P2P Information Retrieval and Filtering

Christos Tryfonopoulos
Databases and Information Systems Department
Max-Planck Institute for Informatics
trifon@mpi-inf.mpg.de

1. Introduction

Today's content providers are naturally distributed and produce large amounts of new information every day, making peer-to-peer (P2P) data management a promising approach that offers scalability, adaptivity to high dynamics, and failure resilience. Although there exist many P2P data management systems in the literature, most of them focus on providing only *information retrieval* (IR) [BMT+05b, LC05, SCL+05, TXD03, SEA+04, BMT+05b] or *filtering* (IF) [TX03, AT06] functionality (also referred to as *publish/subscribe* or *alerting*), and have no support for a combined service. Querying in such scenarios is unarguably the most popular user activity, however subscribing with a *continuous query* is of equal importance as it allows the user to cope with the high rate of information production and avoid the cognitive overload of repeated searches. In an IF setting users, or services that act on users' behalf, specify continuous queries, thus subscribing to newly appearing documents that satisfy the query conditions. The IF system is responsible for notifying the user automatically whenever a new matching document is published.

The work presented here, tries to bridge the gap between these two important querying paradigms and support both IR and IF in a unifying P2P framework. In the following, two different approaches that demonstrate the use of structured overlays as a routing substrate for two types of data management systems are presented. DHTrie [TIK05a] is an *exact* IR and IF system that stresses retrieval effectiveness, while MAPS [ZTB08, ZTW07] provides *approximate* IR and IF by relaxing recall guarantees to achieve better scalability. In [TZK+07] a comparison between the two different system designs is presented, and the trade-offs between the two approaches are highlighted. Documents and (continuous) queries in both systems are expressed using a well-understood attribute-value model that is based on named attributes with free text as value, interpreted under the Boolean and VSM (or LSI) models [KST06].

2. Exact Information Retrieval and Filtering

Most of the research in P2P data management has focused in providing exact retrieval functionality over both structured (e.g., *distributed hash tables* - DHTs [RD01, SMK+01]) and unstructured overlays (e.g., Gnutella). In DESENT [DNV07], peers are clustered by virtue of containing similar documents, and these clusters are organized in hierarchies to support a DL application [DNV06]. Clusters use peers that act as cluster gateways to forward queries and groups of clusters may form super-clusters with their own gateways. In a similar spirit, iCluster [RP08] organizes peers in an unstructured overlay into communities sharing similar content, and allows them to have multiple and dynamic interests by utilizing unsupervised clustering methods (e.g., k-means [SKK00]) to identify these interests. An extension of the iCluster protocols, that supports both IR and IF functionality in a DL domain, has been presented in [RPT+08]. Other approaches that focus on IR over unstructured overlays include the summarization of a peer's content through specialized data structures to facilitate routing of user queries to appropriate peers [KP04, KP08].

The DHTrie architecture follows a different route from the approaches presented earlier by utilizing a structured overlay as the routing substrate. It provides protocols based on DHTs for efficient and adaptive data management, and centralized algorithms for handling document and queries in each peer. To achieve this is it employs two levels of indexing documents (for the IR task) and continuous queries (for the IF task). A prototype system that presents the ideas behind the DHTrie protocols in the context of digital libraries is presented in [TIK05b].

The first level of the DHTrie indexing scheme corresponds to the partitioning of the global index to different peers using the DHT as the underlying routing infrastructure. Each peer is responsible for a fraction of the submitted continuous queries through a mapping of attribute values to peer identifiers. The DHT infrastructure is used to define the mapping scheme and also manages the routing of messages between different nodes. The set of DHTrie protocols extends the basic functionality of the DHT to offer retrieval and filtering functionality in a dynamic peer-to-peer environment [TIK05a].

The second level of the DHTrie indexing mechanism is managed locally by each peer, and is used for indexing the documents and the continuous queries the peer is responsible for. To be able to scale up to large numbers of documents and continuous queries, specialized data structures and local indexing algorithms are of paramount importance. The idea behind the centralized indexing mechanism is to use trie-based data structures to capture common elements between indexed queries, and exploit this clustering at filtering time [TKD04].

3. Approximate Information Retrieval and Filtering

All approaches to P2P data management taken so far, focus on exact retrieval [BMT+05b, LC05, SCL+05, TXD03, SEA+04, BMT+05a] or filtering [TX03, AT06, TIM05a, TIK05b] by using the P2P network as a decentralized index for both documents and continuous queries. To facilitate this indexing, appropriate protocols that disseminate documents and queries in a deterministic way, depending on the terms contained in them are employed. These document and query indexing protocols lead to filtering effectiveness that is exactly the same as that of a centralized system. This, however, creates an efficiency and scalability bottleneck, while in certain applications this design might not even be desirable (e.g., in applications like news of blog filtering, where the user is not interested in *all* relevant items, but rather in the most interesting ones).

Contrary to approaches that provide exact IR and IF functionality by utilizing per-document indexing, in MAPS [ZTB+08, ZTW08, ZTW07] the concept of approximate IR and IF is introduced; publications are processed locally and peers query or subscribe to only a few, selected information sources that are most likely to satisfy the user's information demand. In this way, per-peer (rather than per-document) indexing is employed and efficiency and scalability are enhanced by trading a small reduction in recall for lower message traffic.

The MAPS system utilizes a structured overlay to support publisher selection and ranking necessary for both IR and IF scenarios. This selection is driven by statistical summaries stored in a distributed P2P directory built on top of the Pastry DHT [RD01]. For scalability, summaries have publisher and not document granularity, thus capturing the best publisher for certain keywords but not for specific documents. Both approximate IR and IF services utilize the same conceptually global, but physically distributed directory of statistical metadata to derive information provider rankings. To support the IR functionality, MAPS utilizes well-known resource selection techniques for P2P query routing such as tf.idf based methods, CORI, or language models (see [NF03] for an overview) to route the user query to a carefully selected subset of information sources. Resource selection in such an autonomous and dynamic environment can be improved by taking into account the overlap in the document collections of different content providers [BMT+05b].

To support P2P IF in a scalable and efficient way, MAPS ranks sources, and delivers matches only from the best ones, by utilizing novel publisher selection strategies. Thus, the continuous query is replicated to the best information sources and only published documents from these sources are forwarded to the subscriber. This approximate IF relaxes the assumption, which holds in most IF systems, of potentially delivering notifications from every producer and amplifies scalability. To select the most appropriate publishers to subscribe to, a subscriber computes scores that reflect the past publishing behavior and utilises them to predict future peer behavior. This score is based on a combination of resource selection (i.e., tf.idf based) and behavior prediction to deal with the dynamics of publishing. Behavior prediction uses time-series analysis with double exponential smoothing techniques [C04] to predict future publishing behavior, and adapt faster to changes in it. In addition, correlations among keywords in multi-term continuous queries can be exploited to further improve publisher selection. In [ZTW08], two such strategies based on statistical synopses are described in detail. In this way, approximate IF achieves higher scalability by trading faster response times and lower message traffic for a moderate loss in recall.

References

[AT06] I. Aekaterinidis and P. Triantafillou. PastryStrings: A Comprehensive Content-Based Publish/Subscribe DHT Network. In ICDCS, 2006.

[BMT+05a] M. Bender, S. Michel, P. Triantafillou, G. Weikum, and C. Zimmer. MINERVA: Collaborative P2P Search. In VLDB, 2005.

[BMT+05b] M. Bender, S. Michel, P. Triantafillou, G. Weikum, and C. Zimmer. Improving Collection Selection with Overlap-Awareness. In SIGIR, 2005.

[C04] C. Chatfield. The Analysis of Time Series - An Introduction. CRC Press 2004.

[DNV06] C. Doulkeridis, K. Nørvåg, M. Vazirgiannis: Scalable Semantic Overlay Generation for P2P-Based Digital Libraries. In ECDL, 2006.

[DNV07] C. Doulkeridis, K. Nørvåg, M. Vazirgiannis: DESENT: Decentralized and Distributed Semantic Overlay Generation in P2P Networks. IEEE Journal on Selected Areas in Communications, 2007.

[KP04] G. Koloniari and E. Pitoura. Content-based Routing of Path Queries in Peer-to-Peer Systems. In EDBT, 2004.

[KP08] G. Koloniari and E. Pitoura. A Clustered Index Approach to Distributed XPath Processing. In ICDE, 2008.

[KST06] M. Koubarakis, S. Skiadopoulos, and C. Tryfonopoulos. Logic and Computational Complexity for Boolean Information Retrieval. In TKDE, 2006.

[LC05] J. Lu and J. Callan. Federated Search of Text-based Digital Libraries in Hierarchical Peer-to-Peer Networks. In ECIR, 2005.

[NF03] H. Nottelmann and N. Fuhr. Evaluating Different Methods of Estimating Retrieval Quality for Resource Selection. In SIGIR 2003.

[RD01] A. Rowstron and P. Druschel. Pastry: Scalable, Distributed Object Location and Routing for Large-Scale Peer-to-Peer Systems. In Middleware, 2001.

[RP08] P. Raftopoulou and E.G.M. Petrakis. iCluster: a Self-Organizing Overlay Network for P2P Information Retrieval. In ECIR, 2008.

[RPT+08] P. Raftopoulou, E.G.M. Petrakis, C. Tryfonopoulos, and G. Weikum. Information Retrieval and Filtering over Self-Organising Digital Libraries. In ECDL, 2008.

[SCL+05] J. Stribling, I. Councill, J. Li, M. Kaashoek, D. Karger, R. Morris, and S. Shenker. Overcite: A Cooperative Digital Research Library. In IPTPS, 2005.

[SEA+04] O. Sahin, F. Emekci, D. Agrawal, and A. Abbadi. Content-based Similarity Search over Peer-to-Peer Systems. In DBISP2P, 2004.

[SKK00] M. Steinbach, G. Karypis, and V. Kumar. A Comparison of Document Clustering

Techniques. In TextDM, 2000.

[SMK+01] I. Stoica, R. Morris, D. Karger, M.F. Kaashoek, and H. Balakrishnan. Chord: A Scalable Peer-to-peer Lookup Service for Internet Applications. In SIGCOMM, 2001.

[TIK05a] C. Tryfonopoulos, S. Idreos, and M. Koubarakis. Publish/Subscribe Functionality in IR Environments using Structured Overlay Networks. In SIGIR, 2005.

[TIK05b] C. Tryfonopoulos, S. Idreos, and M. Koubarakis. LibraRing: An Architecture for Distributed Digital Libraries Based on DHTs. In ECDL, 2005.

[TKD04] C. Tryfonopoulos, M. Koubarakis, and Y. Drougas. Filtering Algorithms for Information Retrieval Models with Named Attributes and Proximity Operators. In SIGIR, 2004.

[TX03] C. Tang and Z. Xu. pFilter: Global Information Filtering and Dissemination Using Structured Overlays. In FTDCS, 2003.

[TXD03] C. Tang, Z. Xu, and S. Dwarkadas. Peer-to-Peer Information Retrieval Using Self-Organizing Semantic Overlay Networks. In SIGCOMM, 2003.

[TZK+07] C. Tryfonopoulos, C. Zimmer, M. Koubarakis and G. Weikum. Architectural Alternatives for Information Filtering in Structured Overlay Networks. In IEEE Internet Computing, 2007.

[ZTB+08] C. Zimmer, C. Tryfonopoulos, K. Berberich, M. Koubarakis, and G. Weikum. Approximate Information Filtering in Peer-to-Peer Networks. In WISE, 2008.

[ZTW07] C. Zimmer, C. Tryfonopoulos, and G. Weikum. MinervaDL: An Architecture for Information Retrieval and Filtering in Distributed Digital Libraries. In ECDL, 2007.

[ZTW08] C. Zimmer, C. Tryfonopoulos, and G. Weikum. Exploiting Correlated Keywords to Improve Approximate Information Filtering. In SIGIR, 2008.

Greek Institutions Involved

Part of the work described in this document is research conducted in the following Greek institutions[1]:

- Intelligent Systems Lab, Dept. of Electronic and Computer Engineering, Technical University of Crete.
- Department of Informatics and Telecommunications, National and Kapodistrian University of Athens.
- DB-NET group, Department of Informatics, Athens University of Economics and Business.
- Network-Centric Information Systems (NetCINS) Lab, Department of Computer Engineering and Informatics, University of Patras.
- Distributed Management of Data (DMOD) Lab, Computer Science Department, University of Ioannina.
- Software and Database Systems Lab, Department of Computer Science and Technology, University of Peloponnese
- Research Academic Computer Technology Institute (CTI).

[1] The institutions were obtained using the affiliation of the authors at the time of the publication.