
WORKSHOP REPORT

The CIKM 2006 Workshop on Information Retrieval in Peer-to-Peer Networks

Ivana Podnar Žarko
FER, University of Zagreb, Croatia
ivana.podnar@fer.hr

Fabrizio Silvestri
ISTI-CNR, Pisa, Italy
fabrizio.silvestri@isti.cnr.it

1 Introduction

Information Retrieval (IR) in distributed and decentralized environments has become an active field of research during the last years. Recently, Peer-to-Peer (P2P) networks have emerged as an attractive architectural paradigm for IR, both for technical and economic reasons. P2P networks are distributed and self-organizing systems that support resource sharing. For this reasons, they are promising building blocks for next-generation search engines that will have to deal with huge amounts of heterogeneous (e.g. textual, multimedia, audiovisual) and continuously changing data. Moreover, P2P search is appealing from an economic perspective since it requires minimal in-place infrastructure and maintenance, yet facilitating higher diversity in contents and search methods. However, P2P retrieval methods still pose a lot of research challenges. Search methods are typically limited to simple keyword queries and the use of advanced retrieval models is quite constrained. There is an ongoing debate within the research community on the efficiency and scalability of such approaches, as well as effectiveness and applicability of traditional models in this new environment.

P2PIR 2006, the third Workshop on Information Retrieval in Peer-to-Peer Networks was held on November 11, 2006 in conjunction with CIKM 2006. This year's workshop was a team effort continuing the tradition of past workshops, the Workshop on Information Retrieval in Peer-to-Peer Networks which was collocated with SIGIR in 2004 and with CIKM in 2005, and the Workshop on Heterogeneous Distributed Information Retrieval collocated with SIGIR in 2005. The workshop attracted around 15 researchers from the domains of IR and databases working on distributed and decentralized IR systems.

The call for papers attracted 13 submissions from Asia, Europe, and the United States. The program committee accepted 7 scientific papers presenting original solutions for resource selection and query routing, indexing techniques in decentralized environments, and implementations of P2P retrieval engines. The workshop was opened by Fausto Rabitti, research director at ISTI-CNR, Pisa, Italy, who gave an invited talk on the future of P2P audiovisual search and provoked a lively discussion that continued throughout the paper presentations and the closing panel discussion. Paper presentations were divided into the following sessions: *Resource selection*, *Similarity search*, and *Systems and applications*. The panel discussion was organized around the usage scenarios, and economic and social factors steering the adoption of the P2P technology in the field of IR. The invited talk slides and workshop program are available online from <http://lsirwww.epfl.ch/p2pir2006/>, and all published papers are accessible online from the ACM Digital Library.

2 Presentations

The invited talk given by Fausto Rabitti on *The Future of P2P Audiovisual Search* motivated the need for effective and efficient search methods that can deal with the growing volume of digital content most of which

is in audiovisual form. Fausto gave an overview of existing P2P applications and their search capabilities, and commented on the limitations of the state-of-the-art audiovisual web search engines. Similarity search was proposed as one of the promising techniques for audiovisual search, and P2P systems as the underlying technology enabling dynamic search scenarios. Search expressiveness, scalability, and load balancing have been identified as further research challenges.

The first technical session on Resource selection was opened by Matthias Bender with *Size Doesn't Always Matter: Exploiting PageRank for Query Routing in Distributed IR*. Matthias presented a method for distributed authority scoring and a strategy for incorporating global PageRank scores into the query routing process which shows the potential to improve peer selection compared to CORI. Gleb Skobeltsyn presented a query-driven full-text indexing strategy in structured P2P networks with his talk on *Distributed Cache Table: Efficient Query-Driven Processing of Multi-Term Queries in P2P Networks*. Presented results show that by caching specially selected and profitable query answers in structured P2P networks under the assumption of limited storage per peer, a large fraction of future queries can be efficiently answered.

Three papers were presented in the session on Similarity search. Ruj Akavipat introduced a method to evaluate a peer search network based on the concept of semantic similarity in his talk on *Emerging Semantic Communities in Peer Web Search*. Ruj gave a demo of 6S, an unstructured P2P search engine for collaborative web search which was used to investigate peer clustering by monitoring the quality of the peers neighbors over time. The analysis has shown the spontaneous formation of semantic communities through local peer interactions. Alessandro Linari presented the work on *Efficient Peer-to-Peer Semantic Overlay Networks based on Statistical Language Models* and proposed the formation of P2P network in such a way that each peer becomes a neighbor of a small number of peers holding content most similar to the one it hosts. Similarity is determined using the square root of the Jensen-Shannon divergence. In his talk on *M-Grid: Similarity Searching in Grids*, Vlastislav Dohnal proposed the Metric Grid (M-Grid), a distributed index structure for similarity searching that exploits a Grid infrastructure.

The last session was related to systems and applications. In his talk on *Applying Information Retrieval Techniques to Peer-to-Peer File Sharing Systems* Ophir Frieder gave an overview of research directions related to P2P information retrieval at the Illinois Institute of Technology and illustrated how IR techniques can improve the search efficiency of P2P file sharing systems. The last talk was delivered by Martin Rajman who presented *ALVIS Peers: A Scalable Full-text Peer-to-Peer Retrieval Engine*. Martin explained a novel indexing and retrieval model designed for scalability in structured P2P networks, and presented how it has been integrated with a solution for content-based ranking into a fully-functional P2P full-text retrieval engine.

3 Discussions

The panel discussion was motivated by the question on whether there is a place for the P2P technology within the field of IR and which factors and applications will increase the adoption of this technology to produce next-generation search engines.

The volume of produced data and diversity of such data was identified as the major factor driving the need for distributed and decentralized IR solutions. There are immense amounts of privately owned data that will never be indexed by search engines like Google or Yahoo. These engines are extremely good to satisfy the needs of everyday users. However, there was a common agreement that P2P search engines are not playing against state-of-the-art engines on the same field, but are trying to be complementary specifically oriented towards community-based search.

Technology, social impact, and economic issues were listed as the three main factors influencing the adoption of the P2P technology. Someone suggested that intellectual input from volunteers where people invest their knowledge and time is important because P2P is about sharing of resources and data. There is potential to generate new technology and to improve user experience with e.g. online home stores that are getting more popular and that might create a successful business model. There was no common agreement on whether we can define a successful business model because the participants were mainly from academia

focusing on innovation and research. However, it was concluded that new prototypes and interesting applications are needed to attract more users and enable large-scale experiments. Additionally, benchmark tests are necessary to compare different systems. The discussions were closed with a proposal to initiate a wiki page for the P2PIR community and to organize a summer school where participants can further discuss the technology-related issues and define the benchmark tests.

4 Acknowledgments

We thank the authors and presenters, the invited speaker and the members of the program committee, and all workshop participants for their contributions to the presented material and productive discussions. We would like to remember and acknowledge the efforts of our dear colleague Henrik Nottelmann in appreciation of his enthusiastic efforts invested in organization of the previous P2PIR workshops.