Chapter 3.8
Machine Learning Approach to Search Query Classification

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

Search query classification is a necessary step for a number of information retrieval tasks. This chapter presents an approach to non-hierarchical classification of search queries that focuses on two specific areas of machine learning: short text classification and limited manual labeling. Typically, search queries are short, display little class specific information per single query and are therefore a weak source for traditional machine learning. To improve the effectiveness of the classification process the chapter introduces background knowledge discovery by using information retrieval techniques. The proposed approach is applied to a task of age classification of a corpus of queries from a commercial search engine. In the process, various classification scenarios are generated and executed, providing insight into choice, significance and range of tuning parameters.

INTRODUCTION

Machine learning for text classification is an active area of research, encompassing a variety of learning algorithms (Sebastiani, 2002), classification systems (Barry et al., 2004) and data representations (Spink and Jansen, 2004). Classification of search queries 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

DOI: 10.4018/978-1-60960-818-7.ch3.8

weak source for traditional machine learning. This chapter focuses on two specific areas of machine learning: short text classification problems and using a small set of labeled documents. We examine the issues of non-hierarchical (Cesa-Bianchi et al., 2006) classification and introduce 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.

For this discussion we view query logs as sets of textual data on which we perform classification (Jansen, 2006). 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*. The approach described in this chapter classifies a corpus of search queries from the Excite search engine, by retrieving from the Web a set of background knowledge to learn additional features that are indicative of the 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, $C = \{c_1, c_2, ...c_j...c_m\}$. We present an approach that allows classification or the assignment of a class from the set C to many of the documents in the set D. 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 investigate the impact of the new terms' selection algorithms on the effectiveness of the classification process.

BACKGROUND

Text classification (or alternatively, text categorization) can be defined as follows: Given a set of documents D and a set of m classes (or labels) C, define a function F that will assign a value from

Figure 1. Steps in a classification process