

Node Behavior Prediction for Large-Scale Approximate Information Filtering

Christian Zimmer¹, Christos Tryfonopoulos¹, Klaus Berberich¹, Gerhard Weikum¹, and Manolis Koubarakis²
¹Max-Planck Institute for Informatics, Saarbrücken, Germany | ²National and Kapodistrian University of Athens, Greece

Contact Author and Poster by: Christian Zimmer, Max-Planck Institute for Informatics, czimmer@mpi-inf.mpg.de

Introduction to Information Filtering

Information Filtering (IF), also referred to as [publish/subscribe](#) or [continuous querying](#), is equally important to one-time querying. In an IF scenario, a user posts a [subscription](#) (or [continuous query](#) or profile) to the system to receive [notifications](#) whenever certain events of interest take place (e.g., when a paper on P2P becomes available).

„Win Scalability and Efficiency“

Our Approach - MAPS

We forward MAPS ([Minerva Approximate Publish/Subscribe](#)), a novel architecture to support approximate information filtering functionality in a Peer-to-Peer (P2P) environment: While most information filtering approaches taken so far have the underlying hypothesis of potentially delivering notifications from every information producer, MAPS relaxes this assumption by monitoring only selected sources that are likely to publish. The user query is replicated to these sources and only published documents from these sources are forwarded to him.

Peer Selection: Critical Decision

To select which publisher nodes should be monitored, the subscription process utilizes a scoring function to rank nodes. In the MAPS approach, the querying node computes a node score by using the directory service to collect per-term statistics of each query term. The node score combines a resource selection score and a node behavior prediction score. [Resource Selection](#) uses well-known techniques in the Information Retrieval field to identify authorities (e.g., CORI). [Behavior Prediction](#) uses [time-series analysis](#) of IR (e.g., document frequencies) metrics to predict the future behavior. MAPS applies [double exponential smoothing](#) as prediction technique and recognizes trends in the node behavior.

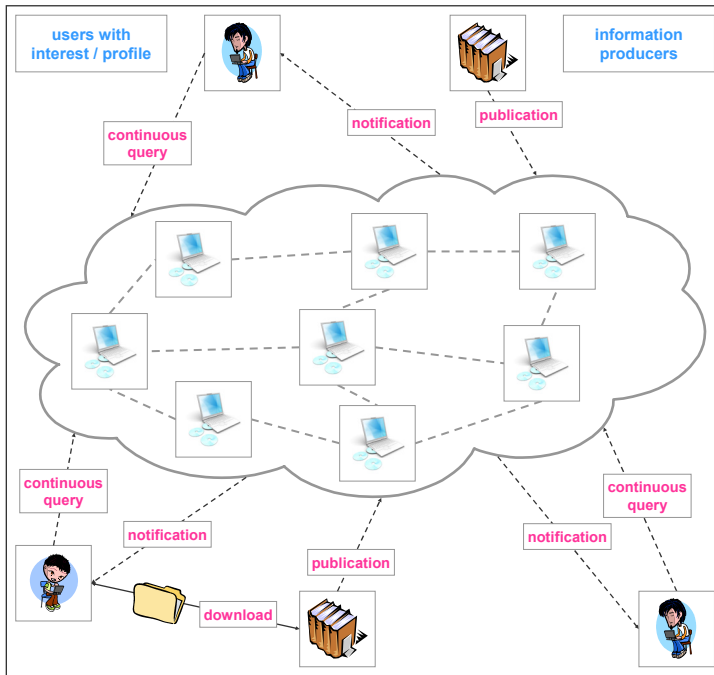


Figure 1: MAPS Network

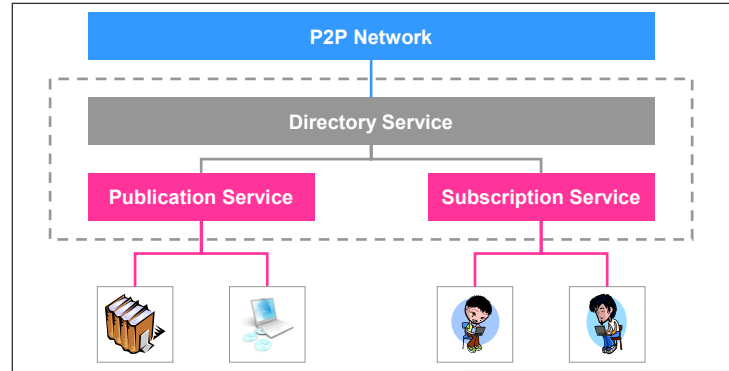


Figure 2: MAPS Architecture

Node Behavior Prediction

To predict the exact IR statistics in the Behavior Prediction, a global parameter setting of double exponential smoothing is not ideal. For different behaviors, there are different parameter combinations that yield to satisfying predictions. The choice of the parameters needs to be adapted to the observed value behavior.

MAPS [Selective Method](#) approach always selects the most appropriate parameter setting concerning the last observed prediction value. This way, the double exponential smoothing parameters are individually selected depending on the publisher node behavior for a given continuous query. There is no additional communication needed, and only local computations are performed.

Future Research Aspects

- The consideration of [correlated termsets](#) can improve recall quality by selecting only the most appropriate publisher nodes for a given multi-term query.
- To reach a satisfying level of recall, the [number of subscribed information sources can be selected automatically](#).
- Subscribed queries need to be re-positioned after certain time periods. The [automatic adaptation of these time periods](#) can improve recall and benefit/cost ratio.
- A [detailed message cost analysis](#) (including directory costs and filtering costs) will show the advantages of approximate IF in comparison to exact IF.
- [Information Retrieval and Information Filtering can be integrated](#) into one single system with one global directory, and two functionalities.

References

- [1] M. Bender, S. Michel, C. Zimmer, P. Triantafillou, and G. Weikum: Improving Collection Selection with Overlap-Awareness. *SIGIR*, 2005.
- [2] C. Tryfonopoulos, S. Idreos, and M. Koubarakis. Publish/Subscribe Functionality in IR Environments Using Structured Overlay Networks. *SIGIR*, 2005.
- [3] S. Michel, M. Bender, N. Ntarmos, P. Triantafillou, G. Weikum, and C. Zimmer. Discovering and Exploiting Keyword and Attribute-Value Co-Occurrences to Improve P2P Routing Indices. *CIKM*, 2006.
- [4] I. Stoica, R. Morris, D. R. Karger, M. F. Kaashoek, and H. Balakrishnan. Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications. *SIGCOMM*, 2001.
- [5] J. Callan. Learning While Filtering Documents. *SIGIR*, 1998.
- [6] C. Zimmer, K. Berberich, K. Manolis, C. Tryfonopoulos, and G. Weikum. MAPS: Approximate Publish/Subscribe Functionality in Peer-to-Peer Networks. *ADPUC*, 2006.
- [7] C. Zimmer, C. Tryfonopoulos, and G. Weikum. MinervaDL: An Architecture for Information Retrieval and Filtering in Distributed Digital Libraries. *ECDL*, 2007.
- [8] C. Tryfonopoulos, C. Zimmer, M. Koubarakis, and G. Weikum. Architectural Alternatives for Information Filtering in Structured Overlay Networks. *IEEE Internet Computing*, 2007.
- [9] C. Zimmer, C. Tryfonopoulos, K. Berberich, G. Weikum, and M. Koubarakis. Node Behavior Prediction for Large-Scale Information Filtering. *LSDS-IR*, 2007.

LSDS-IR 2007

First Workshop on Large Scale Distributed Systems for Information Retrieval
July 27, 2007, Amsterdam, Netherlands

Poster by
Christian Zimmer



The workshop is co-located with the 30th Annual International ACM SIGIR Conference 2007
July 23-27, 2007, Amsterdam, Netherlands
Hotel Krasnapolsky & University of Amsterdam

