Efficient Text Proximity Search

Ralf Schenkel[1], Andreas Broschart[1], Seungwon Hwang[2], Martin Theobald[3], and Gerhard Weikum[1]

[1] Max-Planck-Institut für Informatik, Saarbrücken, Germany
{abrosch,schenkel,weikum}@mpi-inf.mpg.de
[2] POSTECH, Korea
swhwang@postech.ac.kr
[3] Stanford University
theobald@stanford.edu

Abstract. In addition to purely occurrence-based relevance models, term proximity has been frequently used to enhance retrieval quality of keyword-oriented retrieval systems. While there have been approaches on effective scoring functions that incorporate proximity, there has not been much work on algorithms or access methods for their efficient evaluation. This paper presents an efficient evaluation framework including a proximity scoring function integrated within a top-k query engine for text retrieval. We propose precomputed and materialized index structures that boost performance. The increased retrieval effectiveness and efficiency of our framework are demonstrated through extensive experiments on a very large text benchmark collection. In combination with static index pruning for the proximity lists, our algorithm achieves an improvement of two orders of magnitude compared to a term-based top-k evaluation, with a significantly improved result quality.

1 Introduction

Techniques for ranked retrieval of text documents have been intensively studied including relevance scoring models such as tf*idf, Okapi BM25, and statistical language models [13]. Most of the models in these families are based on the (multinomial) bag-of-words representation of documents, with consideration of term frequencies (tf) and inverse document frequencies (idf) but without considering term proximity. However, there are many queries where the best results contain the query terms in a single phrase, or at least in close proximity.

To illustrate the importance of proximity, let us consider the query "*surface area* of *rectangular pyramids*". Schemes that do not take proximity into account return general mathematical documents in which all the four terms *surface*, *area*, *rectangular* and *pyramid* are individually important, but the document does not necessarily contain information about the surface area of rectangular pyramids (for example, it may discuss the volume of pyramids and the area of rectangular prisms. On the other hand, an exact phrase match "*surface area* of *rectangular pyramids*" would most certainly ensure that the document retrieved is of

N. Ziviani and R. Baeza-Yates (Eds.): SPIRE 2007, LNCS 4726, pp. 287–299, 2007.

the desired type, but strictly enforcing such phrase matchings in a boolean way would exclude many relevant results. A good proximity-aware scoring scheme should give perfect phrase matches a high score, but reward also high proximity matches such as "*surface area* of a *rectangular*-based *pyramid*" with good scores. There has been a number of proposals in the literature for such proximity-aware scoring schemes [5,6,9,10,16,18,20]; however, none of these proposals considered efficiently finding the best results to queries in a top-*k* style with dynamic pruning techniques. This paper shows that integrating proximity in the scoring model can not only improve retrieval effectiveness, but also improve retrieval efficiency by up to two orders of magnitude compared to state-of-the-art processing algorithms for purely occurrence-based scoring models.

2 Related Work

Using phrases is a common means in term queries to restrict the results to those that exactly contain the phrase and is often useful for effective query evaluation [7]. A simple way to efficiently evaluate phrases are *word-level indexes*, inverted files that maintain positional information [24]. There have been some proposals for specialized index structures for efficient phrase evaluation that utilize term pair indexes and/or phrase caching, but only in the context of boolean retrieval and hence not optimized for top-*k* style retrieval with ranked results [8,22,23]. There are proposals to extend phrases to window queries, where users can specify the size of a window that must include the query terms to favor documents containing all terms within such a window [15,17,4]. However, this line of works has treated term proximity only as an afterthought after ranking, i.e., proximity conditions are formulated as a simplistic Boolean condition and optimized as separate post-pruning step after rank evaluation.

More recently, some scoring models were proposed that integrate content and proximity scores for ranking results [5,6,9,10,16,18,20]. These scoring models can be categorized into the following two classes. First, linear combination approaches attempt to reflect proximity in the scoring by linearly combining a proximity score with a text-based content score [5,6,16,18]. Monz quantified the proximity score based on the size of the minimum window containing all query keywords occurring in the document [16]. Rasolofo et al. consider term pairs that occur together in a small window in a document, and use a distance-based proximity score for these term pairs [18]. Büttcher et al. extend on this work by considering adjacent query term occurrences without a limit on the window size and use a proximity score similar to BM25 for text [5,6]. Second, holistic approaches have more tightly integrated proximity metrics and content scoring [9,10,20]. De Kretser and Moffat [10] and Clarke et al. [9] proposed scoring methods that reward the density of query terms in documents, and Song et al. [20] integrate a similar term density score within a BM25-based scoring model. However, none of the proximity proposals we are aware of has been designed to be used within a top-*k* style evaluation.