Chapter XV

Taxonomy Based Fuzzy Filtering of Search Results

S. Vrettos, National Technical University of Athens, Greece
A. Stafylopatis, National Technical University of Athens, Greece

ABSTRACT

Our work proposes the use of topic taxonomies as part of a filtering language. Given a taxonomy, we train classifiers for every topic of it. The user is able to formulate logical rules combining the available topics, e.g., (Topic1 AND Topic2) OR Topic3, in order to filter related documents in a stream of documents. Using the classifiers, every document in the stream is assigned a belief value of belonging to the topics of the filter. These belief values are then aggregated using logical operators to yield the belief to the filter. In that framework, we are concerned with the operators that provide the best filtering performance for the user. In our study, Support Vector Machines (SVMs) and Naïve Bayes (NB) classifiers were used to provide topic probabilities. Fuzzy aggregation operators were tested on the Reuters text corpus and showed better results.
than their Boolean counterparts. Moreover, the application of Ordered Weighted Averaging (OWA) operators considerably improved the performance of fuzzy aggregation, especially in the case of NB classifiers. Finally, we describe a filtering system to exemplify the use of fuzzy filtering.

**INTRODUCTION**

The primary way of interactively finding information on the Web is to make a query in a search engine and then browse a ranked list of possibly related web pages. Alternatively, we can browse a manually organized topic taxonomy to find pages related to the query that we have in mind. Although web taxonomies may be very large, they cover a small portion of the Web relative to search engines, primarily because they rely on human effort.

Text/Hypertext categorization (see Yang, 1999; Yang et al., 1999; Chen, 2000) promises to help maintain updated and large web taxonomies and also to improve query-based (Dumais, 2001) retrieval. The idea is to use topic classifiers, which have been trained using the portion of the well-structured web taxonomy, to organize the results of a query to the much larger, but unclassified, web portion indexed by a search engine. Basically, as regards the interface used to include topic information in the query results, it can be topic-oriented or list-oriented. In topic-oriented interfaces, results are organized in a flat or hierarchical taxonomy; in list-oriented interfaces, the original query list is enriched with topic meta-data.

Our work proposes the use of topic taxonomies as part of a filtering language. The user is able to formulate logical rules combining the available topics, e.g., (Topic1 AND Topic2) OR Topic3, in order to retrieve or filter related documents. In that framework, we are concerned with the operators that provide the best filtering performance for the user.

Typically, classification is a YES/NO assignment, so the Boolean model is a good candidate for the filtering task. Nevertheless, Boolean filtering provides no ordering, which is a drawback to both retrieval effectiveness and man-machine interaction. If perfect classifiers were available, Boolean filtering would be enough because all the true positive documents of the stream, and only them, would be retrieved. In that case, Boolean filtering would yield recall and precision equal to 1. Unfortunately, no perfect classifiers are available yet, and even the best performing classifiers in laboratory text corpora might have poor results in real, noisy environments such as the Web. In such cases, ranking
Related Content

Analysis of Textual Data Based on Inductive Learning Techniques
[www.igi-global.com/article/analysis-of-textual-data-based-on-inductive-learning-techniques/100040?camid=4v1a](www.igi-global.com/article/analysis-of-textual-data-based-on-inductive-learning-techniques/100040?camid=4v1a)

XKFitter: A Keyword Filter on XML Stream
[www.igi-global.com/article/xkfitter-keyword-filter-xml-stream/53123?camid=4v1a](www.igi-global.com/article/xkfitter-keyword-filter-xml-stream/53123?camid=4v1a)

Development of the Leader Integrity Assessment
[www.igi-global.com/chapter/development-leader-integrity-assessment/69743?camid=4v1a](www.igi-global.com/chapter/development-leader-integrity-assessment/69743?camid=4v1a)
Tweet Sentiment Analysis with Latent Dirichlet Allocation
[www.igi-global.com/article/tweet-sentiment-analysis-with-latent-dirichlet-allocation/127002?camid=4v1a](www.igi-global.com/article/tweet-sentiment-analysis-with-latent-dirichlet-allocation/127002?camid=4v1a)