Chapter 3.14
Non-Topical Classification of Query Logs Using Background Knowledge

Isak Taksa
Baruch College, City University of New York, USA

Sarah Zelikovitz
The College of Staten Island, City University of New York, USA

Amanda Spink
Queensland University of Technology, Australia

ABSTRACT

Background knowledge has been actively investigated as a possible means to improve performance of machine learning algorithms. Research has shown that background knowledge plays an especially critical role in three atypical text categorization tasks: short-text classification, limited labeled data, and non-topical classification. This chapter explores the use of machine learning for non-hierarchical classification of search queries, and presents an approach to background knowledge discovery by using information retrieval techniques. Two different sets of background knowledge that were obtained from the World Wide Web, one in 2006 and one in 2009, are used with the proposed approach to classify a commercial corpus of web query data by the age of the user. In the process, various classification scenarios are generated and executed, providing insight into choice, significance and range of tuning parameters, and exploring impact of the dynamic web on classification results.

DOI: 10.4018/978-1-60960-818-7.ch3.14

INTRODUCTION

Text classification in the framework of machine learning is an active area of research, encompassing a variety of learning algorithms (Ensuli et al., 2008), classification systems (Banerjee, 2008) and data representations (Wu et al., 2008). Three non-standard issues in machine learning are the

focus of the research in this chapter: short text classification problems, limited labeled data, and non-topical classification. This chapter studies the classification of search queries, which is one example of text classification that is particularly complex and challenging. Typically, search queries are short, reveal very few features per single query and are therefore a weak source for traditional machine learning (Gabrilovich at el., 2009).

We examine the issues of non-hierarchical (Cesa-Bianchi et al., 2006) classification and investigate a method that combines limited manual labeling, computational linguistics and information retrieval to classify a large collection of search queries. We discuss classification proficiency of the proposed method on a large search engine query log, and the implication of this approach on the advancement of short-text classification. We also compare results of two classification tasks executed in 2006 and 2009 to examine the impact of the growing internet collection on consistency of classification results.

We start with a search engine query log which is viewed as a set of textual data on which we perform classification (Jansen et al., 2009; Zimmer and Spink, 2008). Observed in this way, each query in a log can be seen as a document that is to be classified according to some pre-defined set of labels, or *classes*. Viewing the initial log with the search queries as a document corpus $D = \{d_1, d_2, ...d_i,...d_n\}$, we create a set of classes that indicate a personal demographic characteristic of the searcher, $C = \{c_1, c_2, ...c_j,...c_m\}$. Using Web searches, our approach retrieves a set of background knowledge to learn additional features that are indicative of the classes, C. This allows for the categorization of the queries. This approach consists of the following five steps:

1. Select (from the print and the online media) a short set of manually chosen terms $T_{init} = \{t_1, t_2, ...,t_j, ...,t_m\}$ consisting of terms t_j that are known a priori to be descriptive of a particular class c_j;
2. Use this initial set T to classify a small subset of (search queries) set D thereby creating an initial set of classified queries $Q_{init} = \{q_1, q_2, ...q_j...q_l\}$;
3. Submit these queries q_j to a commercial search engine and use the returned search results to build a temporary corpus of background knowledge $B_{temp} = \{b_1, b_2, ...b_j...b_{l*10}\}$;
4. Use an algorithm to select from B more class related terms T;
5. Use this newly created set T to classify more documents (search queries) in corpus D thereby adding more classified queries to set Q.

While steps 1 and 2 are executed only once, steps 3 through 5 are repeated continuously until the classification process is terminated (Figure 1).

We focus on validating our approach to the classification of a set of short documents, namely search queries. This approach uses a combination of techniques: we first look at developing a method to obtain relevant background knowledge for a set of web queries; then we build the background knowledge to acquire ranked terms for improved information retrieval; we then investi-

Figure 1. Steps in a classification process