinguish them properly having similar centroïds (similar class issue). On the other hand, distinct classes like (rec.sport.hockey, rec.sport.baseball) are well distinguished by all classifiers. Finally, through this diagram we can observe that the similarity measure Kullback-Leibler outperforms other similarity measures. In addition, Cosine and Pearson similarities show similar behaviors.

After detailed results analysis, it is observed that at least (50%) of incorrectly classified documents (False Negative) are classified in a similar class. Indeed, similar classes, using similar vocabularies, usually have their centroïds close to each other in the feature space. This implies some classification difficulties in order to distinguish classes' boundaries affecting overall performance. In addition, document contents might be related to multiple classes making classifier's task tricky.

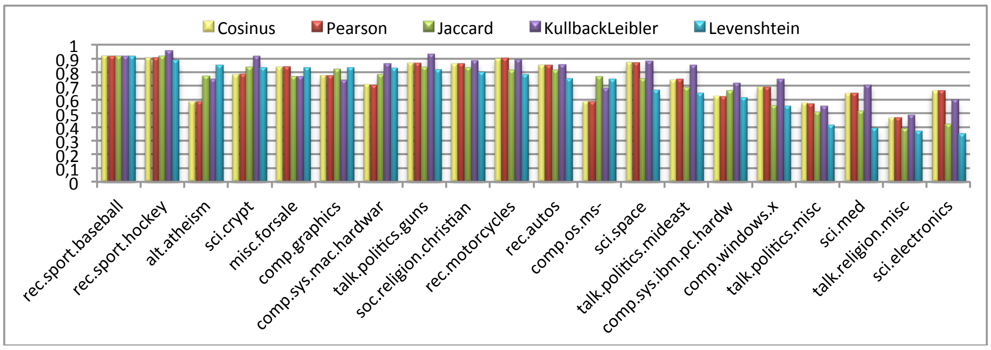

Fig. 1. Evaluating five similarity measures on the original 20NewsGroups corpus (F1-Measure)

In order to support our observations, we present next two supplementary experimentations, using in the first six classes chosen from the original corpus, and in the second the original corpus reorganized in six meta-classes.

Experimentations on six chosen classes

In these experimentations, six classes, relatively distinct, of the twenty classes of the original corpus are chosen for both training and test. Classifier is first trained and then tested on the following classes: comp.windows.x, misc.forsale, rec.auto, sci.med, soc.religion.christian, talk.politics.mideast.

Even though (sci.med) is isolated from other scientific classes, classifier's performance is still relatively poor, compared to other classes. This is due to the large distribution of medical documents in feature space so no learned centroïd can be sufficiently representative. For other classes, classifier shows better performance as they are distinct, and therefore positions of their centroïds are well dispersed in the feature space. Kullback-Leibler seems to outperform other similarity measures in these experimentations as well. Results are illustrated in Fig. 2**.** where columns follow the same order of legends from the left to the right.

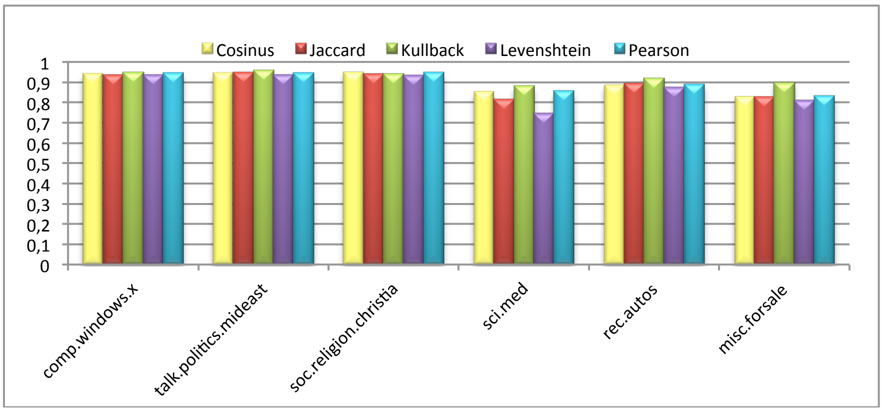

Fig. 2. Evaluating five similarity measures on six classes of 20NewsGroups corpus (F1-Measure)

Experimentations on the corpus after reorganization

Original 20NewsGroups corpus classes are reorganized depending on initial class similarities, so documents of similar classes are gathered in a more general class resulting in six new classes in total (comp, rec, science, forsale, politics, religion) as presented in Table 1. Rocchio based classifier is trained on the training set and then learned model is tested using different similarity measures on the test set.

According to results illustrated in Fig. 3**.** (columns follow the same order of legends from the left to the right), classifier's performance is relatively high for most classes applying at least for one of the similarity measures. In fact these classes assemble similar original classes like (religion) or well specified classes like (rec). Classifier shows some difficulties classifying (science) as the classes it assembles contain diverse information (heterogeneous class issue). In fact, one centroïd for such heterogeneous class is not very representative justifying the relatively poor value of f-measure for this class in **Fig. 3.**

Conclusion

Throughout previous experimentations, several limitations seem to affect Rocchio's performance particularly in dealing with similarities among classes, general classes and heterogeneous classes. These limitations are mainly related to class representation and similarity calculations. Limitations observed with similar classes can be overcome by means of semantic resources. Indeed, redefining centroïds using concepts instead of terms might limit intersections between spheres of similar classes in concept space. Consequently, ambiguities between classes using similar vocabulary can be resolved at representation level through vector conceptualization using semantic resources or ontologies. Next section introduces some semantic solutions.

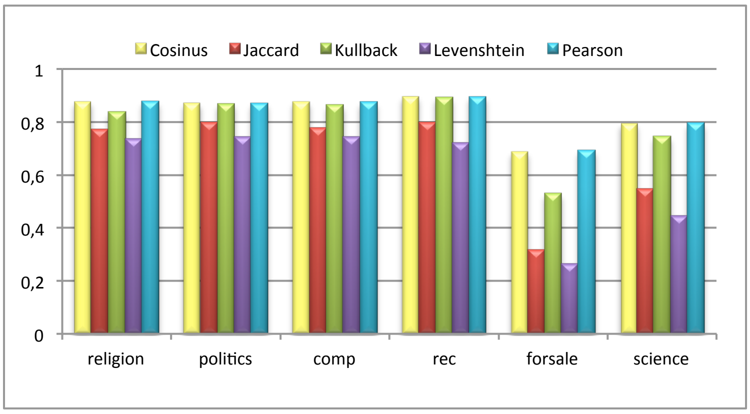

Fig. 3. Evaluating five similarity measures on 20NewsGroups corpus reorganized in six classes (F1-Measure)

3 Towards A Rocchio-Based Semantic Classification Of Text

In spite of being considered the most popular text representation method, VSM suffers from certain limitations [18, 19] especially for processing composed words, synonyms, polysemy, etc. In order to overcome these limitations, semantic resources (like thesaurus & ontologies) can be used to replace term-based representation by concept-based one.

This section presents first different strategies for vector conceptualization. Then, the use of conceptualized vectors during classification process is discussed according to the chosen strategy for conceptualization. We aim at combining both conceptualization and semantic similarities in order to extend the original Rocchio for a more efficient semantic classification.

3.1 Vector Conceptualization

Conceptualization is the process of mapping terms literally occurring in treated text to their semantically corresponding concepts or senses in semantic resources that might permit better classification. As an example of semantic resources that might be used for conceptualization: WordNet, Wikipedia and other domain specific ontologies usually called domain ontologies.

In general, vector conceptualization is realized in two steps: search for corresponding concepts related to vector's terms and then the integration of these concepts in the vector producing the final conceptualized vector. Integrated concepts are assigned new scores derived from the frequencies of their related terms. Three different strategies have been proposed for conceptualization [11] :

1. *Adding Concepts*: Where the original vector is extended and corresponding concepts are added.
2. *Partial Conceptualization*: Where terms are substituted by corresponding concepts. Terms having no related concepts are held in the vector.
3. *Complete Conceptualization*: Similarly to Partial Conceptualization, terms are substituted by concepts whereas remaining terms are eliminated from the final vector.

Second strategy seems to be the most appropriate as it removes no term without replacing it with a related concept, so no original feature is removed from the vector (compared to the third one), and no extra feature is added (compared to the first one) resulting in minimized efficiency effects. Yet, the classifier has to be adapted to hybrid (concepts + terms) representation.

While searching concepts corresponding to a polysemic term in semantic resources, multiple matches are detected introducing some ambiguities in final document representation. For example: the term "Book" signifies in English a book and also a reservation (Ticket, accommodations…). According to [11], three strategies for disambiguation deploying WordNet can be used:

1. *All*: Accept all candidate concepts as matches for the considered term.
2. *First*: Accept the most frequently used concept among candidates using document language statistics.
3. *Context*: Accept the candidate concepts having the most similar semantic context compared to the term's context in the document.

First strategy, despite being the simplest, is the least reliable as it accepts all candidate concepts without choosing a specific meaning for the term. In cases where a term is used in the document signifying its rarely used sense, second strategy chooses an incorrect concept. Despite its complexity, last strategy seems to be more accurate as concept context can be derived from semantic resources using: its definition, its descriptive terms or from text corpus.

3.2 Using Conceptualized Vectors in Rocchio

Vectors resulting from the conceptualization step are applied to decision-making criteria in order to classify corresponding documents. Most semantic classification systems neglect semantic relations between relevant concepts and apply the same pure mathematic similarity measures for decision-making. Indeed, the exact concept must occur in both compared vectors to influence decision-making through its score. Similar or related concepts are rarely taken into consideration.

Using semantic resources, it is important to incorporate related concepts to those already adopted through conceptualization step enriching document representation and eventually improving classification results. For example, some works propose integrating super concepts to the conceptualized vector and demonstrate some ameliorations in classification results as multiple levels of ontology are considered [11].

New semantic similarity measures considering semantic relations among ontology concepts are also being developed making more advances towards a semantic classification. These measures permit similar concepts to be compared and contribute to vector comparison beside common concepts, [12]. For an overview on different semantic similarity measures see [20].

Integrating similarity measures in semantic classification depends on adopted conceptualization strategy. Precedent measures can be directly applied to vectors resulting from "Complete Conceptualization". Considering other conceptualization strategies that produce hybrid (concepts+terms) document representation, both mathematic and semantic similarity measures are applied to terms and concepts respectively.

New representation models using parts of ontology hierarchy are also proposed. These parts constitute semantic trees or forests where each concept is assigned an importance score. As for decision making in systems using these models, similarity measures between semantic hierarchies is considered as the accumulation of similarities between their concepts peer to peer [21]. Semantic similarity between two concepts is related to their scores and their positions in the hierarchy.

4 Conclusion and Perspectives

In the past decade, publishing millions of pages has resulted in the explosive growth of the web. The performance of many search engines is still limited to meet the needs of users leading to a challenging need for new filtering and ranking techniques. In this context, we presented in this paper and compared four traditional methods for text classification that can be applied to web page classification as well.

We retain Rocchio that shows a good performance compared to its relatively minimal complexity. Moreover, it may provide feedback on the relevance of classification results permitting better result understanding and so potential classification improvements. Rocchio method has been tested using different similarity measures on the 20newsGroups corpus. However, Rocchio results illustrate several limitations in text classification, which could be surpassed by means of semantic resources taking meaning into consideration in text classification.

In this work, we propose a new method for semantic classification of web pages based on Rocchio. Indeed, Rocchio allows integrating semantics in conceptualization and during decision making through different semantic similarity measures. Taking into account information stored in HTML tags, document vectorization followed by the conceptualization using a knowledge base, helps to complete VSM approach with semantics. Appearing ambiguities can be resolved using the context strategy. Finally, semantic similarity measures can also be used implying similar concepts in calculating vector similarities. The next step of our work consists in developing this method and then in evaluating it on classification tasks using other web pages corpus.

References

1. Asirvatham, A.P. and K.K. Ravi, Web page classification based on document structure. Awarded second prize in National Level Student Paper Contest conducted by IEEE India Council., in 2001.

2. Pierre, J.M., On the Automated Classification of Web Sites. Linköping Electronic Articles in Computer and Information Science, 2001. **6**(1).
3. Qi, X. and B.D. Davison, Web page classification: Features and algorithms. ACM Comput. Surv., 2009. **41**(2): p. 1-31.
4. Manning, C.D., P. Raghavan, and H. Schtze, Introduction to Information Retrieval2008, New York, NY, USA: Cambridge University Press. 496.
5. Lewis, D.D., Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval, in Proceedings of the 10th European Conference on Machine Learning1998, Springer-Verlag. p. 4-15.
6. Vapnik, V., Statistical learning theory1998 NY: Springer-Verlag.
7. Vapnik, V.N., The nature of statistical learning theory1995, New York, NY, USA: Springer-Verlag New York, Inc.
8. Burges, C.J.C., A Tutorial on Support Vector Machines for Pattern Recognition. Data Min. Knowl. Discov., 1998. **2**(2): p. 121-167.
9. Soucy, P. and G.W. Mineau, A Simple KNN Algorithm for Text Categorization, in Proceedings of the 2001 IEEE International Conference on Data Mining2001, IEEE Computer Society. p. 647-648.
10. Han, E.-H. and G. Karypis, Centroid-Based Document Classification: Analysis and Experimental Results, in Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery2000, Springer-Verlag. p. 424-431.
11. Hotho, A., S. Staab, and G. Stumme, Text clustering based on background knowledge, 2003.
12. Guisse, A., K. Khelif, and M. Collard, PatClust : une plateforme pour la classification sémantique des brevets, in Conférence d'Ingénierie des connaissances2009: Hammamet, Tunisie.
13. Huang, A., Similarity measures for text document clustering. Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008), Christchurch, New Zealand, 2008: p. 49-56.
14. Salton, G., The SMART Retrieval System-Experiments in Automatic Document Processing1971: Prentice-Hall, Inc.
15. Lan, M., et al., Supervised and Traditional Term Weighting Methods for Automatic Text Categorization. IEEE Trans. Pattern Anal. Mach. Intell., 2009. **31**(4): p. 721-735.
16. Rennie, J. Home Page for 20 Newsgroups Data Set. Mon Jan 14 2008; Available from: http://people.csail.mit.edu/jrennie/20Newsgroups.
17. Sokolova, M. and G. Lapalme, A systematic analysis of performance measures for classification tasks. Information Processing &amp; Management, 2009. **45**(4): p. 427-437.
18. Bloehdorn, S. and A. Hotho, Boosting for text classification with semantic features, in Proceedings of the 6th international conference on Knowledge Discovery on the Web: advances in Web Mining and Web Usage Analysis2006, Springer-Verlag: Seattle, WA. p. 149-166.
19. Wang, P. and C. Domeniconi, Building semantic kernels for text classification using wikipedia, in Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining2008, ACM: Las Vegas, Nevada, USA. p. 713-721.
20. Pirro, G., A semantic similarity metric combining features and intrinsic information content. Data Knowl. Eng., 2009. **68**(11): p. 1289-1308.
21. Lee, J.-J., et al., Novel web page classification techniques in contextual advertising, in Proceedings of the eleventh international workshop on Web information and data management2009, ACM: Hong Kong, China. p. 39-47.