Project Number 001907

DELIS
Dynamically Evolving, Large-scale Information Systems

Integrated Project
Member of the FET Proactive Initiative Complex Systems

Deliverable D6.1.3

Start date of the project: January 2004

Duration: 48 months

Project Coordinator: Prof. Dr. math. Friedhelm Meyer auf der Heide
Heinz Nixdorf Institute, University of Paderborn, Germany

Due date of deliverable: December 2006
Actual submission date: January 2007
Dissemination level: PU – public

Work Package 6.1: Models and Strategies for Collaborative Web Information Search

Participants: Max Planck Institute for Computer Science (MPII), Saarbrücken, Germany
Computer Science Institute (CTI), Patras, Greece
Telenor Communication AS (Telenor), Oslo, Norway

Authors of deliverable: Gerhard Weikum (weikum@mpi-sb.mpg.de)
Holger Bast (bast@mpi-sb.mpg.de)
Geoffrey Canright (geoffrey.canright@telenor.com)
Peter Triantafillou (peter@ceid.upatras.gr)
1 Introduction and Overview

A major goal of DELIS SP6 is to develop foundations for collaborative Web information search in an Internet-scale peer-to-peer (P2P) system. We are aiming at a P2P system where each peer has a full-fledged Web search engine, including a crawler and an index manager. The crawler may be thematically focused or crawl results may be postprocessed so that the local index contents reflects the corresponding user’s interest profile. Each of the peer-specific local search engines could leverage advanced methods such as spectral analysis, linguistic tagging, ontology-based concept search, and other techniques that cannot be applied to a centralized Web-scale corpus. With such a highly specialized and personalized “power search engine” most queries should be executed locally, but the user may occasionally not be satisfied with the local results and would then want to contact other peers. A “good” peer to which the user’s query should be forwarded would have thematically relevant index contents, which could be measured by statistical notions of similarity between peers. Query routing could greatly benefit from collective human inputs in addition to standard statistics about terms, links, etc.: knowing the bookmarks and query logs of thousands of users would be a great resource to build on.

While the first two years of this package have mostly focused on the P2P architecture and collaboration strategies for keyword-search functionality and basic but efficient and scalable query routing, we are now pursuing, in the second half of DELIS, advanced functionality for statistically enhanced query routing, concept-based search, new forms of link analyses for authority and trust assessment, and personalization strategies based on user or community behavior. This report gives an intermediate snapshot of our results so far and the ongoing work. Our main achievements in this line of research are the following:

- Improvements of our earlier work [11] on overlap-aware query routing for combined quality-novelty in P2P networks, by better estimation of overlap measures among peers and global measures of term specificity like network-wide document frequencies. These results are joint work by MPII and CTI, and have been published in the highly selective EDBT conference 2006 [74] and the well regarded WebDB workshop 2006 [16]. Further aspects of this line of research have been published in additional refereed workshops [15, 81, 62].

- A new way of detecting and exploiting correlations among different terms, at local and global scale and for queries as well as peers’ corpora. This joint work by MPII and CTI has been published in the P2P research community’s flagship conference IPTPS 2006 [14] and in the highly selective CIKM conference 2006 [75]. The latter paper won the CIKM 2006 Best Interdisciplinary Paper Award (the conference’s best-paper award, as its mission is to foster integration of DB and IR methods), out of 518 submitted and 81 accepted full papers.

- Extending our earlier work at MPII [7] on understanding LSI and related spectral analysis methods by an algorithm that can effectively extract synonymy, hyponymy, and hyponymy relations from a term-document matrix in a completely unsupervised manner [6].

- Developing a new method for extracting binary relations from Web pages and other text documents, by combining advanced linguistic analyses with statistical learning in an almost-unsupervised manner. This method, developed at MPII and implemented in the open-source software tool LEILA, outperforms all prior methods on various tasks for information extraction [91].
• Extending our earlier work at MPII [96, 97] on efficient algorithms for top-k query processing by new methods for further query acceleration by judiciously scheduling sequential and random index accesses. These results have been published in the prestigious and highly competitive VLDB conference 2006 [8].

• Completing and extending our ongoing work on distributed link analysis with asynchronous collaboration among peers. This joint work by Telenor and UniBo has been submitted for publication [27]. An extended collaboration towards comparing different approaches to distributed link analysis and integrating them in the Minerva testbed has been initiated among Telenor, UniBo, and MPII.

• Extending our initial work [80] on the JXP method for P2P computation of PageRank authority measures, by new techniques for making the computation more light-weight and converge faster. This joint work between MPII and URDLS, which appeared in the prestigious and highly competitive VLDB 2006 conference, also includes a proof of JXP convergence based on Markov-chain state-lumping arguments [82]. Very recently, additional enhancements for combatting and compensating the impact of cheating peers have been developed and incorporated into an extended TrustJXP method [84].

• Extending our initial work at MPII [67] on leveraging query logs and click streams for personalized search result ranking, by the QRank method, which is based on a Markov reward model and can incorporate also negative assessments, both at the level of individual users and entire user communities [68]. In addition, new methods for mining query logs based on IP-level Internet traffic measurements, have been jointly developed by TUM and MPII and published in the highly reputable IMC conference 2006 [55].

2 Main Results

2.1 Query Routing

2.1.1 State of the Art

For P2P search over Web data with ranked retrieval, we are developing the Minerva system [12, 17]. This is one of several projects world-wide (e.g., [1, 34, 85, 92]; see also the book [90] and the references given there) that are pursuing ways of making search-engine technology decentralized, so as to make it ultra-scalable and self-organizing, exploit advanced linguistic and statistical learning models, benefit from individual-user and social-community inputs, and make Web search resilient to manipulation or censorship.

In Minerva, each peer has a full-fledged Web search engine, including a crawler and an index manager. The crawler may be thematically focused or crawl results may be postprocessed so that the local index contents reflects the corresponding user’s interest profile. For collaborative search, the peers are connected by an overlay network based on a distributed hash table (DHT). The DHT also forms the basis of a conceptually global but physically decentralized and scalable directory that contains metadata and statistics about the peers’ contents and quality. Note that, for scalability, the directory is not designed as a page-granularity global Web index, but is limited in size to the number of indexed features (e.g., keywords or topics) times the number of peers. This avoids the pitfalls outlined in [60]. Also note that the pursued P2P Web search includes ranked retrieval and is thus fundamentally much more difficult than Gnutella-style file sharing or simple key lookups via DHTs.

With a user’s highly specialized and personalized “power search engine” most queries should be executed locally, but once in a while the user may not be satisfied with the local results and would then...
want to contact other peers. This is the query routing (or peer-selection) problem, the cornerstone
of the P2P search engine. Although the problem is related to earlier work on metasearch engines
and distributed information retrieval [26, 72], the P2P setting is much more challenging because of
larger scale and high dynamics. A “good” peer to which the user’s query should be forwarded would
have thematically relevant index contents, which could be measured by statistical notions of similar-
ity between peers. To this end, state-of-the-art methods employ term-frequency-based models [26],
statistical language models [65, 66], or cost-based decision models [78, 79]. On the other hand, each
additional target peer for a query should yield novel results that are not yet provided by previously
selected peers or even the local index of the query initiator itself. To this end, we have developed an
overlap-aware query routing strategy that combines content-quality measures with Bloom-filter-based

2.1.2 Statistically Enhanced Query Routing

Our initial work [11] on overlap-aware query routing has been further extended and significantly
enhanced in [74]. Our method aims to optimize a weighted combination of search result quality and
novelty. As the query routing decision made for execution planning is in the critical path of user-
perceived response time, fast estimation of quality-novelty measures is crucial. We have developed
new methods for this purpose, utilizing compact synopses like Bloom filters [20], hash sketches [45]
and min-wise independent permutations [23] in combination with the underlying DHT.

The above considerations show that distributed management of statistical information about peers
and their data collections is a key issue. This involves efficient gathering and dissemination of statistics
as well as estimations for specific purposes. A global measure of particular interest (for query routing
and query result merging) is the document frequency (df) of a keyword, i.e., the total number of
distinct documents in the entire network that contain the keyword. Estimating this number is difficult
because of duplicate documents at different peers. We have developed a highly efficient and accurate
method for this problem, by combining hash sketches and the DHT-based overlay network [16].

An important network-wide statistical estimation problem is to efficiently determine pairs of highly
correlated or anti-correlated keywords, either in queries or in the data. In [14, 75] we have developed
a technique that utilizes the DHT-based infrastructure for an efficient solution, which can piggyback
all necessary message exchanges on the network traffic that is needed for standard query routing and
execution anyway. Awareness of keyword correlations is very useful for more effective query routing
decisions.

The problem that mandates correlation awareness is illustrated by the following example. Con-
sider two- or three-keyword queries such as “Anna Kournikova”, “native American music”, or “PhD
admission”. A standard approach would decompose each query into individual keywords such as
“native”, “American”, and “music”, identify the best peers for each of the keywords separately, and
finally combine them (e.g., by intersection or some form of aggregating the summary scores) in order
to derive a candidate list of peers to which the query should be forwarded. This approach may lead
to mediocre query results as the best peers for the entire query may not be among the top candidates
for the individual keywords. In a worst case scenario, these peers might not have a single data item
that matches all keywords at once. Hence, we miss out on the fact that, for example, “PhD” and
“admission” are statistically correlated in the underlying corpora and that the best matches for the
entire query should exhibit a higher-than-average frequency of both keywords.

We have developed two alternative approaches to this problem: sk-STAT, using the already ex-
isting single-keyword statistics to estimate a peer’s quality for key sets, and mk-STAT, enhancing
the distributed directory that guides the query routing decisions to explicitly include also statistical
information about judiciously chosen sets of multiple keys. Our methods can be used with a large
variety of P2P overlay networks, including DHTs but also arbitrary graph topologies with requests
being routed among peers based on peer-local routing indices. In the DHT case, the statistical in-
formation that drives our query routing covers. While mk-STAT in principle is the more powerful method, it faces the necessity to identify those valuable key sets that are most likely to enable improvements, as it is practically infeasible to build and disseminate statistics for all possible key sets for combinatorial complexity. Instead, the discovery of interesting key sets is initiated by mining locally gathered query logs, to improve the performance of frequently queried key combinations. This discovery phase can optionally trigger an in-depth statistical analysis of the correlations within the peers’ data collections. One of the paper’s novel key contributions is how to make this analysis efficient and scalable. The approach is highly scalable by piggybacking all network communication for gathering and disseminating statistical information on messages that need to be sent between peers anyway (for their regular query traffic). We employ hash sketches (HS) [45] as compact synopses for capturing key- and key-set-specific collection quality, that we combine efficiently for different keys and from different peers in a distributed setting. The information gained is harnessed by the query routing process, utilizing the DHT infrastructure for efficiency, and leads to significantly better peer selection decisions for subsequent queries.

Query routing makes a peer-selection decision when it sees a query, so it contributes to the user-perceived query run-time. This limits the efforts that query routing can afford. For example, it is out of the question to perform extensive computations for comparing statistical language models of different peers. On the other hand, this consideration motivates pre-computing some of the query-routing decisions and encoding them in a “semantic overlay network” (SON) [33, 104]. The SON is an additional overlay network that connects frequently and positively interacting peers as immediate neighbors. For example, a peer that often provides good answers to the queries posed by some other peer should become that other peer’s neighbor. This precomputed SON topology can accelerate query routing by preferring SON neighbors. However, the SON does not necessarily pre-determine the query routing decisions. It should still be possible that a peer explores alternatives to its SON neighbors for a particularly important and difficult query. Also, the SON topology itself should, of course, be dynamically maintained as the data, workload, and behavior of the peers in the network evolve over time.

In our recent work [62] and [83] we have developed new kinds of SON architectures, one based on comparing statistical language models between peers and the other based on peer affinity in terms of the Web pages and links known at different peers. The first approach [62] uses the Kullback-Leibler (KL) divergence between the term-frequency distributions of two peers as a measure of peer similarity and aims to connect the most similar peers as SON neighbors. This is a very natural model, as it results from the standard multinomial generative model for statistical language models. However, it faces the difficulty that the KL divergence is not a metric, so there are no obvious ways of avoiding a comparison of all peer pairs for deriving the best SON topology. We exploit recent results from information theory [41] and embed peers into a metric space with appropriate pruning capabilities. The second approach [83] aims to connect peers such that they are effective neighbors for exchanging complementary information about the Web graph. More specifically, a peer \( y \) is viewed as a “good” neighbor of peer \( x \) if \( y \) has many Web pages that point to pages that are held by \( x \). We speed up the corresponding comparison by encoding Web pages and the outgoing neighbors of Web pages into synopses based on min-wise independent permutations [23].

Yet another approach for building an effective SON it utilize an effective construction for random graphs with certain quality guarantees. Such an approach is being pursued in Workpackage WP6.2 [70, 69], and we plan to explore this option as a means for more flexible query routing, too.

---

2 SON has become an established technical term in the literature. Perhaps, a better characterization would be “statistical overlay network” or “social overlay network”.

3 This approach has immediate benefits for the decentralized link analysis method JXP that will be explained in Subsection 2.4.2.
2.2 Concept-based Search

2.2.1 State of the Art

Lifting search from keywords to semantics concepts and relations can be pursued in various ways. Latent-semantics methods detect dependencies and strong correlations between terms, by spectral decomposition or unsupervised clustering on the term-document matrix of the underlying corpus. A certain degree of information reduction is desired for eliminating noise; this is achieved by mapping the original data onto a lower-rank matrix. LSI [38, 5], pLSI [51, 52], LDA [19], and some forms of statistical language models with background models [64] fall into this category of fully automated methods. Queries are also mapped into the lower-rank space whose dimensions correspond to latent concepts. Search result quality can often be improved this way, but the scalability of these methods is limited, they often require careful tuning of the rank of the latent-concept space, and the quality improvements highly depend on the data and query characteristics, with gains ranging from significant to very marginal.

Explicit knowledge about concepts and semantic relations between concepts is a major alternative, but it requires human work to construct the corresponding knowledge base. Existing thesauri and ontologies such as WordNet [44] can be used to build or at least bootstrap the concept space. This approach yields explicit relations like synonymy, polysemy, hypernymy/hyponymy (aka Is-A relation), holonymy/meronymy (aka Part-Of relation), etc. Co-occurrence statistics gathered on large corpora like Wikipedia or the Web can be used to quantify the “strength” of relations between different concepts. This empowers a search engine to employ various forms of query expansion: adding synonyms to the keywords in a query, adding hyponyms or other strongly related concepts to a query with low recall, or even disambiguating keywords that can have multiple meanings. For these techniques, additional information sources such as relevance feedback, query logs, and click streams can be incorporated and combined with knowledge about semantic concepts. Query expansion has been shown to boost the result quality of difficult queries, but it is also known as a very subtle technique that may possibly lead to result degradation when applied without sufficient care and fine-tuning [18]. Our work [97] has developed a novel, self-tuning way of making query expansion more robust and efficient by adding expansion terms incrementally at query run-time on demand.

Methods that leverage explicit knowledge can only be as good as the quality of their underlying knowledge base. WordNet, as the most prominent example, is very rich and fine-grained in modeling synonymy, polysemy, and hypernymy of abstract concepts such as “corruption”, “capital”, “state”, or “process”. But it is sparse on concrete entities such as geographic locations and organizations, important people like scientists, contemporary politicians, sports and movie stars, music bands, etc. To obtain such knowledge and also relationships between these entities, information extraction (IE) methods can be applied to text documents, Web forms and tables, and other already existing sources. IE combines pattern matching with linguistic analyses, statistical learning, and lookups in existing lexicons (e.g., so-called gazetteers with city names). The results may include false positives, but the accuracy of these methods is often acceptable, depending on the difficulty of the extraction task. Thus, such automatically gathered concepts and relations can be added into a knowledge base and tagged with their corresponding confidence values. Recent work on IE at a larger scale is reported in [4, 42]; broader surveys include [31, 36].

2.2.2 Unsupervised and Almost-Unsupervised Extraction of Semantic Relations

In our work [7] we developed a new framework for explaining how spectral decomposition and particularly LSI achieves its latent-concept behavior. This framework characterizes the relatedness of two terms by varying the rank of the LSI target space and studying the term-term similarity when transformed into the low-rank space. Different shapes of the relatedness curves, as functions of the rank, allow us to identify, with high confidence, synonyms and other pairs of highly related terms.
This enables techniques for document expansion so as to virtually include, with lower weight, such related terms in a document although the term does not literally appear.

The recent paper [6] generalizes our prior work by showing that the spectral decomposition can be used to extract even a term taxonomy with hierarchical hypernymy/hyponymy relations in an unsupervised manner. A mathematical criterion is developed for distinguishing between between four kinds of term-term relations in a given document collection: unrelated (e.g., car - fruit), symmetrically related (e.g., car - automobile), asymmetrically related with the first term being more specific than the second (e.g., banana - fruit), and asymmetrically related with the first term generalizing the second (e.g., fruit - banana). The method has been tested on the ODP (open directory project) corpus, and has been able to reconstruct the OPD hierarchy of category names without any training input.

An alternative way towards concept-based semantic search capabilities is to to exploit explicitly represented knowledge in the form of ontologies and thesauri. The latter can be constructed, to some extent, from hand-crafted knowledge such as WordNet, but a larger-scale and self-maintaining, promising approach is to automatically learn the concepts and relations for an ontology directly from rich text corpora such as Wikipedia and other Internet sources. This requires a combination of linguistic analysis, pattern matching, and statistical learning to identify, for example, person names or locations and to extract instances of binary relations such as located-at (city, river), born-in (person, place), plays-instrument (person, instrument), or the generic is-instance-of (entity, concept).

For finding and extracting information on binary relations we pursue a novel approach based on a link-grammar representation of natural-language sentences and an SVM-based statistical learner for determining robust, generalizable linguistic patterns. This approach is implemented in the LEILA prototype system [91]. The method is almost unsupervised by starting with merely a small set of user-provided positive examples such as (Paris, Seine), (Calcutta, Ganges), (London, Thames) and either explicit or token-based negative examples such as (New York, Mississippi) or (<proper noun>, <number>). The system then automatically finds candidate patterns such as <river> flows through <city>, <city> is located on the banks of the <river>, or <city> not only offers many cultural attractions, but visitors can also enjoy tubing or swimming in the <river>. These patterns cannot be directly applied as they do not generalize well and would lead to many false positives. Instead, we run a dependency parser on the individual sentences, compute a characteristic feature vector from the resulting graph representation of each sentence, and feed these vectors into an SVM classifier. For higher recall, we also perform anaphora resolution across neighboring sentences. The method outperforms competitors in terms of the F1 measure (i.e., the harmonic mean of precision and recall). We are currently working on further extensions and better scalability.

Automatic classification of text documents is another important building block for many forms of intelligent data organization, for example: assigning Web pages to topics of a hierarchical directory, filtering news feeds or thematic subscriptions in digital libraries, steering a focused crawler towards Web regions that are more likely to contain pages that fall into the user's specific interest profile. Classifiers are models over high-dimensional feature spaces that are derived from supervised learning, with positive and negative examples as training input. The underlying methods for statistical learning such as SVM or Bayesian methods have become fairly mature, and the practical bottleneck is typically the scarceness of training data, because compiling training samples is a laborious and time-consuming human activity.

We have investigated richer feature models and contextual features of training samples so as to improve classification accuracy with relatively small training sets. One approach is to consider neighbors of text documents in environments with many cross-references such as Web links (but covering also settings such as book, music, or blog recommendations or citation graphs for publications). The graph-based classifier presented in [3] builds on the theory of Markov Random Fields, and additionally develops new techniques for enhanced robustness. This method outperforms purely text-based classifiers and also all previously proposed link-aware classifiers. Finally, we investigate new models
that map text features onto semantic concepts in an ontology and leverage the concepts as additional information. In [53] we have developed a generative probabilistic model that uses concepts as a latent-variable layer and an efficient EM-based parameter-estimation procedure. In contrast to previous work on latent semantic models, concepts are explicit and directly interpretable by humans and greatly simplify the model-selection problem of choosing the right number of latent dimensions. In combination with a transductive learning procedure that leverages the richer feature space of unlabeled background corpora, we have built a highly effective classifier that excels especially when training data is very scarce. As an example, we can train a text classifier for a topic such as biology on a few dozen training documents from the Web, use Wikipedia as an unlabeled background corpus, and a subset of WordNet to construct the ontology; this methodology yields a classifier that outperforms the very best purely text-based methods by a substantial margin.

2.3 Top-k Query Processing
2.3.1 State of the Art

For Web search, ranking query results by their estimated relevance (according to an appropriate scoring model) is a standard technique for satisfactory user experience. It is also an efficiency issue: the query processor only needs to compute the $k$ best results (e.g., with $k$ set to 10 or 100) ideally without having to scan all candidate results that match the query keywords (i.e., in sublinear time).

Such top-$k$ queries operate on index lists for a query's elementary conditions and aggregate scores for result candidates. One of the best implementation methods in this setting is the family of threshold algorithms (TA) [43, 47, 77], which has been shown to be asymptotically instance-optimal when the number of index lists is viewed as a constant. TA methods aim to terminate the index scans as early as possible based on lower and upper bounds for the final scores of result candidates. Advanced methods from this family [29, 39, 48, 61, 71, 76, 96] perform sequential disk accesses for sorted index scans, but also have the option of performing random accesses to resolve score uncertainty.

When lifting Web search from keywords to concepts, top-$k$ queries may also become more complex and thus pose efficiency challenges. For example, when we use query expansion based on explicit concept spaces, queries have to operate over a much larger set of index lists; when using document expansion, the average length of index lists increases significantly. These complications do not necessarily affect the asymptotic time complexity of TA, but they do cause major differences in practical run-times. Our methods outlined below have addressed these performance issues.

2.3.2 Index-Access Optimized Algorithms

Advanced TA-style algorithms require scheduling for two kinds of accesses: 1) the prioritization of different index lists in the sequential accesses, and 2) the decision on when to perform random accesses and for which candidates. The prior literature has studied some of these scheduling issues, but only for each of the two access types in isolation [29, 71]. Our recent work [8] has taken an integrated view of the scheduling issues and developed a novel method, coined IO-Top-$k$ (index-access optimized top-$k$ algorithm), that outperform prior methods by a large margin. Our main contributions are new, principled, scheduling methods based on a Knapsack-related optimization for sequential accesses and a cost model for random accesses. The methods can be further boosted by harnessing probabilistic estimators for scores, selectivities, and index list correlations.

Compared to a straightforward implementation of the best TA variants, our algorithms and implementation techniques – priority queue management, judicious scheduling, and probabilistic pruning – yield a performance improvement of two to three orders of magnitude. For example, keyword queries like "kyrgyzstan united states relation" on the 25-million-documents TREC Terabyte corpus can be answered in 10 milliseconds and need to scan only 2 percent of the entries of the 4 relevant index lists. In extensive performance experiments with a variety of large datasets we have often observed that
our method is at most 25 percent above a lower bound for the execution cost of any top-$k$ algorithm; so IO-Top-$k$ is empirically near-optimal.

2.4 Extended Link Analyses

2.4.1 State of the Art

Authority analysis on link graphs typically builds on spectral methods, i.e., computing Eigenvalues and Eigenvectors and characterizing the corresponding decomposition of a graph’s adjacency matrix. PageRank [22] and HITS [58] are the seminal methods for link analysis on the Web graph, which can be used to derive authority scores for Web pages that reflect query-independent importance and community endorsements. [21] and [59] provide good in-depth surveys of follow-up and related work, including their mathematical foundations.

PageRank computes the principal Eigenvector of a matrix that is derived from the Web link graph with an additional feature of uniform random jumps. PageRank vectors, characterizing the prestige of Web pages, are usually computed by the iterative Jacobi method (aka. power iteration) with fast convergence. The same mathematical algorithm has been applied to computing TrustRank [49], SpamRank [10], and other methods of social authority and (positive or negative) impact (e.g., [9, 57, 50]). A general framework for different types of trust and distrust propagation in a graph of Web pages, sites, or other entities is introduced in [46].

A P2P network is a natural habitat for “social search” that leverages community assessments. A simple form of such community input are link analysis methods like PageRank or HITS. But these are centralized algorithms with very high memory demand, and their distributed variants assume that the underlying Web graph can be nicely partitioned among sites. Recently, various distributed versions of link analysis methods have been proposed (e.g., [2, 24, 37, 56, 87, 98, 102]), but they all assume that the global graph is partitioned into disjoint fragments across Web hosts or peers.

2.4.2 Decentralized Methods for Authority and Trust Assessment

In [27] we have developed a comprehensive framework for fully distributed versions of a wide variety of link analysis methods including PageRank but also variants with undirected graphs or un-normalized transition matrices. When propagating authority mass from a node to its successor, our distributed methods use message passing whenever the two connected nodes reside on different computers. These novel methods do not require any global coordination among peers, other than an agreed updating rate. In particular, global synchronization of the individual updates is not needed. The methods have been shown to be robust against high levels of noise, such as message delay, message delay jitter, and message loss. The fine-grained message passing of our approach is appropriate, and an intriguing alternative to centralized link analysis, when peers are willing to collaborate and the underlying network has relatively short latency and very good connectivity.

A different setting arises when we emphasize the autonomy of peers, like in the Minerva prototype where peers can crawl Web fragments and gather their own local content at their discretion. For such an environment, we have developed the JXP [82] algorithm, a new method for dynamically computing, in a decentralized P2P manner, global authority scores when the Web graph is spread across many autonomous peers. The peers’ graph fragments may overlap arbitrarily, and peers are (a priori) unaware of other peers’ fragments. With JXP, each peer computes the authority scores of the pages that it has in its local index, by locally running the standard PageRank algorithm. A page may be known and indexed by multiple peers, and these may have different scores for that same page. A peer gradually increases its knowledge about the rest of the network by meeting with other, randomly chosen, peers and exchanging information, and then recomputes the PageRank scores for its pages of interest. The local computations are very space-efficient (as they require only the local graph and the authority-score vector), and fast (as they operate on much smaller graph fragments.
than a server-side global PageRank computation). We have proven, using the theory of Markov-chain aggregation/disaggregation, that the JXP scores do indeed converge to the same values as a global PageRank computation on the full Web graph [82].

The open and anonymous nature of P2P networks opens doors to manipulation of information and quality ratings. For example, in the JXP algorithm we may face cheating peers that aim to distort the global authority values by providing manipulated data in the peer meetings. In [84] we have developed novel techniques that aim to detect and compensate such cheating. Our approach to this problem enhances JXP with a variety of statistical techniques for detecting suspicious behavior, and we integrate the JXP rankings with reputation-based techniques. Our method, coined TrustJXP, is again completely decentralized, and we demonstrate its viability and robustness in experiments with real Web data.

2.5 Personalized Search and Ranking

2.5.1 State of the Art

Information about user behavior is a rich source to build on. This includes relatively static properties like bookmarks or embedded hyperlinks pointing to high-quality Web pages, but also dynamic properties inferred from query logs and click streams. For example, suppose a user clicks on a specific subset of the top 10 results returned by a search engine for a query with several keywords, based on having seen the summaries of these pages. This implicit form of relevance feedback establishes a strong correlation between the query and the clicked-on pages. When a user does not click on any of the top 10 results for a given query and rather chooses to rephrase the query using different keywords, this may be interpreted as negative feedback on the relevance or quality of the initial results. Exploiting this kind of user-behavior information can help to improve search result quality for individuals as well as entire communities by adding more "cognitive" elements to the search engine.

The literature contains interesting work in the outlined direction (e.g., [99, 35, 54, 89, 93]), but is still far from a comprehensive and fully convincing solution. [99] applied clustering to query logs in order to identify frequently asked questions regardless of keyword variations. [35] exploited knowledge about similar queries for query expansion. [54] analyzed the statistical significance of interpreting non-clicked top-10 results as negative feedback and found that only non-clicks that are followed by a click-on document at a lower rank should be viewed as relative user preferences. [86] used statistical learning to identify query chains of reformulated queries for the same information demand by the same user. [89, 93] enhanced statistical language models for query-result ranking by background "knowledge" obtained from a user's long-term query log and click stream.

Link analysis methods such as Google's PageRank are often based on a random-walk model over the Web graph. Our prior work [67] generalizes the notion of a "random surfer" into a "random expert user" by enhancing the underlying Markov chain to incorporate also query nodes and transitions from queries to query refinements as well as clicked-on documents. Transition probabilities are derived from the statistical analysis of query logs and click streams. In addition, virtual links can be created from textual references using named-entity recognition techniques. The resulting Markov chain converges to stationary authority scores that reflect not only the link structure but also the implicit feedback and collective human input of a search engine's users. We experimentally showed that this approach could significantly improve the user-perceived quality of search results for advanced queries.

2.5.2 Exploiting User and Community Behavior for Personalized Ranking

Our approach is based on a Markov-chain model with queries as additional nodes, additional edges that capture query refinements and result clicks, and corresponding transition probabilities. Similarly to Google's PageRank, the computation of stationary visiting probabilities yields the behavior-aware
authority ranking. Our original model could not express negative feedback, as probabilities are non-negative and L₁-normalized. To capture and exploit also negative assessment such as assigning trust levels to Web pages (e.g., marking a Web page as spam, low-quality, out-of-date, or untrusted), we are pursuing several extended approaches, one of which is based on a Markov reward model [95] where the assessment part is uncoupled from the random walk in the extended Web graph. A page receives a specific lump reward each time a transition is made to it, and this reward depends on the transition's source and target and will be derived from the query-log and click-stream information as well as explicit page assessments. The reward itself can be positive or negative. The most interesting measure in such reward models is the expected earned reward per time-step. Preliminary results with this approach are presented in [68].

The evaluation of such personalization models requires experiments with real users' behavioral traces. To this end, TUM and MPII have jointly developed a novel way of gathering query logs and click streams by monitoring and analyzing network traffic at the IP level. This work has also led to a finite-state transition model that captures the user behavior in query sessions [55].

3 Dissemination of Results

The research in this WP has led to 19 publications [3, 8, 6, 14, 15, 16, 17, 27, 53, 55, 62, 68, 74, 75, 81, 82, 83, 84, 91] in international conferences and refereed workshops in the year 2006 (including a few papers that will appear in early 2007). The publication list includes 7 papers in the prestigious, highly selective first-rate conferences VLDB [8, 82], SIGIR [3], EDBT [74], CIKM [75], KDD [91], and PKDD [53], all of which have acceptance ratios of 1:5 or higher. The paper [75] won the best-paper award at CIKM 2006. An overview of the research in SP6, jointly co-authored by the five partners MPII, CTI, UPB, Telenor, and UBologna, has also been presented at the European Conference on Complex Systems in November 2005 in Paris [100, 101]. All referenced papers are available on request from the coordinating site UPB.


Selected results are implemented in the Minerva software, a prototype testbed for P2P Web search. A first release of Minerva has been made available as open source software as Deliverable D.6.5.2 in July 2005. Minerva has been demonstrated at two major conferences, VLDB 2005 in Trondheim and Middleware 2005 in Grenoble, and received great attention and very positive feedback [12, 13]. A new, significantly extended, version of Minerva will be demonstrated at the CIDR 2007 conference in Asilomar [17]. This version is also released as DELIS software deliverable D.6.5.3 in January 2007.

References


[8] Holger Bast, Debapriyo Majumdar, Ralf Schenkel, Martin Theobald, Gerhard Weikum: IO-Top-k: Index-access Optimized Top-k Query Processing. International Conference on Very Large Data Bases (VLDB), Seoul, Korea, 2006. (DELIS-TR-0324)


[53] Georgiana Ifrim, Gerhard Weikum: Transductive Learning for Text Classification using Explicit Knowledge Models, 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), Berlin, Germany, 2006. (DELIS-TR-0444)

[54] Thorsten Joachims, Laura A. Granka, Bing Pan, Helene Hembrooke, Geri Gay: Accurately Interpreting Clickthrough Data as Implicit Feedback. ACM SIGIR International Conference on Research and Development in Information Retrieval, Salvador, Brazil, 2005


[74] Sebastian Michel, Matthias Bender, Peter Triantafillou, Gerhard Weikum: IQN Routing: Integrating Quality and Novelty for Web Search. Int. Conf. on Extending Database Technology (EDBT), Munich, Germany, 2006. (DELIS-TR-0290)


[77] Shuang Nepal, M.V. Ramakrishna. Query processing issues in image (multimedia) databases. International Conference on Data Engineering (ICDE), Sydney, Australia, 1999.


[91] Fabian Suchanek, Georgiana Iftim, Gerhard Weikum: Combining Linguistic and Statistical Analysis to Extract Relations from Web Documents, ACM International Conference on Knowledge Discovery and Data Mining (KDD), Philadelphia, USA, 2006. (DELIS-TR-0453)


