Global and Local Resources for Peer-to-Peer Text Retrieval

Zusammenfassung der Dissertation
Abstract of PhD dissertation

Hans Friedrich Witschel

When compared to traditional centralised solutions for document storage and retrieval, peer-to-peer (P2P) systems offer a number of potential advantages. For instance, they offer greater ease of publishing and significantly reduce maintenance costs and risk of failure.
However, in order to become really attractive, peer-to-peer text retrieval systems must become both efficient and effective. Currently, there are still a number of unsolved issues that prevent efficient systems from being effective and vice versa. This thesis studies some of these issues in detail.

In a theoretical part of the work, a formal and graph-based framework is developed that represents the most important aspects of information retrieval (IR) in a unified way. It serves as a means to extend algorithms and ideas from one field of IR onto others. This is exemplified by embedding distributed and peer-to-peer IR within the field of traditional IR.

Second, an empirical part of the thesis is devoted to answering two concrete IR research questions:

- Global knowledge and results merging: Some components of traditional IR systems will not work without knowledge of global collection characteristics, e.g. computing document scores w.r.t. queries. When each peer computes these scores on the basis of statistics derived from its local document collection only, the scores returned by different peers are generally not comparable. Since there is no global view on the data in a P2P network, the central question is: can global collection statistics be replaced with something else, e.g. with external sources or statistics gathered from collection samples?
- Profiles and query routing: Search in P2P networks works by query messages being forwarded from one peer to the next. In order to make this forwarding effective, it is important to develop a mechanism that allows any peer to distinguish useful peers from others. Here, we study a mechanism where each peer stores profiles of its neighbours and makes forwarding decisions by matching queries against profiles. Since profiles are often sent through the network, they need to be compact. The question is thus: how many items can we prune from a profile and still have acceptable results? Further, are there any techniques for learning either better queries or better profiles that can improve forwarding decisions?

Experimental results indicate that when replacing global collection statistics with generic external sources, retrieval effectiveness will be degraded significantly. However, mixing statistics from an external source with very small samples of the target collection yields good results.
As far as the second question is concerned: pruning words from a peer's profile does not seem to significantly complicate the task of query routing. Learning better queries is much harder than learning better profiles. The latter can be done by boosting the influence of a word within a peer's profile when the peer has successfully answered a query containing that word, a technique that yields substantial improvement in terms of retrieval effectiveness.