P2P Information Retrieval and Filtering with MAPS

Christian Zimmer, Johannes Heinz, Christos Tryfonopoulos, and Gerhard Weikum
Max-Planck Institute for Informatics, 66123 Saarbrücken, Germany
{czimmer, johheinz, trifon, weikum}@mpi-inf.mpg.de

Abstract

In this demonstration paper we present MAPS, a novel system that combines approximate information retrieval and filtering functionality in a peer-to-peer setting. In MAPS, a user is able to submit one-time and continuous queries, and receive matching resources and notifications from selected information sources. The selection of these sources in the retrieval case is based on well-known resource selection techniques for peer-to-peer query routing, while in the filtering case a combination of resource selection and novel behavior prediction techniques using time-series analysis of publisher statistics is used. The integration of the two functionalities is done in a seamless way utilizing the same machinery: a conceptually global, but physically distributed directory of statistics about information sources based on distributed hash tables.

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 (e.g., [4, 1]), most of them focus on providing only information retrieval (IR) or filtering[1] (IF) functionality, and have no support for a combined service. Querying in P2P systems 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 automatically notifying the user whenever a new matching document is published. To bridge the gap between these two important querying paradigms, we have developed the MAPS (Minerva Approximate Publish/Subscribe) prototype that builds on existing and novel techniques to support both IR and IF functionality in a unifying P2P framework.

Contrary to approaches like [4, 1] that provide *exact* IR and IF functionality (e.g., by utilizing per-document indexing), in MAPS we introduce the concept of *approximate* IR and IF; 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, we employ per-peer (rather than per-document) indexing and enhance efficiency and scalability by trading a small reduction in recall for lower message traffic.

To illustrate the necessity of supporting both IR and IF in a single system, consider an application scenario where John, who is a professor in computer science, is interested in data mining and wants to follow the work of prominent researchers in the area. He regularly uses a search engine and the digital library of his department to search for new papers. Even though searching for interesting papers today turned up nothing, a search next week may turn up new papers. Clearly, John would benefit from accessing a system that is able to not only provide a search functionality that integrates a lot of sources, but also capture his long term information need. This system would be a valuable tool, beyond anything supported in current information management systems, that would allow John to save time and effort.

2 System Overview

Figure 1 shows a high-level view of the MAPS architecture and the different types of services implemented.

The P2P directory. 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. 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.

IR in MAPS. To support the IR functionality, we use well-known resource selection techniques for P2P query

[1] Also known as publish/subscribe, or continuous querying.

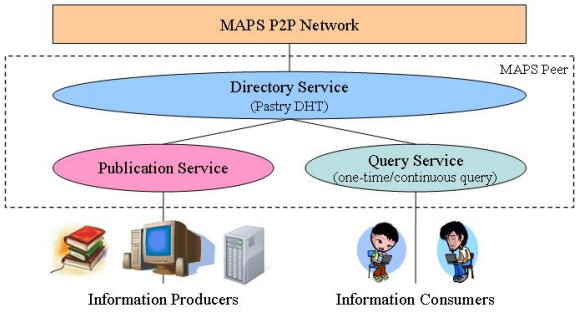

Figure 1. High-level architecture of MAPS.

routing, e.g., CORI [3], to route the user query to a carefully selected subset of information sources. Resource selection in such an autonomous and dynamic environment is improved by taking into account the overlap in the document collections of different content providers [2].

IF in MAPS. To support P2P IF in a scalable and efficient way, we have introduced the concept of approximate IF. Most approaches on IF taken so far have the underlying hypothesis of potentially delivering notifications from every information producer to subscribers (e.g., [1]). This exact IF model imposes a cognitive overload on the user in the case of applications like blog or news filtering, and creates an efficiency and scalability bottleneck. Contrary to this, our approximate IF approach 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 utilizes them to predict future peer behavior. This score is based on a combination of resource selection (e.g., CORI [3]) and behavior prediction to deal with the dynamics of publishing [5]. Behavior prediction uses *time-series analysis* with *double exponential smoothing* techniques 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 [7], 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.

3 Demonstrator Setup

In our demonstration, we will present the MAPS prototype system. All peers in MAPS implement the directory service and the subscription or the publication service.

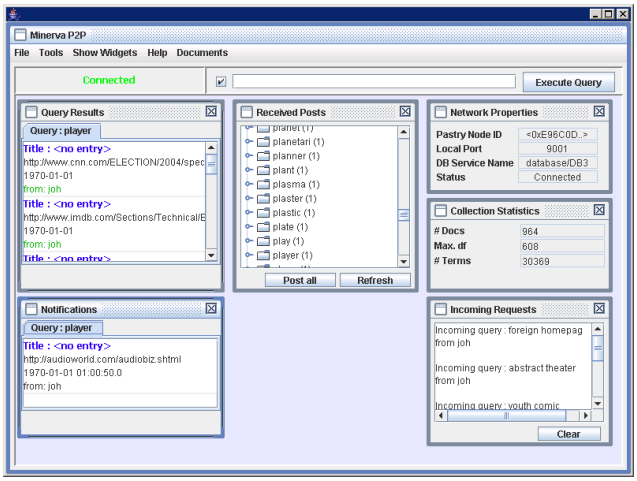

Figure 2. MAPS graphical user interface.

Through an appropriate user interface shown in Figure 2, a user utilizing the subscription service can pose multi-keyword one-time queries and receive search results from the top-ranked peers. Additionally, he can subscribe with continuous queries and receive notifications when new documents matching his information demand are published in the future by some other peer in the network. Finally, a user utilizing the publication service is able to manage its own document collection in a local database and choose when to publish a document (i.e., make it available) to the rest of the peers in the network. The demonstrator presents a use case of the implemented prototype including the following steps: initialization and metadata dissemination, one-time query execution, continuous query subscription, document publication, and notification delivery.

The MAPS prototype is a novel P2P system developed to integrate the search and filtering paradigms by emphasizing peer autonomy. Contrary to exact IR and IF systems, documents in MAPS are not disseminated in the network, but are kept by their owners, who may charge for their content. For more details on the architecture, the behavior prediction mechanisms used in IF and a comparison against exact IR and IF systems the interested reader is referred to [6, 7, 5].

References

[1] I. Aekaterinidis and P. Triantafillou. Internet Scale String Attribute Publish/Subscribe Data Networks. In *CIKM*, 2005.

[2] M. Bender, S. Michel, P. Triantafillou, G. Weikum, and C. Zimmer. Improving Collection Selection with Overlap Awareness in P2P Search Engines. In *SIGIR*, 2005.

[3] James P. Callan, Zhihong Lu, and W. Bruce Croft. Searching Distributed Collections with Inference Networks. In *SIGIR*, 1995.

[4] G. Skobeltsyn, T. Luu, I. Podnar Zarko, M. Rajman, and K. Aberer. Web Text Retrieval with a P2P Query-Driven Index. 2007.

[5] C. Tryfonopoulos, C. Zimmer, G. Weikum, and M. Koubarakis. Architectural Alternatives for Information Filtering in Structured Overlays. *IEEE Internet Computing*, 2007.

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

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