Applying Text Classification in Conference Management: Some Lessons Learned

© Andreas Pesenhofer, Helmut Berger, Michael Dittenbach, Andreas Rauber*

iSpaces Group, eCommerce Competence Center – ec3,
Donau-City-Straße 1, A–1220 Wien, Austria
{andreas.pesenhofer, helmut.berger, michael.dittenbach}@ec3.at
*Department of Software Technology and Interactive Systems
Vienna University of Technology,
Favoritenstraße 9–11/188, A–1040 Wien, Austria
rauber@ifs.tuwien.ac.at

Abstract

In this paper we report on experiments with automatic text classification systems in combination with a conference management system (TCeReview). Furthermore we describe how the automatic text classification system was trained, where it was integrated into the conference management system and what the advantages of such a system for conference organizers and scientists are.

1 Introduction

Automatic text classification (ATC) - the task of automatically assigning natural language documents to predefined categories - is a convenient way for handling and organizing document collections. A wide range of machine learning algorithms has been adopted for the task of document classification and has proven to be successful. Sebastiani mentions in [14, 13] application domains such as genre identification, authorship attribution, survey coding, document organization and text filtering. One emerging area during the last years is filtering of unsolicited e-mail [1].

The research side of text classification has been widely published via conferences and journals in Information Retrieval, Natural Language Processing, Machine Learning, and other fields. This research has emphasized on the effectiveness of supervised machine learning techniques on experimental data sets (e.g. OHSUMED[5], Reuters-21578[7], RCV1[8] or 20 Newsgroups[6]). In contrast, the amount of research in the area of real-world application is sparse especially in the domain of text classification. In the area of spam detection most of the widely-used spam detection products incorporate some kind of personalized text classifier with the majority using incrementally trained Naïve Bayes classifiers.

In this paper we will describe the experiments of two different real-world corpera, which were carried out with an ATC enhanced conference management system. Generally speaking, a conference management system is a digital library of a particular conference, where users provide, route and seek for information. In traditional conference management systems problems can occur in the phase where the author has to choose the research topic under which the paper should be filed. The author can be irresolute and uncertain in selecting the topic of the paper. This feeling can be reinforced when the categories are not precisely described. In our system the classifier is trained with examples from previous conferences and so the author should be guided to the correct category.

The remainder of this paper is structured as follows. Section 2 gives an introduction to the used methods. Section 3 describes the experimental setup. The results are presented and discussed in Section 4. The related work in this area is presented in Section 5 and finally, a conclusion is given in Section 6.

2 Methods

2.1 Naïve Bayes

The Naïve Bayes classifier (NB) is a probabilistic classifier where the assumption is made that texts can be represented by different probability distributions (c.f. [10]). For a new document represented through a vector $\vec{d_j}$ the posterior probability is calculated for each class c_i, based on the Bayes Theorem:

$$P(c_i|\vec{d_j}) = \frac{P(c_i) \cdot P(\vec{d_j}\,|\,c_i)}{P(\vec{d_j})} \qquad (1)$$

The document $\vec{d_j}$ is assigned to the class where the highest value is obtained. $P(\vec{d_j})$ is the probability that any document $\vec{d_j}$ will be observed and $P(c_i)$ is the probability that the hypothesis c_i holds. The estimation of $P(\vec{d_j}\,|\,c_i)$ and $P(\vec{d_j})$ is problematic because of the possible high number of vectors $\vec{d_j}$. For reducing this problem the assumption is made that all terms (w_{kj}) of a document vector are pairwise

Proceedings of the 7th Russian Conference on Digital Libraries RCDL'2005, Pushchino, Russia, 2005

class-ID	class description	1998	1999	2000	2001	2002	2004	sum
1	Concepts of Digital Libraries, Concepts of Documents and Metadata	5	6	9	6	5	3	34
2	System Architectures, Open Archives, Collection Building, Integration and Interoperability	7	0	3	3	15	12	40
3	Information Retrieval, Information Organization, Search and Usage	22	7	12	11	5	10	67
4	User Studies, System Evaluation, Personalization, User Interfaces and User Centered Design	14	5	11	4	3	13	50
5	Digital Preservation, Web Archiving and Long Term Access	2	0	1	1	7	1	12
6	Digital Library Applications and Case Studies	23	3	22	7	9	1	65
7	Multimedia, Mixed Media, Audio, Video, 3D and non-traditional Objects	6	6	12	6	6	7	43
sum over the selected abstracts		79	27	70	38	50	47	311

Table 1: Research topics of the ECDL 2005 & selected abstracts of previous conferences

different, thus they are statistical independent. This assumption is revealed by the following formula:

$$P(\vec{d_j}|c_i) = \prod_{k=1}^{|\mathcal{T}|} P(w_{kj}|c_i) \quad (2)$$

The "naïve" character is due to the fact that usually this assumption is not verified in practice.

2.2 Information gain

In the high-dimensional vector spaces only these terms should be used that are descriptive for a document or more general spoken for a category. One solution for this task is to compute the Information gain (IG) for each unique term [17]. After ranking the terms so that the highest is at the first position we can remove terms that are below a predefined threshold.

IG measures the number of bits of information obtained for category prediction by knowing the presence or absence of a term in a document [17]. It is frequently employed as a term goodness criterion in the field of machine learning. The information gain of term t is defined as:

$$\begin{aligned} IG(t) &= -\sum_{i=1}^{m} P(c_i)\log P(c_i) \\ &\quad +P(t)\sum_{i=1}^{m} P(c_i|t)\log P(c_i|t) \\ &\quad +P(\bar{t})\sum_{i=1}^{m} P(c_i|\bar{t})\log P(c_i|\bar{t}) \quad (3) \end{aligned}$$

where $\{c_i\}_{i=1}^{m}$ is the set of categories in the target space and $P(c_i)$ is the probability of the category c_i. $P(t)$ is the probability that t occurs in the collection, $P(c_i|t)$ is the probability that a category is c_i, given the term t appears, and $P(c_i|\bar{t})$ is the probability that a category is c_i, given the term t does not appear.

2.3 Performance Measures

In order to evaluate the effectiveness of text classification algorithms the standard precision π, cf. Equation (4), and recall ρ, cf. Equation (5), measures were used:

$$\pi = \frac{TP}{TP+FP} \quad (4)$$

$$\rho = \frac{TP}{TP+FN} \quad (5)$$

TP (true positives) is the number of positive test documents correctly classified; TN (true negatives) is the number of negative test documents, that are correctly classified. FP (false positives) is the number of positive test documents incorrectly classified and FN (false negatives) are defined accordingly.

The F_1-measure, cf. Equation (6), as described in [16] combines the standard π and ρ with an equal weight as shown in Equation (6).

$$F_1 = \frac{2\cdot\pi\cdot\rho}{\pi+\rho} \quad (6)$$

The percentage of correctly classified instances is assessed by the *Accuracy* measure. It calculates the proportion of the number of correctly classified instances on the total number of instances in the collection. Formally,

$$\mathit{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

Because we deal with a single-label TC task, i.e. each document belongs exactly to one category, the accuracy can be used as a measurement of the effectiveness as stated in [14].

Additional to that we will use an *Accuracy* based measurement called *One-of-best-n* (cf. [18]), where n determines how large the window should be. For example *One-of-best-1* is equal to the *Accuracy*. In the case *One-of-best-2* the classes with the first and second highest probabilities are taken into consideration. The value of n is limited by the amount of available classes, so that a value of 1 for the measurement infer from the maximum value for n.

3 Experimental Setup

The experiments will focus on two different corpera where TCeReview was used. The two scenarios can be described in the following way:

- The European Conference on Research and Advanced Technology for Digital Libraries (ECDL) is organized by the Vienna University of Technology this year. In order to ease the paper submission procedure of the conference we integrated an ATC into the conference management system. The combined system will be

class-ID	class description	2003	2004
1	Abdominal and Gastrointestinal	160	119
2	Breast	80	59
3	Cardiac	70	70
4	Chest	60	70
5	Computer Applications	30	30
6	Contrast Media	40	39
7	Genitourinary	70	60
8	Head and Neck	40	40
9	Interventional Radiology	130	117
10	Musculoskeletal	90	80
11	Neuro	90	99
12	Pediatric	30	40
13	Physics in Radiology	40	40
14	Radiographers	10	10
15	Vascular	69	70
sum		1009	943

Table 2: Topics of the ECR corpus & training documents

called TCeReview in this paper. We focused on the paper registration and abstract submission step. Based on the abstract TCeReview determines the category of the paper and returns the suggestion to the user, subsequently she or he has the possibility to change the category to a different one, which will be tracked for the evaluation.

- TCeReview is evaluated in an offline study with abstracts from the European Congress for Radiology (ECR). Here the correct classes for the test set documents are known from the beginning.

3.1 ECDL Corpus

The data used comprises English abstracts of previous ECDL conferences. In order to build the training set, all available data from previous conferences were downloaded via the on-line service of Springer Online[1]. The abstracts of the years 1998 to 2002 and of the year 2004 were available on-line at the time of generating the corpus. Based on the topics stated in the conference call we defined seven categories. Table 1 details the different categories and corresponding class identifiers. The abstracts of previous conferences are grouped into sessions in the table of contents (e.g. Digital Library Architecture, Evaluation and Usability, Web Archiving, ...). Some of these sessions can be found in multiple proceedings, others are only present a single time. We manually mapped the abstracts into one of the seven categories. The results are shown in Table 1, where only the relevant topics are taken into consideration. In sum over all years and categories 311 abstracts which are not equally distributed over the different categories were selected. The largest category consists of 67 documents and the smallest contains only 12 documents. Furthermore, some years contribute more than others.

Before any of the before described methods can be applied the documents have to be pre-processed and transformed into a representation that is understandable by the algorithms. Therefore we used the well known bag-of-words approach to index the documents and as weighting scheme *tfidf* [12] was used. *tfidf* is based on the term frequency (tf) in the given document and the inverse document frequency (idf) of the term in the whole collection. Pre-processing steps in form of removing all numbers, punctuation marks and special characters were applied. The remaining characters were transformed to lower case and then indexed with the Rainbow library [9]. Note that the blank character was used as word separator. We did not apply any stop word list nor a stemmer. In the end, the corpus consisted of 4,141 unique terms.

Based on this corpus we calculated the information gain for each term. Based on the resulting ranked list we empirically determined the best threshold. This was performed by selecting the top n terms and performing 10 single Naïve Bayes classification runs with this amount of terms where the trainset-testset split was 70:30. The training documents were randomly picked from the corpus. The selection of the 3,460 top ranked terms produced the best model, where the average accuracy over all category is 58.60% as shown in Figure 1.

3.2 ECR Corpus

This corpus consists of the abstracts of the European Congress for Radiology (ECR)[2] from three years, where the first two years (2003 and 2004) were used as training set and the year 2005 as test set. Altogether these are 2,876 English documents which were presented during the scientific sessions of the congress and each belongs to one of the 15 different topics (c.f. Table 2). Every abstract is assigned exactly to one topic. The distribution of the documents over the different categories is comparable to

[1]www.springerlink.com

[2]www.ecr.org

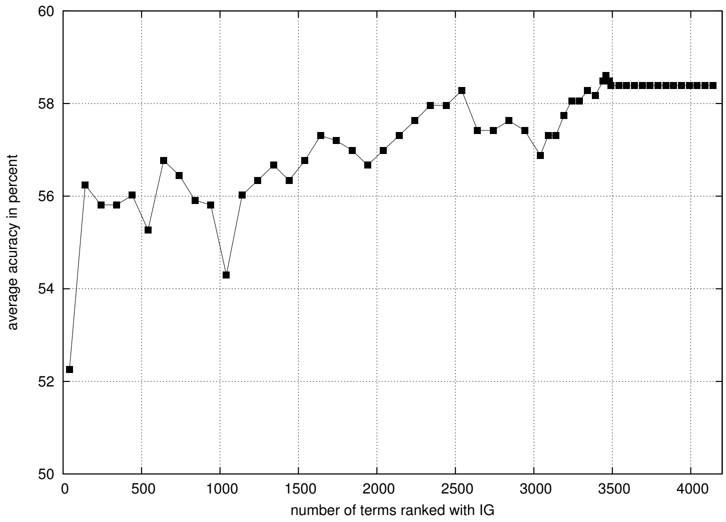

Figure 1: Comparison of feature selection levels for the ECDL Corpus

the ECDL corpus described before. The main difference is that more training documents are present in the specific classes. We applied the same preprocessing steps as in Section 3.1. In the end, the trainings set consisted of 14,887 unique terms. The selection of the best model was also made conforming to Section 3.1. The selection of the 5,720 top ranked terms produced the best model with an average accuracy over all category of 73.57% as shown in Figure 2.

3.3 Integration into the conference management system

Next we used the top 3,460 terms based information gain ranking and trained a Naïve Bayes classifier on the complete training set of the ECDL corpus. This classifier was integrated in the first phase of the web-based conference management system MyReview [11] where the authors submitted their contributions to the conference. Based on the abstract the system decides under which category the paper should be filed and stores the decision. We separated the task of submitting a paper into the stages registration of an abstract and the submission of the final paper. In the first stage the authors had to provide their contact information, their email address, the title and the abstract of the paper. They had no possibility to choose the topic. After submitting these details the authors received the upload code for the final version. On the webpage for the second stage, which could only be accessed with the upload code, the automatically determined topic was pre-selected in a drop-down-box. The user did not know that the selection of the topic has been generated automatically. She or he had the possibility to change it to another topic, if she or he was not satisfied with the automatic estimated topic. We tracked the action of the user so that we could evaluate the performance of the classifier. Furthermore, the abstracts of the first stage were stored for the evaluation, because the user could access the second stage as often she or he wanted to modify of their submissions. Only the last version of the submission was kept in the system.

In the ECR offline study we used the 5,720 top ranked terms and trained a Naïve Bayes classifier on the training set of the corpus. The remaining 924 test documents were submitted by a script to TCeReview. The categories for the documents were present, so no human interaction was necessary.

In our work we performed for the ECDL corpus an ex-post evaluation which means that we measured the disagreement and not the agreement with a special class. This can lead to different results when a paper can be seen in more than one classes. Furthermore, the user can be positive influenced with the preselected class. This evaluation differs compared to other studies [8, 15] and the evaluation of the ECR offline study, where the membership of the test documents is available from the beginning.

4 Results and Discussion

4.1 ECDL Corpus

While the submission site was open, 132 abstracts were received. In seven cases the authors did not complete the second stage, so these papers could not be considered for the evaluation of the system. For the remaining 125 papers TCeReview proposed the correct category in 61.60% of all cases. In-depth examination showed that 3 abstracts were shorter than 100 characters. In two cases the authors posted only the title in the abstract field and in the third case the abstract consisted only of one word. After removing these abstracts the accuracy improved to 65.57% which is depicted in Table 3. This confusion matrix obtained with the classifier, where rows give the class assignments and columns correspond to the prediction of the classifier gives an impression where the authors and the classifier agree and dissagree. The classifier obtains in the classes 2, 3, 4 and 6 recall values from 66.67% to 90% and the precision lies in the interval from 66.67% to 71.79%. In class 7 all abstacts were correct (recall of 100%); users changed the pre-selected class 4 times, which

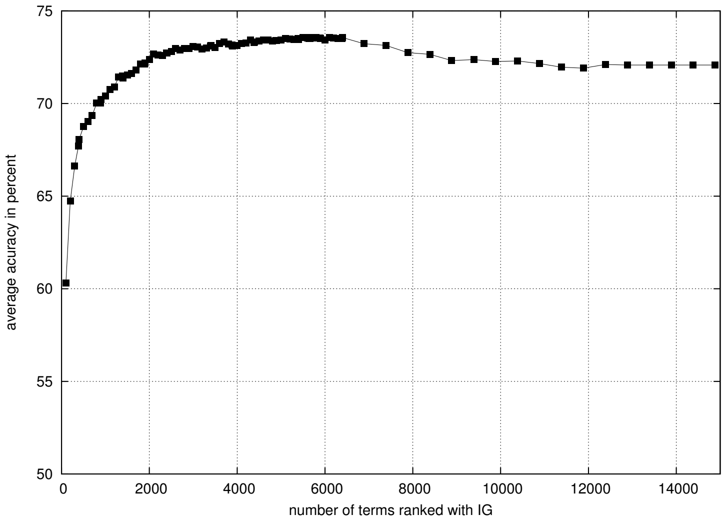

Figure 2: Comparison of feature selection levels for the trainings set of the ECR corpus

class-ID	**1**	**2**	**3**	**4**	**5**	**6**	**7**	total	recall	F_1
1	1	1	2	2	.	1	1	8	0.13	0.17
2	1	17	1	.	.	.	.	19	0.89	0.77
3	1	3	26	6	.	2	.	38	0.68	0.69
4	.	.	4	21	.	2	1	28	0.75	0.71
5	1	1	3	.	.	1	1	7	0.00	0.00
6	.	3	1	2	.	12	1	19	0.63	0.65
7	.	.	.	.	.	.	3	3	1.00	0.60
precision	0.25	0.68	0.70	0.68	0.00	0.67	0.43			

Table 3: Assignments of the classifier for the ECDL corpus

precipitated to a precision of 42.86%. The classes 1 and 5 perform poorly; the recall and precision of class 5 is 0%. This poor performance can be attributed to the small amount of training data for class 5 and the fact that the class deals with topics that are also present in other classes.

In our implementation we presented the user the category determined by the classifier with the highest probability, by preselecting this class in the interface. We did not return the ranked category list, starting with the highest probability, to the user. For evaluation purpose we looked up the 42 cases where the users changed the pre-selected class, if the classifier was also taken into consideration this class. The aggregated results, by summing up the correct classifications, for the first, second and third place are presented in Table 4. The *One-of-best-2* and *One-of-best-3* measurements are calculated class-specific and on the complete test-set. In particular we found out that for class 2 and 4 all changes were found on the second or third place - One-of-best-3 is 100%. In class 5 the changes were on the sixth and seventh place. None of them was under the top 3.

The bad performance of class 5, which deals with digital preservation, web archiving and long term access issues, was quite astonishing, so that we had a look at the content of each paper. The document that was assigned to class 1 deals with a new concept for digital preservation, that was not used before in this context. The classification to class 2 describes an architecture for digital preservation. The three papers that are assigned to class 3 describe e.g. an e-evidence generating system for web content, a system for managing duplicates in a web archive and a mutual deposition model between open archival information systems. The document that was assigned to class 6 deals with the description of key technologies for the creation of digital information. The abstract about an effective access to digital interview archives was assigned to class 7. Alltogether it shows that in most of the cases the classifier gets an impression about one of the content stream of a document which differs from the topic the author wants to see the document.

Figure 3 gives an impression of the probability levels of the first-ranked decisions over the different classes. Additionally, we marked correct classifications with a filled square and the false ones are marked with a star. For class 7 this means, that the three correct classifications were made with a probability of 95.44%, 99.13% and 100%. 5 of the seven false classification of class 5 fall in the interval from 98,45% to 99,98%, and the other two reach a value of 90.49% and 48.44%. In general it is not possible to say that all correct classifications are above a threshold of e.g. 80%.

class-ID	1.	2.	3.	4. to 7	total	best-of-2	best-of-3
1	1	0	3	4	8	12.50%	50.00%
2	17	1	1	0	19	94.74%	100.00%
3	26	4	1	7	38	78.95%	81.58%
4	21	3	4	0	28	85.71%	100.00%
5	0	0	0	7	7	0.00%	0.00%
6	12	2	2	3	19	73.68%	84.21%
7	3	0	0	0	3	100.00%	100.00%
best-of-n	65.57%	73.77%	74.59%				

Table 4: Ranking of the classification correctness

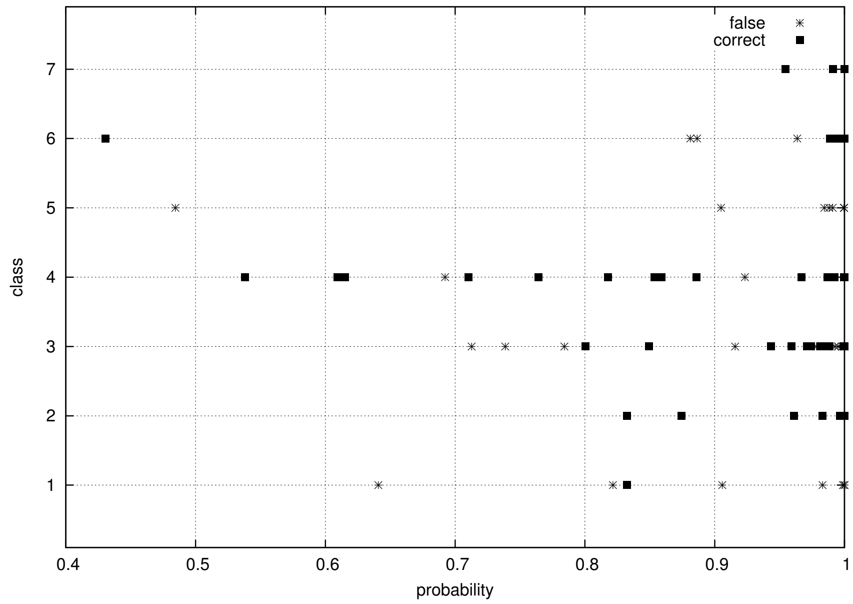

Figure 3: Classification probability

4.2 ECR Corpus

The evaluation of the 924 test documents from the ECR corpus showed that in 74.46% the correct class was determined. The confusion matrix and the recall and precision for the specific classes can be found in Table 5. The recall values of the classifiers are above 70% except for six classes (5, 6, 7, 8, 12 and 14). Only in four classes (5, 11, 13 and 15) the precision values are below 70%.

5 Related Work

Previous work in the area of assigning conference papers to reviewers had approached the problem as one of content-based information retrieval. Dumais and Nielsen used data provided by 15 members of the reviewing committee for the HYPERTEXT'91 conference [4]. These reviewers not only submitted abstracts of their papers and/or interests, but also provided complete relevance assessments for the 117 submitted papers. They used information retrieval principals and latent semantic indexing to generate the automatic assignments for each reviewer. So they achieved an average improvement of 48% with this method compared to the random assignment of articles to reviewers.

Yarowsky and Florian [18] focused on the classification of every paper to exactly one of six conference committees. They used 92 papers which were submitted to the ACL conference in electronic form and additionally requested committee members to provide representative papers so that a reviewer profile could be created. When the number of papers returned by these members was insufficient, they augmented the collection with other papers downloaded from online sources. The main algorithm first computed a centroid for each reviewer and then computed a centroid for each committee as the sum of its reviewer centroids. Then for each paper the cosine similarity was computed and compared with the committee centroids where the highest rank was the selection criteria. They also experimented with a Naïve Bayes classifier where their results outperformed the simple unsupervised model. Furthermore, they compared their systems with the performance of human judges on the same task. They concluded that the automatic methods could be as effective as human judges, especially in the case where the judges may be less experienced.

Paper recommendation is an other solution to handle the problem of assigning papers to reviewers. Basu et al. described a content-based system for

class-ID	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	total	recall	F_1
1	111	1	.	1	2	2	2	.	2	1	2	.	1	.	1	126	0.88	0.79
2	1	61	.	.	.	.	.	.	1	.	.	.	6	.	.	69	0.88	0.87
3	1	.	73	.	.	.	.	.	.	1	.	.	3	.	2	80	0.91	0.86
4	6	.	5	49	1	.	.	.	3	.	1	.	.	.	5	70	0.70	0.77
5	2	2	.	3	10	.	.	.	.	.	3	.	7	.	3	30	0.33	0.43
6	12	.	.	1	.	26	2	.	1	2	1	.	1	.	3	49	0.53	0.61
7	5	.	.	.	.	1	38	.	5	3	2	.	3	.	1	58	0.66	0.73
8	4	.	.	1	.	2	4	8	2	2	4	.	2	.	1	30	0.27	0.39
9	2	4	2	.	1	3	.	.	99	2	2	.	.	.	5	120	0.82	0.81
10	2	2	.	1	1	1	.	.	2	60	5	1	2	.	1	78	0.77	0.78
11	1	.	1	1	.	.	.	1	4	.	64	2	1	.	4	79	0.81	0.73
12	4	.	1	.	.	.	.	1	1	1	10	11	.	.	1	30	0.37	0.50
13	.	1	3	.	.	.	.	.	.	1	2	.	39	.	2	48	0.81	0.68
14	2	.	.	.	1	.	.	.	.	3	.	.	.	2	.	8	0.25	0.40
15	2	.	4	.	.	1	.	1	3	.	.	.	1	.	37	49	0.76	0.64
precision	0.72	0.86	0.82	0.86	0.63	0.72	0.83	0.73	0.80	0.79	0.67	0.79	0.59	1	0.56			

Table 5: Assignments of the classifier for the ECR corpus

technical paper recommendation based on different information sources [3]. They treated the problem as one of decomposing reviewer interest and paper content into information sources, and then of combining the information sources using different query formulations. In their experiments they compared two ways of formulating the queries, first the content-based information retrieval and second the collaborative based approach. The algorithm using conjunctive queries outperformed the other approaches. Furthermore they noticed a general performance increase if they used more information.

For automatically establishing semantic similarities among papers and allocating them into common themes a prototype matching system for conference papers was presented in [2]. Furthermore, the system supports the attendees to retrieve the papers from the conference proceedings based on their content similarities The user can take an abstract or a paragraph from an interesting paper and use it as a prototype.

6 Conclusion

The paper presented an approach to integrate an automatic text classification system into a conference management system. Results showed that in categories where enough good training examples were present the user did not change the automatically pre-selected category that often. Another implication is that classes which overlap with other classes or are subclasses of others perform quite poorly. In general 2/3 of the classification results were not changed by the users in the first scenario. In the second scenario, where more documents per class were available, the average amount of correct classification over all classes reached a value of 74.46%.

The classification results can also be taken into consideration when the scheduling of the sessions for the conference is prepared. This means if a session about "multimedia, mixed media, audio, video, 3D and non-traditional objects" is planned the organizer should keep in mind that four papers were submitted where the classifier has chosen this class.

References

[1] I. Androutsopoulos, J. Koutsias, K.V. Chandrinos, and C. D. Spyropoulos. An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages. In *Proceedings of the Annual Int'l ACM SIGIR Conference on R&D in Information Retrieval*, Athens, Greece, 2000.

[2] Kloptchenko Antonina, Back Barbro, Vanharanta Hannu, Toivonen Jarmo, and Visa Ari. Prototype-matching system for allocating conference papers. In *Proceedings of the 36th Annual Hawaii International Conference on System Sciences (HICSS'03) - Track 3*, page 70, Washington, DC, USA, 2003. IEEE Computer Society.

[3] Chumki Basu, Haym Hirsh, William Cohen, and Craig Nevill-Manning. Technical paper recommendation: A study in combining multiple information sources. *Journal of Artificial Intelligence Research*, 14:231–252, 2001.

[4] Susan T. Dumais and Jakob Nielsen. Automating the Assignment of Submitted Manuscripts to Reviewers. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '92)*, pages 233–244, Copenhagen, Denmark, 1992. ACM Press.

[5] William R. Hersh, Chris Buckley, T. J. Leone, and David H. Hickam. OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (Special Issue of the SIGIR*

Forum), pages 192–201, Dublin, Ireland, 1994. ACM/Springer.

[6] Ken Lang. NewsWeeder: learning to filter netnews. In *Proceedings of the 12th International Conference on Machine Learning (ICML-95)*, pages 331–339. Morgan Kaufmann Publishers, San Mateo, US, 1995.

[7] David D. Lewis. Reuters-21578, distribution 1.0. (HTML page) `http://www.daviddlewis.com/resources/testcollections/reuters21578/`, 2004.

[8] David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. RCV1: A New Benchmark Collection for Text Categorization Research. *The Journal of Machine Learning Research*, 5:361–397, 2004.

[9] Andrew McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. (HTML page) `http://www.cs.cmu.edu/~mccallum/bow`, 1996.

[10] Andrew McCallum and Kamal Nigam. A Comparison of Event Models for Naive Bayes Text Classification. In *Proceedings of AAAI-98 Workshop on Learning for Text Categorization*, pages 41–48. AAAI Press, 1998.

[11] Philippe Rigaux. The MyReview System. (HTML page) `http://myreview.lri.fr/`, 2004.

[12] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[13] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[14] Fabrizio Sebastiani. Text Categorization. In Alessandro Zanasi, editor, *Text Mining and its Applications to Intelligence, CRM and Knowledge Management*, pages 109–129. WIT Press, Southampton, UK, 2005.

[15] Yiming Yang. An Evaluation of Statistical Approaches to Text Categorization. *Information Retrieval*, 1(1/2):69–90, 1999.

[16] Yiming Yang and Xin Liu. A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '99)*, pages 42–49, Berkley, August 1999. ACM Press.

[17] Yiming Yang and Jan Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of the 14th International Conference on Machine Learning (ICML-97)*, pages 412–420. Morgan Kaufmann Publishers, San Francisco, US, 1997.

[18] David Yarowsky and Radu Florian. Taking the load off the conference chairs: Towards a digital paper-routing assistant. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in NLP and Very-Large Corpora. (EMNLP/VLC-99)*, 1999.