# Global Resources for Peer-to-Peer Text Retrieval

Hans Friedrich Witschel
University of Leipzig
P.O. Box 100920
D-04009 Leipzig
witschel@informatik.uni-leipzig.de

## ABSTRACT

The thesis presented in this paper tackles selected issues in unstructured peer-to-peer information retrieval (P2PIR) systems, using world knowledge for solving P2PIR problems. A first part uses so-called reference corpora for estimating global term weights such as IDF instead of sampling them from the distributed collection. A second part of the work will be dedicated to the question of query routing in unstructured P2PIR systems using peer resource descriptions and world knowledge for query expansion.

**Categories and Subject Descriptors:** H.3.3 [Information Search and Retrieval]: Selection Process, C.2.4 [Distributed Systems]: Distributed Applications.

**General Terms:** Algorithms, Experimentation.

**Keywords:** Peer-to-peer information retrieval, term weighting, query routing.

## 1. INTRODUCTION

My thesis aims at solving problems in P2PIR for both precision- and recall-oriented retrieval, the guiding question being to what extent global "world" knowledge can be applied in this context, i.e. data that is independent of the collection shared by peers in a particular P2PIR system; it is assumed that this knowledge can be gathered – once and for all – from sources such as the WWW and then used in many different P2PIR systems.

## 2. GLOBAL TERM WEIGHTS

In a first series of experiments [2], I investigated the question whether one can globally estimate collection or document frequencies of terms – independent of a given collection – well enough in order not to degrade retrieval performance seriously when using them for computing e.g. IDF.

The results indicate that weights estimated from reference corpora slightly degrade retrieval results, but this is often not statistically significant. Weights can also be improved by mixing them with estimates derived from very small samples of the retrieval collection. Finally, a large fraction of infrequent terms can be pruned (i.e. treated as if they had not occurred) from the resulting term lists without any ill effects. All in all, this indicates that what is really needed for global term weighting is just an "extended stop word list" which can be robustly sampled from a reference corpus. Some collection or domain-specific "stop words", however, cannot be found that way and need to be sampled from the retrieval collection.

## 3. QUERY ROUTING

For further experiments, I propose a new P2PIR testbed which has a realistic distribution of content and queries (similar to [1]): Given that the data shared by nodes in a peer-to-peer network should reflect the interests of persons running peers, I propose to identify peers with authors of documents using a data set where authoring and citation information is available (e.g. the CiteSeer database). To generate queries, each peer will ask for keywords of papers that are referenced in its own papers.

Since there are no relevance judgments for these queries, I propose to measure the performance of distributed retrieval algorithms using a new evaluation measure which – roughly – tells us how high the best $k$ documents that the distributed search finds are ranked – on average – by a centralised search engine.

Using this experimental setup, I would like to address the question if there is a way to keep peers' resource descriptions very compact and still guarantee good recall when matching queries against compressed profiles. To this end, I will propose various query expansion strategies based on world knowledge, including the WWW, other large collections or thesauri, similar in spirit to what has been done in [3], but using the new evaluation framework and more data sources. Expansion using these global knowledge sources will then be compared to other expansion strategies such as local feedback on peers.

## 4. REFERENCES

[1] I. A. Klampanos, J. M. Jose, V. Poznanski, and P. Dickman. A Suite of Testbeds for the Realistic Evaluation of Peer-to-Peer Information Retrieval Systems. In *ECIR 2005*, pages 38–51, 2005.

[2] H.F. Witschel. Estimation of global term weights for distributed and ubiquitous IR. In *Proc. of UKDU'06*, 2006.

[3] H.F. Witschel and T. Böhme. Evaluating Profiling and Query Expansion Methods for P2P Information Retrieval. In *Proc. of the P2PIR Workshop at CIKM*, 2005.


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