A Platform for Search in the Big Web 2.0

Emiran Curtmola
UC San Diego
9500 Gilman Dr.
La Jolla, CA 92093
ecurtmola@cs.ucsd.edu

ABSTRACT

The recent explosion of the amount of different types of information being generated from so many different places under different social types of interactions between users has made search a hot topic for many research communities. While the traditional web search focused on simple keyword search and on references between pages, nowadays getting the right information at the right time is getting harder all the time posing a critical need for expressive, efficient, relevant and flexible search tools.

We study the search in large-scale social systems by capturing logically the natural way people search and discover information: the relevance of keywords relative to the document structure, the importance of references between pages and the associations generated by the online social context. We argue that the key for successful search is to provide a strong theoretical basis to enable the development of theory and practical optimization algorithms. We are the first to show how to transfer the well-established relational world expertise into keyword search. The thesis of this research is to build a prototype based on this formalism and to demonstrate how we can leverage it to address these search challenges.

1. INTRODUCTION

According to a recent emarketer report [28], the number of search engine users in US is increasing and it is expected to reach 56.1% of the US population by 2010. This result is in line with the worldwide ascending trend in the number of web users [33] as well as in the amount of information created and replicated which is predicted to surge by 2010 more than six times than 2006's to an estimated total of 988 billion GB [22]. At the same time we note the multitude and diversity of collaborative web applications, in particular of community-based online content cooperation and sharing services such as del.icio.us [16], slideshare [45], Flickr [20], YouTube [51], MySpace [37], Facebook [18]. The increasing popularity of such forums, web blogs, wikis, mash-ups and social networking applications where users provide content, share and organize it, and search it creates new synergies that characterize the new Web 2.0. In this arena *search* is and will be a major player. Currently, we are experiencing a shift from searching based only on keywords to integrating more social semantics in the search process itself by taking advantage of the properly structured underlaying semantics. Web 2.0 brings different requirements: with more online users, comes more heterogeneous data that need to be searched efficiently and effectively in a way that reflects the users interactions.

Proceedings of SIGMOD2007 Ph.D. Workshop on Innovative Database Research 2007 (IDAR2007), June 10, 2007, Beijing, China.

While the search on traditional databases where data are highly structured with a clear notion of schema has been extensively studied, web search where the data tend to be loosely structured, or worse, unstructured has received a lot of attention lately. The problem resides in finding expressive ways to access the new type of information by combining state-of-the-art Database XML retrieval (DB) and Information Retrieval (IR) techniques. Within this research area, we identify three key challenges that we discuss next.

Scenario 1.1. **(Search over structured, semi-structured and unstructured data)**. *Alice wants to use a search system that helps her find the best* locations *in* South America *where she can spend her 14-day* vacation *with her family. She wants to be close to the* water *where she can* sail *(keywords* water *and* sail *should appear in close proximity in the text -no more than 15 keywords of each other). Alice prefers to rent* sailboat *classes she knows to sail with, either a* C&C 99 *or a* J102 *yacht.*

Challenge 1.1. *Support for efficient full-text search evaluation over individual XML data sources while still guaranteeing consistent scoring and consistent ranking methodologies independent of the query processing.*

Scenario 1.1 shows how a person is inclined to search based on keywords and also to give hints on the possible contexts where the keywords appear. We assume that applications support the now very popular eXtensible Markup Language (XML) data format for its flexible and self-describing nature. Under this assumption, *South America* is a *location* and it is to be searched anywhere nested under a *location* tagname. Similarly, *J102* should appear under the *sailboat* tagname.

First, let's note that there is no agreed language for querying XML data on both the structure of the document and the

textual part. Instead, there is a variety of competing proposals with variable expressive power, scoring methods and often fuzzy semantics coming from both communities, DB and IR. In this class of languages we distinguish the most expressive of them which is a W3C standard proposal, XQuery Full-Text [53] query language (XQFT). However, there is no obvious approach for the efficient evaluation of such expressive languages as stated in Challenge 1.1. In addition, in the context of query optimization, it is non-trivial to guarantee the same set or the same order of scored-results of two query plans for the same full-text expression and for the same data. In this direction, we pioneer formal techniques for global optimization of full-text queries with scores.

Scenario 1.2. **(Social search)**. *In taking her decision, Alice simultaneously wants to use the search system to find valuable information about the local attractions, about nearby hotels and about the schedule of existing flights to and from each location. She prefers to use travel stories from her circle of* friends *(people in the social networks Alice participates in) and their profiles, her work mates opinions, and also advices from travel experts. Alice wants to make her decision based on the top-10 most enjoyable experiences reported by other travelers and friends, based on their votes.*

Challenge 1.2. *Support for effective search evaluation in the context of Web 2.0, which takes into account the social search phenomenon.*

As illustrated in Scenario 1.2, the second research topic pushes the relevance of social networking information within the full-text search. Let us note that people do not only post, search and read information, but they also define explicit online-relationships the same way as interactions are defined in real life, e.g., define friendships, voting, tagging or saving relationships and then, they surf from the list of friends to find friends of friends of friends etc. We are moving from what the traditional search on collection of documents was considered to be towards pushing more semantics in it. Under the social networking umbrella we are mostly interested in social search to analyze how people interact, exchange information, learn and influence one another. The challenge consists in exploiting the new types of available contextual information (e.g., prevalent data sources based on friendships) and in supporting infrastructures for the management of such information. Our approach is to use query optimization techniques to push the social context into the query evaluation.

Scenario 1.3. **(Search over highly decentralized data)**. *Alice also wants to keep up to date with the political situation and the entertainment news in the selected locations for travel. She wants to see the most interesting local photos and video clips posted by other travelers and be notified when new information matching her interests is added. Occasionally, she posts messages and uploads on different travel community websites her video clips from past trips. The system she connects to and which helps her access and post this information is expected to do all the above for Alice. Figure 1 depicts the types of interaction the producers and the consumers of information are involved in.*

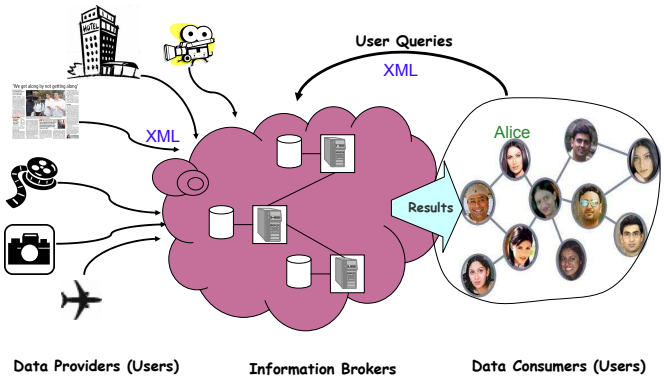

Figure 1: Web 2.0 Search in P2P.

Challenge 1.3. *The ability to efficiently deliver relevant information to and from a large and dynamic group of completely decentralized data consumers and producers.*

So far we addressed such issues as "how to ask" queries expressively and "how to answer" them efficiently under the assumption that the data are centralized at one location. However, the web data may be stored or provided by several distributed data sources within a collaborative environment: the news information can have a different provider than the photos and video clips sharing service or than the travel blog service. The overlay network is only locally aware of data producers and data consumers and guarantees storage and search capabilities for the existing data. However, there is nobody in the system with global knowledge about the overlay as the data is completely decentralized among the participating nodes. The last topic, as illustrated in Scenario 1.3, looks at "whom to ask" queries among the data sources, and "whom to tell" in the network when sources publish new data. The challenge is to adapt the search process from a centralized setting into the decentralized arena where the data may be highly distributed and queried across a large network of potentially hundred to thousand of users. In this new setting we want to design better algorithms and tools for information discovery, evaluation and ranking of complex full-text queries.

Contributions. My research over the past two of years has been concerned with addressing some of these open problems and is oriented towards designing algorithms and developing tools for efficient and effective keyword search. We have looked at search on structured and unstructured data in both settings, centralized as well as decentralized (e.g., Grid and P2P networks). In this thesis, we expect the following results

- A rigorous approach to describe and reason about XML full-text search in terms of semantic specifications and global score-aware optimizations. In [3] we open the way to formal techniques for XML keyword search. Our future work is to push for global query optimizations that account for consistent scoring and ranking methodologies.
- A characterization of the relevance of full-text search in social networking to enable user profile optimiza-

tions. This is ongoing work in which we use query optimization techniques to push the user profile into the search.

- The design of distributed access methods for efficient full-text content-based dissemination and querying. Our preliminary results show that we identify data sources quickly providing high throughput in the network and good load balancing. We propose a fine grained characterization of the data combined with a partitioning strategy on it and a query routing methodology that leverages multicast dissemination trees in a way that avoids checking the queries redundantly at each node.
- A search prototype which can be used to prove the validity of our methods.

We will elaborate more on each of these challenges in the following sections. Section 2 surveys related work. Section 3 presents our solution and briefly discusses several main technical challenges. We conclude in Section 4.

2. RELATED WORK

Recent research has addressed the problem of efficiency and relevancy in keyword search on structured, semi-structured and unstructured data.

Free-form keyword search over relational databases has attracted research interest as in BANKS [5], DBXplorer [1], DISCOVER [30], Hristidis et al. [29] and Liu et al. [35]. All these works view a database as a graph where nodes are the database tuples and edges reflect the application-specific relations. They answer AND/OR keyword queries where all keywords or a partial set of keywords appears in the answer represented by a joining tree of tuples from the input. Techniques used to compute and prune the set of answers include various heuristics to find the Steiner trees, the use of schema information and IR-ish relevance methods to score the joining tuples.

At the same time, there has been extensive research in evaluation of keyword search for structured search on XML coming from the DB community (Florescu et al. [21], XKeyword [31]), as well as from the IR community (XQuery/IR [7], XSearch [12], XIRQL [24], XXL [47]). The idea of computing the most specific elements for conjunctive queries has been actively explored mostly by the DB community using the lowest common ancestors (LCAs) as in XRank [27], Schema-free [34], Schmidt et al. [43] and XKSearch [50]. However, unlike our approach, these works develop algorithms for specific keyword search primitives in isolation. The work based on LCAs does not account for full-text primitives evaluation. In addition, none of these works account for full-text composable primitives with each other nor with search primitives on the document structure.

Several full-text algebras and query evaluation algorithms have been proposed in the past [2, 13, 25, 32, 41, 49]. Yet, they come with limited expressive power and none of them has really addressed the problem of global query optimizations of full-text queries with scores. We open this research direction by exploiting the fact that full-text search can generally be captured in terms of tuples of keyword matches in the text and in terms of their relevance.

A related research topic where the data model is a graph of objects together with the associations between them is represented by Personal Information Management (PIM) systems such as Semex [8], MyLifeBits [26], LifeStreams [23] and the Google Desktop search toolkit. Here, the keyword search is enhanced with the various automatically discovered or manually specified associations between objects. Our work aims to augment the search with the user's dynamic social profile knowledge while still guaranteeing efficient evaluation. As a result, each user will get a different set of answers depending on his social profile.

Finally, a large body of work gravitates towards distributed content access and dissemination systems where timely access to information is critical. While this work started with topic-based approaches where the search was done via predefined coarse-grained sets of topics, the focus has shifted towards full content-based approaches, at a finer granularity, based on the actual content of the published data. Most related to our work are systems as XTreeNet [19], Semcast [39] and ONYX [17]. These works incur high network control overhead either by maintaining a large amount of state, by matching data with aggregate subscriptions redundantly at multiple nodes or by expensive global optimizations targeted to balance the load in the network. We build on the XTreeNet approach and, in addition, we fix all these limitations.

3. OUR SOLUTION

In this section we address several key technical issues and propose solutions for expressive ways to efficiently search in the big Web 2.0.

3.1 Global optimizations of full-text queries

Together with the increase in the existing volume of XML data, be that available in public repositories or generated during data exchange and collaboration, we witness an increasing demand of query languages for unstructured data management. Currently, there are several approaches for XML full-text query languages and scoring methods [12, 21, 24, 34, 47, 48, 53] proposed by both the database and the information retrieval communities.

The need to optimize these approaches in the presence of scores, but also the need to formally understand their semantics, calls for a unified treatment of all these language formalisms. So far, their diversity ruled it out. In addition, full-text search poses a challenge to efficient evaluation. Due to the interplay of the expressive full-text predicates and the element nesting in the XML documents, evaluating predicates on each element independently may result in redundant work. Moreover, if we look at the class of full-text search languages we note their functional language style approach, e.g. XQFT. At first, this looks very different from our notion of a database query language. Therefore, it is non-trivial to reason about their efficient evaluation, especially in a relational approach.

Our breakthrough result is the ability to express these semantics relationally in terms of tuples of matches of the keywords into the documents the same way the relational world manipulates sets of tuples. This result enables to leverage the tried-and-true relational-style optimization techniques

(e.g., relational algebraic rewritings such as join reordering and selection pushing).

Our first step in this direction is to provide a novel score-aware unified framework [3] for expressing the different existing XML full-text search languages. At the same time, we bridge the gap between DB and IR by a successful marriage of the search on the document navigation (e.g., Lowest Common Ancestor stack-based computation techniques) and traditional relational tuple-at-time manipulation techniques with keyword search and full-text predicate evaluation.

The ability to express XML keyword search in the relational world enabled the design of an XML full-text algebra (XFT) and efficient evaluation algorithms for each operator. This allows for concise and clean formal query semantics specification for all the XML full-text query languages by translation to the XFT algebra, along with a uniform treatment of their optimization. A user can benefit from writing an XFT expression, optimize and evaluate it regardless of the underlying full-text language. In particular, XFT can also accommodate NEXI [48] and XQFT, for which we did the first complete reference implementation, GalaTex [14][1].

For the next steps we want to dive deeper into foundational problems such as deciding semantic equivalence of two full-text expressions. We target optimizations based on full-text query rewritings including query minimization and query rewriting by using our previous results from rewriting with materialized views [38]. Moreover, we believe that XFT can be fully integrated with algebras for structured XML search in order to enable optimizations that go across both the text and the navigation part of the documents.

However, the previously mentioned equivalence rewritings break when intermediate results are scored, because we can either get same results but with different scores or worse, get different results. We propose to study the property of the scoring functions for XML search such that certain equivalence query rewritings hold (e.g., relational algebraic rewritings), thus enabling consistent scoring and consistent ranking methodologies. By consistent methodologies we understand procedural-independent ways to compute the scores such that given two different plans for the same query they will compute the same set of answers with either the exact same score per answer or same top-K answers (answers appear in the same ranking order). Our approach is a generic one by abstracting away the scoring functions and by trying to catalog all existing ones in order to identify those properties that enable certain optimization results.

Having described our research agenda on foundational issues of full-text query evaluation, we examine next how moving towards Web search can enrich the search with real-world user-driven relevance.

3.2 Effective Web 2.0 search: role of structure

Following the rapid success of highly popular user-centered applications, forming what is called today the Web 2.0, we note the natural desire of the online user for enhanced interactivity with the data as well as with other users. These applications generate a huge amount of social media ranging from text to rich multimedia. Assuming a popular data format suitable for structured, semi-structured and unstructured data search such as XML, finding quickly information relevant to the user leads to a variety of interesting challenges. Currently, most of the data on the Web are manipulated as unstructured just because it appears to be very weakly structured. However, the user expects the data to be searched not only as flat text, but he also wants to see the impact of his relation with the online community in the query answer. Building on the ideas presented in Section 3.1, we identify and focus on two more interrelated directions: scoring and ranking user generated content, and contextual data analysis.

Scoring user generated content. One of the forces that drives the Web 2.0 is the tendency of online users to express their intentions and interactions as blogs and annotations in forms of links to other pages, tags, recommendations, comments and voting systems. The user generated content is usually semi-structured containing a mix of both textual information and some structure to it. The content can be captured as an XML extension where XML-based descriptors are used to enrich the web to navigate on documents. Hence, full-text search is a convenient tool to search these data.

In addition to this, different user communities are being formed based on users sharing similar interests. This is a rich information source to help users find more relevant data, a topic that will be detailed next.

Contextual data analysis. The current web search applications have already started to change the way people interact online. Users seek not only the content itself, but also want to implicitly benefit from the hidden community-based relationships. However, the current technologies are not mature enough to offer the full benefits. The questions to be asked here are how does one systematically exploit the new types of contextual information such as the network context, the social context and the system context, and how does one enable infrastructures to support management of this contextual information. Thus, in order to help people use information more effectively, we propose to understand the community structure and dynamics, which can be used to enrich the Web 2.0 search by integrating the user's profile. In this way, for the same query and same data, users will get different answers based on their profile interest.

The challenges consist in modeling the social networking information in such a way that it can be leveraged for effective query answering in a formal framework.

We look at ways to characterize user interests by using a graph-like structure. The graph captures semantic links between the community members (e.g., friendships), the user interests for pages (e.g., people tagging pages, voting systems, saving systems and other forms of user feedback) and the internal dependence between pages (e.g., references across pages).

To answer queries on this data model, we propose the follow-

[1] http://www.galaxquery.com/galatex/

ing solution. Accounting for the social context can be viewed as a special class of scoring functions that can be interpreted as a relaxed query which accommodates the new relations (friendship, tagging) besides the hyperlinks relation. From an algorithmic point of view one can ask interesting questions about possible optimizations of these scoring functions. We are looking for efficient *walks* in this space to find the set of answers. However, the graph traversal order might generate different algebraic plans. For the purpose of query optimization, it is non-trivial to analyze the different possible equivalent execution plans of a query which deals with joining all the different search data sources and combining them with selections on text, on structure and on contextual relationships.

Our work is inspired by relational keyword search techniques for aggressive candidate join-plans pruning (BANKS [5], DBXplorer [1], DISCOVER [30], XKeyword [31]) and also for forms of score aggregation during recursive traversals [10]. We want to address the notion of a scoring function for social networks: define it, compile it into scoring plans, optimize it and ultimately, evaluate it efficiently. In this context, we believe that the first identified research topic in Section 3.1 fits perfectly by bridging the gap between formal query semantics specification and reasoning, and social web-search. This can lead to an efficient, effective and flexible generic search with a semantics independent of the underlying query processing.

3.3 Efficient distributed content search: discovery, evaluation and ranking

Going one step beyond the social networks we identify a more important chunk of Internet traffic, the peer-2-peer (P2P) generated traffic which according to some estimates represents $50\% - 70\%$ of the Internet traffic [11, 36]. The P2P includes traffic from distributed data storages, newsgroups, file sharing and online social networks, community-based data exchange, internet TV broadcasting and telecommunications (e.g., Skype [44], cellular networks).

To match up this large amount of both structured and unstructured data, the system moves from a centralized world to a decentralized one by distributing the information in the network. For both the query-answering and the publisher-subscriber paradigms, it is now crucial to understand how to search the web when the data are distributed in a large scale network.

There is a class of applications for which it is infeasible to apply the current Google or Yahoo! data aggregator approach. In the aggregator approach, the data are fully materialized at a centralized entity, also called a data warehouse. Several reasons support this inability. We mention a few of them: the data freshness, the data privacy, the data coverage, the dynamically generated documents or simply the traffic congestion and the processing bottleneck in a single point of data access. We are looking at efficient ways to discover information in P2P and we want to design tools that can quickly identify which sources are relevant to a user's interests.

While designing access methods for XML in distributed systems we explore and address different tradeoffs on network load balancing, latency, efficient search data structures, estimation methods for query selectivity and incremental updates. The main goal is to achieve optimum overall throughput. Our design approach is a scalable overlay network connecting XML-based information producers and consumers for efficient content dissemination and querying. We propose to route queries in the network from users to data sources. To cope with the system scalability and with the full distributed nature of the overlay network, matching the published data against the users' queries leverages the concept of content descriptors [19]. Intuitively, XML-based content descriptors model the information that users want to receive and that producers generate. We consider that the users issue conjunctions of content descriptors, also called conjunctive queries. To ensure efficient routing we combine the decoupling between data consumers and producers via the content descriptors with the manipulation of multicast distribution trees [9].

Several multicast tree solutions based on content-based trees [4] or on distributed hash tables [46, 40, 42] are obvious. First, let us notice that using only one tree to disseminate all the data in the overlay from producers to users is appropriate for routing conjunctive queries as all the conjuncts can be used to filter during the routing. However, this creates a tremendous traffic bottleneck in the upper level nodes of the tree. On the other hand, using one tree for each content descriptor is much too granular which makes it infeasible in practice. The latter solution leads to a high control overhead in the overlay when there are millions of trees. Moreover, it does not help to resolve conjunctive queries efficiently. We believe that partitioning the overlay traffic into a reasonable number of blocks, where each block controls the traffic in its distribution tree, can give better control over balancing the overall load, thus achieving better throughput.

Currently, we are experimenting query routing with Boolean query workloads on a subset of the XML Wikipedia dataset [15]. For fast information filtering at each node we use Bloom Filter structures [6] which prove to be feasible data summaries incurring good overall performance at the cost of very small fixed false positive error rates. The preliminary experimental results show that our method achieves high throughput and load balances the traffic in the network. As future work we want to design and use access methods to support a richer class of queries ranging from Boolean of XPath [52] expressions to more complex, full-text queries, while still guaranteeing good performance for the distributed search.

4. CONCLUSION

The growth and the heterogeneous character of available data call for ever more sophisticated techniques to manage information, and in particular to search. In the new Web 2.0 paradigm the main problem is to efficiently expose the unstructured data together with its relationship to the users in order to ease their web navigation and to provide them with relevant personalized web experience. We believe that the more expressivity and structure (e.g., document navigation or social networks) are pushed into the search process, the more satisfied the users will be.

We present a new vision on how to support enhanced search for the next generation of large-scale community-based on-

line search. We are looking for a solution which has strong foundations and which can be seamlessly integrated with full-text search. A key idea in our research is to support expressive querying for which the formal grounds enable the development of theory and practical optimization algorithms. For the first time this will allow the transfer of the vast body of expertise from relational querying (DB) into keyword search (IR). By taking the best of both worlds and at the same time integrating the social aspect we intend to shape a new, better solution to the Web 2.0 search paradigm!

5. REFERENCES

[1] S. Agrawal, S. Chaudhuri, G. Das. DBXplorer: A system for keyword-based search over relational databases. ICDE 2002.

[2] S. Al-Khalifa, C. Yu, H. V. Jagadish. Querying Structured Text in an XML Database. SIGMOD 2003.

[3] S. Amer-Yahia, E. Curtmola, A. Deutsch. Flexible and Efficient XML Search with Complex Full-Text Predicates. SIGMOD 2006.

[4] T. Ballardie, P. Francis, J. Crowcroft. Core based trees (CBT). SIGCOMM 1993.

[5] G. Bhalotia, C. Nakhe, A. Hulgeri, S. Chakrabarti, S. Sudarshan. Keyword Searching and Browsing in databases using BANKS. ICDE 2002.

[6] B.H. Bloom. Space/time trade-offs in hash coding with allowable errors. Commun. ACM, v.13 n.7, p.422-426, July 1970.

[7] J.M.Bremer, M. Gertz. XQuery/IR: Integrating XML Document and Data Retrieval. WebDB 2002.

[8] Y. Cai, X. L. Dong, A. Halevy, J. M. Liu, J. Madhavan. Personal information management with SEMEX. SIGMOD 2005.

[9] M. Castro, P. Druschel, A-M. Kermarrec and A. Rowstron. SCRIBE: A large-scale and decentralised application-level multicast infrastructure. JSAC 2002.

[10] K. Chakrabarti, V. Ganti, J. Han, D. Xin. Ranking Objects by Exploiting Relationships: Computing Top-K over Aggregation. SIGMOD 2006.

[11] Cisco Systems, Inc. Managing Peer-to-Peer Traffic with Cisco Service Control Technology. White paper, February 2005.

[12] S. Cohen, J. Mamou. Y. Kanza, Y. Sagiv. XSEarch: A Semantic Search Engine for XML. VLDB 2003.

[13] M. P. Consens, T. Milo. Algebras for Querying Text Regions: Expressive Power and Optimization. J. Comput. Syst. Sci. 57(3): 272-288 (1998).

[14] E. Curtmola, S. Amer-Yahia, P. Brown, M. Fernández. GalaTex: A Conformant Implementation of the XQuery Full-Text Language. XIME-P 2005.

[15] L. Denoyer, P. Gallinari. The Wikipedia XML Corpus. SIGIR 2006.

[16] del.icio.us. http://del.icio.us/.

[17] Y. Diao, S. Rizvi, M. Franklin. Towards an internet-scale XML dissemination service. VLDB 2004.

[18] Facebook. http://www.facebook.com/.

[19] W. Fenner, M. Rabinovich, K. K. Ramakrishnan, D. Srivastava and Y. Zhang. XTreeNet: Scalable overlay networks for XML content dissemination and querying. WCW 2005.

[20] Flickr. http://www.flickr.com/.

[21] D. Florescu, D. Kossmann, I. Manolescu. Integrating Keyword Search into XML Query Processing. WWW 2000.

[22] A Forecast of Worldwide Information Growth Through 2010. J.F. Gantz, D. Reinsel, C. Chute, W. Schlichting, J. McArthur, S. Minton, I. Xhenti, A. Toncheva, A. Manfrediz. IDC White paper, March 2007.

[23] E. Freeman, D. Gelernter. Lifestreams: a storage model for personal data. SIGMOD Bulletin 1996.

[24] N. Fuhr, K. Grossjohann. XIRQL: An Extension of XQL for Information Retrieval. SIGIR 2000.

[25] N. Fuhr, T. Rölleke. A Probabilistic Relational Algebra for the Integration of Information Retrieval and Database Systems. ACM TOIS 15(1), 1997.

[26] J. Gemmell, G. Bell, R. Lueder, S. Drucker, C. Wong. Mylifebits: Fulfilling the memex vision. ACM Multimedia 2002.

[27] L. Guo, F. Shao, C. Botev, J. Shanmugasundaram. XRANK: Ranked Keyword Search over XML Documents. SIGMOD 2003.

[28] D. Hallerman. Search Marketing: Players and Problems. http://www.emarketer.com/. April 2006.

[29] V. Hristidis, L. Gravano, Y. Papakonstantinou. Efficient IR-style keyword search over relational databases. VLDB 2003.

[30] V. Hristidis and Y. Papakonstantinou. DISCOVER: Keyword search in relational databases. VLDB 2002.

[31] V. Hristidis and Y. Papakonstantinou and A. Balmin. Keyword Proximity Search on XML Graphs. ICDE 2003.

[32] E. Hung, Y. Deng, V. S. Subrahmanian. TOSS: An Extension of TAX with Ontologies and Similarity Queries. SIGMOD 2004.

[33] B. Macklin. Worldwide Internet Users: 2005-2011. http://www.emarketer.com/. February 2007.

[34] Y. Li, C. Yu, H. V. Jagadish. Schema-Free XQuery. VLDB 2004.

[35] F. Liu, C. Yu, W. Meng, A. Chowdhury Effective keyword search in relational databases. SIGMOD 2006.

[36] A. Madhukar, C. Williamson. A Longitudinal Study of P2P Traffic Classification. MASCOTS 2006.

[37] MySpace http://www.myspace.com/.

[38] N. Onose, A. Deutsch, Y. Papakonstantinou, E. Curtmola. Rewriting Nested XML Queries Using Nested Views. SIGMOD 2006.

[39] O. Papaemmanouil, U. Centintemel. Semcast: Semantic multicast for content-based data dissemination. ICDE 2005.

[40] A. Rowstron, P. Druschel. Pastry: Scalable, decentralized object location and routing for large-scale peer-to-peer systems. IFIP/ACM Middleware 2001.

[41] A. Salminen, F. Tompa. PAT Expressions: an Algebra for Text Search. Acta Linguistica Hungar. 41 (1-4), 1992.

[42] S. Ratnasamy, P. Francis, M. Handley, R. Karp, S. Shenke. A Scalable Content-Addressable Network. SIGCOMM 2001.

[43] A. Schmidt, M. Kersten, M. Windhouwer. Querying XML Documents Made Easy: Nearest Concept Queries. ICDE 2001.

[44] Skype. http://skype.com/products/explained.html.

[45] slideshare. http://www.slideshare.net/.

[46] I. Stoica, R. Morris, D. Karger, M.F. Kaashoek, H. Balakrishnan. Chord: A scalable peer-to-peer lookup service for internet applications. SIGCOMM 2001.

[47] A. Theobald, G. Weikum. The Index-Based XXL Search Engine for Querying XML Data with Relevance Ranking. EDBT 2002.

[48] A. Trotman and B. Sigurbjrnsson. NEXI, Now and Next. INEX 2004.

[49] J.N. Vittaut, B. Piwowarski, P. Gallinari. An Algebra for Structured Queries in Bayesian Networks. INEX 2004.

[50] Y. Xu, Y. Papakonstantinou. Efficient Keyword Search for Smallest LCAs in XML Databases. SIGMOD 2005.

[51] YouTube. http://www.youtube.com/.

[52] The World Wide Web Consortium. XPath 2.0. http://www.w3.org/TR/xpath/.

[53] The World Wide Web Consortium. XQuery 1.0 and XPath 2.0 Full-Text. Working draft. http://www.w3.org/TR/xquery-full-text/.