An Evaluation Framework for Semantic Search in P2P Networks

Florian Holz Hans-Friedrich Witschel Gregor Heinrich Gerhard Heyer
Sven Teresniak

Abteilung ASV, Institut für Informatik, Universität Leipzig
{holz|witschel|heyer|teresniak}@informatik.uni-leipzig.de
gregor@arbylon.de

Abstract:

We address the problem of evaluating peer-to-peer information retrieval (P2PIR) systems with semantic overlay structure. The P2PIR community lacks a commonly accepted testbed, such as TREC is for the classic IR community. The problem with using classic test collections in a P2P scenario is that they provide no realistic distribution of documents and queries over peers, which is, however, crucial for realistically simulating and evaluating semantic overlays. On the other hand, document collections that can be easily distributed (e.g. by exploiting categories or author information) lack both queries and relevance judgments.

Therefore, we propose an evaluation framework, which provides a strategy for constructing a P2PIR testbed, consisting of a prescription for content distribution, query generation and measuring effectiveness without the need for human relevance judgments. It can be used with any document collection that contains author information and document relatedness (e.g. references in scientific literature). Author information is used for assigning documents to peers, relatedness is used for generating queries from related documents. The ranking produced by the P2PIR system is evaluated by comparing it to the ranking of a centralised IR system using a new evaluation measure related to mean average precision. The combination of these three things – realistic content distribution, realistic and automated query generation and distribution, and a meaningful and flexible evaluation measure for rankings – offers an improvement over existing P2PIR evaluation approaches.

1 Introduction

Semantic peer-to-peer information retrieval (P2PIR) offers important advantages to the design of information infrastructures, among them scalability, localisation of information and localisation of access control. However, for evaluation and comparison of P2PIR systems, no sufficiently objective methodologies exist yet as they do for “classical” information retrieval (IR) where (1) test collections are available (e.g., TREC [TRE]) and (2) methods to evaluate retrieval performance are scientifically well-established [VH06].

The classical evaluation method in information retrieval is to train the IR system with a test collection and determine the precision and recall values for a set of queries and relevance judgements defined by the test collection. Transferring this method to P2PIR is not a straightforward task, however. Although in principle, precision and recall can be defined on a P2PIR result set analogous to a centralised IR system, the features of the test collections available are not sufficient for practical evaluations. The main problems arise from the distributed character of the required test collection: There are no data available that provide realistic distributions of the test collection over peers, which – in order to be sufficient data for a methodology analogous to classical IR – must include distributions of documents, queries and associated relevance judgements over the P2P network.

This lack of available data has become a serious obstacle in the advancement of P2PIR research. To alleviate this problem, an evaluation framework is proposed in this paper that draws from two approaches: First, a plausible distribution of documents and queries is elaborated for the case of P2PIR, and second, an evaluation method is proposed to overcome the lack of relevance judgements for queries. Using these two contributions, we present an evaluation plan for a P2PIR system.

We continue this article by reviewing the technical background on P2PIR in Section 2 and review the state of the art in evaluation methods in Section 3. Subsequently, we present the parts of our framework in Section 4 and finally describe how to parameterise the testbed in Section 5.

2 Background

Since the evaluation framework presented in this paper is designed for unstructured, "non-flooding" P2P systems with semantic overlays, it is necessary to describe the characteristics of such systems:

A P2P system is called *unstructured* if the topology of the overlay network is not fixed in any way and if content can be stored anywhere in the system. This is in contrast to structured systems such as distributed hash tables [RFH$^+$01, SMK$^+$01, ZKJ01] where each peer is responsible for storing data that corresponds to a certain range of hash values. Unstructured systems are more flexible than structured ones because no control over data placement is assumed.

On the other hand, the systems we are interested in avoid flooding the network with queries. Flooding imposes a great load on the underlying network (cf. [Gnu]).

In order to avoid flooding and still guarantee good recall, such systems have to provide solutions to the following three tasks:

- *Peer description*: For making predictions about which peer is likely to be capable of contributing to a certain topic, there must be descriptions or *profiles* of peers' contents.
- *Query routing*: Assuming that peers have both the address and the profile of their neighbours, these profiles can be used for *informed search* (rather than flooding): they must be matched against queries in order to decide where the queries should

be forwarded. Queries that are sent to the wrong peers will fail to retrieve relevant documents.

- *Neighbour selection*: in order to facilitate query routing, peers should choose their neighbours in a way that maximises the probability of always finding a good neighbour to forward queries to. Networks in which peers choose neighbours according to semantic criteria are called *semantic overlay networks* [CGM02b].

Regarding the last task, there is an increasing agreement in the research community that peers should be organised in clusters of semantic similarity [SMZ03]. Additionally, some random shortcuts have been shown to be beneficial, resulting in a small world graph [Kle00] structure of the overlay [AWMM06, Sch05, WB05, LLS04, ZGG02].

In the following, we will shortly describe our own approach to semantic search and overlay structuring. It has been developed as part of a project supported by DFG[1] and was described and evaluated in more detail in [Wit05, WB05].

The algorithm for *building overlay structure* is based on gossiping: by searching for its own profile – i.e. sending out queries that consist of the profile – and receiving answers from other peers, each participant in the network fills up a part of its routing table with addresses and profiles of neighbours that offer content similar to its own. Peers thus organise into clusters of semantic similarity.

Another part of the routing table is reserved for some arbitrarily chosen neighbours (random shortcuts) that provide links between different clusters of peers. The combination of these two neighbour selection strategies is intended to result in a small world network structure (see [Kle00] for a theoretical model) that can be exploited for an efficient search algorithm.

The *search mechanism* works as follows: each peer that receives a query, first scans its local index for matching documents and then forwards the message to just one (or at least only a small number) of its neighbours: the one whose profile best matches the query, i.e. the one deemed most likely to have an answer to it. This continues until the time-to-live (TTL) of the query expires.

Since peers are organised into clusters of semantic similarity, queries will find many relevant results once they have reached the right cluster because peers that can contribute answers to a given query are likely to know others that also can.

Because in this and most other systems with semantic overlays, structure is built based on peers' shared content, it should be clear that the evaluation of such systems requires realistic models of content and query distribution. These problems are tackled in the remainder of this paper.

[1] DFG project grant nr. 255712

3 Related Work

In order to choose a feasible approach for getting a *testbed*, one might have first a look at the requirements a P2PIR test collection must satisfy: it has to provide the usual features of a IR test collection, i.e. documents, queries and relevance judgments, and in addition, there needs to be a prescription of how to distribute documents and queries among peers. Distribution of documents is either done in a way that springs naturally from the collection, e.g. via author information [BMR03] or built-in categories [CGM02a, SCK03], or it is established in less natural ways via clustering [NBMW06], latent concept analysis [HTW06], or domains of web pages [LC03, KJPD05]. In [Coo04], documents are even artificial which facilitates their distribution, but doesn't result in a real testbed.

In [BMR03], a number of different testbeds is proposed. In each, authors are identified with peers and all the papers an author has written are assigned to the corresponding peer. One testbed uses TREC, the other two (Reuters and CiteSeer) do not have queries or relevance judgments. Therefore, queries are generated randomly from documents and documents are considered relevant w.r.t. a query if they contain all query terms.

In [LC03, KJPD05], web pages are used as a test collection and the prefixes of their URLs define the peers. In [LC03], the TREC WT10g web test collection is used, together with queries generated from the documents. Evaluation is done by comparing results to that of a centralised system using plain precision and recall. Klampanos et al. [KJPD05] also use the WT10g collection and define and compare various testbeds. In addition to domains of URLs, they also use link information and textual similarity to enlarge the document set on a peer. Queries and relevance judgements are taken from the original WT10g collection (100 queries).

The approach of [CGM02a, SCK03] uses freely available copora (like OpenDirectory or CiteSeer) which contains classified documents. The distribution of documents over peers follows the topic structure in the corpus.

Without a preclassification of documents one can get a structuring by clustering the corpus. Therefore, Neumann et al. [NBMW06] use Wikipedia, with a clustering based on the internal linking structure between the wikipedia articles. To distribute the documents among the peers, the clusters are seperated into chunks and documents are chosen from these chunks using a sliding window in order to get overlap. Neumann further proposes the usage of google zeitgeist archive for generating queries, which, however, come without relevance judgements. In addition, the construction of overlapping content between peers is not as natural as e.g. a fuzzy clustering would be.

An approach based on statistics is proposed in [Coo04]. There, histograms over the counts of relevant documents per query, the degree of document replication, and the mean count of relevant documents for a query on one peer are used to generate an artificial testbed, which refactors real statistical relations. The disadvantage of this approach is that one doesn't get a real testbed, but one which merely behaves similar in certain circumstances, and that real data of P2P communities and therefore feasible relevance judgments are hard to get.

Heinrich et al. propose in [HTW06] to use a latent concept analysis to estimate relevant parameters of the distribution of topics over peer and documents and to use these parameters to distribute another document collection over peers in the same manner. That means that it is only necessary to have a little statistics of a P2P network which can be scaled up to a much bigger set of documents and peers. An alternative to create P2P evaluation data without data from a real P2P community was proposed in [Hei06] based on corpora with author and citation data.

One can see that often a natural content distribution means that there are no queries and relevance judgments, in which case artificial queries are generated from the collection's documents [LC03, BMR03] or taken from other sources [NBMW06]. In that case, relevance judgments are not available and performance is compared to a centralised setting via simple precision and recall measures. The next section presents a new idea of designing a P2PIR testbed and a new evaluation measure.

4 The proposed evaluation framework

Our framework for constructing a concrete testbed consists of two distinct parts: a well-defined distribution of content (documents and automatically generated queries) among peers, which can be applied to various existing corpora, and an evaluation measure, which respects the ranking of the retrieved documents and has no need for given relevance judgements.

4.1 Distributing content and queries

Since nodes in peer-to-peer networks correspond most often to single persons, a P2PIR testbed should provide a content distribution that realistically reflects the interests of persons running peers. One way to achieve this is to identify peers with authors of documents, as done in [BMR03].

We propose to use corpora that provide relations between documents and authors (authoring relation) as well as between documents and other documents (citation relation). In [Hei06] various freely available examples of such corpora are analysed, e.g. the CiteSeer corpus [GBL98][2] or the Cora corpus [MNRS00][3].

We propose to map the entities and relations available in these corpora to those in a P2PIR testbed in the following way:

- *Peers* map to *authors* in a relational corpus.
- *Documents* associated with a peer trivially map to documents within a relational corpus via the authoring relation. As an extended approach, the documents cited in authors' documents can be additionally associated with the peer. After all, people

[2] http://citeseer.ist.psu.edu/oai.html

[3] http://www.cs.umass.edu/∼mccallum/code-data.html

are likely to be competent in the field of expertise that their citations are concerned with. Using this extended document mapping allows a higher degree of overlap between the peers. The plain set of authored documents restricts document overlap between peers to co-authorship relations.

- *Queries* associated with a peer map to documents that are cited by the associated author's documents. For this, only citations are valid that refer to documents included in the corpus. This makes sense because, in our opinion, it is realistic to assume that people ask questions concerning issues which will extend their knowledge of the things within their focus of interest (as reflected by their documents).

Generating queries automatically out of the collection is considered necessary in order to have a sufficient number of queries, which are semantically associated to and issued by the persons running the peers. However, this means that there will be no relevance judgments for these queries. The next section presents an approach to evaluation without the need for any human relevance judgements in this setting.

4.2 Evaluation measure

Since there are no relevance judgments for queries, the performance of distributed retrieval algorithms will be measured by comparing it to a centralised setting. This means that the optimal value of the evaluation measure is reached if the P2P system retrieves the same documents and in the same order as a centralised system. It is assumed that both systems use the same retrieval function for ranking documents and the same global basis for estimating term weights (cf. [Kro02, WB05]) so that the difference between rankings is only attributable to the failure of the P2P system to retrieve certain documents. This, in turn, depends on the routing strategy: a good strategy routes queries to those peers that can contribute the most relevant documents (i.e. those ranked most highly by the centralised system) and leaves aside peers that can only contribute documents ranked lowly by the centralised system.

For this, we would like to propose a new measure that reflects the capability to retrieve the highest ranked documents, and does so better than naïve approaches that only use simple precision and recall on some defined sets of returned documents [LC03, NBMW06]. Thus, a system that retrieves the 20 documents ranked highest by the centralised system receives a better evaluation score than a system that retrieves the 200 lowest ranked documents.

The measure is closely related to mean average precision (cf. [VH06], chapter 3) and can be computed as follows:

- We assume that a query in the P2PIR system returns a ranked list A of the best k documents it can find (after merging results, that is). The value of k is assumed to be set by the user who tells the system how many (namely k) documents he/she is willing to look at.

- This will be compared to the ranked list C of *all* documents returned by a centralised search engine.

- We now mark the positions of all documents in A within C. As an example, let's assume that the user has requested to view the best $k = 3$ documents and that $A = [L, M, O]$ and $C = [K, \mathbf{L}, \mathbf{M}, N, \mathbf{O}, P]$.
- Now we compute

$$\sum_{i=1}^{k} \frac{m(A_i) prec(A_1, ..., A_i)}{min(k, |C|)} \tag{1}$$

where $m(D)$ is 1 if D is marked (see above), else 0. This means that at each document found in the distributed case, we calculate precision and we average this over $min(k, |C|)$, i.e. over the k documents the user requested or $|C|$, if even the centralised system retrieves less than k documents. In our example, this yields $\frac{1}{3}(\frac{1}{2} + \frac{2}{3} + \frac{3}{5}) = 0.59$. In contrast, if the system retrieves $B = [N, O, P]$, then we get $\frac{1}{3}(\frac{1}{4} + \frac{2}{5} + \frac{3}{6}) = 0.38$

Note that if we use a reference corpus for weight estimation as proposed in [WB05], then rankings are global and hence the scores of documents within A, B, and C will not differ. The measure then tells us how high the best k documents that the distributed search finds are ranked – on average – by the centralised search engine.

5 Parameterising the system

The above evaluation framework leaves a number of parameters free, which need to be chosen in order to make it operational:

- *Which corpus to use*: see [Hei06] for a comparison of freely available corpora.
- *Details of document distribution*: as mentioned above, there are two main possibilities for choosing the documents of a peer: either assigning only the documents the author has written himself, or additionally assigning those that are referenced in these papers. The advantage of the second approach is that it is maybe more realistic to assume that people store more documents from their field of expertise than their own. On the other hand, with the proposed query generation mechansim, this would mean that peers search for a subset of their shared documents.
- *Query generation*: the exact way to form queries was not detailed above because there may be different views on how to realistically choose a query's length and the exact key terms. For example, a librarian is likely to phrase longer and more detailed queries than a casual internet user. A simple possibility would be to use the titles of referenced documents as queries. Alternatively, one can choose the most relevant keywords from the entire document, the distribution of query lengths being

determined by examining e.g. query logs of a web search engine. In general, the query length distribution also is linked to the retrieval function: Given a boolean retrieval more key terms lead to less retrieved documents, but given a vector space retrieval so more key terms lead to more retrieved documents.

- k: to fix the number of documents that should be retrieved, one might again use an empirical study of user behaviour: how many result documents do users request on average from e.g. a web search engine? This question can be easily answered from query logs and the distribution can be used for varying k in a guided way.
- *Retrieval function*: the exact measure of similarity between queries and documents, but also between other entities such as peer profiles, can be chosen individually for each experiment and varied for comparison.

As a conscious decision, we decided not to include network dynamics into the evaluation framework. However, these may be easily be added at a later stage. These aspects of dynamics include, for example:

- *Churn at several time scales*: short-term churn – i.e. peers joining and leaving the network during the day – or long-term churn – i.e. new peers being introduced into the system or leaving it terminally. The latter may require not to use the full set of author-generated peers in the beginning of a simulation, but let the network grow.
- *Document distribution*: one may take the testbed only as a starting setup for the simulation and model the download behaviour of users.
- *Querying activity*: determine in what time intervals peers should ask which of their queries. The query sets associated to the peers may also change over time according to changes in the document distribution.

Because it leaves a great deal of freedom for modeling different aspects of P2PIR, the framework allows to be used in many contexts, but still guarantees comparable results within the P2PIR community.

6 Conclusions

In this paper, we have presented an evaluation framework, providing a strategy for the construction of a P2PIR testbed, that consists of a realistic prescription for content distribution, query generation and distribution, and for measuring a retrieval system's effectiveness in ranking documents. We rely only on a freely availbale relational corpus and there is no need for hand-crafted queries or human relevance judgments. The framework offers a great amount of flexibility via free parameters but still ensures comparability of results.

In the future, the framework will be used in our research, in order to evaluate the quality of the peer description, query routing and neighbour selection strategies that have been described above in Section 2.

References

[AWMM06] R. Akavipat, L.-S. Wu, F. Menczer, and A.G. Maguitman. Emerging semantic communities in peer web search. In *P2PIR '06: Proceedings of the international workshop on Information retrieval in peer-to-peer networks*, pages 1–8, 2006.

[BMR03] M. Bawa, G. S. Manku, and P. Raghavan. SETS: search enhanced by topic segmentation. In *Proc. of SIGIR '03*, pages 306–313, 2003.

[CGM02a] A. Crespo and H. Garcia-Molina. Semantic Overlay Networks for P2P Systems. Technical report, Computer Science Department, Stanford University, 2002.

[CGM02b] Arturo Crespo and Hector Garcia-Molina. Semantic Overlay Networks for P2P Systems, 2002.

[Coo04] B. F. Cooper. A content model for evaluating peer-to-peer searching techniques. In *ACM/IFIP/USENIX 5th International Middleware Conference*, Toronto, 2004.

[GBL98] C. Lee Giles, Kurt Bollacker, and Steve Lawrence. CiteSeer: An Automatic Citation Indexing System. *Proc. 3rd ACM Conf. on Digital Libraries*, pages 89–98, June 23–26 1998.

[Gnu] Gnutella. www.gnutella.com, last visited 04/03/2007.

[Hei06] Gregor Heinrich. Free text corpora and their application to community retrieval evaluation. Technical report, Arbylon & University of Leipzig, http://www.arbylon.net/publications/corpora.pdf, 2006.

[HTW06] G. Heinrich, S. Teresniak, and H. F. Witschel. Entwicklung von Testkollektionen für P2P Information Retrieval. In Christian Hochberger and Rüdiger Liskowsky, editors, *Workshop P2P Information Retrieval, 36. Jahrestagung der Gesellschaft für Informatik*, volume P-93 of *Lecture Notes in Computer Science (LNI)*, pages 20–27, 2006.

[KJPD05] I. A. Klampanos, J. M. Jose, V. Poznanski, and P. Dickman. A Suite of Testbeds for the Realistic Evaluation of Peer-to-Peer Information Retrieval Systems. In *27th European Conference on IR Research, ECIR 2005*, pages 38–51, 2005.

[Kle00] J. Kleinberg. The Small-World Phenomenon: An Algorithmic Perspective. In *Proceedings of the 32nd ACM Symposium on Theory of Computing*, 2000.

[Kro02] A. Z. Kronfol. FASD: A Fault-tolerant, Adaptive, Scalable, Distributed Search Engine, 2002.

[LC03] J. Lu and J. Callan. Content-based retrieval in hybrid peer-to-peer networks. In *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, pages 199–206, 2003.

[LLS04] M. Li, W.-C. Lee, and A. Sivasubramaniam. Semantic Small World: An Overlay Network for Peer-to-Peer Search. In *Proceedings of the International Conference on Network Protocols (ICNP)*, pages 228–238, 2004.

[MNRS00] Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Automating the Construction of Internet Portals with Machine Learning. *Information Retrieval Journal*, 3:127–163, 2000. www.research.whizbang.com/data.

[NBMW06] T. Neumann, M. Bender, S. Michel, and G. Weikum. A Reproducible Benchmark for P2P Retrieval. In *Proc. of the First International Workshop on Performance and Evaluation of Data Management Systems, ExpDB 2006 at ACM SIGMOD*, pages 1–8, 2006.

[RFH$^+$01] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Schenker. A Scalable Content Addressable Network. In *Proceedings of the ACM SIGCOMM*, 2001.

[Sch05] C. Schmitz. Self-Organization of a Small World by Topic. In *Proceedings of 1st International Workshop on Peer-to-Peer Knowledge Management*, 2005.

[SCK03] M. Schlosser, T. Condie, and S. Kamvar. Simulating A File-Sharing P2P Network. In *Proc. of 1st Workshop on Semantics in Grid and P2P Networks*, 2003.

[SMK$^+$01] I. Stoica, R. Morris, D. Karger, F. Kaashoek, and H. Balakrishnan. Chord: A Scalable Peer-To-Peer Lookup Service for Internet Applications. In *Proceedings of the 2001 ACM SIGCOMM Conference*, pages 149–160, 2001.

[SMZ03] K. Sripanidkulchai, B. Maggs, and H. Zhang. Efficient Content Location Using Interest-Based Locality in Peer-to-Peer Systems. 2003.

[TRE] Text REtrieval Conference. http://trec.nist.gov/.

[VH06] E.M. Voorhees and D.K. Harman. *TREC – Experiment and Evaluation in Infromation Retrieval.* The MIT press, Cambridge, Massachusetts, 2006.

[WB05] H.F. Witschel and T. Böhme. Evaluating Profiling and Query Expansion Methods for P2P Information Retrieval. In *Proc. of the 2005 ACM Workshop on Information Retrieval in Peer-to-Peer Networks (P2PIR)*, 2005.

[Wit05] H.F. Witschel. Content-oriented Topology Restructuring for Search in P2P Networks. Technical report, University of Leipzig, http://wortschatz.uni-leipzig.de/˜fwitschel/papers/simulation.pdf, 2005.

[ZGG02] H. Zhang, A. Goel, and R. Govindan. Using the Small-World Model to Improve Freenet Performance. In *Proc. of IEEE Infocom, 2002. 14*, 2002.

[ZKJ01] B. Y. Zhao, J. D. Kubiatowicz, and A. D. Joseph. Tapestry: An infrastructure for fault-resilient wide-area location and routing. Technical Report UCB//CSD-01- 1141, U. C. Berkeley, 2001.