The SIGIR Peer-to-Peer Information Retrieval Workshop

Jamie Callan
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15241, USA
callan@cmu.edu

Norbert Fuhr
Institute of Informatics and Interactive Systems
University of Duisburg-Essen
47048 Duisburg, Germany
fuhr@uni-duisburg.de

Introduction

Peer-to-peer (P2P) systems have emerged as a popular way to share huge volumes of data. The P2P paradigm holds many promises: it fits naturally with the Internet, the universal knowledge and service exchange medium; it favors scalability, by allowing the seamless plugging of data, services and computational resources into the global system; it increases system resilience, by avoiding unique points of failures; and it can speed up global access by distributing the indexing and query processing tasks to multiple computing nodes. However, retrieval methods for P2P systems are still in their infancy. P2P networks are prone to congestion when messages are not routed intelligently. Many of the most effective routing or data placement methods developed recently rely on relatively simple retrieval methods and homogeneous network environments.

This workshop focused on new methods of resource representation, resource selection, and data fusion in peer-to-peer networks. The workshop particularly encouraged papers that addressed heterogeneous peer-to-peer networks (e.g., a variety of data types and service providers), as well as papers about methods that cope with partial and uncertain information. However, more broadly, papers were solicited on any topic related to information retrieval in peer-to-peer networks. The goals of the workshop were to discuss current retrieval methods of P2P systems, as well as the adaptation of distributed IR methods for P2P systems, and to involve both researchers from the P2P area interested in IR methods as well as IR researchers aiming at extending their methods for P2P systems.

The workshop was held on July 29, 2004 in Sheffield, the United Kingdom, immediately following the SIGIR 2004 conference. Although the workshop was intended to appeal to a wide range of IR and P2P researchers, it primarily drew people with a background in IR. About 20 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.

Presentations

Below we provide very brief descriptions of the workshop presentations, to give a sense of the range of themes and topics covered. The complete workshop proceedings are available in electronic form at http://p2pir.is.informatik.uni-duisburg.de/programme.html.

Christos Tryfonopoulos opened the workshop with a talk on *Implementing Publish/Subscribe Systems with Languages from Information Retrieval on Top of Structured Overlay Networks*. The goal of this project is to support both ad-hoc and publish/subscribe (information filtering, or SDI) within a single P2P network. The network is hierarchical, and is based upon ideas from self-organizing networks and distributed hash tables. Hub nodes specialize in particular topics, and they are where topic-specific filtering occurs. Message grouping reduces message traffic to subscribers, and message traffic grows logarithmically. The system was evaluated initially using documents from ResearchIndex.com, and synthetic filtering queries; more thorough evaluations are underway.

Paul-Alexandru Chita's *Knowing Where to Search: Personalized Search Strategies for Peers in P2P Networks* described a distributed version of the popular Page Rank algorithm, and how it could be personalized to represent an individual's specific interests. The algorithm has been tested on P2P networks of up to 150 hubs and 215,000 leaf nodes containing synthetic data; the retrieval model was exact match, and documents are returned until the desired number of documents is reached. The personalized Page Rank method was found to be better than alternatives such as Best-neighbor search.

Atip Asvanund's *Content-Based Community Formation in Hybrid Peer-to-Peer Networks* addressed the problem of organizing peer-to-peer networks into virtual communities (or "clubs", in the economic sense) so that nodes with similar content are grouped under the same hubs. The nodes in his network are autonomous text-based digital libraries. The regional directory services (hubs) each optimize their own utility functions, but in so doing the network gradually becomes more organized, message traffic decreases, and Recall increases. The algorithm was tested on a testbed of 2,500 text-based digital libraries that is based on the TREC wt10g data.

Fabius Klemm talked about *An Architecture for Peer-to-Peer Information Retrieval*. His work is a formal, abstract architecture that can be operationalized to model specific P2P systems. It consists of four layers: Transport/communication; structured overlay/network maintenance; document and content management; and retrieval models. The architecture is intended to help isolate and clarify the roles and contributions of P2P and IR research in a hybrid system. The architecture has been verified by mapping several well-known P2P systems onto it. The next steps involve greater focus on retrieval models designed for use within P2P networks.

Sebastian Michel presented *Bookmark-driven Query Routing in Peer-to-Peer Web Search*. A peer's contents are assumed to be topically similar, for example, acquired by focused crawling from bookmarks. Peers post summary information about their contents to a global directory. Individual users can also post their bookmark lists, which are routed to peers with similar content. When a new query is received, a query attempts to satisfy it locally; if it cannot, it acquires information from the centralized directory service about nodes with similar content, and then routes the query using this information; it may also route the query to nodes that have previously published similar bookmark lists. A prototype has been created, and the system is currently being evaluated.

Jie Lu's *Federated Search of Text-Based Digital Libraries in Hierarchical Peer-to-Peer Networks* continues her prior research on adapting text-based distributed IR techniques for P2P networks. Each leaf node is an autonomous text-based digital library that uses best-match search. Hub nodes are topically-organized directory services that route messages and merge ranked-lists returned by leaf nodes. Each hub node makes a local, query-specific decision about how many leaf nodes it will forward the query to. Leaf nodes, hub nodes, and network neighborhoods are all represented by terms and term frequencies. Her research was evaluated with an extensive set of experiments on a testbed of 2,500 full-text search engines based on TREC wt10g data, 25 hub nodes, and 2 sets of queries. Her results suggest that good, full-text distributed IR techniques in a large P2P network have Precision that is 15-20% lower than centralized search over the same data.

Henrik Nottelmann described *A Logic-based Approach for Computing Service Execution Plans in Peer-to-Peer Networks*. This research models a P2P heterogeneous network in which different nodes offer different services (e.g., search, query expansion, query translation, …). The network is dynamic, so a dynamic method is required to form query-specific execution plans. Henrik's approach is to use DAML-S profiles to describe services, to use pDatalog to describe facts, and to chain together services into execution plans. Cost models can be introduced, for example describing actual costs or preferences, to select optimal plans.

Mario Nascimento presented *Taxonomy-based Routing Indices for Peer-to-Peer Networks*. When the network is based on a taxonomy, assigning documents to locations in the network, and routing queries, become text categorization problems. Network nodes are assumed to be myopic, insofar as the amount of detail they know about other nodes depends on the nodes' distance in the network. At each hop one level of detail in the taxonomy is lost; a node knows nothing beyond a horizon of 4 hops (the depth of the taxonomy). Given a new query, a node rates its neighbors based on the likely number of hops to matching documents (h-score) and an estimate of the number of documents down that path (d-score). The query is routed along the top-scoring path. The approach was evaluated with synthetic data and shown to be effective.

Conclusions

One clear gap between the presentations and the workshop audience was the degree of realism in the network designs and the evaluations. Most of the papers presented in the workshop studied static networks, in which nodes don't enter, leave, or change their contents over time. Most of the papers had either no evaluation, evaluation with small networks, or evaluation with synthetic data that is easily challenged. It is not unreasonable to want techniques that are tested on large, dynamic peer-to-peer networks, using real data and real query streams, but these requirements are difficult for most research groups to meet. This research area lacks a good method for modeling large dynamic networks with reasonable hardware resources. It also lacks good datasets; although Web data is easy to obtain, realistic networks require large query streams and relevance assessments.

One problem for the P2P IR community is that its task models are still very much undefined. The most well-known uses of peer-to-peer networks are file-sharing networks, in which file naming conventions are known or easily guessed, content is highly replicated across the network, and exact- or partial-match retrieval is sufficient. The database research community studies an abstract version of this task, in which the representation is a controlled-vocabulary, exact- or partial-match is sufficient, and ranking isn't much of an issue.

When asked about the advantages of peer-to-peer search for information retrieval tasks, the workshop participants thought that text retrieval in P2P networks would be very appropriate for small-office environments, small family-and-friends networks, community-based search, and small, ad-hoc networks; in other words, environments where a loosely organized group, each with moderate information technology resources, wants to share a "small" number of documents. It isn't clear that these applications require research that is tested immediately on large-scale, very dynamic networks. There was also discussion about using peer-to-peer search for "hidden Web" type problems, e.g., to integrate search engines that don't reveal their contents, or to provide more current information than can be provided easily by crawling. This scenario use would clearly require evaluation on large networks, but even in this setting the degree of dynamism in the network is unclear; many large information providers have a relatively stable presence on the network.

We do not advocate one scenario over the other. We merely argue that because there is not yet agreement on what the task models are for peer-to-peer Information Retrieval, it is difficult to define evaluation methodologies that are widely accepted.

Perhaps because of significant research attention from the database community, peer-to-peer research is an area where many battles from the past are being refought. There continues to be disagreement about full-text vs. controlled-vocabulary representations; exact-match vs. best-match retrieval; and ranking vs. not ranking results. Perhaps most jarring to an IR researcher, evaluation on artificial datasets is common; real documents, queries, and relevance judgements aren't required.

This workshop was a first attempt to bring together a diverse set of peer-to-peer researchers, particularly from the IR and DB communities, to begin exchanging ideas. Although the audience was primarily an IR audience, the papers spanned the set of issues that are currently of interest in peer-to-peer research, thus partially achieving our goals. They indicate a few areas of agreement, many areas of disagreement, the serious need for more clearly defined and accepted task models, and the need for widely-accepted, reusable peer-to-peer testbeds. The workshop was a beginning, but we hope that it is only the beginning, of a continuing dialog between the many corners of the peer-to-peer research community.

Acknowledgements

We thank Henrik Nottelmann for his help in all aspects of preparing the workshop, including setting up the web site, providing technical support to the PC, and compiling the workshop proceedings; it would have been a much less professional effort without his help. We also thank the organizers of the SIGIR conference for their support of the workshop. We especially thank them for locating the workshop above the top floor served by the paternoster lifts; going "over the top" was an experience that most of the workshop participants will not forget.