Supporting ECDL'05 using TCeReview

Andreas Pesenhofer[1], Helmut Berger[1], Andreas Rauber[1,2]

[1] iSpaces Group, eCommerce Competence Center – ec3,
Donau-City-Straße 1, A–1220 Wien, Austria
[2] Vienna University of Technology,
Favoritenstraße 9–11/188, A–1040 Wien, Austria

Abstract. Conference Management constitutes a field in Digital Libraries including tasks such as paper to reviewer assignment and session compilation. These tasks depend on the paper to topic assignment. TCeReview addresses the automatic organization of text documents and enhances conventional conference management applications by incorporating a text classification module. This paper presents the results obtained during the empirical evaluation of the TCeReview applied at ECDL'05.

1 Introduction

One task authors have to complete when submitting a paper to a conference is to select a research topic that identifies their submission. Subsequently, this topic is used by conference organizers to determine appropriate reviewers and, in case of acceptance, compile sessions. However, authors might be uncertain about selecting an appropriate topic for the paper. Authors confusion might be even greater when research topics cannot be clearly described in a few words. TCeReview (Text Classification Enhanced Review) addresses this issue by incorporating a text classification module into the conference management application *MyReview* (http://myreview.lri.fr/). The classification module was trained with accepted submissions from previous conferences and automatically suggests the most likely topic to the author. TCeReview is currently being evaluated in different conference settings, starting with small and medium sized events up to a challenging medical conference with about 3,000 submissions. In this paper we report in a quasi-recursive manner first results using ECDL'05 as live example.

Data from previous ECDL conferences were downloaded from Springer Online (http://www.springerlink.com) in order to build the training set for TCeReview. 311 abstracts related to conference call topics were manually selected and assigned to research topics (cf. Table 1, first column). After preprocessing, the abstracts were indexed with the Rainbow library [1]. 4,141 unique terms were obtained and further reduced to 3,460 terms using Information Gain as feature selection metric. The classification task was carried out by means of a Naïve Bayes classifier [2].

At the first step of submission authors were asked to register their paper and supply meta-data including an abstract. Later, authors completed their

class name	ID	1	2	3	4	5	6	7	recall
Concepts of Digital Libraries, Concepts of Documents and Metadata	1	1	1	2	3	.	1	1	0.11
System Architectures, Open Archives, Collection Building, Integration and Interoperability	2	1	18	1	.	.	.	.	0.90
Information Retrieval, Information Organization, Search and Usage	3	1	3	28	6	.	2	.	0.70
User Studies, System Evaluation, Personalization, User Interfaces and User Centered Design	4	.	.	4	22	.	2	1	0.76
Digital Preservation, Web Archiving and Long Term Access	5	1	1	3	.	.	1	1	0.00
Digital Library Applications and Case Studies	6	.	3	1	2	.	14	1	0.67
Multimedia, Mixed Media, Audio, Video, 3D and non-traditional Objects	7	.	.	.	.	.	.	3	1.00
precision		0.25	0.68	0.72	0.67	0.00	0.70	0.43	

Table 1. Classifier confusion matrix

submission by uploading their final paper. Based on the abstract TCeReview suggested a research topic for the paper. In case of disagreement the author had the possibility to change the topic. Author's agreement or disagreement was tracked in order to evaluate the system.

2 Results and Conclusion

While the submission site was open 129 abstracts were received. Note that abstracts being shorter than 100 characters were discarded. We performed an ex-post evaluation of the author's disagreement with the suggested research topic. The results of the classifier are depicted in the confusion matrix (cf. Table 1). Rows give the class assignments and columns correspond to the prediction of the classifier. The obtained results showed that precision of the classes 2, 3, 4 and 6 is about 70%. These classes were represented by 40 to 67 instances per class in the training set. In case of the class "Concepts of Digital Libraries, Concepts of Documents and Metadata" 25% are correctly classified. The poor performance of this class is attributed to the few training examples (12). Overall, TCeReview archieved an accuracy of 66.67%. This indicates that the system helps the author in assigning the appropriate topic to the submitted abstract.

A similar approach might be taken in order to assign papers to reviewers. Moreover, the system assists in session compilation by applying clustering algorithm on final submissions.

References

1. Andrew McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. Internet: http://www.cs.cmu.edu/ mccallum/bow, 1996. (last visit: 27.4.2005).
2. 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.