The CIKM 2005 Workshop on Information Retrieval in Peer-to-Peer Networks

Henrik Nottelmann[1], **Karl Aberer**[2], **Jamie Callan**[3]
and **Wolfgang Nejdl**[4]

[1]University of Duisburg-Essen, Germany
nottelmann@uni-duisburg.de

[2]EPFL Lausanne, Switzerland
karl.aberer@epfl.ch

[3] Carnegie Mellon University, USA
callan@cs.cmu.edu

[4]L3S and University of Hannover, Germany
nejdl@l3s.de

1 Introduction

Peer-to-peer (P2P) networks have emerged as a popular way to build large scale information systems by using the principle of resource sharing. The P2P paradigm holds many promises, e.g. scalability, failure resilience and increased autonomy of nodes. For these reasons P2P seems also to be an interesting architectural paradigm for realizing large scale information retrieval systems. However, search methods in P2P networks are still mostly limited to simple keyword queries and the use of advanced retrieval models is in its infancy.

Researchers from different areas, including database systems, distributed systems, networking and information retrieval, have recently started to work on efficient, yet semantically powerful search mechanisms in peer-to-peer systems. An important factor for making this research successful will be an intensive exchange among researches from the relevant disciplines.

The call for papers for the second workshop on information retrieval in peer-to-peer networks attracted 15 submissions from Asia, Canada, the United States, Australia and Europe, from which we accepted 6 papers for presentation. The workshop was held on November 4, 2005, in Bremen, Germany, immediately following the CIKM 2005 conference. We opted for this conference in contrast to SIGIR last year in order to bring together the different communities. About 15 people participated. The size was small enough to allow an interactive format and discussion between and during presentations. Our impression is that most of the participants found it very productive.

More details about the workshop can be found online at http://p2pir.is.informatik.uni-duisburg.de/2005/.

2 Paper presentations

The first presentation was given by Hans Friedrich Witschel and dealt with *Evaluating Profiling and Query Expansion Methods for P2P Information Retrieval*. Various methods are employed for modelling peer descriptions (called "profiles"), namely categories from a classification scheme, concepts from latent semantic indexing (LSI) and TF.IDF vectors of the most significant keywords. For the latter, IDF values are computed based on a fixed, globally available reference corpus. These profiles are used for query expansion, e.g. by providing local (pseudo-relevance) feedback. Another approach employs the reference corpus to create a co-occurrence matrix; additional keywords are extracted from that matrix and added to query in later phases. Experiments showed that rather compact profiles (48 terms) yield effective results.

Massimo Melucci presented *An Evaluation of a Recursive Weighting Scheme for Information Retrieval in Peer-to-Peer Networks*. Within a coherent framework, weights are assigned to informative resources on

multiple levels: to terms in documents, to documents in peers, and to peers in ultra-peers. The same TF.IDF-like weighting scheme is used on all levels, where the weights on one level depend on the weights of the lower levels. As a consequence, the same IR infrastructure can be applied to document retrieval, peer ranking and ultra-peer ranking. As the approach assigns high weights to peers and ultra-peers which contains few documents (peers, respectively) with highly relevant documents, it is also reduces bandwidth.

A Queueing Theory Based Analysis of An Agent Control Mechanism in Peer-to-Peer Information Retrieval Systems by Haizheng Zhang follows a different perspective. The major objective is not to optimize efficiency and effectiveness when each query is processed in isolation, but to optimize the throughput of the overall system with concurrent search sessions. Peers are modelled by a local query queue, which stores incoming queries for resource selection and local search, and message queries for contacting neighbor peers. Congestion is minimized by a probabilistic load balancing approach. The work employs a two-phase routing algorithm for efficient concurrent query processing which takes advantage of a hierarchical peer organization. In the first phase, the query is forwarded to relevant peers on the same level, while the second phase only involves local neighborhood of selected peers.

Gudrun Fischer tackled another search access paradigm in *Towards Scatter/Gather Browsing in a Hierarchical Peer-to-Peer Network*. Scatter/gather clustering has been proposed in the past as a flexible tool for interactive collection exploration. In multiple iterations, the system generates a clustering and presents it to a user. The user selects one or multiple clusters, which are merged again, re-clustered, and presented to the user again. This allows for user-guided top-down browsing. In hierarchical peer-to-peer networks with leaf peers (storing the documents) and hubs (ultra-peers), each leaf peer precomputes and stores its own cluster hierarchy. The top levels of these peer clusters are combined in an overall cluster hierarchy in the hubs.

Search Strategies for Scientific Collaboration Networks were investigated by Paul-Alexandru Chirita. Co-authorship of papers as well as the regularly exchange of scientific articles and references is considered as signs for collaboration. As one result of the paper, both frequency distribution of the number of co-authors and the distribution of collaboration links follow a power law. Somehow surprisingly, similarity-based peer selection approaches outperforms the connective-based one, leaving room for future research.

Felix Heine presented his approach for *Processing Complex RDF Queries over P2P Networks*. One application area is the grid, where RDF can be used for semantically describing resources like computers, their hardware, the operating systems it is running, or their capabilities and limitations. Distributed hash tables (DHT) are then employed for storing the RDF graph in a decentralized way, by using subject, predicates and objects of RDF triples as separate keys for maximum flexibility. An efficient algorithm is proposed for distributed query evaluation.

3   Discussions

To foster exchange of ideas and collaboration, the workshop programme left room for two long discussions phases.

One discussion about P2PIR algorithms started with advantages, disadvantages and application scenarios of distributed hash tables (DHTs). One drawback is the load in cases of skewed distributions of key lookups. In addition, DHTs only support exact lookups. However, this is sufficient not only for RDF query processing but also for the IR task or distributing inverted lists (with terms as keys, after stemming and stop word removal) of document indexes or statistical metadata ("resource descriptions" or "summaries"). Privacy is another issue for DHTs: Document providers have to publish all documents, while people searching for documents protect their privacy, as their queries are split among the peers.

A second topic was comparability of approaches and the software, and message/event-driven software vs. data-dependent applications. Some discussion arose around simulations and prototypes. 10 participants use simulations, 7 use both prototypes and simulations. Among them, 3 have completely disjoint code bases.

For 4 participants, the prototype provides functionality beyond the simulation.

The other discussion block was centered around the evaluation of peer-to-peer networks, and the problems (and potential solutions) connected to them. Evaluation in information retrieval is typically based on massive test beds in the range of gigabytes nowadays. As standard collections like TREC or INEX are used, different approaches can be compared rather easily. On the other hand, no such standard ways of evaluating P2PIR has been established so far. Thus, a significant amount of papers in this area base their evaluation on small number of peers or documents, yielding only preliminary results and making comparison of approaches difficult if not impossible.

One result of the workshop's discussion phase is a three-dimensional classification scheme, which distinguishes between the application scenario (e.g. enterprise search, networks of scientists or web search), the task (e.g. recall-, precision- or efficiency-oriented), and the techniques employed (e.g. retrieval, clustering, filtering). Concrete approaches for information retrieval in peer-to-peer networks should be put into this frame, and then can be compared to competing approaches in the same class. Questions like the size of the network or the test bed to be used in evaluations can be answered then based on this classification scheme.

A lively discussion was centered about dynamics of a P2P network. Unless peer joins and departures are ignored completely, current research relies on artificial churn models. The participants believed that a better evaluation requires realistic scenarios. The idea of a small project was born which would aim at creating such a test bed: Interested researchers from the P2PIR community would install a small program (which has to be created) which monitors the online times and queries posed, and make them available in an anonymous version. Queries, e.g. to CiteSeer, should be logged by some kind of internal proxy. Of course, all collected data would be anonymized, and submission to the test bed has to be initiated explicitly by the user. This project is currently under discussion on a mailing list we set up for this purpose after the workshop. Please feel free to participate and join the list under http://www.is.informatik.uni-duisburg.de/post-p2pir2005, every help is appreciated.

4   Conclusions

This workshop was a second attempt to bring together a diverse set of peer-to-peer researchers, particularly from the IR and DB communities. This has been reflected by the submissions and the audience. The papers spanned the set of issues that are currently of interest in peer-to-peer research, and showed new approaches like scatter/gather browsing or overall system throughput optimization for concurrent query.

Similarly to last year's workshop, the participants agreed in the need for more clearly defined and accepted task models, and for a widely-accepted, reusable peer-to-peer test beds. The first issue has been partially covered by a draft three-dimensional classification scheme of application scenario, task and techniques, although further research is required here. The second issue lead to the setup of the project for acquiring real user online data. We expect that this project will be highly influential for evaluating realistic peer-to-peer network settings.

We want to thank the organizers of the CIKM conference for their support. Special thanks go the members of program committee, which did a great job on reviewing the submissions. Finally, we also thank the authors and the participants for inspiring this research field.