Chapter 100
Proximity-Based Good Turing Discounting and Kernel Functions for Pseudo-Relevance Feedback

Ilyes Khennak
University of Science and Technology Houari Boumediene (USTHB), Algeria

Bab Ezzouar
University of Science and Technology Houari Boumediene (USTHB), Algeria

ABSTRACT

During the last few years, it has become abundantly clear that the technological advances in information technology have led to the dramatic proliferation of information on the web and this, in turn, has led to the appearance of new words in the Internet. Due to the difficulty of reaching the meanings of these new terms, which play an essential role in retrieving the desired information, it becomes necessary to give more importance to the sites and topics where these new words appear, or rather, to give value to the words that occur frequently with them. For this purpose, in this paper, the authors propose a new robust correlation measure that assesses the relatedness of words for pseudo-relevance feedback. It is based on the co-occurrence and closeness of terms, and aims to select the appropriate words that best capture the user information need. Extensive experiments have been conducted on the OHSUMED test collection and the results show that the proposed approach achieves a considerable performance improvement over the baseline.

INTRODUCTION

Over the years, many different retrieval models, such as vector space models (Salton et al., 1975; Salton & Buckley 1988), classic probabilistic models (Robertson et al., 1995; Turtle & Croft, 1991; Fuhr, 1992), and statistical language models (Ponte & Croft, 1998; Lavrenko & Croft, 2001; Zhai & Lafferty, 2001a), have been proposed and studied in order to fix the issue of searching relevant documents in a large data

DOI: 10.4018/978-1-5225-5191-1.ch100

source that satisfy the users' information needs (Van Rijsbergen, 1979). Nevertheless, it remains a great challenge to develop Information Retrieval Systems (IRSs) that are robust, effective, and efficient.

The reason for the ineffectiveness of IRSs is predominantly caused by the ambiguity, incompleteness and imprecision of keywords that are used to express the genuine user's information need. One well-known technique to bypass this shortcoming is to expand the original user query with extra terms that best characterize the actual user intent. In this regard, various approaches dealing with the proximity and the interdependence of words have been implemented and tested to assess the strength of the relationship between an extra word candidate and the user query in order to find the most important terms to be used as extra terms, or rather, as expansion features. (Carpineto & Romano, 2012)

In this sense, the main goal of this work is to propose a robust correlation measure that evaluates the relatedness of words based on the co-occurrence and closeness of terms. This principle gives importance to words that frequently occur in the same context during the search process. For example, the term 'IJIRR' is often found in the same sites where the words 'Journal,' 'IGI Global', and 'Retrieval' occur. Relying on this concept was not a coincidence but rather came as a result of the researches conducted recently about the growth of the World Wide Web. All of these researches have demonstrated an exponential growth of the Web and rapid increase in the number of new pages created. In his study, Ranganathan (2011) estimated that the volume of online data indexed by Google had increased from 5 exabytes in 2002 to 280 exabytes in 2009. According to Zhu et al. (2009), this volume is expected to double in every 18 months. Ntoulas et al. (2004) interpreted these statistics in terms of the number of new pages created and indicated that their number is increasing by 8% a week. The work of Bharat and Broder (1998) went further and estimated that the World Wide Web pages are growing at the rate of 7.5 pages every second. This revolution, that the Web is witnessing, has led to the appearance of two points:

- The first point is the entry of new words into the Web which is estimated, according to Williams and Zobel (2005), at about one new word in every two hundred words. Studies by (Williams and Zobel, 2005; Eisenstein et al., 2012; Sun, 2010) have shown that this invasion is mainly due to: neologisms, acronyms, abbreviations, emoticons, URLs and typographical errors.
- The second point is that the users use these new words during the search. Chen et al. (2007) indicated in their study that more than 17% of query terms are non-dictionary words, 45% of them are E-speak (lol), 18% are companies and products (Google), 16% are proper names, 15% are misspellings and foreign words (Subramaniam et al., 2009; Ahmad & Kondrak, 2005).

Out of these two points which the web is facing and due to the difficulty of exploring the meanings of these words, we proposed a method that based on finding the sites and topics where these words appear, and then employing the terms which neighbor and co-occur with the latter in the search process. We will use the two well-known Pseudo-Relevance Feedback techniques: Rocchio's method, and Robertson/ Sparck Jones' term-ranking function as the baseline for comparison; and evaluate our approach using OHSUMED test collection. The main contributions of our work are the following:

- The adoption of an external correlation measure in order to evaluate the co-occurrence of words with respect to the query features.
- The determination of an internal correlation measure in order to assess the proximity and closeness of words relative to the features of the query.