Project Number 001907

DELIS
Dynamically Evolving, Large-scale Information Systems

Integrated Project
Member of the FET Proactive Initiative Complex Systems

Deliverable D6.1.1

Intermediate Report on Models of Collaborative Peer Behavior and Strategies for Global Search
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 2004
Actual submission date: January 2005
Re-submission date: April 2005
Version: 2
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

One of the major goals 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. With such a 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. 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. Both query routing and the formation of “semantic overlay networks” 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. Note that this notion of Web search includes ranked retrieval and thus is fundamentally more difficult than Gnutella-style file sharing or simple key lookups via distributed hash tables. Further note that, although query routing in P2P Web search resembles earlier work on metasearch engines and distributed information retrieval, it is much more challenging because of the large scale and the high dynamics of the envisioned P2P system with thousands or millions of computers and users.

So far the following results have been achieved on the research issues outlined above:

- A novel query routing strategy for P2P Web search has been developed at MPII. It utilizes peer bookmarks to identify promising peers that are likely to contribute relevant results for a given query but have small overlap with the local index contents of the query originator [6].

- An efficient algorithm for top-k query processing, coined Prob-sorted, has been developed at MPII. It utilizes score distribution statistics for score predictions to prune result candidates early, thus outperforming the so far best known methods by an order of magnitude with small and acceptable loss in precision and recall [33].

- A distributed top-k algorithm, coined KLEE, has been developed in joint work of MPII and CTI. KLEE includes optimizations for network latency and bandwidth consumption and gains dramatic performance advantages over the prior work [25].

- Retrieval methods that make use of the concepts underlying large collections of text have been investigated and improved at MPII. The concepts are derived from the raw data by spectral analysis or other matrix decomposition techniques, without training or dictionaries. Novel insights have been found into what makes such approaches work in practice. Various new methods are derived, which improve search results, have a better theoretical foundation, and are well-suited for a P2P setting [3, 4, 5].

- A new method, coined QRank, for incorporating query-log and click-stream information into Web page ranking has been developed at MPII. It is based on a Markov model that extends the traditional Page-Rank approach by query nodes and edges for query refinement, query-result clicking, and node-pair similarity [24].

- Link analysis techniques have been studied at Telenor, with new insights on the subtle but important differences between directed vs. undirected graphs and fully connected vs. multi-component graphs. (A paper on these results is in preparation.)
2 Main Results

2.1 Query Routing

2.1.1 State of the Art

Query routing aims to select the most promising peers among a large set of candidates, for executing a given query. This problem is also known as database selection or resource selection in the information retrieval (IR) literature. However, collaborative P2P search is substantially more challenging than the traditional setup for distributed IR over a small federation of sources such as digital libraries, as these prior approaches mediate only a small and rather static set of underlying nodes, as opposed to the large scale and high dynamics of a P2P system.

The literature on P2P request routing has mostly focused on simple key lookups. Even techniques that consider multi-dimensional keyword queries and the formation of “semantic overlay networks” (e.g., [35, 14, 1]) disregards the need for ranked retrieval based on relevance scoring (as opposed to Boolean retrieval where data items either satisfy search conditions or not but are not further discriminated regarding their relevance). On the other hand, prior research on distributed information retrieval and metasearch engines [10, 26] has addressed the ranking of data sources and also the reconciliation of search results from different sources. relevant, too. GIOSS [18] and CORI [9] are the most prominent distributed IR systems, but neither of them aimed at very-large-scale, highly dynamic, self-organizing P2P environments (which were not an issue at the time these systems were developed).

Recent approaches to query routing [23, 28, 15] consider larger federations of data sources, but they use computationally intensive techniques based on statistical language models for assessing a peer’s quality with regard to a given query. It is an open issue to what extent such techniques scale and can copy with high dynamics.

2.1.2 New Approaches

The rationale for the query routing strategy developed in DELIS is based on the following three observations:

1. The query initiator should prefer peers that have similar interest profiles and are thus likely to hold thematically relevant information in their indexes.

2. On the other hand, the query should be forwarded to peers that offer complementary results. If the remote peer returns more or less the same high-quality results that the query initiator already obtained from its own local index, then the whole approach of collaborative P2P search would be pointless.

3. Finally, all parties have to be cautious that the execution cost of communicating with other peers and involving them in query processing is tightly controlled and incurs acceptable overhead.

We address the first two points by defining the benefit that a remote peer offers for the given query to be proportional to the thematic similarity of that peer and the query initiator and inversely proportional to the overlap between the two peers in terms of their local index contents. To limit the overhead of estimating these measures, we use the Kullback-Leibler divergence between the bookmark documents of the two peers and the overlap in their bookmarks as the basis for estimating benefit. Here we view the index contents of a peer as being generated by the peer’s bookmarks, which served as seeds for the peer’s Web crawls and possibly also as training data for a thematically focused crawler [13, 30]. We reconcile this notion of benefit with the third of the above observations by considering the benefit/cost ratio of peers, where cost is estimated based on tracking the utilization and resulting response time of different peers.
The outlined strategy as well as various alternative strategies from the prior literature have been implemented in the Minerva testbed developed at MPII [7, 8]. Minerva uses a Chord-style distributed hash table (DHT) [31] as an overlay network. Each peer has its own search engine, along with a crawler and a local index. Peers post statistical summaries to the overlay network; they can also post bookmarks and information about their query logs and click streams at their discretion. All this information is managed as a decentralized directory, using the DHT, and is utilized by the query routing strategies. Minerva is implemented in Java and fully operational; preliminary measurements have already been performed and we plan on more comprehensive experimentation. The system will serve as a backbone for testing and evaluating algorithms developed by all DELIS partners in SP6.

2.2 Efficient Top-k Query Processing

2.2.1 State of the Art

Among the strong results achieved in DELIS SP6 so far are highly efficient algorithms for top-k queries that are suitable for a P2P environment. Top-k query processing is a fundamental cornerstone for similarity search on multimedia data, ranked retrieval on text and semi-structured documents in digital libraries and on the Web, network and stream monitoring, collaborative recommendation and preference queries, and ranking of query results on structured data sources in general. It aggregates scores for different search terms or attribute values using a monotonic aggregation function such as weighted summation, and returns the top-ranked data items as the query result. Scores are usually precomputed features of different aspects of a data item, e.g., color distributions in images, access frequencies in Web server logs, or word occurrence statistics in text documents. The state-of-the-art algorithm for top-k queries on multiple index lists, each sorted in descending order of relevance scores, is the Threshold Algorithm, TA for short [17, 19, 27]. It is applicable to both structured data such as product catalogs and text documents such as Web data. In the latter case, the fact that TA performs random accesses on very long, disk-resident index lists (e.g., all URLs or document ids for a frequently occurring word), with only short prefixes of the lists in memory, makes TA much less attractive, however.

In such a situation, the TA variant with sorted access only, coined NRA (no random accesses), stream-combine, or TA-sorted in the literature, is the method of choice [17]. TA-sorted works by maintaining lower bounds and upper bounds for the scores of the top-k candidates that are kept in a priority queue in memory while scanning the index lists. The algorithm can safely stop when the lower bound for the score of the rank-k result is at least as high as the highest upper bound for the scores of the candidates that are not among the current top-k. Unfortunately, albeit theoretically instance-optimal for computing a precise top-k result [17], TA-sorted tends to degrade in performance when operating on a large number of index lists, which happens when user queries are automatically expanded based on ontologies, user profiles, or relevance feedback.

2.2.2 The Prob-sorted Algorithm

Statistics about the score distributions in the various index lists and some probabilistic reasoning help to overcome this efficiency problem and gain performance. In TA-sorted a top-k candidate \(d\) that has already been seen in the index lists in \(E(d) \subseteq [1..m]\), achieving score \(s_j(d)\) in list \(j\) \((0 < s_j(d) \leq 1)\), and has unknown scores in the index lists \([1..m] - E(d)\), satisfies:

\[
\text{lowerb}(d) = \sum_{j \in E(d)} s_j(d) \leq s(d) \leq \sum_{j \in E(d)} s_j(d) + \sum_{j \notin E(d)} \text{high}_j = \text{upperb}(d)
\]

where \(s(d)\) denotes the total, but not yet known, score that \(d\) achieves by summing up the scores from all index lists in which \(d\) occurs, \(\text{lowerb}(d)\) and \(\text{upperb}(d)\) are the lower and upper bounds of \(d\)'s score, and \(\text{high}_j\) is the score that was last seen in the scan of index list \(j\), upper-bounding
the score that any candidate may obtain in list \(j\). A candidate \(d\) remains a candidate as long as \(\text{upperb}(d) > \text{lowerb}(\text{rank-k})\) where \(\text{rank-k}\) is the candidate that currently has rank \(k\) with regard to the candidates’ lower bounds (i.e., the worst one among the current top-k). Assuming that \(d\) can achieve a score \(\text{high}_j\) in all lists in which it has not yet been encountered is conservative and, almost always, overly conservative. Rather we could treat these unknown scores as random variables \(S_j\) \((j \not\in E(d))\), and estimate the probability that \(d\)'s total score can exceed \(\text{lowerb}(\text{rank-k})\). Then \(d\) is discarded from the candidate list if

\[
P[\text{lowerb}(d) + \sum_{j \not\in E(d)} S_j > \text{lowerb}(\text{rank-k})] < \delta
\]

with some pruning threshold \(\delta\). Technically, this score prediction requires computing the convolution of the score distributions in the yet to be scanned parts of the index lists. This can be implemented, for example, using histograms, fitting appropriate parametric distributions such as Poisson mixes, or using Laplace transforms of the underlying score distributions and Chernoff-Hoeffding bounds for the tail of the convolution. Figure 1 illustrates the probabilistic score predictor for early candidate pruning.

![Inverted Index](image)

**Figure 1:** Probabilistic score predictor for early candidate pruning

This probabilistic interpretation makes some small, but precisely quantifiable, potential error in that it could dismiss some candidates too early. Thus, the top-k result computed this way is only approximate. However, the loss in precision and recall, relative to the exact top-k result using the same index lists, is stochastically bounded and can be set according to the application’s needs. A value of \(\delta = 0.1\) seems to be acceptable in most situations. Details of this \(\text{Prob-sorted}\) method can be found in [33]. Experiments with the TREC-12 .Gov corpus and the TREC-13 Terabyte corpus (http://trec.nist.gov/), the INEX benchmark for XML information retrieval (http://inex.is.informatik.uni-duisburg.de:2003/), and the IMDB data collection (www.imdb.com) have shown that such a probabilistic top-k method gains about a factor of ten (and sometimes more) in run-time compared to a highly tuned version of TA-sorted.

The algorithm for approximate top-k queries with probabilistic guarantees is a versatile building block for ranked retrieval on text and semistructured data such as Web or XML documents. In combination with ontology-based query relaxation, for example, expanding a phrase query like \(\sim \text{“top-k query”}\) into \(\text{“top-k query” or “ranked retrieval” or “score aggregation”}\), it can add index lists dynamically and incrementally, rather than having to expand the query upfront based on thresholds. To this end, the algorithm considers the ontological similarity \(\text{sim}(i,j)\) between concept \(i\) in the original query and concept \(j\) in the relaxed query, and multiplies it with the \(\text{high}_j\) value of index list \(j\) to obtain an upper bound for the score (and characterize the score distribution) that a candidate can obtain from the relaxation \(j\). This information is dynamically combined with the probabilistic
prediction of the other unknown scores and their sum. The incremental algorithm outperforms the traditional techniques for query expansion by a factor of 3 to 50 in run-time; at the same time, it avoids the danger of topic drifts caused by over-expansion and eliminates the need for tuning expansion thresholds. The Prob-sorted algorithm has been implemented in the Minerva testbed for P2P Web search that is being developed within DELIS SP6 [6, 7].

2.2.3 Distributed Top-k Algorithms for P2P Systems

Minerva can run distributed versions of both TA-sorted and the Prob-sorted algorithm. A query is collaboratively processed by a set of peers each of which holds one or more index lists for the search terms or attribute values in a query. The query-initiating peer serves as a per-query coordinator and aggregates information about top-k candidates. While such algorithms are efficient in terms of the peers’ local resource consumption, they do not pay sufficient attention to the communication costs of the computation.

In joint work of CTI Patras and MPII Saarbrücken, a new family of algorithms, coined KLEE [25], has been developed to address the networking costs of top-k query algorithms. KLEE aims to minimize network latency, network bandwidth consumption, and the local work of the participating peers. To this end, it proceeds in a fixed number of phases to ensure bounded latency. In this regard, we follow the recent work of [12], but we differ significantly in the phases themselves and introduce various novel considerations.

The first phase of KLEE gathers an initial set of top-k candidates from the peers’ index lists and derives a crude threshold for the final top-k result. Along with the candidates, peers send summary information in the form of score-distribution histograms and Bloom filters for data items that locally fall into high-score histogram cells. In the second phase the coordinator performs a benefit/cost estimation for a possible additional message round that would collect further Bloom-filter information to improve the knowledge about top-k candidates. In the optional third phase this information is sent to the coordinator, which in turn prepares a refined list of candidates. In the last phase the peers are requested to send the missing scores for all data items in the candidate list above some lower bound of relevant scores. Because of the approximate nature of histograms and Bloom filters, KLEE computes an approximate top-k results, but similarly to the Prob-sorted algorithm describe above, the loss in precision and recall is small and controllable.

KLEE has been implemented in the Minerva testbed and intensively evaluated on various real-life datasets and query benchmarks. It outperforms both the distributed version of TA-sorted with batching and the TPUT method of [12] by one to two orders of magnitude when three or more peers participate in a query. This impressive performance gain is achieved by reducing both the network bandwidth consumption and the local work at the index-scanning peers. Precision and recall are 80 percent or higher, and the score and rank error measures indicate that the approximate top-k results are as good as the exact ones from a user acceptance viewpoint. This demonstrates the advantages of KLEE’s design for flexible control over different cost and query-result quality metrics, with excellent performance in terms of the quality/cost ratio.

2.3 Concept-based Retrieval

2.3.1 State of the Art

Top-k queries usually operate on what is known as the bag-of-words or vector space model, where documents as well as queries are represented as vectors, with each dimension corresponding to a word or term that occurs in the document collection. This space may be enhanced by semantic information, as given, for example, by XML tags, but it typically stays on a term-centric basis.

One effective way to further boost search result quality is to switch to a representation of queries and documents in terms of the semantic concepts underlying a collection. Concept-based search is able
to detect similarities that are beyond direct term matches, e.g., a query on “large-scale information systems” would also return a document entitled “DELIS research results”, even when it mentions none of the query words explicitly. The methods we consider could easily detect this similarity from the observation that in sufficiently many other documents the acronym DELIS will be mentioned together with its verbose form. Our research is concerned mainly with such unsupervised concept-based search, where concepts are derived from the raw data itself, without training or dictionaries.

A major obstacle in integrating concept-based search in a P2P scenario is that its foundations are still little understood even in the centralized case. A confusing variety of methods exist, each of which works well on particular kinds of collections, but none of which can be called a clear winner. For example, latent semantic indexing (LSI), the grandmother of many concept-based retrieval schemes, has just recently been adapted for a P2P setting [32], although its usefulness for large document collections is questionable already in the centralized case.

Our research goal within DELIS is hence twofold [3]: (i) to develop practically useful algorithms for concept-based search that have a strong theoretical foundation, and (ii) to bring these algorithms to the P2P world.

2.3.2 Concept-Based Retrieval Based on Spectral Analysis

One of the most prominent approaches to unsupervised concept-based retrieval is to derive the concepts from a spectral analysis of the term-document matrix, that is, from some sort of eigenvector computation. These include LSI and its many variants, and Google’s PageRank can be viewed as an eigenvector analysis, too. But even for these methods, despite their mathematical flavor, it it still little understood under which circumstances they work well and why.

We have made major advances in this respect. In [4], we work out the essence of spectral retrieval with a number of surprising insights. One result is a simple parameterless scheme that outperforms all of the previous spectral retrieval schemes that commit themselves to a fixed number of dimensions (the concepts). We achieve this result, even when giving the fixed-dimension schemes the unfair advantage that they may try every possible number of concepts and every one from a broad selection of standard term or document normalizations, and then pick the one which gives their best result.

Our work is built on the observation that all spectral retrieval schemes, no matter which variant, essentially work by performing a document expansion: each term pair is (implicitly) assigned a relatedness score and a document will match a query well to the extent that the document contains terms which are, by these scores, related to terms in the query. Note that if each term were related only to itself, this would result in retrieval according to the standard vector space model.

Our main trick is the following: instead of looking at the relatedness scores for all term pairs for a fixed number of concepts — this is what all previous methods implicitly do — look at the relatedness score of a fixed term pair for all possible number of concepts. Plotting these scores gives us, for each term pair, what we call its synonymy graph. These synonymy graphs turn out to be key to understanding how and why spectral retrieval works.

We find that what distinguishes intuitively related from intuitively unrelated terms is not any relatedness score for a particular fixed dimension, but the shape of the synonymy graph. Graphs for related terms have a positive gradient up to some point, then drop, and finally go up again. Graphs for unrelated terms fluctuate around zero, with most (but not necessarily all) of their absolute values at least an order of magnitude below those reached by graphs for related terms. We are able to derive these properties from a mathematical model, and these insights are confirmed by extensive experiments.

The synonymy graphs reveal that no fixed choice for the number of concepts can be good for all term pairs. An extreme, but not untypical example is given in Figure 2(a) and (b). Both term pairs are highly related — indeed, one pair being the singular and plural of the same word and the other pair forming a common phrase, they are related in the strongest sense possible — and this
relatedness indeed shows in the form of the synonymy graphs. But any fixed choice of dimension can give a high relatedness score to only one of the two term pairs.

The synonymy graphs give us new insights into the effects of the various (heuristic) data normalizations, which are known to be crucial for good retrieval performance. From all these insights, we finally derive the aforementioned parameterless algorithm that is superior to all of the previous fixed-dimensions schemes even at their best parameter settings. Together with our theoretical findings this gives strong evidence that we indeed identified the major principle that makes spectral retrieval work.

2.3.3 Insights from Viewing Ranked Retrieval as Rank Aggregation

In [5], we take a broader perspective and study the common principles behind ranked retrieval in general. We view various methods in a unifying manner, namely as processes of combining query-independent rankings that were precomputed for certain attributes. In particular, we apply this view to various established concept-based retrieval schemes. While the research described in the previous subsection assumes linear mappings from term to concept space, here also non-linear methods are included. We pay particular attention to one of the most prominent and promising non-linear concept-based retrieval schemes, named probabilistic latent semantic indexing (PLSI) [20].

Non-linear methods have fundamental advantages over linear schemes; for example, they have the ability to deal with polysems, which schemes based on a linear mapping from term to concept space can only do to a limited extent. This has been vaguely commented on by various authors; in [5], we turn it into a precise statement.

It seems that the extended capabilities of non-linear methods do not come for free. Indeed, and not surprisingly, the underlying optimization problems are significantly harder than for linear methods — for a characterization of the differences see our research within SP3, the subproject on large-scale optimization. The preprocessing time required for a method like PLSI is an order of magnitude higher than for a method like LSI. For both approaches, the preprocessing is an inherently global computation, which on one hand is gives these methods their power, but on the other hand makes an integration into a distributed P2P setting difficult.

In [5], we replace PLSI’s compute-intensive part by a simple set of document rankings, which need not even be precomputed. The trick lies in an unusual choice for the attributes. On a number of test collections, we find the retrieval performance for this new method to be at least as good and sometimes even better than that of PLSI and a number of other non-linear matrix decomposition schemes. Another advantage of the new method is that it is much more suitable for a distributed P2P setting. We are currently working on the integration of our method into the Minerva Testbed,

Figure 2: The synonymy graph for two pairs of related terms (a) + (b), and for a pair of unrelated terms (c), all from a collection of computer science abstracts. On the x-axis the number of concepts is varied, the y-axis gives the relatedness score.
which is being developed in DELIS WP6.6. Preliminary experiments have shown promising results.

2.4 Exploiting Query Logs and Click Streams

Information about user behavior is a rich source to build on. This could include 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, Google’s PageRank views a Web page as more important if it has many incoming links and the sources of these links are themselves high authorities. This rationale can be carried over to analyzing and exploiting entire surf trails and query logs of individual users or an entire user community. These trails, which can be gathered from browser histories, local proxies, or Web servers, capture implicit user judgements. 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. Further suppose that the user refines a query by adding or replacing keywords, e.g., to eliminate ambiguities in the previous query. Again, this establishes correlations between the new keywords and the subsequently clicked-on pages, but also, albeit possibly to a lesser extent, between the original query and the eventually relevant pages.

Observing and exploiting such user behavior could be a key element in adding more “semantic” or “cognitive” quality to a search engine. The literature contains interesting work in this direction (e.g., [16, 34]), but is rather preliminary at this point. Perhaps, the difficulties in obtaining comprehensive query logs and surf trails outside of big service providers is a limiting factor in this line of experimental research. Our own recent work 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. 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 [24].

Complementary to the above kind of extended authority analysis models are fundamental studies of link analysis techniques, particularly, the subtle but important differences between different methods and their robustness to perturbations of the input data. Telenor has started investigating this area systematically [11].

3 Result Dissemination

The research in this WP has led to 6 publications [3, 6, 8, 11, 24, 33] in international conferences and workshops, and another 4 papers [4, 5, 7, 25] are under submission for publication. All referenced papers are available on request from the coordinating site UPB.

MPII organized a summer school on information retrieval (ADFOCS 2004, September 6–10 in Saarbrücken, Germany); this was attended by 70 participants including several researchers from DELIS partners.

References


