DIP3: The 3Ps of Distributed Information Delivery: Preferences, Privacy and Performance

Today there is an abundance of data on line. The grand challenge is turning this huge amount of data to knowledge useful to the individual users of the Internet. DIP3 will address this challenge by tackling one form of data processing, often referred to as "push" data delivery. In push data delivery, instead of explicitly searching for information, users get notified when relevant information becomes available. Examples of such systems include RSS feeds, news alerts and aggregators. The scientific objective of the proposal is to derive models, algorithms and techniques to control both the amount and quality of information received by users. To this end, we propose incorporating user preferences in data delivery to rank data items based on their relevance to the users. Although preference specification has been extensively studied, there is little previous research work on incorporating preferences in Internet-scale data delivery. Furthermore, DIP3 will exploit the inherent social connections between users in Web 2.0 as expressed through social networks, social tagging, and other community-based features to enhance preference specification and ranked information delivery. Summarizing, the overall objective of DIP3 is delivering to the users the most relevant and interesting information.

Research Work and Results

The specific goal of DIP3 is to produce research results of high quality and subsequently publish them in top international conferences and journal. During the first year, the scientific results achieved are as follows:

The researcher continued her work on preferences in databases. In this aspect, she benefited from a close cooperation with researchers at Georgia Tech. The focus of this specific scientific work is on preferences in conjunction with keyword-based search in relational databases. Whereas keyword search allows users to discover relevant information without knowing the database schema or using complicated queries, keyword search often produces an overwhelming number of results, often loosely related to the user intent. To address this problem, personalizing keyword database search by utilizing user preferences was proposed. Query results are ranked based on both their relevance to the query and their preference degree for the user. To further increase the quality of results, two new metrics were introduced that evaluate the goodness of the result as a set, namely coverage of many user interests and content diversity. The efficiency and effectiveness of this approach was evaluated through extensive experiments. Results of this line of research have been published in [SDP10]. Furthermore, a demo of the implementation of an extension of this research towards extending databases with a recommendation functionality was presented in [KSDP10].

New research results were attained in the context of database selection for XML document collections. The database selection problem is defined as follows. Given a set of databases and a user query, how to rank the databases based on their goodness to the query. Goodness is determined by the relevance of the documents in the collection with the query. The fellow cooperated on this issue with PhD students at Georgia Tech. The focus of this research is on keyword queries with Lowest Common Ancestor (LCA) semantics for defining query results, where the relevance of each document to a query is determined by properties of the LCA of those nodes in the XML document that contain the query keywords. To avoid evaluating queries against each document in a collection, a pre-processing phase was introduced, during which information about the LCAs of all pairs of keywords was calculated and then used to approximate the properties of the LCA-based results of a query. To improve performance, both in terms of storage and processing efficiency, appropriate summaries of the LCA information based on Bloom filters were used. Results of this line of research have been published in [KP10].

Two new lines of work with regards to privacy were initiated. The theoretical underpinning related to preferences and privacy preservation in large scale distributed systems was the main focus of both.

- The first line of work refers to the problem of privacy through data anonymization in a distributed setting. The large majority of the related literature has focused in the centralized case. This work addresses the problem in a distributed setting were the data is horizontally partitioned at a number of data providers. A method for achieving k-anonymity for a horizontally partitioned relation is proposed that guarantees k-anonymity both for the final result and the individual subsets of it that are transferred from the data

providers to the central site that performs the anonymization. Details of this line of research can be found in [BSVPP10].

- The second line of research refers to the problem of enforcing privacy in a topic-based push based system. The main idea is to model the problem using item-set related privacy. Details of this ongoing line of research can be found in [GP10].

Final Results and Impact

The overall research in DIP3 will advance the state-of-the-art in the following ways.

- Whereas traditional pub/sub systems rely on a binary, match/no-match model for sending relevant data to users DIP3 proposes non binary matching, where items are assigned degrees of relevance. This is a novel idea.
- Although there is extensive work on preference models preference models and algorithms for push data delivery have not been addressed yet. To this end, DIP3 will explore both quantitative and qualitative models for expressing preferences. Another novel feature of DIP3 is exploiting social networks for enhancing pub/sub systems.
- DIP3 will address the interplay between system quality (as achieved through ranking) and system performance (especially in terms of response time). Enforcing performance guarantees in this setting is challenging mainly because of the large scale of the system, its diversity and dynamicity. DIP3 will exploit caching and replication at different system levels and at various granularities. Caching and replication are well studied; however, this new setting introduces new requirements. DIP3 will also develop new data structures and algorithms for clustering and indexing subscriptions and preferences that will take into account ranking and preferences.
- Finally, although there is a lot of research on privacy, privacy-preserving push-based delivery has not been explicitly addressed by previous research. We expect that the output of DIP3 will include a set of new personalized privacy models and mechanisms.

Privacy and large-scale Internet systems are central in the digital economy and areas of potential competitiveness for Europe. Many research labs (most notably Yahoo and Microsoft) are now hosting offices in Europe. In particular privacy has been an important concern in modern society and a major consideration with regards to the widespread use of Internet services. The goal of DIP3 is to provide a novel perspective for distributed information delivery by exploiting preferences, respecting privacy and increasing efficiency thus making it suitable for modern large-scale internet systems.

References

[SDP10] K. Stefanidis. M. Drosou and E. Pitoura. “PerK: Personalized Keyword Search in Relational Databases through Preferences”, Proceedings of the 13th International Conference on Extending Database Technology (EDBT), Lausanne, Switzerland, March 2010

[KSDP10] E. Koletsou, K. Stefanidis. M. Drosou and E. Pitoura. “YMAL: Recommendation for Relational Database Systems”, Demo in the 9th Hellenic Database Conference, Ayia Napa, Cyprus, June/July 2010

[KP10] G. Koloniari and E. Pitoura. “LCA-based Selection for XML Document Collections”, Proceedings of the 19th International Conference on World Wide Web (WWW), Raleigh, North Carolina, USA, April 2010

[BSVPP10] V. Bourgos, D. Souravlias, P. Vassiliadis, E. Pitoura, and E. Papapetrou, “Distributed Incognito: Minimizing the Risk of Privacy Breach for Distributed Data Sources”, Submitted

[GP10] A. Goulalas-Divanis and E. Pitoura, “Privacy in Publish/Subscribe Systems”, In Preparation.