SINAI at INFILE 2009: Experiments with Google News

Arturo Montejo-Ráez, José M. Perea-Ortega, Manuel Carlos Díaz-Galiano, L. Alfonso Ureña-López
SINAI research group. Computer Science Department. University of Jaén
Campus Las Lagunillas, Ed. A3, E-23071, Jaén, Spain
{amontejo,jmperea,mcdiaz,laurena}@ujaen.es

Abstract

This paper describes the SINAI team participation in the INFILE routing and filtering track of the CLEF campaign. This is the first participation of the SINAI research group in the INFILE task. We have participated in the batch filtering subtask and submitted two experiments: one using the topics' text as learning data to train a classifier, and another one where training data has been constructed from Google News pages. Our results show that our use of Google News did not improved the classification obtained using only topics description.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software

General Terms

Algorithms, Experimentation, Languages, Performance

Keywords

Text classification, Information Retrieval

1 Introduction

The INFILE (Information Filtering Evaluation) track is a cross-language adaptive filtering evaluation campaign that takes place within the CLEF (Cross-Lingual Evaluation Forum) campaign [1].

This is the first participation of the SINAI research group in the INFILE task, at CLEF campaign. We have participated in the batch filtering subtask and payed attention only on English texts. Two runs have been submitted: one using the topics' text as learning data to train a classifier, and another one where training data has been constructed from Google News pages. Our results show that our use of Google News did not improved the classification obtained using only topics description.

The Web has been extensively used as resource when dealing with different text mining problems [2, 4]. We have applied it in video retrieval and classification [3]

In our experiments, we have generated a training corpus by querying Google News on the different topics, generating for every topic an small corpus of related news. This corpus has served as learning data for building a model which was, thus, used for classifying every document into one of the fifty different classes proposed. The results obtained have been compared to those where

only topics texts were used a learning data. The learning algorithm was Support Vector Machines [5] on both cases.

2 Experiment Description

A traditional supervised-learning scheme has been followed. The difference between the runs submitted rely on the training corpus used. One was on Google News entries and another on the descriptions of the topics provided.

The Google News corpus was generated by querying Google News on each of the topics keywords. For example, topic 101 contains the keywords *doping*, *legislation doping*, *athletes*, *doping substances* and *fight against doping*. For each of these keywords, 50 links were retrieved, downloaded and their HTML cleaned out. In this way, about 200 documents existed per topic.

Once each corpus was generated, a SVM model was trained on it. This is a binary classifier turned into a multi-class classifier by training a different SVM model per topic. The topic with the highest confidence was selected as label for the incoming document and, therefore, the document was routed to that topic. It is important to note here that a label was proposed for all of the incoming documents, that is, no document was left without one of the 50 labels (topics).

3 Results and Discussion

Overall results are displayed in Table 1. The results obtained are discouraging: few relevant assignations are made. In fact, the use of Google News as learning source leads to very poor results.

Topics descriptions		**Google News**	
Retrieved	100000	Retrieved	100000
Relevant	1597	Relevant	1597
Relevant Retrieved	940	Relevant Retrieved	196

Table 1: Overall results for both runs

Both precision and recall are low, as can be seen in figures 3 and 2. From these graphs we can observe a random behavior of the performance when Google News corpus has been involved.

4 Conclusions and Further Work

Google News as a source of information for generating learning corpus has shown quite bad results. After inspecting these problematic corpus, we found that huge amount of useless text was not filtered. Therefore, we plan to improve the quality of the data extracted from the web in order to avoid undesirable side effects due to noisy content.

Although the results obtained in this task are really very low in terms of performance, it represents a challenge in text mining, as real data has been used, compared to previous too controlled corpora. We expect to continue our research on this data, and analyze in depth the effect of incorporating web content in filtering tasks.

5 Acknowledgements

This work has been supported by the Andalusian Regional Government (Spain) under excellence project GeOasis (P08-41999), under project on Tourism (FFIEXP06-TU2301-2007/000024), the Spanish Government under project Text-Mess TIMOM (TIN2006-15265-C06-03) and the local project RFC/PP2008/UJA-08-16-14.