Information Processing / Management

Manolis Koubarakis *Timos Sellis*

Department of Informatics and Telecommunications
National and Kapodistrian University of Athens

Computer Science Division School of Electrical and Computer Engineering
National Technical University of Athens

koubarak@di.uoa.gr timos@dblab.ece.ntua.gr

TABLE OF CONTENTS

Towards RDF/S-based Data Management Systems

Vassilis Christophides
Department of Computer Science, University of Crete, Greece
Institute of Computer Science, Foundation for Research and Technology-Hellas, Greece
christop@ics.forth.gr

Introduction

RDF/S [1] represents nowadays the most popular language for exchanging resource descriptions on the Semantic Web (SW). In the RDF data model [23], a universe of

discourse to be modeled is a set of resources (essentially anything that can have a universal resource identifier, URI). Resources are then described through a set of properties (i.e. binary predicates) while descriptions are statements (i.e. *triples*) of the form subject-predicate-object: a *subject* denotes the described resource, a *predicate* denotes a resource property, and an *object* the corresponding property value. The predicate and object are in turn resources (or literal values) . A vocabulary (called namespace) of the properties but also of the classes employed to describe resources can be defined in the RDF Schema (RDFS [24]) language also represented in the core RDF model. An *RDF/S graph G* is *a set of RDF/S triples T.*
RDF/S graphs need to be efficiently stored into relational databases in order to be *queried*, *restructured* and *updated* via high-level declarative languages. RDF/S graphs need additionally to be *integrated* and *shared* by distributed peers/agents, but also to be *compared* for synchronizing independently evolved *versions*. Finally, as the size of the SW is expected to further grow in the coming years [12], *scalability* and *performance* of RDF/S-based Data Management Systems becomes increasingly important. These challenges are briefly presented in the rest of this paper.

Representing and Storing RDF/S Graphs in Databases

The RDF/S modeling primitives are reminiscent of knowledge representation languages. Compared to traditional object or relational database models, RDF/S blurs the distinction between schema and instances. RDFS schemas are *descriptive* (and not prescriptive, designed by DB experts), *interleaved with the instances* (i.e. crossing abstraction layers when a resource is related through a property with a class) while may be *large* (compared to the size of instances). In particular, unlike objects (or tuples) RDF/S resources *are not strongly typed.* The need for capturing RDF/S primitives using database models [21] and leverage relational technology becomes self-evident. It is also worth noticing that RDF/S has substantial differences from recent semi-structured and XML data models [26,27].
The most popular representations [16] for shredding RDF/S descriptions into relational tables are: the *schema-oblivious* (also called generic or vertical), the *schema-aware* (also called specific or binary) and a *hybrid* representation, combining features of the previous two. In the schema-oblivious, a single table is used for storing both RDF/S schemata and resource descriptions under the form of triples. In the schema-aware, each property (or class) defined in an RDF/S schema is represented by a separate table. In the hybrid, a table is created for RDF class and property instances with different range values (i.e., resource, string, integer, etc.). Several variations (e.g., with explicit or implicit database representation of subsumption relationships, use of resource URIs vs IDs, etc.) of these three core storage schemes have also been implemented in existing RDF stores. In terms of inferring triples from schema information there are two approaches: either to precompute them (at compile-time) or to compute them on demand (at run-time). The schema-oblivious, as well as, approaches using materialized views adopt the former approach, while schema-aware and hybrid adopt the latter. Computations can be performed either in main or secondary memory on demand. Their experimental evaluation [16,29] has proven that the most space/time efficient representation for voluminous RDF/S graphs (> 1G) was the hybrid

one supported by RDFSuite [28] based on a labeling schema for encoding entire subsumption relations [20] avoiding thus costly transitive closure computations on demand. Of the rest, schema-aware exhibits overall better performance [9] than schema-oblivious supported by the majority of existing RDF/S stores.

RDF/S Querying and Updating using Declarative Languages

Although voluminous RDF/S graphs have already appeared (e.g. in digital libraries and museums), expressive declarative languages for querying and updating both RDF descriptions and schemas are still in their infancy. We pay particular attention to the declarative nature of RDF/S manipulation languages for the old and good reasons: logical/physical data independence, efficient query/update optimization and transaction management.

RQL [26] is still the only declarative language [13] for querying both explicitly stated triples of RDF/S graphs and inferred ones by transitivity of subsumption and type relationships [25]. It is a typed, functional language with limited recursion which relies on a formal model for RDF/S graphs permitting the interpretation of instances by means of one or more schema vocabularies. *RQL* adapts the functionality of semistructured/ XML query languages to the peculiarities of the RDF/S data model but, foremost, it integrates smoothly RDF/S reasoning with querying (called *semantics-aware querying*). To this end, RQL provides sophisticated pattern matching facilities. Given an RDF/S graph G, an RQL query consists of one or several class and property patterns which are matched against schema and/or instance triples of G, and the variable bindings obtained from this matching are further processed (filtered, projected or grouped) to form the answer. The expressiveness of RQL is completed by *RVL* [19], a view definition language capable of creating not only virtual RDF instances, but also virtual RDFS schemas from (meta)classes, properties, as well as, related instances available on the SW. RVL exploits the functional nature and type system of RQL in order to navigate, filter and restructure complex RDF/S graphs. Unlike the tabular form of variable bindings returned by RQL, the output of an RVL view is a new RDF/S graph G', valid instance of the RDF/S data model. It is worth noticing that RQL/RVL queries/views can be arbitrary composed while their semantics is formalized using standard database machinery [18]. Semantics-aware RDF/S querying is not actually supported by the recently published SPARQL standard [5] while its formal foundations are still under investigation [14,3].

Last but not least, based on the RQL/RVL paradigms , RDF/S graphs can also be updated using declarative languages such as RUL [17]. RUL ensures that the execution of the Insert, Delete, Modify primitives on class and property instances violates neither the semantics of the RDF model (e.g. insert a property as an instance of a class) nor the semantics of the involved RDFS schemata (e.g. modify the subject of a property with a resource not classified under its domain class). This is the main design choice of RUL given that *type safety* for updates is even more important than for queries: the more errors we can catch at update specification time the less costly runtime checks (and possibly expensive rollbacks) we need. The rest of the design choices concern (a) the *granularity* of the supported update primitives at class or property triples; (b) the *semantics* of the update primitives in terms of their effects and side-effects (w.r.t. RDFS

reasoning and redundancy elimination); (c) the deterministic behavior of *set-oriented updates*; (d) the *smooth integration* with the pattern matching of the underlying RDF/S query language.

Peer-to-Peer Integration and Sharing of RDF/S Graphs

One of the main objectives of SW technology is to facilitate the integration of data sources spread worldwide. In their vast majority, data sources still rely on relational databases (RDB) published on the Web or corporate intranets as virtual XML. [10] advocates a first-order logic framework for mediating high-level queries to relational and/or XML sources using community ontologies expressed in RDF/S. Then, *GLAV mappings* can be employed to match any fragment (i.e. a view) of the RDF/S ontology to any view over the relational sources or any subtree of the XML documents. By capturing mappings as well as the semantics of RDF/S, XML and Relational data models as a set of constraints under the form of disjunctive embedded dependencies, we can devise a sound and complete algorithm for finding an optimal *maximally-contained rewriting* of the initial query (the one addressed to the mediated ontology) in terms of the local sources. Intuitively, such a rewriting, *approximates the best possible answer* of the original query, which can be obtained given the set of the available rules that map the local sources to the mediated ontology. The quality of such approximation is determined by the quality of both mapping assertions (i.e., if they are lossy or lossless) and the mediator's knowledge about the meaning of data in the local sources.
Such data exchanging techniques can be actually exploited by Peer Data Management Systems (PDMS) that scale to very large numbers of participating peers, while ensuring high autonomy and fault-tolerance. More precisely, in [6] routing algorithms and protocols have been proposed for RDF/S graph queries in a PDMS where peers advertise their local bases using fragments of community RDF/S schemes (i.e., views). This setting is particularly suited for supporting large-scale autonomous organizations (e.g. e-learning) for which neither a centralized warehouse nor an unlimited data migration from one peer to another are feasible solutions due to societal or technical restrictions. To this end, an original *encoding for peer schema fragments* is proposed allowing us to efficiently check whether a peer RVL view is subsumed by an RQL query . Then, an *RDF/S view lookup service* is built featuring both a *statefull* and a *stateless* execution over a DHT-based PDMS infrastructure. The conducted experiments have proven that the proposed lookup service scales gracefully for very large number of peers (up to 20000) while it requires less than half of the routing hops required by the original Chord protocol. These performance guarantees are crucial for building sophisticated *RDF/S query planning and execution services* which take into account the data distribution in peer bases [11].

RDF/S Change Management and Deltas

Ontologies (e.g. in e-science) may change for a variety of reasons, such as when the domain itself or our understanding of it changes, when applying modeling corrections, or expanding the domain representation [15]. These changes may cause serious

problems to its data instantiations, the applications and services that might be dependent on the ontology, as well as any ontologies that import that changed ontology. Hence, ontology change management is required to cope with ontology evolution in order to minimize its negative manners on the deployed applications from the beginning of their life-cycle. The most critical part of an ontology evolution algorithm is the determination of *what* can be changed and *how* each change should be implemented.

[2] devises a general-purpose algorithm for determining the *effects* and *side-effects* of a requested change, including *atomic* and *complex* update operations (for classes, properties and their instances). The algorithm is inspired by belief revision principles (i.e. validity, success and minimal change) and allows us to handle any change operation in a provably rational and consistent manner compared to existing approaches dealing with change operations on a per-case basis. The algorithm, after detecting all the invalidities that any given change operation could cause upon the updated ontology, determines the best way to overcome potential invalidity problems in an automatic way, by exploring the various alternatives for restoring validity and comparing them using a selection mechanism based on an ordering relation on potential side-effects. It is worth noticing that the proposed algorithm is highly parameterizable, both in terms of the employed ontology language (e.g. fragments of RDF/S such as [18]) and in terms of the ordering relation of update operators cost. Such techniques are useful in controlled environments (e.g. editing tools) where changes can be tracked on the fly.

However, ontologies and their instances in the SW usually change without any notification. Thus, to build advanced ontology *synchronization* and *versioning* services, appropriate snapshot differential relations (*Deltas*) need to be defined for core SW languages, such as RDF/S. By considering Deltas as sets of update operations, [7] investigates various ways to compare and transform two RDF/S graphs. The main challenges are related to the existence of inferred triples, as well as the semantics of the Delta update operations such as insert and delete (i.e. their effects and side-effects). Then, pairs of Delta relations and update semantics are identified to *correctly* transform one RDF/S graph to the other. Clearly, the storage requirements of the employed Deltas between consecutive RDF/S graph versions is crucial. *Small sized* Deltas yielding as less as possible update operations are quite beneficial in various backward and forward versioning policies. In extremis, Deltas should not report any change between two *semantically equivalent* RDF/S graphs (i.e. with the same set of explicit and inferred triples) or when executed, produce duplicates on *redundant-free* RDF/S graphs. Additionally, when we want to propagate changes across distributed RDF/S graph versions (i.e. synchronization), we also need Deltas that can be *reversed* and *composed* without materializing the involved intermediate RDF/S graph versions.

Benchmarking RDF/S Data Management Systems

As the size of the SW is expected to further grow in the coming years, scalability and performance of SW systems becomes increasingly important. Typically, such systems provide services for storing, querying, updating and reasoning over large volumes of RDF/S triples. Although in many related domains, such as databases and theorem proving, standard benchmarks exist and are ready to guide research on optimization

techniques, in the SW area there is not yet a commonly agreed benchmark covering the complexity of real RDF/S schemas and related instances.

[4] measures and analyzes the graph features of RDF/S schemas by focusing on the *power-law degree distributions* of the various class and property relationships. As a matter of fact, the majority of RDF/S schemas with a significant number of properties (resp. classes) approximate a power-law for total-degree (resp. number of subsumed classes) distribution. Moreover, the analysis of [4] revealed some emerging conceptual modeling practices of SW schema developers, namely: a) each schema has a few focal classes that have been analyzed in detail (i.e., having numerous properties and subclasses) which are further connected with focal classes defined in other schemas, b) the class subsumption hierarchies are mostly unbalanced (i.e., some branches are deep and heavy, while others are shallow and light), c) most properties have domain/range classes that are located highly at the class subsumption hierarchies and d) the number of recursive/multiple properties is significant. The knowledge of these features is essential for guiding synthetic RDF/S schema generation, as an important step towards benchmarking SW systems. An RDF/S schema generator which takes into account such morphological features is presented in [8]. In particular, using linear programming techniques, we can generate synthetically the two core components of an RDF/S schema, namely the *property* and the *subsumption* graphs, whose distributions follow a given power-law exponent with a confidence ranging between 90-98%. The conducted experiments demonstrate the scalability of the approach for generating typical schemas with 300 classes and 1000 properties in 10 secs and 16M memory.

REFERENCES

1. V. Christophides. "Resource Description Framework (RDF) Schema (RDFS)". Encyclopedia of Database Systems. Liu, Ling and Özsu, M. Tamer (eds). Springer-Verlag, 2008.
2. G. Konstantinidis, G. Flouris, G. Antoniou, V. Christophides, "A Formal Approach for RDF/S Ontology Evolution". In Proc. of the 18th Conference on Artificial Intelligence (ECAI'08), July 2008, Patra, Greece
3. M. Arenas, C. Gutierrez, J. Perez. "An Extension of SPARQL for RDFS". In Post Proc. of the Joint Workshop on Semantic Web, Ontologies, Databases (SWDB-ODBIS'07), Vienna, Austria, Springer-Verlag, 2008.
4. Y. Theoharis, Y. Tzitzikas, D. Kotzinos, V. Christophides, "On Graph Features of Semantic Web Schemas". IEEE Transactions On Knowledge And Data Engineering (TKDE), 20(5), May, 2008.
5. E. Prud'hommeaux, A. Seaborne. "SPARQL Query Language for RDF". W3C Recommendation 15 January 2008. Available at http://www.w3.org/TR/rdf-sparql-query
6. L. Sidirourgos, G. Kokkinidis, T. Dalamagas, V. Christophides, T. Sellis, "Indexing Views to Route Queries in a PDMS". Journal of Distributed and Parallel Databases (DPD), 23(1), pp. 45-68, February, 2008.
7. D. Zeginis, Y. Tzitzikas, V. Christophides, "On the Foundations of Computing Deltas between RDF Models". In Proc. of the 6th International Semantic Web Conference (ISWC'07) and the 2nd ASWC, Busan, Korea, 2007.
8. Y. Theoharis, G. Georgakopoulos, V. Christophides, "On the Synthetic

Generation of Semantic Web Schemas". In Post Proc. of the Joint Workshop on Semantic Web, Ontologies, Databases (SWDB-ODBIS'07), Vienna, Austria, Springer-Verlag, 2008.

9. D. J. Abadi, A. Marcus, S. R. Madden, and K. Hollenbach. "Scalable Semantic Web Data Management using Vertical Partitioning". In Proc. of the VLDB'07, Vienna, Austria, 2007.
10. I. Koffina, G. Serfiotis, V. Christophides, V. Tannen, "Mediating RDF/S Queries to Relational and XML Sources", International Journal on Semantic Web & Information Systems (IJSWIS), 2(4), pp 68-91, 2006.
11. G. Kokkinidis, L. Sidirourgos, V. Christophides, "Query Processing in RDF/S-based P2P Database Systems" at Semantic Web and Peer-to-Peer, S. Staab, H. Stuckenschmidt (eds.), Springer-Verlag, 2006.
12. L. Ding and T. Finin. "Gauging Ontologies and Schemas by Numbers". In Proc. of the 5th International Semantic Web Conference (ISWC'06), Athens, GA, USA, 2006.
13. T. Furche, B. Linse, F. Bry, D. Plexousakis, G. Gottlob. "RDF Querying: Language Constructs and Evaluation Methods Compared". Reasoning Web, Second International Summer School 2006, P. Barahona et al., (Eds.), LNCS 4126, pp. 1-52, Springer-Verlag, 2006.
14. J. Pérez, M. Arenas, C. Gutierrez. "Semantics and Complexity of SPARQL". In Proc. of the 5th International Semantic Web Conference (ISWC'06), Athens, GA, USA, 2006.
15. G. Flouris, Z. Huang, J.Z. Pan, D. Plexousakis, H. Wache. "Inconsistencies, Negations and Changes in Ontologies". In Proc. of the AAAI'06, Boston, Massachusetts, USA, 2006.
16. Y. Theoharis, V. Christophides, G. Karvounarakis, "Benchmarking Database Representations of RDF/S Stores", In Proc. of the Fourth International Semantic Web Conference (ISWC'05), Galway, Ireland, 2005.
17. M. Magiridou, S. Sahtouris, V. Christophides, M. Koubarakis, "RUL: A Declarative Update Language for RDF", In Proc. of the Fourth International Semantic Web Conference (ISWC'05), Galway, Ireland, 2005.
18. G. Serfiotis, I. Koffina, V. Christophides, V. Tannen, "Containment and Minimization of RDF/S Query Patterns", In Proc. of the Fourth International Semantic Web Conference (ISWC'05), Galway, Ireland, 2005.
19. A. Magkanaraki, V. Tannen, V. Christophides, D. Plexousakis, "Viewing the Semantic Web Through RVL Lenses". Journal on Web Semantics: Science, Services and Agents on the World Wide Web (JWS), Vol. 1(4), pp. 359-375, 2004.
20. V. Christophides, G. Karvounarakis, D. Plexousakis, Michel Scholl, S. Tourtounis, "Optimizing Taxonomic Semantic Web Queries Using Labeling Schemes". Journal of Web Semantics: Science, Services and Agents on the World Wide Web (JWS), Vol. 1(4), 2004.
21. C. Gutierrez, C. Hurtado, and A. Mendelzon. "Foundations of Semantic Web Databases". In Proc. of the 23 ACM Symposium on Principles of Database Systems (PODS), 2004.
22. N. Athanasis, V. Christophides, D. Kotzinos, "Generating On the Fly Queries for the Semantic Web: The ICS-FORTH Graphical RQL Interface (GRQL)", In Proc. of the Third International Semantic Web Conference (ISWC'04), November, 2004, Hiroshima, Japan.

23. G. Klyne, J. Carroll. "Resource Description Framework (RDF): Concepts and Abstract Syntax". W3C Recommendation 10 February 2004. Available at http://www.w3.org/TR/rdf-concepts
24. D. Brickley, R.V. Guha. "RDF Vocabulary Description Language 1.0: RDF Schema". W3C Recommendation 10 February 2004. Available at http://www.w3.org/TR/rdf-schema
25. P. Hayes. "RDF Semantics". W3C Recommendation 10 February 2004. Available at http://www.w3.org/TR/rdf-mt
26. G. Karvounarakis, A. Magkanaraki, S. Alexaki, V. Christophides, D. Plexousakis, M. Scholl, K. Tolle. "Querying the Semantic Web with RQL". Journal of Computer Networks and ISDN Systems, Vol. 42(5), pp 617-640, 2003.
27. J. Robie, L. M. Garshol, S. Newcomb, M. Biezinski, M. Fuchs, L. Miller, D. Brickley, V. Christophides, G. Karvounarakis. "The syntactic Web: Syntax and semantics on the Web". Journal of Markup Languages: Theory and Practice, 3(4), pp. 411-440, 2001.
28. S. Alexaki, V. Christophides, G. Karvounarakis, D. Plexousakis, K. Tolle. "The ICS-FORTH RDFSuite: Managing Voluminous RDF Description Bases". In Proc. of the 2nd International Workshop on the Semantic Web. In conjunction with WWW10, 2001, Hong Kong,
29. S. Alexaki, V. Christophides, G. Karvounarakis, D. Plexousakis, K. Tolle. "On Storing Voluminous RDF Descriptions: The case of Web Portal Catalogs". Proc. of the 4th International Workshop on the the Web and Databases (WebDB'01). In conjunction with ACM SIGMOD/PODS, 2001, Santa Barbara, CA.

WEB SEARCH

Alexandros Ntoulas
Microsoft Research
1065 La Avenida
Mountain View CA 94043, USA
antoulas@microsoft.com

Michalis Vazirgiannis
Dept. of Informatics, Athens Univ.
of Economics & Business
Patision 76, 10434, Athens
Greece
mvazirg@aueb.gr

INTRODUCTION

The Web today plays a central part in the cultural, educational and commercial life of millions of users. Due to the astonishing amount of information available on the Web, users typically rely on the Web search engines in order to locate relevant and useful information. A Web search engine's task is to find the most relevant content on the Web, given a user's query. To achieve this, different search engines may follow different approaches, but, in general, major search engines such as Google [14], Live Search [19] and Yahoo! [37] follow the architecture outlined in Figure 1.

The *crawlers* are programs that "browse" the Web by following links in a manner similar to the way human users visit different pages by following links. Their job is to download the pages from the Web and store them locally in a *page repository* database. From there, every page is processed and indexed. During the *analysis of its content* the search engine records various useful metadata for each Web page. Such metadata may include the set of outgoing links from a given Web page (which are also fed back to the crawler to download more content), the geo-location of every page, the graph structure around a given page etc. The metadata recorded for every page may also include information from external sources (e.g. the search engine query logs), such as the number of times a page has been clicked or viewed. The *indexer module* extracts all the words from each page and generates a data structure capable of answering very quickly which Web pages contain a set of given words. The *query processing module* is responsible for receiving a user's query and identifying the relevant pages by consulting the metadata, the index and the page repository. Finally, once all relevant results are identified, the *ranking module* sorts them so that the results that the user is most likely interested in appear on top, and presents them to the user.

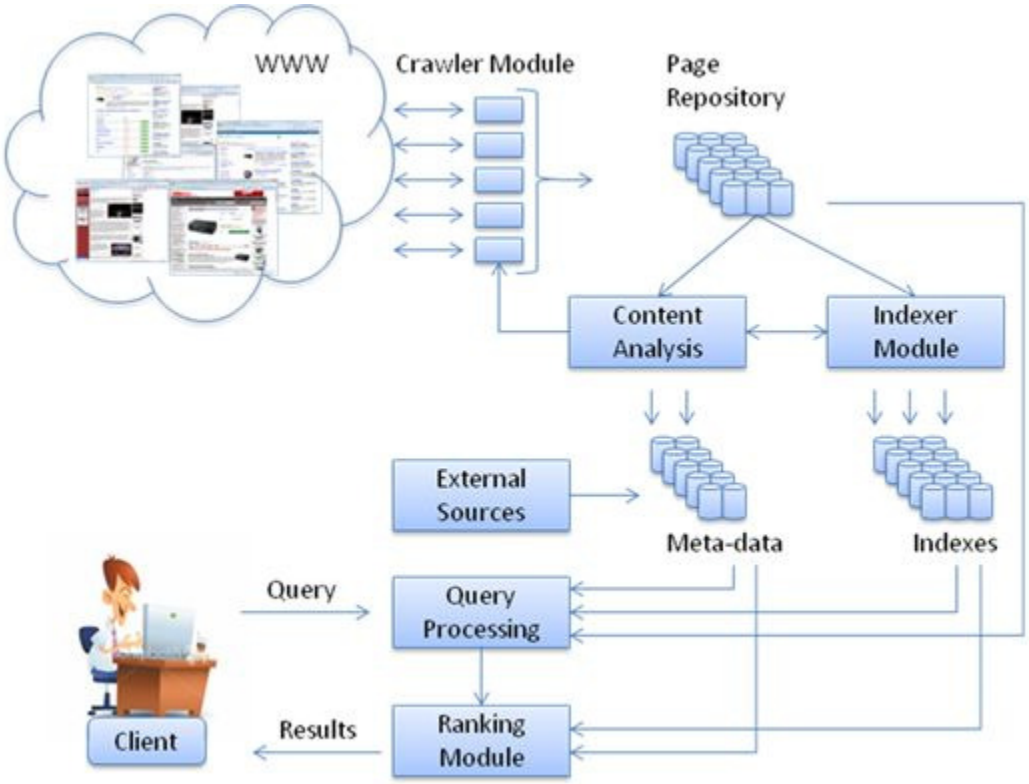

Figure 1. The architecture of a Web search engine

During this process, the most important pieces are the crawler, indexing and ranking modules. If the crawler cannot retrieve the right content from the Web, the search engine will not have an appropriate answer for some of the user's queries. In addition, since the Web is highly dynamic and evolving, the crawlers need to ensure that the search engine always has fresh information in order to serve the users in the best way. The indexing module needs to be capable of answering very quickly and efficiently

which pages contain a given set of keywords coming from the user's query. Given the size of the Web, this task can prove to be very costly and challenging. Finally, if the ranking module cannot surface the right pages to the top of the results the users will be dissatisfied and may turn away from the search engine.
In this chapter we present some of the challenges in the area of Web search and an overview of the research performed by Greek researchers to address these challenges. We start by presenting previous work on the crawling, indexing and ranking areas in the next three sections. We discuss approaches on distributed and peer-to-peer search in Section 5, and we present existing Web search systems in Section 6. We conclude in Section 7.

CRAWLING WEB PAGES

Typically, the crawler of a search engine operates as follows: given an initial set of URLs, the crawler downloads their content and stores it locally in the page repository for later processing. Next, the crawler proceeds iteratively (usually in a breadth-first manner) following all the outgoing links from the already downloaded set of URLs in order to retrieve more content. This process is repeated until all resources (in terms of, e.g., time, bandwidth or storage space) are exhausted.
Since the crawler operates within some resource constraints, it is of paramount importance to select pages to download that are of good quality and of interest to many users. If the crawler downloads pages that are not of good quality the search engine will be unable to return the expected results to the users. To this extend, search engines may choose to crawl a page based on its popularity (captured for example by the number of incoming links), based on the distance of the page from the Web site's root page, based on the expected topic of the page, etc.
A large and important part of the Web, however, is not directly accessible through links but rather is hidden within text databases with content covering a multitude of topics (see [16] for automatically detecting the topic of a text database). This content is oftentimes of great value to the users since it contains well-organized and well-edited information. Unfortunately, the crawler cannot find direct links to such pages (which are collectively called Hidden or Deep Web) because such links are only available after a user types a query in a search interface. Examples of Hidden Web sites are PubMed [28], or the US Patent and Trademark Office [34]. In [25] the authors present algorithms for generating queries in order to download the quality content from the Hidden Web. In certain cases, their algorithm was able to download more than 95% of a Hidden-Web site after issuing about 100 queries. For the cases where a Web site provides access to its content both through links and through a search interface, the authors in [15] present a rigorous framework of estimating the cost of each approach (link-based or query-based) and deciding which one to apply.
Once the crawler has downloaded a set of pages from the Web, it needs to periodically revisit them in order to detect changes and refresh the local repository. Coping with the dynamic nature of the Web is very critical to search engine crawlers. Unless they are able to capture the latest version of the data on the Web the search engines will present the users with obsolete results. In [25] the authors show that 50% of the pages on the Web do not change over the period of one year. Out of the remaining 50% that change

the authors found that previous changes are good predictors for future changes [25]. Given these results, the crawlers should mainly focus on refreshing the previously changed pages in order to make good use of their download resources.
In [11] the authors present a sampling-based technique for downloading as many changed pages as possible from the Web. The main idea is to sample a few pages from every Web site and then download more pages based on how many changes where detected in each sample. In a similar line of work, in [17] the authors use survival analysis techniques to model the changes in text databases (often supporting Hidden-Web sites) and determine which ones to update. In [3], the authors compute a set of features that characterize behaviour of a set of Web sites based on their national domain (including the Greek domain). Such features can prove very useful in refreshing Web pages.

INDEXING

Once the crawler downloads and creates a local copy of a set of pages from the Web, the search engine then proceeds to analyze and index these pages. During this phase, all the links, words and HTML tags are extracted from every page and a set of indexes is generated. Such indexes may contain, for example, the link structure around a given page which is useful during ranking as we will discuss in Section 4, the set of pages within a given domain, etc. The most important index that a search engine builds during this phase is the one that helps identify which pages contain the user query words. Given the enormous size of the Web, the number of queries served by a search engine every day and the user expectations (users expect search results in less than 1 second), this textual index has to be highly scalable and highly efficient.
The Web search engines, typically utilize a data structure called *inverted index* in order to locate the documents relevant to a user's query very quickly. If we consider a collection of pages
D = {D1, ..., DM} that the crawler has downloaded from the Web and the set of all words
T = {t1, ..., tn} in this collection, the search engine maintains a list I(ti) of page IDs that contain ti. Every entry in I(ti) is called a posting and can be extended to include additional information, such as how many times ti appears in a Web page, the positions of ti in the Web pages, whether ti appears in bold/italic etc. The set of all the lists I = {I(t1), ..., I(tn)} is our inverted index. Whenever a query comes in, each query word is matched against the corresponding inverted list in order to identify the set of relevant pages.
Search engines are accepting millions of queries every day from eager users searching for information. In order to cope with this huge query load, one way is to replicate fully the inverted index across a large cluster of machines, but given the size of the Web, this approach can prove very costly. To alleviate this problem, search engines sometimes use a small cache where they put frequently accessed pages. The idea is that the cache contains the desired result most of the time, and in this way the fully replicated indexes are accessed less. In this approach however, it is crucial for the search engine to select

the right pages to put in the cache.
In [4] the authors explore the tradeoffs between static and dynamic caching as well as whether the search engines should cache based on the query results or the inverted lists of the index. They showed that dynamic caching, in general, has limited effectiveness and they propose new static caching algorithms that demonstrate very good performance. In addition, they showed that caching query results is more preferable in cases where the network communication time between machines is big and they study how the effectiveness of caching changes over time given that the observed query load evolves.
In [24], the authors present a replication and caching scheme appropriate for inverted indexes that has the nice property that it can guarantee the results of the scheme to be identical to a scheme that only uses replication, but this is done by using about 50% less space. The paper discusses methods for selecting which postings to put in each tier, one by selecting inverted lists and one by selecting individual postings. It also discusses how we can determine the optimal size of the cache in order to minimize the storage space required without sacrificing performance in terms of speed or result relevance.
For the cases where computing the complete answer from an inverted index is either prohibitively expensive or not required, [2] presents a method for sampling the search results and compute a summary from an inverted index. This method can be useful in identifying the representative topics within the search results without paying the price to calculate the full answer, for providing feedback to the user of what to expect before the full answer is computed, or even for estimating the total number of search results.

RANKING OF RESULTS

Once the search engine determines the relevant pages it needs to present them to the user. During this process the results are ranked in terms of relevance to the user's query so that the most relevant appears on top, followed by the second most relevant and so on. In order to determine the relevance of a page to the user's query the search engine, at a high level, depends mainly on two factors: a) how close the content of the page is to the query (textual relevance); b) how "popular" or "important" the page is overall on the Web (page relevance). For the textual relevance, the most widely used method is the Okapi BM25 [30].
The overall "popularity" or "importance" of Web page (page relevance), is usually captured by link-based ranking algorithms, such as PageRank [28] which is currently implemented within Google. The main idea of PageRank is to consider that a page is important if it is linked by other important pages. Since the first appearance of PageRank, there have been several variations of link-based ranking algorithms which have shown their own merits and shortcomings in different settings. In [6, 9, 33] the authors present a theoretical framework that can be applied in the studying and analysis of link-based ranking algorithms. The authors discuss the properties and differences of a number of link-based ranking algorithms and they present comparative studies of their performance. In [1], the authors present extensions to another popular link-based ranking algorithm called HITS [18] by using multiple eigenvectors instead of a single one.

Computing the link-based ranking for a set of pages typically involves extracting the graph structure from these pages and then performing iterative computations over this graph. Given the enormous size of the Web and its dynamic nature, such computations may become prohibitely expensive to be done on a regular basis. To this extend, the authors in [37] study the rank changes within the Web graph. In [36] the authors present a method for predicting the ranking of pages in the future based on their ranking in the past. The method is based on Markov model learning and was shown capable of predicting the ranks with accuracy of 90% on real-Web datasets. The method consists of the following phases: (a) computing the ranking trends from the rank evolution of individual pages (b) computing of the states of a Markov model based on an equi-probable partitioning and (c) using the resulting Markov models for predicting the next ranking of a page whose ranking history is known at the given point in time. The authors of [36] are also working on alternative prediction methods such as regression and spectral pre-processing techniques.

Given the variety of interests and needs that the users around the world have while performing Web searches, several search engines have tried to personalize their search results to the individual users. Personalization normally involves a phase where the profile of a user is first identified before it can be applied during the ranking phase in order to boost results that are likely to be closer to the user's interests. In [31] the authors consider clicked pages to be a good representative of a user's profile and they use a topical taxonomy to categorize the incoming queries and pages and personalize the search results. In [13] the authors discuss an approach of performing personalization of the results through categorization on the client side, i.e. after the search engine has computed the results.

Finally, given the high potential monetary value of the search traffic, some Web site operators pollute the Web with spam pages (i.e. pages meaningless to the users but existing only to trick search engines), in the hope of improving their ranking. In [26], the authors present a number of features that can capture whether a Web page is spam or not. Features such as number of words in the title, fractions of links in the page, and redundancy of the page's content show high promise and when combined they can help a search engine detect about 87% of the spam with 91% accuracy. In [10] the authors use the query logs to identify potential spam pages as well as query words that are very likely targeted by spammers. Their method yields an accuracy of about 78%. The work in [5] discusses spam removal and ranking for the blogosphere.

PEER-TO-PEER (P2P) APPROACHES TO WEB SEARCH

In our discussion so far, we have assumed that the Web search experience of the users is fully controlled by a search engine which builds a centralized index where the information is stored, processed and retrieved at a user's whim. In this centralized approach the search engines can monopolize the information flow, while the content owners on the Web have no control over how their information is processed, stored, ranked and presented to the users. In addition, although the Web content is already

stored in an enormous number of servers across the Web, it has to be downloaded by the search engine crawlers before the users can actually access it, thus introducing issues of freshness and coverage of the returned results. In order to take advantage of these observations and the need of the content providers to have more control over their content, researchers have worked on creating peer-to-peer (P2P) search systems.
In a P2P search engine each peer contains parts of the overall content available in the system. Once a user issues a query to her local machine, the query is routed to the appropriate peers, the results are collected and they are presented to the users. There are of course several challenges in P2P Web searching, ranging from how each peer organizes its local index to ensuring that all the information is available at the network at any one time. One of the most important challenges though, is the problem of *query routing*. Given the size of information and the potentially enormous number of peers available, the system cannot possibly contact each and every peer and retrieve a response within a reasonable amount of time, but instead it needs to identify which peers are most likely to contain relevant information and *route* the user's query only to those peers.
A good overview of the different approaches in organizing a peer-to-peer Web search engine is presented in [20]. The authors investigate a number of design challenges, mostly regarding partitioning schemes for the index and data storage technologies. They report on the pros and cons and the cost/performance ratios of each presented approach. Overall, a design strategy (named Anakin) that employs term partitioning for the index and a combined storage of hard disk and flash-ram is described as the most promising.
A system for organizing Web content in a distributed and decentralized way is presented in [12]. The main contribution of this work is the creation of distributed semantic overlay networks that are used for query routing in a semi structured P2P Web content architecture. In order to achieve the desired performance and scalability, Semantic Overlay Networks (SONs) connecting peers storing semantically related information are employed. The lack of global content/topology knowledge in a P2P system is the key challenge in forming SONs. The authors in [12] describe an unsupervised approach for decentralized and distributed generation of SONs (DESENT). Through simulations and analytical cost modeling they verify the claims regarding performance, scalability, and quality.
In [6] the authors discuss two methods of improving query routing in a P2P Web search setting. The first one is based on single-keyword statistics collected in each peer, while the second one is based on multi-keyword statistics. These statistics are represented in terms of hash-sketch synopses and they are used to build routing indices that help a peer determine which of the peers are likely to be the best ones to contact. In particular, the work in [6] aims at exploiting potential correlations among query keywords in order to improve the overall query routing within the Minerva system [7].

EXISTING WEB SEARCH SYSTEMS

In this section we briefly discuss some of the existing Web search systems that are based on the work presented in the previous sections.

Blogscope [5] is a system that facilitates the online analysis of contents from the blogosphere. Blogscope currently tracks and indexes some 10 million blogs and is capable of handling several thousands of updates in its collection every day. Blogscope allows for spatio-temporal analysis of blogs, detection of information busts, identification of correlated keywords and a ranking function applied to blogs. The system is currently available at http://www.blogscope.net.
Infocious [23] is a fully-functional Web search engine that is currently indexing more than 2.5 billion Web pages. It implements variations of the technologies presented in [11,20,24,27] and it applies Natural Language Analysis to the downloaded pages in order to understand their content in a better way. It allows the users to perform advanced searches (such as specifying whether a query keyword should be a verb or a noun), and it presents the results topically categorized in the Dmoz hierarchy (similar to [30]). This enables the user to drill-down in a category of interest to retrieve more topically-focused results and it provides the basis for the personalization of results. Infocious is publicly available at http://search.infocious.com.
Minerva [6,20] is a distributed search engine capable of handling a large set of autonomous peers. Within Minerva, each peer maintains a local collection of pages which can be either independently crawled from the Web or imported from external sources. Each peer maintains a local inverted index in order to answer queries quickly and efficiently, with the words stored within the index being stemmed for increased recall. The sets of terms are partitioned among peers with each peer maintaining statistics for every word that it stores. Query routing is performed by exploiting keyword and attribute-value correlations similar to [6]. The project is currently available at http://www.minerva-project.org.
A system called **THESUS**, that supports retrieval of Web content relevant to an ontology is presented in [35]. This line of work puts emphasis on the use of a page's incoming links as means of classification. The authors present the tools needed in order to manage the links, and their semantics. They further process these links using a hierarchy of concepts, akin to an ontology, and a thesaurus. The work results in prototype system, called THESUS that organizes thematic Web documents into semantic clusters. The main contributions of this effort are: (a) a model and language to exploit the information within link semantics (b) the THESUS prototype system, (c) its innovative aspects and algorithms, and in particular the novel similarity measure between Web documents that can be applied to different clustering schemes (DB-Scan and COBWEB).

CLOSURE AND REMARKS

Web Search is a cornerstone of today's industry and economy. Millions of users perform searches on the Web in an attempt to locate useful information for their everyday lives. In the last few years, Web search is starting to bear a strong social aspect as well, especially for collaborative tasks such as booking a trip or arranging for vacation. Greek scientists are remarkably active and successful in this competitive field of research covering the aspects of this process ranging from the ranking of results to P2P architectures. Here, we have presented an overview of the challenges in Web search

and the work done by Greek researchers. Our goal is to give a high-level summary of the past and ongoing work in this area and not provide an exhaustive list. It is our hope that the covered material will stimulate further intriguing research and fruitful discussions so that work in this area will continue with more exciting and ground-breaking results.

REFERENCES

1. D. Achlioptas, A. Fiat, A. Karlin, F. McSherry, *Web search through hub synthesis*, In Proceedings of the 42nd Foundation of Computer Science (FOCS). Las Vegas, NV, 2001.
2. A. Anagnostopoulos, A. Z. Broder, D. Carmel, *Sampling Search-Engine Results*, World Wide Web Journal, Volume 9, Number 4, pp. 397-429, 2006.
3. R. Baeza-Yates, C. Castillo, E. Efthimiadis, *Characterization of National Web Domains. ACM Transactions on Internet Technology*, 7(2), Article 9, 2007.
4. R. Baeza-Yates, A. Gionis, F. Junqueira, V. Murdock, V. Plachouras; F. Silvestri, *The impact of caching on search engines*, In Proceedings of the ACM International Information Retrieval (SIGIR) Conference, 2007, Amsterdam, Netherlands.
5. N. Bansal, N. Koudas, *Searching the Blogosphere,* In Proceedings of WebDB, 2007.
6. M. Bender, S. Michel, Nikos Ntarmos, P. Triantafillou, G. Weikum, C. Zimmer, *Discovering and Exploiting Keyword and Attribute-Value Co-occurrences to Improve P2P routing Indices*, ACM International Conference on Information and Knowledge Management (CIKM), 2006.
7. M. Bender, S. Michel, P. Triantafillou, G. Weikum, C. Zimmer, *MINERVA: Collaborative P2P Web Search*, International Conference on Very Large Data Bases (VLDB) (demo paper), September 2005.
8. A. Borodin, J. S. Rosenthal, G. O. Roberts, P. Tsaparas, *Finding Authorities and Hubs From Link Structures on the World Wide Web*, World Wide Web Conference (WWW), Hong Kong, 2001.
9. A. Borodin, J. S. Rosenthal, G. O. Roberts, P. Tsaparas, *Link Analysis Ranking: Algorithms, Theory and Experimets*, ACM Transactions on Internet Technologies (TOIT), Vol 5, No 1, February 2005.
10. C. Castillo, C. Corsi, D. Donato, P. Ferragina, A. Gionis, *Query-log mining for detecting spam*, Fourth International Workshop on Adversarial Information Retrieval on the Web, 2008.
11. J. Cho, A. Ntoulas, *Effective Change Detection using Sampling*, In Proceedings of the International Conference on Very Large Databases (VLDB), 2002, Hong Kong, China.
12. C. Doulkeridis, K. Norvag, M. Vazirgiannis, *DESENT: Decentralized and distributed semantic overlay generation in P2P networks*, In Special Issue on Peer-to-Peer Communications and Applications, IEEE Journal on Selected Areas in Communications (J-SAC), Vol. 25, Issue 1, pages 25-34, January 2007
13. J. Garofalakis, T. Matsoukas, Y. Panagis, E. Sakkopoulos, A. Tsakalidis, *Personalization Techniques for Web Search Results Categorization*, EEE 2005:

148-151
14. Google Inc. http://www.google.com
15. P. Ipeirotis, E. Agichtein, P. Jain, and L. Gravano. *To Search or to Crawl? Towards a Query Optimizer for Text-Centric Tasks*, In Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data (SIGMOD), 2006.
16. P. Ipeirotis, L. Gravano, *Distributed Search over the Hidden-Web: Hierarchical Database Sampling and Selection*, In Proceedings of the 28th Intenational Conference on Very Large Databases (VLDB), 2002
17. P. Ipeirotis, A. Ntoulas, J. Cho, L. Gravano, *Modeling and Managing Content Changes in Text Databases*, ACM Transactions on Database Systems (TODS), vol. 32, no. 3, September 2007.
18. J. Kleinberg, *Authoritative sources in a hyperlinked environment*, In Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, 1998.
19. Live Search. http://www.live.com
20. S. Michel, M. Bender, P. Triantafillou, G. Weikum, *Design Alternatives for Large-Scale Web Search: Alexander was Great, Aeneas a Pioneer, and Anakin has the Force*, In Workshop on Large scale Distributed Systems for Information Retrieval, Collocated with SIGIR , July 2007.
21. S. Michel, P. Triantafillou, G. Weikum, *MINERVA∞: A Scalable Efficient Peer-to-Peer Search Engine*, ACM/IFIP/USENIX 6th International Middleware Conference, November 2005.
22. A. Ntoulas, *Crawling and Searching the Hidden Web*, Ph.D. thesis, University of California Los Angeles, 2006.
23. A. Ntoulas, G. Chao, J. Cho, *The Infocious Web Search Engine: Improving Web Searching through Linguistic Analysis*, Journal of Digital Information Management (JDIM) vol. 5, no 5, October 2007.
24. A. Ntoulas, J. Cho, *Pruning Policies for Two-Tiered Inverted Index with Correctness Guarantee*,In Proceedings of the ACM International Information Retrieval (SIGIR) Conference, 2007, Amsterdam, Netherlands.
25. A. Ntoulas, J. Cho, C. Olston, *What's New on the Web? The Evolution of the Web from a Search Engine Perspective*, In Proceedings of the World Wide Web (WWW) Conference, 2004, New York, USA.
26. A. Ntoulas, M. Najork, M. Manasse, D. Fetterly, *Detecting Spam Web Pages through Content Analysis*, In Proceedings of the World Wide Web (WWW) Conference, 2006, Edinburgh, Scotland.
27. A. Ntoulas, P. Zerfos, J. Cho. *Downloading Textual Hidden Web Content through Keyword Queries*, In Proceedings of the Joint Conference on Digital Libraries (JCDL), 2005, Denver, USA.
28. L. Page, S. Brin, R. Motwani, T. Winograd, *The PageRank citation ranking: Bringing order to the Web*, Technical report Stanford Digital Library Technologies Project, 1998.
29. PubMed Medical Library. http://www.pubmed.org
30. K. Sparck Jones, S. Walker, S.E. Robertson, *A probabilistic model of information retrieval: development and comparative experiments*, Information Processing and Management 36, 2000.
31. S. Stamou, A. Ntoulas, *Search Personalization through Query and Page Topical Analysis*, Journal of User Modeling and User-Adapted Interaction (UMAI), 2008.
32. S. Stamou, A. Ntoulas, V. Krikos, P. Kokosis, D. Christodoulakis. *Classifying Web*

Data in Directory Structures, In Proceedings of the 8th Asia Pacific Web Conference (APWeb), Harbin, China, January, 2006.
33. P. Tsaparas, *Link analysis ranking*, Ph.D. thesis, University of Toronto, 2004.
34. United States Patent and Trademark Office. http://www.uspto.gov
35. I. Varlamis, M. Vazirgiannis, M. Halkidi, B. Nguyen, *THESUS: Effective Thematic Selection And Organization Of Web Document Collections Based On Link Semantics*, in IEEE Transactions on Knowledge and Data Engineering Journal, June 2004 (Vol. 16, No. 6), pp. 585-600.
36. M. Vazirgiannis, D. Drosos, P. Senellart, A. Vlachou, *Web Page Rank Prediction with Markov Models*, poster, WWW 2008, Beijing, China, April 2008
37. A. Vlachou, M. Vazirgiannis, K. Berberich, *Representing and quantifying rank - change for the Web Graph*, Fourth Workshop on Algorithms and Models for the Web-Graph (WAW2006), Banff, Canada, November 2006
38. Yahoo! Inc. http://search.yahoo.com

See statistics of SW search engines as http://watson.kmi.open.ac.uk and http://swoogle.umbc.edu

Note that the subject and object of a statement can also be non-universally identified objects, called unnamed resources (or *blank nodes*), whose URI could not be shared across different RDF descriptions.

as well as set operations, aggregate functions and nested queries.

And also exploit intuitive graphical interfaces for constructing on the fly queries/views [22].

As a matter of fact we consider an RQL core fragment [18] including also range restricted schema variables as appearing in RDF/S graph patterns.

WEB MINING

Magdalini Eirinaki
Computer Engineering Department
San Jose State University
magdalini.eirinaki@sjsu.edu
Michalis Vazirgiannis
Department of Informatics
Athens University of Economics and Business
mvazirg@aueb.gr

1. INTRODUCTION

The Web has grown rapidly during the past decade, becoming the largest and most popular communication and information dissemination platform that is publicly available. The Web presents some unique characteristics, such as, the size and diversity of the data and information available, the heterogeneity of the data formats available, the hyperlinked environment, as well as its dynamic nature. Furthermore, the past few years, Web has become a virtual society. It is no longer a mere collection of data, information and services, but also a platform where people can interact and communicate through social networking sites, blogs, and review sites [Liu07].

Web mining aims at discovering useful information from the data available on the Web. Although many of the web mining approaches are based on traditional data mining algorithms, the unique characteristics of the Web have lead researchers invent new algorithms, or new ways to integrate and combine the information available with existing techniques. Web mining is often categorized in three types, based on the type of data input that is being used. Those are Web content mining, Web structure mining and Web usage mining.

Web content mining refers to the process of extracting information from the Web page contents. The Web pages can be clustered or classified into relevant topics, processed in order to extract descriptive summaries. Web search engines broadly use such techniques in order to organize content before presenting it to the users. Community-based content can be mined in order to identify users' trends or sentiments on specific events or products, a process that can be used in marketing research for example. Web structure mining discovers useful knowledge from the hyperlinks between pages. This analysis enables the identification of authoritative, important pages. Such information has formed the basis of most web search engines. Web structure mining has also been used to identify connections between communities of users. Finally, web usage mining discovers knowledge from the usage of the Web, in other words, the navigations of the users across Web pages. This information is stored in Web usage logs, and is being used to discover correlations between users or products/web pages. This information can then be used to re-organize a Web site, improve Web page pre-fetching, or, more commonly personalize a Web site by generating recommendations (of pages or products) to the active users.

Many Web mining approaches often combine usage with content, and/or structure mining techniques, in order to enhance the results of the knowledge discovery process. In this Section we will not refer to works related content and structure mining techniques for crawling, characterizing, clustering/classifying or ranking Web documents, or any other technique related to Web search engines, since this topic is covered in another section of this book. The same holds for efforts related to mining on the semantic Web, since this is another broad topic covered separately. We start by discussing pure Web usage mining efforts, aiming at modeling the navigation behavior of users. We then present research efforts in the area of Web personalization and recommender systems. More specifically, we focus on works that employed a collaborative filtering approach (a very common approach for web recommender systems), and those that used other data and web mining techniques to implement a recommender system. We also present some application-oriented approaches related to personalization. We then discuss some research efforts related to web directories. Finally, we present research related to the societal aspect of the Web, involving web communities and blogs. Our overview includes research outcomes presented in the

period 2000 – 2008. We refer to related work per research group and organize the presentation in chronological order per section (earlier works are presented first).

2. WEB USAGE MINING

In this section we cover research work related to web usage mining, also referred to as web log mining. Those approaches involve extensions of traditional data mining algorithms, or proposal of new ones, both types used to discover navigational patterns from the web access logs. Most of the approaches aim at modeling the users' behavior in order to assist web site administrators in redesigning a web site, or managers as a decision support tool. There exist, however, some approaches aiming at web cache enhancement. We present those approaches organized by research group.

2.1 WEB MINING FOR WEB USAGE ANALYSIS

Nanopoulos and Manolopoulos [NM00] (from Delab – Aristotle University, Greece - http://delab.csd.auth.gr/) addressed the problem of discovering frequently occurring access sequences from web logs. The problem is formulated as follows: if a sequence of web documents appears frequently enough, then this indicates a pattern that can be useful for re-designing the site or increase the system's performance. They proposed an Apriori-like algorithm for finding traversal patterns. Contrary to previous approaches, that discovered patterns not corresponding to the site's structure, or noisy patterns, their approach does not pose the constraint that accesses should be consecutive inside the patterns during the frequency counting procedure. Their proposed algorithm, although based on the general structure of the Apriori algorithm, is based on different data structures and procedures for support counting and candidate generation, and also takes into consideration the site structure.
In a subsequent work, they also addressed the problem of using SQL to query web log data stored in relational databases. Such SQL queries can become very complicated, since multiple nested joins or subqueries are required. This problem arises from the fact the SQL does not contain a sequence search statement. Nanopoulos et. al. [NZ+02] propose an indexing method for the storage and querying of large web log files, by encoding the sequences using signatures. They show that this approach improves the scalability to very large web logs, contrary to previously proposed indexing methods.
M. Spiliopoulou and her group (KMD Group – University of Magdeburg, Germany - http://omen.cs.uni-magdeburg.de/itikmd/Home.33.1.html) proposed a complete environment for web usage analysis, used to assess the quality of a web site integrating information from multiple underlying database servers or archives. The proposed environment is largely based on the Web Utilization Miner (WUM) [BS00]. The authors have investigated the impact of site structure on the quality of constructed sessions. They examined session-izing using several session identification methods that exploit session duration, page stay time and page linkage [BM+02]. The proposed framework also provides the means for mining frequent navigation sequences, by aggregating the usage data into conceptual hierarchies [PS02], and providing a query language to mine them. They also proposed a principle of pattern modeling and presentation, in order to assist the users in interpreting the results. In the same context, the researchers have

investigated the effects of incorporating background knowledge, expressing the preferences and properties of the population under observation, in the data mining process, by applying the proposed framework on web marketing applications [SP02]. Compared to similar web usage mining frameworks proposed in the past, WUM navigation patterns allow for more flexibility, addressing form-based (non-static) web sites generated from multiple database servers, and exploiting the concepts of the pages' topology, content, and linkage in assessing the quality of a web site.

In this context there is an active cooperation among the groups:

- D. Pierrakos, G. Paliouras, V.Karkaletsis, C. Spyropoulos, Institute of Informatics and Telecommunications, NCSR Demokritos - http://www.iit.demokritos.gr/
- C. Papatheodorou, Department of Archive and Library Sciences, Ionian University
- M. Dikaiakos, Dept. Of Computer Science, University of Cyprus

In this work, the researchers used clustering and usage mining methods in order to construct communities including the users of large Web sites. Their uppermost objective is to use the extracted knowledge in both commercial sites, for promotions and personalization, and non-commercial sites in order to assist the improvement of the site. In the first part of their work [PP+00], they evaluate three clustering methods on usage data, namely Autoclass and Self Organizing Maps, which are two well known machine learning methods, and their own Cluster Mining method. This work has been extended in [PP+04], where the authors introduce the so-called Web Community Directories. In this context, they propose a cluster mining algorithm called Community Directory Miner (CDM), which is an extension of the Cluster Mining method. They build a taxonomy of Web pages included in the Web log files using a hierarchical agglomerative approach for document clustering. Thus, they automatically create a hierarchical classification of Web documents using a thematic categorization closely related to the preferences of the users. CDM is then applied, in order to discover topic trees, representing the community models. This research has been partially funded by the Greece-Cyprus Research Cooperation project «Web-C-Mine» [WCM].

Christodoulou et. al. [CDS06], from the *KDBS Lab – National Technical University of Athens* - *http://web.dbnet.ntua.gr/en/home.html*, addressed the problem of mining navigational patterns in a special type of web sites that of portal catalogs. Those web sites organize data in topic hierarchies, and provide querying and browsing capabilities to the users. The authors implemented several statistical analysis and usage mining techniques, including clustering, in the prototype system NaviMoz. This system the web site administrators to discover the navigational patterns of the users browsing those portal sites. They can subsequently use the discovered patterns in order to analyze the users' behavior, extract their preferences, and reorganize the structure of the portal accordingly.

2.2 WEB MINING FOR WEB CACHE

ENHANCEMENT

Nanopoulos et. al. [NKM01], from Delab – Aristotle University, Greece - http://delab.csd.auth.gr/, propose a Web caching scheme based on page pre-fetching. They employ Web log mining methods to effectively predict the forthcoming page accesses of a client and pre-fetch them. Their scheme stores the pre-fetched documents in a dedicated part of the cache, to avoid the drawback of incorrect replacement of requested documents. Compared to existing schemes for buffer management in database and operating systems, their approach takes into consideration the special requirements of the Web. Their scheme is shown to improve the cache performance.

The same problem has been addressed by Georgakis and Li [GL06], from Digital Media Laboratory (DML), Umed University, Sweden. In this work, the researchers proposed a speculative algorithm for web page prefetching. The algorithm relies the browsing habits of the user. They build a profile incorporates the frequency of occurrence for selected elements forming the web pages visited by the user. These frequencies are employed in a mechanism for the prediction of the user's future actions. In order to predict an adjacent action, the anchored text around each of the outbound links is used and weights are assigned to these links. Some of the linked documents are then pre-fetched and stored in a local cache according to the assigned weights. The proposed algorithm was tested against three different prefetching algorithms and was shown to increase performance in terms of cache-hits and achieved recall and precision.

Finally, the problem of short-term prefetching on a Web cache environment has been addressed by Pallis et. al. [PVP07], from PLaSE Lab – Aristotle University of Thessaloniki - http://plase.csd.auth.gr/. This works differs from the aforementioned approaches in that it employs a different data mining approach. In particular, the authors propose a clustering algorithm, ClustWeb, for clustering inter-site Web pages. They also provide the means to validate and visualize the resulting clusters.

2. WEB PERSONALIZATION

Web personalization is the process of customizing the content that is provided to a user based on their past navigational behavior and their explicit interests or ratings. The premise of web personalization systems is that, if any two users are similar, in terms of the pages they visit or the items they rate/purchase, then the content that is interesting for the first user, might be interesting for the second as well. The output of the personalization process can be anything related to presenting different content to different people, for example web page or item recommendations, or targeted advertisements and emails. The main input of a web personalization process is the data describing the users' interaction with the web site, in other words the web usage data. This information, however, is often combined with other sources such as, the content

and/or structure of the web site, or explicit user profiles including demographic information and user ratings. Most of the research work done in the area of web mining falls under this broad topic. In what follows, we overview the work of researchers in terms of input data and algorithms used, as well as some systems built to address specific application needs. We start our overview with two important surveys done in this area by Greek researchers.

2.1 SURVEYS ON WEB PERSONALIZATION

Two important surveys on Web usage mining and its application on Web personalization have appeared in the literature at approximately the same time by Eirinaki and Vazirgiannis [EV03], from DB-NET, Athens University of Economics and Business – http://www.db-net.aueb.gr,and Pierrakos et. al. [PP+03] – joint effort among Institute of Informatics and Telecommunications, NCSR Demokritos (http://www.iit.demokritos.gr/) & the Dept. of Archive and Library Services, Ionian University – (http://www.ionio.gr/tab/eng/index.html) . It is noteworthy that both surveys have received significant interest from researchers in this area, and have been cited in numerous papers as a point of reference for an overview of the area and the related literature.

2.2 DATA AND WEB MINING APPROACHES FOR RECOMMENDER SYSTEMS

The research group DB-NET, Athens University of Economics and Business – http://www.db-net.aueb.gr,has presented several data and web mining techniques that use content semantics and the structural properties of a web site in order to improve the effectiveness of web personalization. In [EVV03], Eirinaki et. al. present the SEWeP (SEmantic Web Personalization) system. SEWeP integrates usage data with content semantics in order to compute semantically enhanced navigational patterns and effectively generate useful recommendations. More specifically, they propose a method for automatically characterizing the Web site's content using terms belonging to an ontology. Using this information, they create the so-called C-Logs (concept logs), which are semantically enhanced Web logs. The C-Logs are subsequently used as input to association rules mining in order to discover the navigation patterns of the Web users. Those patterns are enhanced by expanding the initial recommendation set with semantically similar content. The fact that the content is essentially characterized by ontology terms enables computing similarities between terms that are semantically close, but not the same. In this work they utilize THESUS, a similarity metric previously proposed by this group. In [EL+04], they extend this framework so that it generates category-based instead of page-level recommendations, addressing the problem of

continuously updated content. SEWeP has been among the first Web personalization systems that, contrary to previous works, did not require manually annotation of the Web site's content. The SEWeP prototype has been demonstrated in [LP+04]. The framework has been extended in [EM+06] to show how several similarity measures can be integrated in the system. It has also formed the basis of a project aiming at archiving the Greek Web [LE+04].
Eirinaki et. al. also investigated how the underlying structure of a Web site can enhance the quality of recommendations. In this context, they present UPR (Usage-based PageRank), a PageRank-style algorithm that combines usage data and link analysis techniques for ranking the web pages based on their importance in the web site's navigational graph [EV05]. Using this concept, they introduce l-UPR, which is a localized version of UPR, used to generate recommendations to an active user. They propose an algorithm that, based on the active user's path, generates a personalized sub-graph that includes all potential "next" steps of the user. They then apply l-UPR to rank those pages in order to recommend them to the user. Moreover, they propose a hybrid probabilistic predictive model, based on Markov models and link analysis, as a robust mechanism for determining prior probabilities of page visits [EVK05]. This work has been the first approach of using link analysis in the context of Web site personalization systems. An extended version of this work has appeared in [EV07].
The DB-NET (http://www.db-net.aueb.gr) research group has been involved in the IST project I-Know yoU Mine [IKUM], a project aiming at integrating content semantics in the Web personalization process. The IKUM framework has appeared in [EVA05]. They have also contributed in the NEMIS Network of Excellence [NEMIS] by preparing a survey on Web Mining [E04]. Finally, Dr. Eirinaki's Ph.D. thesis was in the area of Web personalization [E06].

In [OR+04] the group from CTI & University of Patras – http://www.cti.gr/ presents a framework for predicting the next visit of a user in a web site. Their approach involves the use of n-grams. N-grams are used to construct sequential patterns that model the navigations of past users. Since long n-grams tend to increase prediction accuracy but result in loss in applicability, and based on the observation that in real-world scenarios long traversal sequences repeat less frequently than shorter ones, they authors focus on 3rd to 5th order n-grams and their suffices. The algorithm uses those n-grams to generate predictions. The authors also propose the use of the web site topology. Assuming that the web site is categorized based on its content from an expert, they create the so-called site-maps, where page files are inter-connected through links. They then use this information in case in case the prediction algorithm cannot find matching n-grams, by expanding their search to all potential frequent single step transitions.
Dr. Sirmakessis has also served as the moderator of the Web mining Group (WG2) of the NEMIS Network of Excellence [NEMIS].

The research group Delab – Aristotle University, has conducted research in the area of nearest-neighbor collaborative filtering (CF) systems. Nearest-neighbor CF is based either on similarities between users or between items, to form a neighborhood of users or items, respectively. For example, in user-based CF, the users are represented as vectors whose values are the ratings of each user for the respective item. In [SN+06], Symeonidis et. al. investigate several factors involved in the CF process, such as sparsity, neighborhood size, the similarity measure, the prediction process

(recommendation list and threshold), as well as several evaluation metrics available. In their analysis they identify choices that have been incorrectly adopted and new issues not yet considered, proposing several extensions to existing approaches. An extended version of this work has appeared in [SN+08].
In the same thematic area, they have also explored techniques for integrating CF with other data mining algorithms. In [SN+06a] they propose a CF algorithm that detects rating trends among users. This algorithm uses Latent Semantic Indexing and is based on the assumption that user ratings are not expected to be independent.
In [SN+06b], the researchers address the issue of duality existing between users and items. Most CF systems are either user-based or item-based and do not take this into consideration. Their approach, integrating clustering techniques, allows them to augment the two approaches by combining them. More specifically, they propose a nearest-biclusters algorithm that enables them to first form simultaneous clusters of users and items and then employ a similarity measure that achieves partial matching of the users' preferences. Their approach, as compared to related works, improves the scalability and increases the effectiveness of the CF system. An extended version of this work has appeared in [SN+08a], where the authors compare new classes of biclustering algorithms and examine model-based approaches too.
Finally, in [SNM07], the authors propose a hybrid model, based on collaborative and content-based filtering. Content-Based filtering (CB) assumes that each user operates independently. As a result, it exploits only information derived from document or item features. Contrary to previous works, that run one algorithm after the other without considering any dependencies between users and items, they propose a model that discloses the duality between user ratings and item features. They introduce a scheme to weight features according to their impact on user preferences, and therefore compute the similarity between users with respect to the dominant features of their profiles.

Giannikopoulos et. al. [GVE08] (joint work among the University of Peloponnese – http://www.uop.gr, WIM– Athens University of Economics and Business – http://wim.aueb.gr
and the Computer Engineering Dept, San Jose State University, USA – http://www.engr.sjsu.edu, propose a solution to the problem of continuously updated web sites, such as social networking sites, or news portals. In such a dynamic environment, any traditional, usage-based approach that takes as input the navigation paths recorded on the web page level is not as effective. Most predictive models are based on frequent item sets, the more recent a page is, the more difficult it is to become part of the recommendation set; at the same time, such pages are more likely to be of interest for the average user. The authors address this problem by generalizing the page-level navigation patterns to a higher, aggregate level. The proposed algorithm, FPG, is based on the modification and integration of two existing algorithms, namely FP-Growth and GP-Close, results in the creation of generalized frequent patterns. Those patterns can be subsequently used to generate recommendations for newly added content.

2.4 DOMAIN-SPECIFIC APPLICATIONS

In this section we present research works that employ web mining techniques in the context of personalization in specific application domains. We categorize the efforts per domain.

2.4.1 Personalized recommendations for e-commerce applications

The team: P. Markellou, I. Mousourouli, M. Rigou, S. Sirmakessis, A. Tsakalidis from CTI & University of Patras (http://www.cti.gr), has focused its efforts on developing recommendation systems for e-commerce applications. Such applications have unique characteristics, for example the users are usually registered and, as a consequence, much information regarding their profile can be used as input in the personalization process. In [RST04], Rigou et. al. present an algorithm for personalized clustering. The proposed work combines the orthogonal range search with the k-windows algorithm. The system allows the customers to model their preferences and search for products, and then presents them a personalized cluster of products and services. In [MM+05], Markellou et. al. propose a framework that combines Naïve Bayesian classification with association rules mining in order to generate recommendations to the users of an e-shop. They use user profile information including demographics and item ratings, as well as the content of the e-shop organized in an ontology.

2.4.2 Personalized news delivery

In this project, Paliouras et. al., present PNS, which is a Personalized News Service that integrates news from multiple sources on the Web and delivers in a personalized fashion to the reader [PM+06]. This is joint work among the Institute of Informatics and Telecommunications, NCSR Demokritos - http://www.iit.demokritos.gr/, the Dept. of Informatics, Technological Institute of Athens – http://www.cs.teiath.gr and the Dept. of Informatics and Telecommunications, University of Athens – http://www.di.uoa.gr. The system consists of source-specific information extraction programs (wrappers) that extract highlights of news items from the various sources, organize them according to pre-defined news categories and present them to the user. The authors employ the Cluster Mining algorithm proposed by them in a previous work, in order to construct a graph representing the users' interests, which are explicitly provided during registration. They then use this information to personalize the content presented to each user, using content-based and collaborative filtering techniques.

Antonellis et. al. [ABP06], from CTI & University of Patras, http://ru6.cti.gr, present an architecture of a personalized news classification system. In their approach, they assume that they do not have explicit information regarding the users' profiles, except for their level of expertise on different categories. The user specifies the level of his expertise on different topics and the system relies on a text analysis technique in order to achieve scalable classification results. They propose a classification technique that represents documents using the vector space representation of their sentences. Thus, contrary to using the traditional 'term-to-documents' matrix they replace it by a 'term-to-sentences' matrix that permits capturing more topic concepts of every document. This procedure enables the system to capture articles that refer to several topics, while their general meaning is different. They then classify the articles "per-user" based on their level of expertise.

Banos et. al. [BK+06] from the LPIS Group, Aristotle University of Thessaloniki – *http://lpis.csd.auth.gr*, address the problem of personalizing a news site as an information overload problem, contrary to previous approaches that have addressed it as a personalization problem. Moreover, they are taking into consideration current ways of information aggregation and dissemination for news media, such as RSS feeds. The proposed system, named PersoNews, is a web-based machine learning enhanced RSS reader. It allows the user to choose a topic of interest from a thematic hierarchy. Using this information and an incremental Naïve Bayes classifier, it filters uninteresting news for the user. The system is enhanced by a machine learning framework previously proposed by the authors.

2.4.3 Personalizing topic directories

Dalamagas et. al. [DB+07] proposed a methodology for personalizing topic directories. This is joint work among the KDBS Lab, National Technical University of Athens - http://web.dbnet.ntua.gr/en/home.html and the Computer Engineering Dept, San Jose State University, CA – http://www.engr.sjsu.edu. The proposed framework provides a set of mining tasks for discovering user navigation patterns, as well as a set of personalization tasks that can be used to customize the organization of the topic directory based on the discovered patterns. The system identifies, among others, frequent visits of popular categories of the directory and "indecisive users" and uses this information to generate and insert in the directory shortcuts to different topics. The users are automatically clustered into overlapping interest groups, and the mining and personalization tasks are targeting each group individually. Thus, each interest group is presented with a personalized "view" of the directory.

3. MINING THE WEB 2.0

The new collective intelligence technologies introduced in the past few years have led to the slow transition from Web to the so-called Web 2.0. In this new era, the users are actively participating in the information exchange and dissemination, by annotating and posting content, as well as participating in blogs, review sites, social networks etc. Moreover, the rate in which such information is being updated has increased very rapidly. We present here some recent web mining approaches focusing on those new Web 2.0 structures.

The research group *KMD* from – *University of Magdeburg, Germany* - *http://omen.cs.uni-magdeburg.de/itikmd/Home.33.1.html* focuses is on social networks. In [FBS06], the authors study the evolution of subgroups in social networks. The difference from previous efforts on community networks is that social networking communities are continuously evolving, thus techniques proposed for static networks do no longer hold. They propose two approaches to analyze the evolution of two different types of online communities on the level of subgroups. The first method allows for an interactive analysis of subgroup evolutions in online communities through

statistical analyses and visualizations. The second method is designed for the detection of communities in an environment with highly fluctuating members.

Stamou et. al. [SK+07], from the *Database Lab, University of Patras* - http://www.dblab.upatras.gr/gr/index.htm, present HiBO, a bookmark management system allowing users to search, browse, organize and share Web data. HiBO enables that by incorporating an automated hierarchical structuring of bookmarks that are shared across users, providing personalized views to shared files. In this work they propose some web mining techniques that allow them to manage this pool of shared bookmarks across community members and show that their approach's potential in assisting users organize their shared data across different social networks.

Eventually Varlamis et. al. [VVP08], from WIM – Athens U. of Economics & Business, investigated the evolution of interests in the blogoshpere. They present a prototype application, named blogTrust, that allows monitoring changes in the interests of blogosphere participants. They also propose an approach for analyzing the blogosphere by monitoring the convergence or dispersion of blogosphere interests. The proposed process, integrated in blogTrust, classifies weblog posts in predefined categories and generates a feature vector for each weblog from the post classification results. It subsequently clusters together blogs with similar topics/interests, and, via visualization techniques, enables the detection of interest convergence or divergence among bloggers over different time periods, using established data mining techniques. In this work they validate the hypothesis that an abrupt change of results in terms of the variety of blog interests from one period to the next is indicative of a real world event.

4. CLOSURE & REMARKS

Web Mining is a field of research and industrial activity gaining high volume and attention in the last decade. This is due to the vast economic interest of the web users navigation and interaction patterns concluding in financial transactions. Greek scientists are remarkably active and successful in this competitive field of research covering various aspects of the web mining process ranging from web log analysis to personalization/recommendation systems, to specific applications such as personalized news systems, and topic directories. We have presented an overview of the challenges in Web search and the work done by Greek researchers. Our goal is to give a high-level summary of the past and ongoing work in this area and not provide an exhaustive list. We aspire that the covered material will stimulate further intriguing research and fruitful discussions so that work in this area will continue with more exciting research and industrial results.

5. REFERENCES

[ABP06] I. Antonellis, C. Bouras, V. Poulopoulos. "Personalized News Categorization Through Scalable Text Classification", APWeb 2006
[BM+02] B. Berendt, B. Mobasher, M. Nakagawa, M. Spiliopoulou. "The Impact of Site Structure and User Environment on Session Reconstruction in Web Usage Analysis", WEBKDD 2002
[BK+06] E. Banos, I. Katakis, N. Bassiliades, G. Tsoumakas, I. Vlahavas. "PersoNews: A Personalized News Reader Enhanced by Machine Learning and Semantic Filtering", ODBASE 2006
[BS00] B. Berendt, M. Spiliopoulou. "Analysis of Navigation Behaviour in Web Sites Integrating Multiple Information Systems" VLDB J. (VLDB) 9(1):56-75, 2000
[CDS06] E. Christodoulou, T. Dalamagas, T. Sellis. "NaviMoz: Mining Navigational Patterns in Portal Catalogs", PaRMa 2006
[DB+07] T. Dalamagas, P. Bouros, T. Galanis, M. Eirinaki, T. Sellis. "Mining User Navigation Patterns for Personalizing Topic Directories", WIDM 2007
[E04] M. Eirinaki. "Web Mining: A Roadmap", Technical Report IST/NEMIS 2004 (available at http://www.engr.sjsu.edu/meirinaki/research.htm)
[E06] M. Eirinaki. "New Approaches to Web Personalization", Ph.D. Thesis, Athens University of Economics and Business, Dept. of Informatics, May 2006 (available at http://www.engr.sjsu.edu/meirinaki/research.htm)
[EL+04] M. Eirinaki, C. Lampos, S. Paulakis, and M. Vazirgiannis. "Web Personalization Integrating Content Semantics and Navigational Patterns", WIDM 2004
[EM+06] M. Eirinaki, D. Mavroeidis, G. Tsatsaronis, and M. Vazirgiannis. "Introducing Semantics in Web Personalization: The role of Ontologies", in "Semantics, Web, and Mining", LNCS 4289, pp. 147-162, Springer 2006
[EV03] M. Eirinaki and M. Vazirgiannis. "Web Mining for Web Personalization", in ACM Transactions on Internet Technology (TOIT), 3(1):1-27, February 2003
[EV05] M. Eirinaki and M. Vazirgiannis. "Usage-based PageRank for Web Personalization", ICDM 2005
[EV07] M. Eirinaki and M. Vazirgiannis. "Web Site Personalization based on Link Analysis and Navigational Patterns", in ACM Transactions on Internet Technology (TOIT), 7(4), October 2007
[EVA05] M. Eirinaki, Y. Vlachakis, and S.S.Anand. "IKUM: An Integrated Web Personalization Platform Based on Content Structures and User Behavior", in "Intelligent Techniques in Web Personalization", LNCS 3169, pp.272-288, Springer 2005
[EVK05] M. Eirinaki, M. Vazirgiannis, and D. Kapogiannis. "Web path recommendations based on Page Ranking and Markov Models", ACM WIDM 2005
[EVV03] M.Eirinaki, M. Vazirgiannis, and I. Varlamis. "SEWeP: Using Site Semantics and a Taxonomy to Enhance the Web Personalization Process", KDD2003
[GL06] A. Georgakis, H. Li. "User behavior modeling and content based speculative web page prefetching", Data & Knowledge Engineering, 59(3), December 2006
[GVE08] P. Giannikopoulos, I. Varlamis, M. Eirinaki. "Mining Frequent Generalized Patterns for Web Personalization", MSoDa 2008
[FBS06] T. Falkowski, J. Bartelheimer, M. Spiliopoulou. "Mining and Visualizing the Evolution of Subgroups in Social Networks", WI 2006
[IKUM] I Know YoU Mine, IST/I-KnowUMine (IST-2000-31077)
[Liu07] B. Liu. "Web Data Mining. Exploring Hyperlinks, Contents, and Usage Data", Springer-Verlag Berlin Heidelberg 2007

[LE+04] C. Lampos, M. Eirinaki, D. Jevtuchova, M. Vazirgiannis. "Archiving the Greek Web", IWAW 2004
[MM+05] P. Markellou, I. Mousourouli, S. Sirmakessis, A. Tsakalidis. "Personalized E-commerce Recommendations," ICEBE 2005
[NEMIS] NEMIS: Network of Excellence in Text Mining and its applications in Statistics, IST/NEMIS (IST-2001-37574), http://nemis.cti.gr
[NKM01] A. Nanopoulos, D. Katsaros, Y. Manolopoulos. "Exploiting Web Log Mining for Web Cache Enhancement", WEBKDD 2001
[NM00] A. Nanopoulos, Y. Manolopoulos. "Finding Generalized Path Patterns for Web Log Data Mining", ADBIS-DASFAA 2000
[NZ+ 02] A. Nanopoulos, M. Zakrzewicz, T. Morzy, Y. Manolopoulos. "Indexing Web Access-Logs for Pattern Queries", WIDM 2002
[OR04] D. Oikonomopoulou, M. Rigou, S. Sirmakessis, A. Tsakalidis. "Full-Coverage Web Prediction based on Web Usage Mining and Site Topology", WI 2004
[PAV07] G. Pallis, L. Angelis, A. Vakali. "Validation and interpretation of Web users' sessions clusters", Information Processing and Management, 43(5), September 2007
[PM+06] G. Paliouras, A. Mouzakidis, C. Ntoutsis, A. Alexopoulos, C. Skourlas. "PNS: Personalized Multi-Source News Delivery", KES 2006.
[PP+00] G. Paliouras, C. Papatheodorou, V. Karkaletsis, C. Spyropoulos. "Clustering the Users of Large Web Sites into Communities" ICML 2000.
[PP+03] D. Pierrakos, G. Paliouras, C. Papatheodorou, C. Spyropoulos. "Web Usage Mining as a tool for personalization: a survey". User Modeling and User-Adapted Interaction, 13(4):311-372, Springer, 2003
[PP+04] D. Pierrakos, G. Paliouras, C. Papatheodorou, V. Karkaletsis, M. Dikaiakos, "Web Community Directories: A New Approach to Web Personalization", in "Web Mining: From Web to Semantic Web", LNCS 3209, pp. 113 - 129, Springer 2004
[PS02] C. Pohle, M. Spiliopoulou. "Building and Exploiting Ad Hoc Concept Hierarchies for Web Log Analysis", DaWaK 2002
[RST04] M. Rigou, S. Sirmakessis, A. Tsakalidis, "A Computational Geometry Approach to Web Personalization", CEC 2004
[SK+07] S. Stamou, L. Kozanidis, P. Tzekou, N. Zotos, D. Christodoulakis. "HiBO: Mining Web's Favorites", WAIM 2007
[SL+04] S. Paulakis, C. Lampos, M. Eirinaki, and M. Vazirgiannis. "SEWeP: A Web Mining System supporting Semantic Personalization", PKDD 2004
[SN+06] P. Symeonidis, A. Nanopoulos, A. Papadopoulos, Y. Manolopoulos "Collaborative Filtering Process in a Whole New Light", IDEAS 2006
[SN+06a] P. Symeonidis, A. Nanopoulos, A. Papadopoulos, Y. Manolopoulos. "Scalable Collaborative Filtering Based on Latent Semantic Indexing", ITWP 2006
[SN+06b] P. Symeonidis, A. Nanopoulos, A. Papadopoulos, Y. Manolopoulos. "Nearest-Biclusters Collaborative Filtering", WEBKDD 2006
[SN+08] P. Symeonidis, A. Nanopoulos, A. Papadopoulos, Y. Manolopoulos. "Collaborative Recommender Systems: Combining Effectiveness and Efficiency", Expert Systems with Applications, 34(4), 2008
[SN+08a] P. Symeonidis, A. Nanopoulos, A. Papadopoulos, Y. Manolopoulos. "Nearest-biclusters collaborative filtering based on constant and coherent values", Inf. Retr. (IR) 11(1):51-75 (2008)
[SP02] M. Spiliopoulou, C. Pohle. "Modelling and Incorporating Background Knowledge in the Web Mining Process", Pattern Detection and Discovery 2002

[VVP08] I. Varlamis, V. Vassalos, A. Palaios. "Monitoring the evolution of interests in the blogosphere", Data Engineering for Blogs, Social Media, and Web 2.0, 2008.
[WCM] WebC-MINE: Web usage mining from proxy/cache server logs, GSRT/Ministry of Development, Bilateral and International S and T Collaboration

DATA INTEGRATION

Yannis Velegrakis
University of Trento
velgias@disi.unitn.eu

1. INTRODUCTION

Nowadays, we are witnessing an exponential growth in the amount of information that is available online. To fully exploit this information, one needs to be able to collect, combine and generate a unified view of the data of the different sources. Unfortunately, these sources have been typically developed at different times, by different organizations and with different requirements in mind, which naturally generates a large degree of heterogeneity, i.e., different access protocols, domains, structures or semantics. This makes the problem of information integration particularly challenging.

2. MEDIATED INFORMATION SYSTEMS

One of the classical approaches in information integration is the use of mediators. A mediator is a piece of software that obtains information directly from one or more sources or other mediators and provides it to other mediators, users or software applications. It embodies the necessary knowledge that upon receiving a query, it is able to generate and send the right requests to the underlying sources, collect the results, integrate them and send them to the one who requested them. Data does not exist in the mediator, but a mediator is queried as if the data was stored in it. In a sense, each mediator is a logical view of the data found in the underlying sources.
The University of Crete (www.csd.uoc.gr) and the Foundation of Research and Technology (www.ics.forth.gr) have a long tradition in issues related to data and knowledge integration. Among the many related results they have produced, they have developed a mediator model for providing integrated and unified access to multiple sources. The model is based on the idea of using a taxonomy to guide the integration. In particular, each source consists of a taxonomy along with a database that indexes objects under the terms of the taxonomy, and a mediator consists of a taxonomy and a set of relationships between the mediator terms and the source terms. The use of taxonomies assist in achieving an efficient and flexible integration that can accommodate different application needs and provide quality guarantees [TSC05].

3. P2P SYSTEMS

The advent of the web offered a platform on which every organization or individual data provider could publish data online. This generated an enormous growth of the online data and the use of the mediator approach became impractical. A new data exchange model emerged, the Peer-to-Peer (P2P). In a P2P system, sources and applications exchange data on one-to-one basis. They have no centralized authority and no integrated schemas. Sources use bilateral acquaintances to propagate queries and results. This model introduces new challenges such as ways to avoid flooding the network with messages, or methods to specify the relationships among the schemas of the peers.

One of the activities of the Distributed Management of Data Group (http://dmod.cs.uoi.gr/) of the University of Ioannina is the development of a decentralized approach for discovering and routing queries among the nodes of a P2P system [GP04]. The approach is based on summarizing the contents of the sources using an extension of Bloom filters. To allow efficient query processing, histograms were used to exploit the query workloads and arranging nodes that receive similar queries close to each other [KPPT05]. The development of the above solutions has been done in the context of two European projects: Aeolus and Self-Peer. Their aim is to study overlay computers and develop tools and functionalities needed by the programmer to make a P2P system robust, efficient, fault-tolerant and easily scalable.

Similar activities have been initiated by the Web Information Management group (wim.aueb.gr) of the Athens University of Economics and Business. In particular, the group has developed a model for the interaction of peers in an incentive-free P2P network. The model is based on Game Theoretic analysis to calculate Nash equilibria and predict peer behavior in terms of individual contribution [VV08]. Furthermore, their DAIMON project deals with data integration issues in a peer-to-peer network of mobile peers. The goal of this project is to develop a data integration system in a mobile computing environment which needs to deal with issues related to location management, frequent disconnections and highly variable bandwidth.

The DIET project (www.intelligence.tuc.gr/p2pdiet), initiated at the Technical University of Crete and continued at the National and Kapodistrian University of Athens, aims at the development of a P2P service that unifies query and notification capabilities. It supports queries, subscriptions and notifications in the single unifying framework. It contains a fault-tolerance mechanism and a location-independent addressing schema which allows nodes to disconnect and reconnect with a different address and at different parts of the network, without the network loosing consistency [TXK+08].

In the context of a MSc thesis work [SKD+08], RDF schemas have been employed as a mean of advertising the contents of the individual peers in a P2P network. The specific work has demonstrated that RDF-based P2P networks can scale well and can provide good performance in terms of data discovery and retrieval.

4. SEMANTIC INTEROPERABILITY

Of the main challenges in heterogeneous information integration is to understand the semantics of the data. Ontologies have been proposed as a mean to communicate these semantics. An ontology is an explicit specification of a conceptualization, i.e., an abstract view of the world that needs to be represented for some purpose.

Resource discovery

Ontologies are forms of meta-data. As such, they can be used not only to communicate the semantics of the data, but also to provide various other services, such as data quality evaluation or resource discovery. The term resource discovery is used to describe the task of discovering data sources that are related to a given query. The OntoGrid project (www.ontogrid.net) is a European project in which the National and Kapodistrian University of Athens, is one of the participants, and aims to facilitate resource discovery in a Grid framework. The basic assumption in OntoGrid is that resources can be annotated with RDF metadata describing information such as provenance or trust. This information is shown to greatly facilitate resource discovery.

Matching / Mapping

Although ontologies can model the semantics of the data, in order to achieve interoperability one needs to find matchings (or mappings) between the concepts of different ontologies. A matching is an expression that describes how a concept in an ontology is related to a concept in another. This relationship can range from high level, such as, “is-equal”, “is-disjoin” or “overlap”, to very detailed, like the one described by a logical expression. Finding matchings between ontologies requires a lot of effort and good knowledge of the semantics of the modeled data. A large portion of research has been devoted to the development of techniques to automate this task.
The AI lab at the University of Aegean (www.icsd.aegean.gr/ai-lab) is developing an ontology matching tool that integrates many different state-of-the-art matching methods in order to improve accuracy [VSK+07]. The effort is done within the AUTOMS project. The integrated solution involves lexical and structural matching techniques as well as various heuristics.

Query Translation

In order to meaningfully exchange data between different data sources, one needs to be able not only to understand how the concepts of the two data sources relate to each other, but also how to translate queries and data expressed in the “world” of one source into queries or data expressed in the “world” of the other.
The Papyrus (www.ict-papyrus.eu) project aims at studying and exploring solutions to this problem. Papyrus is a European project coordinated by the Athens Technology center (www.atc.gr) and includes the National and Kapodistrian University of Athens

(www.nkua.gr) as one of the main participants of the consortium. It intends to be a dynamic digital library which will understand user queries in the context of a specific discipline, look for content in a domain alien to that discipline and return the results presented in a way useful and comprehensive to the user. It is achieving this through advanced modeling, mapping and query translation techniques between ontologies.
The Web Information Management group (wim.aueb.gr) at the Athens University of Economics and Business has devoted a large part of its research activities into the development of methods related to information integration and querying of XML data. In particular, its Pythagoras research program deals with the integration of the data retrieved from different sources that operate under different semantics [LVM06] and also with the query rewriting of aggregation queries using conjunctive views. Furthermore, the same group is studying ways to make information integration systems full-fledged data management tools by allowing an integration system to accept not only queries, but also updates. To achieve this, the group is employing update-through-views techniques to translate update requests on an integrated view into update tasks at the sources [KSV06].

5. CLOSURE & REMARKS

We have provided a brief reference to the information integration research activities taking place in Greek research and academic institutions. The study of these efforts has indicated that they cover multiple major topics and provide advanced solutions to many of the main challenges.

6. REFERENCES

[GP04] Koloniari G. and Pitoura E. (2004) Content-based Routing of Path Queries in Peer-to-Peer Systems. In EDBT.
[KPPT05] Koloniari G., Petrakis Y., Pitoura E. and Tsotsos T (2005) Query Workload-Aware Overlay Construction Using Histograms. In CIKM.
[KSV06] Kotidis Y., Srivastava D. and Velegrakis Y. (2006) Updates Through Views: A New Hope. In Proceedings of the 22nd International Conference on Data Engineering (ICDE).
[LVM06] Lin V., Vassalos V. and Malakasiotis P. (2006) MiniCount: Effficient Rewriting of COUNT-Queries Using Views (ICDE).
[SKD+08] Sidirourgos L., Kokkinidis G., Dalamagas T., Christophides V., Sellis T. K. (2008) Indexing views to route queries in a PDMS. In Distributed and Parallel Databases 23(1): 45-68.
[TSC05] Tzitzikas Y., Spyratos N., Constantopoulos P. (2005) Mediators over taxonomy-based information sources. In VLDB Journal 14(1): 112-136.
[TXK+08] Triantafillou P., Xiruhaki C., Koubarakis M. and Ntarmos N. (2003) Towards High-Performance Peer-to-Peer Content and Resource Sharing Systems. In First

Biennial Conference on Innovative Data Systems Research (CIDR).
[VSK+07] Valarakos A., Spiliopoulos V., Kotis K., Vouros G. (2007) AUTOMS-F: A Java Framework for Synthesizing Ontology Mapping Methods. I-KNOW Special Track on Knowledge Organization and Semantic Technologies (KOST).
[VV08] Vassalos V., Vassilakis D. (2008) A model and analysis of peer to peer networks with altruistic peers. Submitted for publication to Peer to Peer Networking and Applications (Springer).

DATA MINING

Maria Halkdi
Dept of Technology Education and Digital Systems, University of Piraeus
Email: mhalk@unipi.gr

1. INTRODUCTION

The Data Mining refers to the techniques used for the automated data analysis and extraction of knowledge from large data sets. The incredible growth of the digital universe and the new characteristics of digital information surrounding us in every aspect of our lives, pose new problems in data and knowledge management. A number of novel techniques have recently been proposed aiming to address some of the new challenging issues in data mining. We can classify these techniques based on the types of data under analysis, the types of extracted knowledge, and the different aspects of mining process that they deal with (such as privacy issues, data integrity).

2. DATA MINING ALGORITHMS IN TRATIDIOTAL AREAS

The activities of the DBNET research team (head Prof Vazirgiannis) at Athens Univ. Economics and Business, on Data Mining Systems aim at quality assessment in clustering process in terms of evaluation of clustering results. A new validity index is presented in [HV08] that handles efficiently arbitrarily shaped clusters by representing each cluster with multiple points rather than by a single representative point. Also a semi-supervised framework for learning a weighted Euclidean subspace, where the best clustering can be achieved is proposed in [HGV08]. This approach capitalizes on user-constraints and the quality of intermediate clustering results in terms of its structural properties.
The Delab at Aristotle University (head Prof Y. Manolopoulos) works on methods that improve the efficiency of the mining procedure. In [NPM07] an association-rule mining algorithm is proposed that is scalable with the domain size. The main idea of the proposed algorithm is that the domain of items is partitioned according to their

correlations and then a mining algorithm is described that carefully combines partitions to improve the efficiency. Also a new scheme for memory-adaptive association rule mining is proposed and evaluated in [NM04].

3. STREAM MINING

The Database group at UCR (head Prof D.Gunopulos) works on techniques that allow users to understand the underlying process that controls the changes recorded in the spatiotemporal datasets. In [GGK03] they first address the problem of performing on-line analysis on multidimensional data streams. Then they present techniques for computing temporal aggregates over streams in [ISJ03]. An online mechanism for efficiently determining the overall network status is presented in [HK+06]. It uses a learning phase that clusters the sensors' readings in order to define the states of the network and update them properly, as new measurement arrives.
The research of the database group at CMU (head Prof. C. Faloutsos) in the area of stream mining focuses on the problem of summarizing the key trends in a stream collection. The problem of incrementally discover n-dimensional correlations and hidden variables in a set of n numerical data streams is addressed in [PF05, PF06].
The research activities of the KDBSL at NTUA (head Prof. Sellis and Prof. Vassiliou) related to the stream management are: i) development of efficient methods to create and maintain wavelet synopses over large multidimensional data streams [JSS05]. Also the work in [CGS06] solves the problem of estimating the most significant wavelet coefficients online, ii) In [PS06] an approach for window specifications is proposed that can be used to limit the scope of processing items as a means of providing incremental response to continuous queries
The research activities of Delab at Aristotle University involve techniques for continuous subspace clustering. The work in [KP04] uses the concept of subspace *a*-clusters and provides methods for efficient cluster maintenance as time progresses. In [KPM07] they propose the use of the *Incremental DFT Computation – Index* to deal with the problem of efficient similarity query processing in streaming time series.
Prof M. Spiliopoulou (University Magdeburg) in collaboration with the Data Management & Mining group (head Prof. Y. Theodoridis) works on methods for pattern monitoring over time proposing. In [MN+06] they presented the framework MONIC for modeling and monitoring of cluster transitions.

5. GRAPH MINING-SOCIAL NETWORKS

Data mining for graph is one of the research areas to which the database group at CMU (head Prof. C. Faloutsos) has significant contributions. It introduces the use of tensors (multi-dimensional extensions of arrays) to analyze time-evolving graphs. They used tensors to find correlations and hidden variables in IP traffic matrices, as well as other graphs evolving over time [STF06, SX+07]. Also the group works on the analysis of

influence propagation in blogs. They found several power-law patterns, in real blog data [LG+07]. Another activity is the development of fast algorithms to find the most central node or nodes ('CenterPiece SubGraphs') for a set of query nodes in a given social network. The proposed algorithms are proved to be faster than the naive implementation, and received the 'best paper' award in ICDM [TFP06].

7. SPATIAL AND TEMPORAL DATA MINING

The research work of the Data Management and Mining group (head Prof. Theodoridis) at Univ. of Piraeus focuses on developing special types for data warehouses [MTK08, MF+08] and spatial mining techniques for special data types, such as moving objects [PK07+, NMM08] or medical images [IP+08].
Also the UCR database lab (head Prof Gunopulos and Prof. Tsotras) works on techniques that support efficient data mining operations on geospatial data. In [VHG03] the authors give new indexing techniques for finding similar trajectories. In [PD+03] algorithms for clustering gene expression time sequences are presented while in [LV+04] they present any-time techniques for clustering time series.

8. PRIVACY PRESERVING DATA MINING

The primary research conducted at the DM&PG group at University of Thessaly (head Prof. Verykios) focuses on the development of innovative approaches for the preservation of privacy in the mining of sensitive data. The group is interested in developing state of the art privacy preserving methodologies for the hiding of sensitive knowledge as exemplified by the form of association rules, classification rules and clustering models[VE+04, GV06, GV08].

9. CLOSURE & REMARKS

Since huge amount of data are continuously generated and new types of the digital data emerge, new problems in data and knowledge mining have attracted the interest of the researchers. In this report we summarize the most recent works of Greek researchers in the research area of data mining.

10. REFERENCES

[CGS062] G. Cormode, M. Garofalakis, and D. Sacharidis. "Fast Approximate Wavelet Tracking on Streams". EDBT, 2006.
[GGK03] Sudipto Guha, Dimitrios Gunopulos, Nikos Koudas, "Correlating synchronous and asynchronous data-streams", ACM SIGKDD 2003, vol. , 2003
[GV08] Gkoulalas-Divanis, A., Verykios, V.S. "A Parallelization Framework for Exact Knowledge Hiding in Transactional Databases". International Conference on Information Security (SEC) 2008.
[GV06] Gkoulalas-Divanis, A., Verykios, V.S. "An Integer Programming Approach for Frequent Itemset Hiding". CIKM, 2006.
[HK+06] M. Halkidi, V. Kalogeraki, D. Gunopulos, D. Papadopoulos, D. Zeinalipour-Yazti, M. Vlachos, "Efficient Online State Tracking Using Sensor Networks". MDM, 2006.
[HV08]M. Halkidi, M. Vazirgiannis. "A Density-based Cluster Validity Approach using Multi-representatives", *Pattern Recognition Letters*, Vol. 29, Issue 6, 2008.
[HGV08]M. Halkidi, D. Gunopulos, M. Vazirgiannis, N. Kumar, C. Domeniconi. "A Clustering Framework based on Subjective and Objective Validity Criteria", *ACM TKDD,* Vol. 1, No. 4, 2008.
[IP+08]D.K. Iakovidis, N. Pelekis, E.E. Kotsifakos, I. Kopanakis, H. Karanikas, Y. Theodoridis: "A Pattern Similarity Scheme for Medical Image Retrieval", *IEEE Transactions on Information Technology in Biomedicine,* 2008.
[JSS051] M. Jahangiri, D. Sacharidis, and C. Shahabi. Shift-Split: I/O Efficient Maintenance of Wavelet-transformed Multidimensional Data. ACM SIGMOD, 2005.
[KP04] Kontaki M., Papadopoulos A. "Efficient Similarity Search in Streaming Time Sequences". SSDBM, 2004
[KPM07] Kontaki M., Papadopoulos A., Manolopoulos Y. "Adaptive Similarity Search in Streaming Time Series with Sliding Windows". *Data & Knowledge Engineering,* Vol.63, 2007
[LG+07] Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie Glance and Matthew Hurst Cascading Behavior in Large Blog Graphs (poster presentation). SDM, 2007.
[LV+04]Jessica Lin, Michail Vlachos, Eamonn J. Keogh, Dimitrios Gunopulos, "Iterative Incremental Clustering of Time Series", EDBT, 2004.
[MF+08]G. Marketos, E. Frentzos, I. Ntoutsi, N. Pelekis, A. Raffaeta, Y. Theodoridis: "Building Real World Trajectory Warehouses", Workshop MobiDE'08, Vancouver, Canada, June 2008.
[MKG06]M. Halkidi, V. Kalogeraki, D. Gunopulos, D. Papadopoulos, D. Zeinalipour-Yazti, M. Vlachos, "Efficient Online State Tracking Using Sensor Networks". MDM, 2006
[MN+06] M. Spiliopoulou, I. Ntoutsi, Y. Theodoridis and R. Schult: "MONIC - Modeling and Monitoring Cluster Transitions", KDD, 2006.
[MTK08] G. Marketos, Y. Theodoridis, I. Kalogeras: "Seismological Data Warehousing and Mining", Int'l *Journal of Data Warehousing and Mining,* 4(1):1-16, 2008.
[NMM08]I. Ntoutsi, N. Mitsou, G. Marketos: "Traffic mining in a road-network: How does the traffic flow?", *Int'l Journal of Business Intelligence and Data Mining*, to appear, 2008.

[NPM07] Nanopoulos A., Papadopoulos A., Manolopoulos Y. "Mining Association Rules in Very Large Clustered Domains", *Information Systems,* Vol.32, N.5, pp. 649-669, 2007
[NM04] Nanopoulos A., Manolopoulos Y. "Memory-Adaptive Association Rules Mining", *Information Systems,* Vol.29, pp. 365-384, 2004
[PD+03][24] D. Papadopoulos, C. Domeniconi, D. Gunopulos, Sheng Ma, "Clustering Gene Expression Data in SQL Using Locally Adaptive Metrics", ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, 2003.
[PK+07]N. Pelekis, I. Kopanakis, G. Marketos, I. Ntoutsi, G.L. Andrienko, Y. Theodoridis: "Similarity Search in Trajectory Databases", TIME, 2007.
[PF06] Sun, J., Papadimitriou, S., Faloutsos, C. "Distributed Pattern Discovery in Multiple Streams". PAKDD, 2006
[PF05]Papadimitriou, S., Sun, J., Faloutsos, C. "Streaming Pattern Discovery in Multiple Time-Series". VLDB, 2005
[PS063] K. Patroumpas, T. Sellis. "Window Specification over Data Streams". ICSNW, 2006.
[STF06] Jimeng Sun, Dacheng Tao, Christos Faloutsos. "Beyond Streams and Graphs: Dynamic Tensor Analysis". KDD 2006
[SX+07] Jimeng Sun, Yinglian Xie, Hui Zhang and Christos Faloutsos. "Less is More: Compact Matrix Decomposition for Large Sparse Graphs" (best research paper award) SDM, 2007.
[TFP06] Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan. "Fast Random Walk with Restart and Its Applications". ICDM 2006.
[VE+04] Verykios, V.S., Elmagarmid, A.K., Bertino, E., Dasseni, E., Saygin, Y. "Association Rule Hiding". IEEE TKDE,, vol. 16, no. 4, 2004.
[VHG03[[21]M. Vlachos, M. Hadjieleftheriou, D. Gunopulos, E. Keogh, "Indexing Multi-Dimensional Time-Series with Support for Multiple Distance Measures", KDD, 2003.
[ZG+03] Donghui Zhang, Dimitrios Gunopulos, Vassilis J. Tsotras, Bernhard Seeger, "Temporal and spatio-temporal aggregations over data streams using multiple time granularities.", *Information Systems,* vol. 28, 2003.

RESEARCH GROUPS – LABS

1. **UCR Database lab**, Department of Computer Science and Engineering, University of California Riverside, USA. http://dblab.cs.ucr.edu/
2. **Data Engineering Lab (Delab)**, Department of Informatics, Aristotle University, Thessaloniki, http://delab.csd.auth.gr/
3. **Data Management and Mining group,** Department of Informatics, Univ. of Piraeus, http://isl.cs.unipi.gr/db/index.html
4. **Data Mining and Privacy Group (DM&PG),** Dept. of Computer & Communication Engineering, University of Thessaly (UTH), Volos http://dmpg.inf.uth.gr/ppdm_pub.html
5. **Data and Web Mining research team (DBNET),** Department of Informatics, Athens University of Economics and Business, Athens, http://www.db-net.aueb.gr/
6. **Knowledge and Database Systems Lab (KDBSL),** Electrical and Computer Engineering Department, National Technical Univ. of Athens, http://web.dbnet.ntua.gr/en/home.html
7. **Database group at CMU,** Department of Computer Science, Carnegie Mellon,

http://www.db.cs.cmu.edu/

P2P Information Retrieval and Filtering

Christos Tryfonopoulos
Databases and Information Systems Department
Max-Planck Institute for Informatics
trifon@mpi-inf.mpg.de

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, most of them focus on providing only *information retrieval* (IR) [BMT+05b, LC05, SCL+05, TXD03, SEA+04, BMT+05b] or *filtering* (IF) [TX03, AT06] functionality (also referred to as *publish/subscribe* or *alerting*), and have no support for a combined service. Querying in such scenarios 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 notifying the user automatically whenever a new matching document is published.
The work presented here, tries to bridge the gap between these two important querying paradigms and support both IR and IF in a unifying P2P framework. In the following, two different approaches that demonstrate the use of structured overlays as a routing substrate for two types of data management systems are presented. DHTrie [TIK05a] is an *exact* IR and IF system that stresses retrieval effectiveness, while MAPS [ZTB08, ZTW07] provides *approximate* IR and IF by relaxing recall guarantees to achieve better scalability. In [TZK+07] a comparison between the two different system designs is presented, and the trade-offs between the two approaches are highlighted. Documents and (continuous) queries in both systems are expressed using a well-understood attribute-value model that is based on named attributes with free text as value, interpreted under the Boolean and VSM (or LSI) models [KST06].

Exact Information Retrieval and Filtering

Most of the research in P2P data management has focused in providing exact retrieval functionality over both structured (e.g., *distributed hash tables* - DHTs [RD01,

SMK+01]) and unstructured overlays (e.g., Gnutella). In DESENT [DNV07], peers are clustered by virtue of containing similar documents, and these clusters are organized in hierarchies to support a DL application [DNV06]. Clusters use peers that act as cluster gateways to forward queries and groups of clusters may form super-clusters with their own gateways. In a similar spirit, iCluster [RP08] organizes peers in an unstructured overlay into communities sharing similar content, and allows them to have multiple and dynamic interests by utilizing unsupervised clustering methods (e.g., k-means [SKK00]) to identify these interests. An extension of the iCluster protocols, that supports both IR and IF functionality in a DL domain, has been presented in [RPT+08]. Other approaches that focus on IR over unstructured overlays include the summarization of a peer's content through specialized data structures to facilitate routing of user queries to appropriate peers [KP04, KP08].

The DHTrie architecture follows a different route from the approaches presented earlier by utilizing a structured overlay as the routing substrate. It provides protocols based on DHTs for efficient and adaptive data management, and centralized algorithms for handling document and queries in each peer. To achieve this is it employs two levels of indexing documents (for the IR task) and continuous queries (for the IF task). A prototype system that presents the ideas behind the DHTrie protocols in the context of digital libraries is presented in [TIK05b].

The first level of the DHTrie indexing scheme corresponds to the partitioning of the global index to different peers using the DHT as the underlying routing infrastructure. Each peer is responsible for a fraction of the submitted continuous queries through a mapping of attribute values to peer identifiers. The DHT infrastructure is used to define the mapping scheme and also manages the routing of messages between different nodes. The set of DHTrie protocols extends the basic functionality of the DHT to offer retrieval and filtering functionality in a dynamic peer-to-peer environment [TIK05a].

The second level of the DHTrie indexing mechanism is managed locally by each peer, and is used for indexing the documents and the continuous queries the peer is responsible for. To be able to scale up to large numbers of documents and continuous queries, specialized data structures and local indexing algorithms are of paramount importance. The idea behind the centralized indexing mechanism is to use trie-based data structures to capture common elements between indexed queries, and exploit this clustering at filtering time [TKD04].

Approximate Information Retrieval and Filtering

All approaches to P2P data management taken so far, focus on exact retrieval [BMT+05b, LC05, SCL+05, TXD03, SEA+04, BMT+05a] or filtering [TX03, AT06, TIM05a, TIK05b] by using the P2P network as a decentralized index for both documents and continuous queries. To facilitate this indexing, appropriate protocols that disseminate documents and queries in a deterministic way, depending on the terms contained in them are employed. These document and query indexing protocols lead to filtering effectiveness that is exactly the same as that of a centralized system. This, however, creates an efficiency and scalability bottleneck, while in certain applications this design might not even be desirable (e.g., in applications like news of blog filtering, where the user is not interested in *all* relevant items, but rather in the most interesting ones).

Contrary to approaches that provide exact IR and IF functionality by utilizing per-document indexing, in MAPS [ZTB+08, ZTW08, ZTW07] the concept of approximate IR and IF is introduced; 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, per-peer (rather than per-document) indexing is employed and efficiency and scalability are enhanced by trading a small reduction in recall for lower message traffic.
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 [RD01]. 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. To support the IR functionality, MAPS utilizes well-known resource selection techniques for P2P query routing such as tf.idf based methods, CORI, or language models (see [NF03] for an overview) to route the user query to a carefully selected subset of information sources. Resource selection in such an autonomous and dynamic environment can be improved by taking into account the overlap in the document collections of different content providers [BMT+05b].
To support P2P IF in a scalable and efficient way, MAPS 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 utilises them to predict future peer behavior. This score is based on a combination of resource selection (i.e., tf.idf based) and behavior prediction to deal with the dynamics of publishing. Behavior prediction uses time-series analysis with double exponential smoothing techniques [C04] 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 [ZTW08], 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.

References

[AT06] I. Aekaterinidis and P. Triantafillou. PastryStrings: A Comprehensive Content-Based Publish/Subscribe DHT Network. In ICDCS, 2006.

[BMT+05a] M. Bender, S. Michel, P. Triantafillou, G. Weikum, and C. Zimmer. MINERVA: Collaborative P2P Search. In VLDB, 2005.

Greek Institutions Involved

Part of the work described in this document is research conducted in the following Greek institutions:

1. Intelligent Systems Lab, Dept. of Electronic and Computer Engineering, Technical University of Crete.
2. Department of Informatics and Telecommunications, National and Kapodistrian University of Athens.
3. DB-NET group, Department of Informatics, Athens University of Economics and Business.
4. Network-Centric Information Systems (NetCINS) Lab, Department of Computer Engineering and Informatics, University of Patras.
5. Distributed Management of Data (DMOD) Lab, Computer Science Department, University of Ioannina.
6. Software and Database Systems Lab, Department of Computer Science and Technology, University of Peloponnese
7. Research Academic Computer Technology Institute (CTI).

The institutions were obtained using the affiliation of the authors at the time of the publication.