Pricing Information Goods in Distributed Agent-Based Information Filtering

Christos Tryfonopoulos\textsuperscript{1} and Laura Maria Andreescu\textsuperscript{2}

\textsuperscript{1} University of Peloponnese, Tripoli, Greece
trifon@uop.gr

\textsuperscript{2} APYDOS, Luxembourg
landreescu@apydos.com

Abstract. Most approaches to information filtering taken so far have the underlying hypothesis of potentially delivering notifications from every information producer to subscribers; this exact information filtering model creates efficiency and scalability bottlenecks and incurs a cognitive overload to the user. In this work we put forward a distributed agent-based information filtering approach that avoids information overload and scalability bottlenecks by relying on approximate information filtering. In approximate information filtering, the user subscribes to and monitors only carefully selected data sources, to receive interesting events from these sources only. In this way, system scalability is enhanced by trading recall for lower message traffic, information overload is avoided, and information producers are free to specialise, build their subscriber base and charge for the delivered content. We define the specifics of such an agent-based architecture for approximate information filtering, and introduce a novel agent selection mechanism based on the combination of resource selection, predicted publishing behaviour, and information cost to improve publisher selection. To the best of our knowledge, this is the first approach to model the cost of information in a filtering setting, and study its effect on retrieval efficiency and effectiveness.

1 Introduction

Much information of interest to humans is available today on the Web, making it extremely difficult to stay informed without sifting through enormous amounts of information. In such a dynamic setting, information filtering (IF), also referred to as publish/subscribe, continuous querying, or information push, is equally important to one-time querying, since users are able to subscribe to information sources and be notified when documents of interest are published. This need for content-based push technologies is also stressed by the deployment of new tools such as Google Alert. In an IF scenario, a user posts a subscription (or continuous query) to the system to receive notifications whenever certain events of interest take place (e.g., when a document on Special Olympics becomes available). Since in an IF scenario the data is originally highly distributed residing on millions of sites (e.g., with people contributing to blogs,

\textsuperscript{*} Part of the work was done while the authors were with Max-Planck Institute for Informatics. The authors would like to thank David Midgley for his comments and suggestions on the economic aspects of this work.

© Springer-Verlag Berlin Heidelberg 2011
news portals, social networking feeds), a distributed approach seems an ideal candidate for such a setting.

In this work we put forward ABIS (Agent-Based Information filtering System), a novel agent-based architecture that supports content-based approximate information filtering. While most exact information filtering approaches \cite{32,15,14,34,1,27,8} taken so far have the underlying hypothesis of potentially delivering notifications from every information producer, ABIS relaxes this assumption by monitoring only selected sources that are likely to publish documents relevant to the user’s interests in the future. In ABIS, a user subscribes with a continuous query and monitors only the most interesting sources in the system, to receive published documents from these sources only. The system is responsible for managing the user query, discovering new potential sources and moving queries to better or more promising publishers. Approximate IF improves the scalability issues of exact IF by trading recall for lower message traffic, avoids information overload to the user, by allowing him to receive selected notifications from selected publishers, and proves an interesting business model for pricing information goods delivered by information producers. In approximate IF, each information producer might have its own customer base of interested subscribers, and may charge the delivered content by subscription or per item. Notice that this is not possible in the case of exact IF, since information consumers receive all matching notifications, from all producers, while to facilitate the distribution of the service, no notion of ownership control and publisher quality is employed. Finally, notice that system throughput and notification latency in exact IF depend heavily on publication size (which is usually large for textual IF). On the other hand, approximate IF is not affected by publication size (as there is no notion of information dissemination at publication time) and offers one-hop latency, since each publisher maintains its own database of subscribers. The interested reader is referred to \cite{36} for an insightful comparison of exact and approximate IF.

As possible application scenarios for ABIS consider the case of news filtering (but with the emphasis on information quality rather than timeliness of delivery) or blog filtering where users subscribe to new posts. Not only do these settings pose scalability challenges, but they would also incur an information avalanche and thus cognitive overload to the subscribed users, if the users were alerted for each and every new document published at any source whenever this matched a submitted continuous query. Our approximate IF approach ranks sources, and delivers matches only from the best ones, by utilising novel publisher selection strategies. These strategies take into consideration the quality of the information publisher, based on per-publisher statistics, and the price of information as this is set by the publisher. Despite the utilisation of a Distributed Hash Table (DHT) \cite{31} to maintain publisher statistics, notice that our architecture can also be realised in other settings, like a single coordinator agent monitoring a number of distributed sources, or a cloud-based multi-agent system providing an alerting service.

To the best of our knowledge, this is the first approach that aims at connecting system efficiency and effectiveness with the cost component, and puts economic modelling in the picture of distributed IF. In the light of the above, the contributions presented in this work are threefold: