AUTOMATED BLOG CLASSIFICATION: A CROSS DOMAIN APPROACH

Elisabeth Lex
Know-Center Graz
Inffeldgasse 21a, Austria

Christin Seifert
Know-Center Graz
Inffeldgasse 21a, Austria

Michael Granitzer
Know-Center Graz,
Graz University of Technology
Inffeldgasse 21a, Austria

Andreas Juffinger
Know-Center Graz
Inffeldgasse 21a, Austria

ABSTRACT

The automated classification of blogs is highly important for the relatively new field of blog analysis. To classify blogs into topics or other categories, usually supervised text classification algorithms are applied. However, supervised text classifiers need a sufficient large amount of labeled data to learn a good model. Especially for blogs, data labeled with terms that capture current and actual topics are not available and data labeled in the past is usually not applicable due to topic drifts. Besides, tagged blogs collected from the web exhibit a vocabulary that is rather heterogeneous, diverse and not commonly agreed upon. In our work, we focus on news-related blogs dealing with current events. Our goal is to classify blog posts into given, common newspaper categories. As a baseline, we have high quality labeled data from a German news corpus. Our approach is to exploit the labeled data from the news corpus and use this knowledge to perform cross-domain classification on the unlabeled blogs. We need a solution with high performance, because both our corpora are dynamic and our classifier model needs to be up-to-date. In this work, we evaluated a number of text classification algorithms with different parameter settings by means of accuracy and complexity. Qualitative and quantitative analysis revealed that a recently proposed centroid-based algorithm, the Class-Feature-Centroid classifier (CFC), serves best for our setting because it achieves a comparable accuracy with state-of-the-art text classifiers and outperforms all other algorithms regarding complexity and memory consumption.

KEYWORDS

Blogs, Classification, Cross-Domain

INTRODUCTION

Technorati reports in its "State of the Blogosphere 2008" that *"Blogs are Pervasive and Part of Our Daily Lives"*[1]. People use blogs to express their thoughts, ideas and opinions. Therefore, the analysis of the blog

1 http://technorati.com/blogging/state-of-the-blogosphere

response to news stories and current events might reveal different aspects, opinions and perspectives about these stories. Naturally, most users are interested only in specific topics within the blogosphere. Therefore, the classification of blogs into commonly agreed topics is crucial. Usually, machine learning algorithms, supervised text classification algorithms respectively, are used to assign blogs to a set of categories [Ikeda D., Takamura H. and Okumura M., 2008]. However, supervised text classification algorithms require a large amount of labeled training data to construct a classifier model that reflects the data's most characteristic features [Sebastiani F., 2002]. Especially for blogs, data labeled with terms that capture current and actual topics are not available and data labeled in the past is usually not applicable due to topic drifts. Besides, the vocabulary of tagged blogs collected from the web is rather heterogeneous, diverse and not commonly agreed upon [Farooq U. et al, 2007]. Recent work identified two types of taggers, describers and categorizers [Körner C, 2009], who tag resouces quite differently. The impact of this different tagging motivation on tag-based classification is not clear, but it might be a reason for the heterogeneous tagging vocabulary. Also, usually tagging vocabulary is rather subjective and may change over time.

When no training data is available for a domain, a possible approach is to train on a different domain with similar characteristics and features. This procedure is refered to as *cross-domain classification*. For text classification, documents are investigated using the vector space model (VSM). In the VSM, each document is represented by a vector of terms and each term represents a feature. Therefore, in text classification, the vocabulary is highly important because a classifier model trained for a vector space depends on the vocabulary of the domain.

In previous work, several approaches were proposed to solve this cross-domain classification task. In [Xue G.R. et al, 2008] an algorithm is introduced that uses Latent Semantic Analysis (LSA) [Deerwester S. et al, 1990] to exploit common topics and the semantic relations between different domains. With the information gathered from the common topics and the semantic relations, text classification in the target domain is amplified. This approach reveals good results yet at high complexity costs. In [Wang P., Domeniconi C. and Hu J., 2008] also the latent semantic relationships between domains are analyzed and further investigated using a co-clustering methodology. Also, the main drawback of this approach is the high computational complexity.

In our work, we are interested in news-related blogs dealing with current events from the news world. Our goal is to investigate the response to news in the blogosphere. For this, it is also necessary to categorize the blogs into news categories that are commonly agreed upon. The baseline for our work is a German news corpus [Lex E. et al, 2008]. The articles in the corpus were manually categorized into common newspaper categories by newspaper editors. Our blog corpus exhibits similar terms like our news corpus [Juffinger A., Granitzer M. and Lex E., 2009] and we identified that terms, significant for a topic, remain the same across both corpora. Therefore, our approach is to exploit the labeled data from the German news corpus and to apply cross-domain classification to classify the unlabeled blogs. Due to the fact that both our corpora (news and blogs) are dynamic and hourly growing with a topic drift, we need a fast text classification algorithm. Also, the classifier model must be able to deal with the topic drift.

In this work, we apply several text classification algorithms on this cross-domain classification task and evaluate the performance of all algorithms for different scenarios. We claim that text classifiers are sufficiently generalizable and can be applied to our problem. For a qualitative and quantitative evaluation, we analyze the accuracy and the performance of all algorithms.

The remainder of this paper is structured as follows: Section 2 describes our cross-domain classification task. In Section 3, the results and discussions are given. In Section 4, an additional qualitative analysis is provided. Finally, we conclude our work and give an outlook on future work in Section 5.

CROSS-DOMAIN CLASSIFICATION TASK

We aim at solving a cross-domain multi-class problem with five classes. We are provided with a news corpus that consists of Austrian newspapers. News paper editors manually labeled this corpus and categorized it in five common newspaper categories: politics, economy, sports, culture and science. Our blog corpus was crawled from the World Wide Web.

We implemented and evaluated the following text classifiers: a Support Vector Machine based on LibLinear [R.-E. Fan et al. 2008], a k-NN algorithm [Aha D.W., Kibler D. and Albert M.K., 199] and a centroid-based

text classifier that was recently introduced in [Guan H., Zhou J. and Guo M., 2009], the Class Feature Centroid (CFC). This algorithm is described in more detail in Section 3.1. This algorithm is claimed to be very fast and efficient and to outperform SVM and all other centroid-based text classifiers. For the experiments, we vectorized our documents with an information extraction and vectorization module based on OpenNLP[2]. With the retrieved Part-Of-Speech tags, we constructed a vector space based on nouns only.
For our evaluations, we performed a 10-fold cross-validation on both the news and the blog corpus. Besides, we evaluated the performance of all classifiers on six different scenarios:

- [NN]: We train the classifiers on the news corpus and measure the classifier performance on a news evaluation set which was not part of the training set.
- [BB]: We train the classifiers on the blog corpus and measure the classifier performance on the blog evaluation set which was not part of the training set.
- [NB]: We train the classifiers on the news corpus and measure the classifier performance on the blog evaluation set which was not part of the training set.
- [BN]: We train the classifiers on the blog corpus and measure the classifier performance on the news evaluation set which was not part of the training set.
- [NB-B]: We train the classifiers on both the news AND the blog corpus and measure the classifier performance on solely a blog evaluation set which was not part of the training set.
- [NB-N]: We train the classifiers on both the news AND the blog corpus and measure the classifier performance on solely a news evaluation set which was not part of the training set.

From these different settings, we aim to make statements about the generalization ability of all classifiers. Our assumption is that the settings NN and BB exhibit the best results because they are not cross-domain tasks. Therefore these settings serve as a baseline for the cross-domain tasks. We also perform training on a combination of both corpora injecting different amounts of documents from the particular target domain (settings NB-B and NB-N). The assumption is that classifiers trained on a mixture of both news and blogs documents are more accurate because they are able to better capture the vocabulary of both datasets. From these experiments, comparisons can be made with the results from NB and BB. The results of these experiments are described in Section 3.

2.1 Dataset Description

For our experiments, we used a news corpus that contains around 28k news articles in German language. The corpus contains approximately 237k nouns whereas the average document length is 92.5 nouns. The classes are balanced and all hold around 56k documents. The blog corpus consists of approximately 11k blog posts that come from 56 blogs. Those blogs were manually chosen so their topics match the target newspaper categories:

Table 1. Description of blog corpus

Category	# blogs	# blog posts
politics	10	~2800
economy	10	~2800
sports	10	~2400
culture	11	~1400
science	15	~1100

The blogs are labeled with one of the newspaper categories and therefore all blog posts of a blog are assigned this class. We randomly checked whether these assignments are correct but we did not check all blog posts so there might be some mislabeled data. Clearly, the theoretically achievable accuracy is therefore lower than 100%. The blog corpus contains approximately 110k nouns whereas the average document length is 61.5 nouns. We analyzed the term distribution of both corpora to investigate how many terms both dictionaries

2 http://opennlp.sourceforge.net

share. The sum of terms in the news and the blog dictionary is 347k and when merging both dictionaries to one, this contains 302k terms. Consequently, it can be derived that both dictionaries share only 45k terms.

2.2 Parameter Settings

To weight the document vectors, we used BM25 weighting [Jones K.S., Walker S. and Robertson S.E., 2000] for k-NN and SVM with the standard parameters *k=2* and *b=0.75*. For CFC, we used a TF-IDF weighting, as recommended by the authors. The parameter *k* of the k-NN algorithm was varied from *5*, *10* to *15* and we identified *k=15* to be the best. For the SVM, we used a L2-loss SVM parameterized with standard values (cost *C=1*). We also experimented with various values for parameter *b* in CFC. However, different from findings in the original paper, where *b=e-1.7*, we found that *b=e-1.0* performs equally well for our problem.

RESULTS AND DISCUSSION

For evaluation, we first investigated the statistical differences between the news and the blog corpus by computation of the Kullback-Leibler divergence (KL) [Kullback S. and Leibler R., 1951] and the Bhattacharyya similarity [Bhattacharyya, A., 1943]. The Kullback-Leibler divergence measures the difference between two probability distributions B and N and is given by:

$$KL(B\|N) = \sum_t [P_B(t) \log(P_B(t)/P_N(t))]$$

Whereas $P_B(t)$ denotes the probability of term t in corpus B and $P_N(t)$ the probability of term t in corpus N. The Bhattacharyya coefficient measures the amount of overlap between two statistical samples. The coefficient can be used to determine the relative closeness of the two samples. The Bhattacharyya coefficient is given in the following Equation whereas p and q denote discrete probability distributions in the domain X:

$$BH(p,q) = \sum_{x \in X} \sqrt{p(x)q(x)}$$

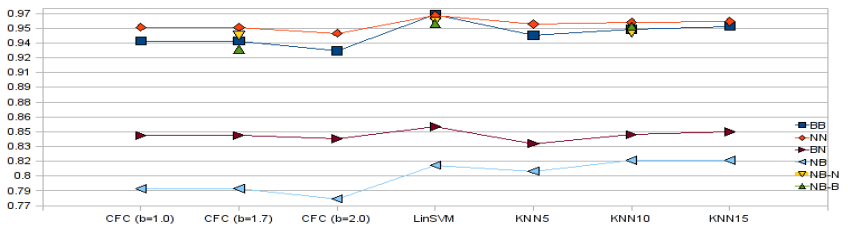
Table 2 shows the Kullback-Leibler divergence and Bhattacharyya similarity for our cross-domain scenarios:

Table 2. Kullback-Leibler divergence and Bhattacharyya Similarity for Blog vs. News Corpus

	BlogNews	NewsBlog
KL Global	0,535	0,430
BH Global	0,335	0,138

The KL divergences 0.535 (BN) and 0.430 (NB) show a clear statistical divergence between the term probability distribution of both corpora. The low values for BH similarity confirm the low statistical overlap.
We then carried out the experiments described in Section 2 and computed the mean accuracy for a 10-fold cross-validation and the standard deviation. First, we evaluated scenarios NN and BB because these experiments determine the maximum achievable performance (due to possibly mislabeled blog posts, see Section 2.1). All algorithms achieve a good performance for NN and BB. We then evaluated the classifier performances on the cross-domain scenarios. The results for our evaluations are shown in Figure 1.

Figure 1: Evaluation results for all scenarios



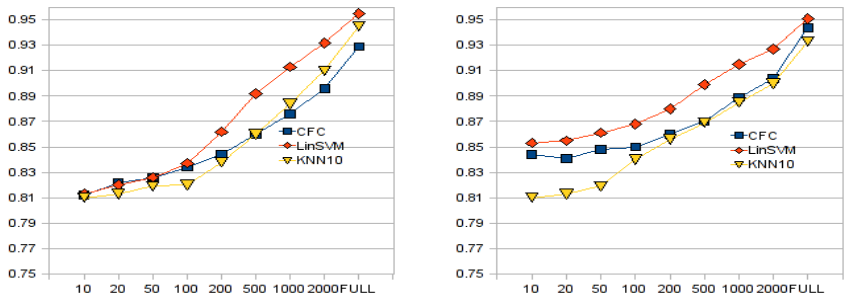
In the cross-domain tasks (NB, BN), the performance of all classifiers significantly decreases. The SVM achieves the best results for scenario BN, the k-NN is second and the CFC algorithm performs slightly worse than k-NN. In scenario NB, which is the most important for our research, the k-NN (mean: 0.82, standard deviation: 0.01) is slightly better than the SVM (mean: 0.81, standard deviation: 0.006) and the CFC algorithm (mean: 0.79, standard deviation: 0.003). The results show that in many cases, the SVM performs best and the k-NN slightly better than the CFC, although all accuracy values are within a range of 0.01. However, when analyzing the computation time (see Table 3), the CFC outperforms both the SVM and the k-NN algorithm.

Table 3. Training and testing time, mean and standard deviation for all algorithms on scenario NB

	Train: mean /std dev.	Test: mean/std dev.
CFC	9.982s / 0.3s	0.197s / 0.1s
SVM	37s / 0.3s	0.18s / 0.003s
k-NN	4.6s / 0.2s	123s / 0.9s

At last we evaluated scenarios NB-N and NB-B. In these scenarios, the training set is step-wise enhanced with documents from the target domain. For instance, we trained on the news data and injected 100 blog posts and evaluated the classifier performance on the remaining blog data.
Consequently, we assume to achieve a higher accuracy for these scenarios than for the pure cross-domain tasks. The aim of these experiments was to estimate, how many blog posts are needed to enrich the news based training set. From these experiments, we wanted to derive whether it is feasible to annotate a few blog posts to significantly improve the performance. The results for this scenario are depicted in Figure 2.

Figure 2. Algorithm performance for scenarios NB-B and NB-N



In the left figure, the amount of blogs is varied whereas in the right figure the amount of news documents is varied. Figure 2 reveals that approximately 200 blog posts are needed to significantly improve the classifier accuracy. Further, the CFC algorithm apparently needs a lower amount of target-domain documents to increase the overall accuracy. For example, the CFC performance increases with 6%, opposite to SVM (4%) and k-NN (2%) when 200 labeled blog posts are added to the news data.
From all experiments, we decided that because we need a classifier that is very fast and whose performance is good, to deeper analyze the CFC due to its outstanding performance (see Section 3.1).

3.1 CFC Algorithm

The CFC algorithm was recently proposed by Guan et al. in [Guan H., Zhou J. and Guo M., 2009]. Generally, centroid-based classifiers achieve good values for accuracy and especially time complexity [Han E.-H. and Karypis G., 2000], although they require centroids with high quality [Lertnattee V. and Theeramunkong T., 2004]. With the CFC algorithm, a novel centroid weight representation is introduced, that considers both the inter-class term distribution and the inner-class term distribution. The formula for the weight representation is given below, whereas *i* represents a term, j denotes a class and *w* corresponds to the term weight of term *i* of class *j*. The value *b* is a constant larger 1. DF^{j}_{ti} represents the terms document frequency in a class, |*Cj*| denotes the number of documents in class *j* and |C| the number of documents of all

classes. The term CF_{ti} represents the number of classes containing term *i*. The first component of following equation represents the inner-class term index and the second the inter-term index:

$$w_{ij} = b^{\frac{DF_{ti}^{j}}{|C_j|}} \times \log\frac{|C|}{CF_{ti}}$$

With this weighting schema, highly discriminative centroids (one for each class) can be derived because terms that occur over all classes are given a low weight and terms that occur in only one class a high weight. This guarantees that the most discriminative terms are assigned a high weight.

To compute the similarity of a test document with a class centroid, a so-called denormalized cosine similarity is applied. The similarity between a document vector and a centroid vector is computed using a standard cosine similarity, which is a common similarity measure when dealing with text. However, the centroids are not normalized in order to preserve their discriminative abilities.

Guan et al. compare the performance of the algorithm with other centroid-based approaches and with variants of SVMs. All experiments are carried out on the 20-newsgroup and the Reuters-21578 corpus. They reveal that the CFC algorithm outperforms the centroid-based approaches as well as the SVMs. Note that, as outlined in [Sebastiani F., 2002], Support Vector Machines belong to the best performing text classification algorithms.

3.2 CFC Analysis

For a deeper CFC analysis, we evaluated the generalization ability of the CFC. We investigated the class centroids, more specifically the terms with weight $w>0$ (note that these are the most discriminative terms). We again computed the KL divergence (cross-validated) between the term distribution of the blog and the news corpus, only based on class centroid terms. The derived values for the KL divergence and the BH similarity are shown in Table 4.

Table 4. KL divergence for CFC centroids

	BN	Std.Dev.	NB	Std.Dev	BH Sim.
KL Local Politics	0,139	0,002	0,202	0,003	1,683
KL Local Economy	0,218	0,002	0,146	0,009	1,633
KL Local Sports	0,132	0,002	0,134	0,002	1,969
KL Local Culture	0,096	0,003	0,075	0,002	2,194
KL Local Science	0,049	0,001	0,093	0,008	2,487

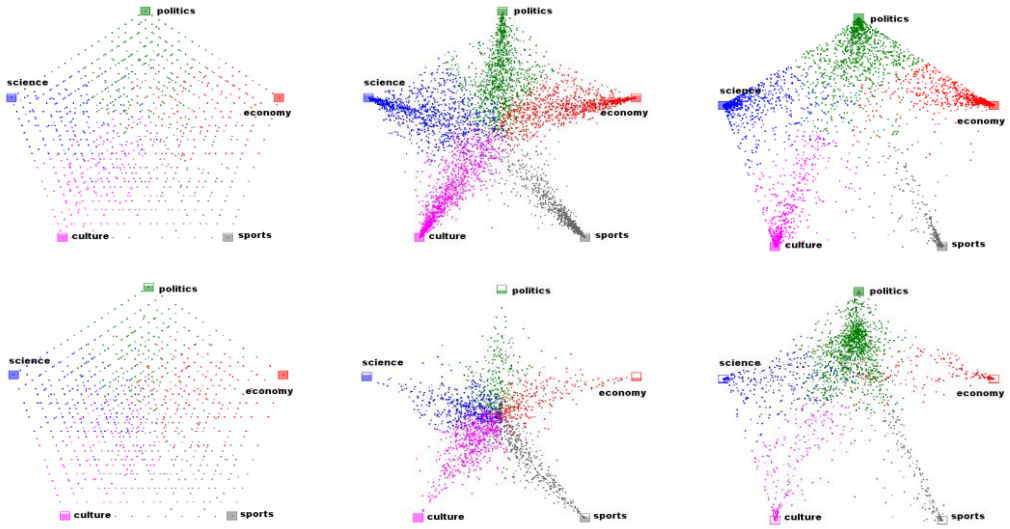
The KL divergence decreases significantly which indicates that the CFC algorithm actually selects only the characteristic terms for a topic and that these terms stay the same across both corpora which indicates a good generalization ability. Again, the BH similarity corroborates this statistical finding. The good CFC generalization results derived from this analysis and the sound performance of the algorithm corroborate that the CFC is well suited for our setting.

QUALITATIVE ANALYSIS

In addition to the numerical evaluations in Section 3, we carried out qualitative analysis with respect to performance and complexity. For a visual performance analysis, we used the classification visualization introduced in [Seifert C. and Lex E., 2009] to investigate the confidence distribution of our classifiers. The classification visualization shows the a-posteriori probabilities for all test documents. Therefore, the outputs of the Support Vector Machine and of the CFC classifier have to be mapped to a probability distribution (see [Platt J.C., 1999] for a description of the procedure). Within the visualization, all classes are placed around the perimeter of a circle and are represented by squares colored in unique colors. The classifier models are generated on the particular training documents and their decisions on the test documents are visualized (they are pictured in the color of the class to which the classifier assigned them). In addition to the visualization, common performance measures like precision and recall are computed and available in a tabular view. The classification visualization enables us to evaluate the classifier performance both numerically and visually and to compare all classifiers in terms of accuracy. We used the visualization to deeper investigate the

performance of the classifiers in our main research scenario NB. The visual results are illustrated in Figure 3, the above figures depict the one-domain scenario NN, and the below figures our research cross-domain scenario NB. It can be seen that the k-NN algorithm is discretely distributed which is clear when taking into consideration that between classes a maximum of 11 discrete confidence steps can occur. From the visualization, it can be derived that obviously the k-NN cannot clearly decide between classes "science" and "culture" and also between "economy" and "sports", as indicated by their color mixture. The visualization also reveals that the SVM more likely assigns documents into the class "politics", because more items are placed near the "politics" class square.

Figure 3: Visualization for test item confidence for k-NN, CFC, and SVM



The distribution of the test documents in the NB scenario is quite similar like for NN which leads to the conclusion that the classifiers perform similarly on a cross-domain task. Regarding the probability distributions, the CFC algorithm is quite balanced and compared to the distribution of the SVM, fewer items are undecided (Note that undecided item are placed rather in the middle of the circle or at the outer boundaries between classes, the nearer an item is placed to a class square, the higher the confidence for the assignment). For example between the categories "politics" and "economy", as well as "politics" and "science", the CFC places fewer items at undecided places than SVM. The visualization also reveals that the CFC does not prefer any class, in contrast to SVM ("politics").

CONCLUSION

In this work, we performed cross-domain classification from the news domain to the blog domain. We applied a novel centroid-based text classifier, the Class-Feature-Centroid classifier (CFC), to our problem setting. We compared the results of this classifier with a fast linear Support Vector Machine and a standard k-NN algorithm. Our experiments revealed that the CFC classifier achieves similar results like SVM and k-NN but it outperforms both algorithms regarding computational speed. Also, in terms of memory consumption, the CFC outperforms SVM and k-NN. In our setting, the CFC has to store only the five centroid vectors, the SVM has to store all support vectors, and the k-NN the whole training set (17k for scenario NN and NB). These results led to our conclusion that the CFC is most suited for our cross-domain task because in our work we need to re-train the classifier model on the dynamic news corpus several times per day and instantly re-classify the hourly growing blog corpus. Furthermore, the experiments indicated that with respect to the generalization ability, the CFC generalizes better than the other two algorithms because its accuracy decreases less from the mono-domain task to the cross-domain task. Besides, the visual evaluation revealed that the CFC exhibits a more trustworthy confidence distribution because it results in less undecidable items.

We also investigated whether it is feasible to label a subset of blogs and to enhance the news data with these labeled items. This experiment showed that for all classifiers, at least 200 labeled blogs are needed to improve classifier accuracy, but at any rate, it is worth to put some additional effort in labeling a few items from the target domain. For future work, we plan to extend our work to cross-language classification because clearly performing a blog analysis over different languages would be of great interest. Besides, we will publish our datasets online to establish a common cross-domain classification dataset.

ACKNOWLEDGEMENT

APOSDLE is partially funded under the FP6 of the European Commission within the IST Workprogramme (project number 027023). The Know-Center is funded within the Austrian COMET Program – Competence Centers for Excellent Technologies – under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

REFERENCES

Aha D.W., Kibler D. and Albert M.K., 1991. Instance-based learning algorithms. *Mach. Learn.*, 6(1):139-149

Deerwester S. et al, 1990. Indexing by latend semantic analysis. *Journal of American Society for Inf. Science*, 41.

Bhattacharyya, A., 1943. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society* **35**: 99–109.

R.-E. Fan et al. 2008. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9(2008) pp. 1871-1874. Software available at http://www.csie.ntu.edu.tw/~cjlin/liblinear

Farooq U. et al, 2007. Evaluating tagging behaviour in social bookmark systems: metrics and design heuristics. *Proceedings of Int. ACM Conference on Supporting Group Work (GROUP)*. New York, USA, pp. 351-360.

Guan H., Zhou J. and Guo M., 2009. A class-feature-centroid classifier for text categorization. *Proceedings of Int. Conf. on World Wide Web (WWW)*. New York, USA.

Han E.-H. and Karypis G., 2000. Centroid-based document classification: Analysis and experimental results. *Proc. of European Conf. on Principles of Data Mining and Knowledge Discovery (PKDD)*. London, UK, pp. 424-431.

Ikeda D., Takamura H. and Okumura M., 2008. Semi-Supervised Learning for Blog Classification. *AAAI*.

Jones K.S., Walker S. and Robertson S.E., 2000. A probabilistic model of information retrieval: development and comparative experiments. *Inf. Process Management*, 36(6).

Juffinger A., Granitzer M. and Lex E., 2009. Blog credibility ranking by exploiting verified content. *Proceedings of Workshop on Information Credibility on the Web (WICOW)*. New York, USA, pp. 51-58.

Körner C, 2009. The Motivation behind Tagging. *ACM Student Research Competition, Hypertext 2009*. Turin, Italy.

Kullback S. and Leibler R., 1951. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79-86

Lertnattee V. and Theeramunkong T., 2004. Effect of term distributions on centroid-based text categorization. *Information Sciences, Informatics and Computer Science*.

Lex E. et al, 2008. A generic framework for visualizing the news article domain and its application to real-world data. *Journal of Digital Information Management*, 6(6): pp. 434-442.

Platt J.C., 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*.

Alippi C., Roveri M., 2007. Reducing Computational Complexity in k-NN based Adaptive Classifiers. *IEEE International Conf. Computational Intelligence for Measurement Systems and Applications (CIMSA)*, pp.68-71.

Shalev-Schwartz, S., Singer Y. and Srebro N, 2007. Pegasos: Primal Estimated sub-Gradient Solver for Svm. *ICML*.

Sebastiani F., 2002. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1) pp. 1-47.

Seifert C. and Lex E., 2009. A Novel Visualization Approach for Data-Mining-Related Classification. *Proc. of IV09*.

Wang P., Domeniconi C. and Hu J., 2008. Using Wikipedia for co-clustering based cross-domain text classification. *Proceedings of IEEE Int. Conference on Data Mining (ICDIM)*. Washington, DC, USA.

Xue G.R. et al, 2008. Topic-bridget PLSA for cross-domain text classification. *SIGIR08*. New York, USA, pp. 627-634.

Yang Y. and Liu X., 1999. A re-examination of text categorization methods. *SIGIR 1999*. New York, USA, pp. 42-49.