Efficient Integration of Proximity for Text, Semi-Structured and Graph Retrieval

Andreas Broschart
Max-Planck-Institut fuer Informatik
Saarbruecken, Germany
abrosch@mpi-inf.mpg.de

ABSTRACT

Search engines rely on scoring functions to rank query results. Content-based scoring functions (e.g. Okapi BM25) are based on the "bag of words" model, and thus refrain from considering *term proximity*, i.e., distances between query term occurrences in a document. Ignoring such proximity information might yield documents in which all query terms are individually important but appear in different paragraphs. As an example, querying for *surface area* of a *rectangular pyramid* might return documents that treat "volume of a pyramid" in the first paragraph and "surface area of a rectangular prism" in the second, likely not satisfying the user's information need. Recently there have been several proposals for proximity enhanced scoring functions that provide the means to execute implicit soft phrase queries such that the user does not have to specify them.

This PhD research aims at (1) leveraging the power of proximity-enhanced scoring models for text, semi-structured and graph retrieval, (2) at the same time treating efficiency as a major concern, and (3) proving the retrieval effectiveness and efficiency of the proposed approaches through extensive experimental studies.

Proximity-enhanced scoring models

Proximity-enhanced scoring models can improve search results by incorporating distance information of query term occurrences. One of the most effective among them is Büttcher's approach[1] that linearly combines a BM25-based content score and a proximity score. In our experiments with TREC data, it improved retrieval effectivity significantly compared to Okapi BM25. Starting from such scoring functions in text retrieval, our goal is to derive proximity-enhanced scoring functions for more sophisticated tasks, including (1) semi-structured retrieval based on (a) keyword queries for XML documents and (b) structural queries for XML documents, and (2) graph retrieval. The complex structure of XML documents does not allow an immediate application of text proximity, as simply using document ordering to derive text positions would ignore structure and especially the semantics of elements that contain the text. Therefore we propose to specify semantically related tag names: while for related element nodes text occurrences are considered order oblivious, for unrelated element nodes they are considered order aware. Referenced documents probably contain semantically unrelated content deserving higher distances than text occurrences in the same document.

For structural queries (e.g., /A[t_3]//B[t_1]) we propose to consider the distance of matching parts and elements on the path that interconnects the matching paths. More complex queries could be handled by computing the compactness of the query graph that contains as nodes all matching elements interconnected according to the element graph. The compactness could be the average edge weight or the sum of the edge weights. If there is more than one result, we plan to return the result having the best integrated score in a scoring model that still has to be determined.

Efficiency as a major concern

Most existing works in text retrieval do not focus on query evaluation of large collections. We propose to tackle the efficiency problem by extensive usage of index structures combined with a top-k style evaluation of queries. Fast query evaluation requires the *creation of appropriate indexes.* We have implemented and evaluated first top-k variants of Büttcher's scoring function. To this end, we have precomputed and materialized several index structures, allowing for sequential and random accesses. ProximityIndexLists contain for each unordered query term pair a variant of the proximity score in Büttcher's approach. Its usage improves retrieval effectiveness as well as efficiency as to mere BM25 score lists. CombinedIndexLists combine the respective ProximityIndexList with the BM25 scores for each of the two query terms. Preliminary experiments showed that we can reduce the (byte-based) transfer cost by an order of magnitude. Motivated by these encouraging results, we intend to investigate other combined index structures. The creation of indexes with limited size seems to be obligatory in order to ensure fast evaluation. We plan to prove retrieval effectiveness and efficiency using commonly used collections and topics (e.g., from the INEX Ad Hoc track for XML data).

Categories and Subject Descriptors: H.2.4 [Systems]: Query processing; H.2.4 [Systems]: Textual databases; H.3.3 [Information Search and Retrieval]: Retrieval models

General Terms: Algorithms, Experimentation, Performance

Keywords: proximity, XML, efficiency

[1] S. Büttcher et al. Term proximity scoring for ad-hoc retrieval on very large text collections. In *SIGIR*, pages 621–622, 2006.

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.
ACM 978-1-59593-597-7/07/0007.