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

http://www.dip-3.eu

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 addresses 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 DIP3 project is to derive models, algorithms and techniques to control both the amount and quality of information received by users. To this end, DIP3 introduces 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 exploits 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 DIP3 facilitates delivering to the users the most relevant and interesting information.

Research Work and Results

To achieve its objectives DIP3 has produced research results of high quality and subsequently published them in top international conferences and journal.

As both the volume of data and the diversity of users accessing them increase, user preferences offer a useful means towards improving the relevance of the query results to the information needs of the specific user posing the query.

However, most often, users have different ***preferences depending on context***. To account for this, a model was introduced for expressing contextual preferences. Context is modelled using a set of hierarchical attributes, thus allowing context specification at various levels of detail. The problem of the context resolution is then defined as the problem of selecting appropriate preferences based on context for personalizing a query. Research work in DIP3 focused on algorithms for context resolution to improve performance. To this end, two special data structures (namely the profile tree and the profile graph) were proposed to index preferences by exploiting the hierarchical nature of the context attributes. The evaluation of the approach has been extended with additional experiments and results. Evaluation was performed from two perspectives: usability and performance. Usability experiments were performed to evaluate the overheads imposed to users for specifying context dependent preferences, as well as their satisfaction from the quality of the results. Performance results have focused on context resolution using the proposed indexes. These results were reported in [SPV11].

Context may express conditions on situations external to the database or related to the data stored in the database. Models for expressing both types of preferences were advanced. Then, given a user query and its surrounding context, selecting related preferences is considered so that the query is appropriately personalized. More details can be found in the invited paper [PSV11].

These results of the DIP3 project were also presented in a tutorial on preferences presented in ICDE 2010 [KPS10]. In addition, a survey article was published with regards to the representation, composition and application of preferences in database systems including important privacy and performance issues [SKP11].

Research has also focused 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].

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].

DIP3 has also advanced research in the area of ***recommendations in databases***. Typically, users interact with database systems by formulating queries that follow a strict form. However, due the abundance of the available information, often, users may not have a clear understanding of their information needs or the exact content of the database, thus, they would like to formulate queries of an exploratory nature. Possible way of assisting users in database exploration have been proposed based on recommending to the users additional items that are highly related with the items in the result of their original query. For instance, when asking for movies directed by F.F Coppola, we guide exploration by recommending movies by other directors that have directed movies similar to those of F.F. Coppola, i.e., with similar characteristics, such as, genre or production year. The original query is also expanded with additional attributes, by finding correlations with other relations. For example, when asking for the title of a movie, its genre or other characteristics are also considered.

The computation of recommended results is based on the most interesting sets of (attribute, value) pairs, called faSets, that appear in the result of the original user query. The interestingness of a faSet expresses how unexpected it is to see this faSet in the result. The computation of interestingness is based on the frequency of the faSet both in the user query result and in the database instance. Since computing the frequencies of faSets in the database instance on-line has prohibitively high cost, statistics are maintained that allow us to estimate those frequencies when needed. More specifically, a novel approach has been proposed that is based on storing a δ-tolerance closed rare faSets representation as a summary of such frequencies and exploit these summaries to estimate the interestingness of the faSets that appear in the result of any given user query. A two-phase algorithm was presented for computing the top-k faSets. In the first phase, the algorithm uses the precomputed statistics to set a frequency threshold that is then used to run a frequent itemset based algorithm on the result of the query. We have evaluated the performance of our approach using both real and synthetic datasets. These results have been published in [DP11].

A prototype was also developed and demonstrated [KSDP10].

Pub/sub systems allow users to stay informed by proactively notifying them of item of potential interest. However, as the volume of data being created increases, some form of ranking of relevant events is needed to avoid overwhelming the users with large amounts of data. To address this problem, a new ranking criterion was proposed where the rank of an event is based on how often the corresponding subscription has been matched in the past. An event is considered novel, if it matches a subscription that has rarely been matched. Some initial results are reported in [SDSP10].

In the context of pub/sub systems, results were also attained towards ***reducing redundancy*** to improve performance and relevance. The growth of online services has created the need for duplicate elimination in high-volume streams of events. The sheer volume of data in applications such as pay-per-click click stream processing, RSS feed syndication and notification services in social sites such Twitter and Facebook makes traditional centralized solutions hard to scale. To this end, a suite of distributed Bloom filters have been introduced that exploit different ways of partitioning the event space. The contribution of this work lies in casting the design space regarding the distribution of Bloom filters and their use for duplicate-free event dissemination. In particular, two fundamental ways of partitioning the filters were proposed: a horizontal and a vertical one. The two partitioning methods were studied both theoretically and experimentally in terms of accuracy, efficiency, load balance and fault tolerance. Vertical partitioning achieves in practice better load balance and fault tolerance but induces higher communication costs. To address the continuous nature of event delivery, the filters are extended to support sliding window semantics. In this case, an event is considered a duplicate, only if it was previously delivered in the same time window. A nice property of this extension is that it allows us to improve

accuracy through a heuristic we call "equal timers". Moreover, locality-related tradeoffs have been studied and a tree-based architecture was proposed that allows for duplicate elimination across geographic locations [KSVP11].

Finally, 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 [GP11].

Lastly, a comparative study on the issue of privacy in online social networks was produced [PL11].

Final Results and Impact

The overall research in DIP3 has advanced 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 [SDSP10]. Furthermore, appropriate data structures were proposed for distributed redundancy elimination [KNPS11]
- To address the abundance of online-data, DIP3 has proposed novel ways to explore databases through recommendations. Database recommendations extend the results of database queries with additional related information highly correlated with these results [DP11, KSDP10].
- The representation, composition and application of preferences in databases were advanced through new data structures that result in improving performance of managing database preferences[KPS10, SKP11, PSV11, SPV11].
- Research has also focused on preferences in conjunction with keyword-based search in relational databases. To increase the quality of the search 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 [SDP10].
- New research results were attained in the context of database selection for XML document collections The focus of this research is on keyword queries with Lowest Common Ancestor (LCA) semantics for defining query results. To improve performance, both in terms of storage and processing efficiency, appropriate summaries of the LCA information based on Bloom filters were proposed [KP10].
- Finally, although there is a lot of research on privacy, privacy-preserving push-based delivery has not been explicitly addressed by previous research. DIP3 has derived a set of new personalized privacy models and mechanisms. [PL11, GP11, BSVPP10]
- Knowledge transfer to University level through offering a new graduate course on Data Privacy in Social Networks, adapting two undergraduate courses on Databases and Data Mining and supervising three MSc students, one PhD student and one undergraduate student. Knowledge transfer to European level through new research collaboration with University of Cyprus and University of Pittsburgh (joint supervision of graduate and PhD students), IBM (joint work on "Privacy in Publish/Subscribe Systems"), University of Bari and University of Patras (joint project under the territorial cooperation program).
- The research results of DIP3 facilitated the submission and approval of: (i) a proposal on developing a tool suite for customized enhancement of regional SMEs (INTERSOCIAL) and (ii) a proposal on developing a platform on Android for location aware dissemination of information with privacy and performance guarantees (Epirus on Androids).

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. DIP3 has provided 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. The research results of DIP3 will be applied in two R&D projects on developing platforms and tools for enhancing, promoting and reorganizing regional SMEs in the region of Epirus (Greece) and Puglia (Italy).

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 for publication.

[PL11] E. Pitoura and L. Liu, "Data Privacy in Online Social Networking", Technical Report, TR-2011-14, Computer Science Department, University of Ioannina

[GP11] A. Goulalas-Divanis and E. Pitoura, "Privacy in Publish/Subscribe Systems", Technical Report, TR-2011-11, Computer Science Department, University of Ioannina

[SDSP10] D. Souravlias, M. Drosou, K. Stefanidis and E. Pitoura: On novelty in publish/subscribe delivery. ICDE Workshops 2010: 20-22

[KPS10] G. Koutrika, E. Pitoura and K. Stefanidis: Representation, composition and application of preferences in databases. ICDE 2010: 1214-1215

[SKP11] K. Stefanidis, G. Koutrika, and E. Pitoura. 2011. A survey on representation, composition and application of preferences in database systems. ACM Trans. Database Syst. 36, 3, Article 19 (August 2011), 45 pages

[PSV11] E. Pitoura, K. Stefanidis and P. Vassiliadis: Contextual Database Preferences. IEEE Data Eng. Bull. 34(2): 19-26 (2011)

[SPV11] K. Stefanidis, E. Pitoura and P. Vassiliadis: Managing contextual preferences. Inf. Syst. 36(8): 1158-1180 (2011)

[DP11] M. Drosou and E. Pitoura, ReDrive: Result-Driven Database Exploration through Recommendations, in Proc. of the 20th ACM Conference on Information and Knowledge Management (CIKM 2011), October 24-28, 2011, Glasgow, Scotland, UK (to appear)

[KNPS11] G. Koloniari, N. Ntarmos, E. Pitoura and D. Souravlias. One is Enough: Distributed Filtering for Duplicate Elimination, in Proc. of the 20th ACM Conference on Information and Knowledge Management (CIKM 2011), October 24-28, 2011, Glasgow, Scotland, UK (to appear)