ABSTRACT

The thesis presented in this paper treats selected issues in unstructured peer-to-peer information retrieval (P2PIR) systems, a guiding principle being the question to what extent global world knowledge can be helpful for solving P2PIR problems. In a first part, where experiments have been conducted already, representative reference corpora were used for estimating global term weights such as IDF instead of sampling them from the collection distributed throughout the P2PIR system. This approach infers very low communication costs and allows for trivial results merging. Experimental results have been encouraging.

A second part of the work will be dedicated to the question of query routing in unstructured P2PIR systems using peer profiles (resource descriptions): is there a way to keep profiles very compact – so that they can be sent around and stored in routing tables – and still guarantee good recall when matching queries against compressed profiles? Solutions and experiments will be proposed, together with a new P2PIR evaluation framework.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval; C.2.4 [Computer Communication Networks]: Distributed Systems—Distributed Databases

General Terms
Algorithms, Experimentation

Keywords
Peer-to-peer information retrieval, term weighting, query routing

1. INTRODUCTION

The place of peer-to-peer technologies within the field of information retrieval is not well determined. On the one hand, the use of P2P networks in multimedia storage and retrieval has shown that this technology has some interesting advantages over centralised search engines: it is more difficult to control by a central instance (which made it popular in the first place), there is no single point of failure, publishing content is easier than on the web and content is available and retrievable as soon as it is inserted, no crawling is needed. On the other hand, there are currently no text retrieval systems based on P2P technology. This is mainly due to the new challenges in this field which are unsolved or solved only at the expense of exceedingly high resource consumption.

For example, flooding-based systems like Gnutella [10] guarantee that a large number of peers is reached by each message and thus provide good recall. However, the exponential growth of message volume generates a high load for the underlying network. Another potential solution is represented by structured P2P systems like CAN, Chord or Tapestry [21, 8, 32]. In these systems, each peer is responsible for a certain range of hash values and a query can be easily answered by hashing it and contacting the peer responsible for the resulting hash value. However, queries are assumed to consist of one key only. Structured systems dealing with multiple-keyword queries in text-based IR have been designed (cf. e.g. [24]), but they require that queries be split and that each constituent term be searched separately which may result in large result sets to be merged afterwards; they also assume that we have complete control over the location of data items.

Therefore, this work concentrates on unstructured, “non-flooding” P2P systems which offer a great deal of flexibility but also pose new problems, some of which remain unsolved. In unstructured systems where flooding is to be avoided, we have to tackle the three tasks which are also at the heart of distributed retrieval [3]:

- Resource description: In order to make predictions about which peer is likely to be capable of contributing to a certain topic, there must be descriptions or profiles of peers. These need to be compact since they are usually stored in routing tables of other peers. The challenge here is to keep loss of information minimal when compressing profiles.

- Resource selection and query routing: Assuming that peers have both the address and the profile of their neighbours, these profiles must be matched against queries in order to decide where the queries should be forwarded. Queries that are sent to the wrong peers...
In the thesis: length. There are two main challenges which are addressed
profiles consisting of the highest weighted terms from their
mechanisms based on peer profiles. A structure can be created reliably using simple gossiping
the overlay [1, 23, 29, 15]. In [27], I have shown that such
beneficial, resulting in a small world graph [12] structure of
should be organised in clusters of semantic similarity. Ad-
onditions of other peers to store in their routing tables.
not fixed in advance, peers can choose which addresses and
descriptions of other peers to store in their routing tables.
The topology of the resulting overlay network has large im-
impact on the performance of search algorithms. There is an
increasing agreement in the research community that peers
should be organised in clusters of semantic similarity. Ad-
ditionally, some random shortcuts have been shown to be
beneficial, resulting in a small world graph [12] structure of
the overlay [1, 23, 29, 15]. In [27], I have shown that such
structure can be created reliably using simple gossiping
mechanisms based on peer profiles.
In my thesis, it is assumed that an unstructured P2P sys-
em as described in [29] is used. In that system, peers have
profiles consisting of the highest weighted terms from their
locally shared documents, much in the spirit of resource de-
scriptions in distributed IR [3], but pruned to an acceptable
length. There are two main challenges which are addressed
in the thesis:
• Precision: Is there a way to compute good weights for
terms in documents without having a global view on
the document collection? This is especially interesting
for weights like idf that normally require such a global
view. Ideally, weights should be globally valid, i.e.
independent of the local contents of a peer, such that
results merging is easy: retrieval status values can be
computed on any peer and are comparable across the
whole system. On the other hand, computing them
should not involve sampling frequency statistics from
peers (in order to reduce message overhead).
• Recall: Is there a way to both keep peer profiles very
compact and still implement a query routing that guar-
antees good recall?
For the first part, experiments have been conducted al-
ready. These are described in [28] and summarized shortly in
section 3.1. The second part is currently being planned and
hence there are many open questions. Ideas and proposed
experiments are presented in sections 3.2 and 3.3. First,
however, the next section summarises some related work.

2. RELATED WORK

Estimating term weights in the face of no or little informa-
tion of the whole collection is interesting in various contexts:
in new-event detection, filtering or routing systems, one of-
ten starts out with an empty document repository, wishing to
judge new, incoming documents. In this case, the collec-
tion is not yet known. A similar case arises when collections
change too quickly in order to keep recomputing global term
weights. Finally, of course, the problem is present in all dis-
tributed scenarios where a global view on the collection is
unavailable.
Work on dynamic and quickly changing collections [26, 4]
found that a training set of 40-50% of the documents in a
collection is normally sufficient for robust IDF estimation.
In new event detection, researchers either use a "reference
corpus" from which IDF estimates are derived [20] or in-
crement IDF estimates when new documents arrive, or combi-
nations of both strategies [31]. Although reference corpora
were used in some systems, there is no evaluation of the
quality of term weights that they yield. Experiments on
this are reported in section 3.1.
In P2P information retrieval, one way to cope with the
missing global information is to use coarser units, e.g. IPF
(inverse peer frequency) instead of IDF [7]. Various result
merging algorithms have also found their way into hierar-
chical P2P networks (e.g. [18]). Alternatively, one may ac-
quire and merge document frequency information from each
database [13], but often one finds that a sample is sufficient
[25, 24]. A reference corpus for IDF estimation has been used in [14].

As indicated in the introduction, a considerable amount
of research has been done on unstructured P2PIR systems
where profiles are used for query routing and often also for
building overlay topologies. Different from distributed IR,
profiles (or resource descriptions) are compressed in most
P2PIR settings, i.e. one either uses data structures like
Bloom filters (e.g. [7, 16]) or prunes unigram language mod-
els in some way [29]. Alternative ways of building compact
resource descriptions rely on categories or ontologies [5, 9],
which requires both peer contents and queries to be classified
according to the categorisation system. This is especially
hard for queries.
When working with pruned unigram language models (i.e.
plain term lists) as profiles, there is inevitably loss of infor-
mation. That is, even if a peer p has relevant information
w.r.t. a given query q, q will not be routed through p if its
terms have all been pruned from p’s profile. An obvious so-
lution to this problem is query expansion using some global
information, which has been shown to be rather successful
for distributed and P2P information retrieval [30, 29]. In
[29], we found that local feedback on peers did not help
much because expansion only takes place when something
has been found already such that queries that have few rele-
vant documents will not benefit. Therefore, section 3.3 is
dedicated to studying possibilities for solving this problem.
Finally, let us look at work on P2PIR testbeds: these have
to provide the usual features of a IR test collection, i.e. doc-
uments, queries and relevance judgments. In addition, there
needs to be a prescription of how to distribute documents
and queries among peers. Distribution of documents is ei-
ther done in a way that springs naturally from the collec-
tion, e.g. via author information [2] or built-in categories
[22] or it is established in less natural ways via clustering
[19] or domains of web pages [17, 11]. Often a natural con-
tent distribution means that there are no queries and rele-
ance judgments, in which case artificial queries are gener-
ated from the collection’s documents [17, 2] or taken from
other sources [19]. In that case, relevance judgments are
not available and performance is compared to a centralised
setting via simple precision and recall measures. Section 3.2 presents a new idea of designing a P2PIR testbed and a new evaluation measure.

3. CONTRIBUTION

As indicated in the introduction, my thesis aims at solving problems in P2PIR on two levels: precision- and recall-oriented retrieval. The guiding principle in this research is the question to what extent global “world” knowledge can be applied in this context. World knowledge refers to data that is independent of the collection shared by peers in a particular P2PIR system; it is assumed that this knowledge can be gathered – once and for all – from sources such as the WWW and then used fruitfully in many different P2PIR systems. This approach will, of course, have to be compared to alternative ones.

3.1 First results

In a first series of experiments [28], I investigated the question whether one can globally estimate collection or document frequencies of terms – independent of a given collection – well enough in order not to degrade retrieval performance seriously. More precisely, I estimated them from the British National corpus (BNC) – which I considered to be a representative reference corpus for English – and used the estimated frequencies for computing IDF weights and smoothing a language model, respectively. Then, I compared the performance resulting from these “global” weights to the performance achieved when deriving weights directly from the test collections that I used: TREC-7 and 8 and a small Medline collection.

The results that I got can be summarised as follows:

- If test collections are small, then using a reasonably large reference corpus for weight estimation improves performance. For larger collections, weights estimated from reference corpora slightly degrade retrieval results, but this is often not statistically significant. All in all, there seem to be some collection- or domain-specific peculiarities (mostly “stop words”) that cannot be predicted by reference corpora, but this effect is very small.

- A large fraction of terms can be pruned from the resulting term lists without any ill effects: in the experiments, removing all terms with frequency ≤ 100 did not degrade results. Pruning low frequency terms from the list means treating them as if they hadn’t occurred in the reference corpus, i.e. as if they were very rare – which earns them high weights (e.g. high IDF). In the last consequence, this tells us that all we need for IR is an extended list of “stop words”: in the experiments, pruning terms with frequency ≤ 100 left over only 40,000 frequent terms.

In a distributed setting, using reference corpora for weight estimation has some outstanding advantages: First, retrieval status values of documents can be computed on any peer that possesses the term list with global weights; hence, results merging is trivial. Second, the full power of retrieval functions can be used without having to resort to coarser measures such as inverse peer frequency\(^1\). Third, there is no communication overhead involved with this solution: each peer will be equipped with a term list that does not need to be updated. If pruning is applied as indicated above, this list will be only some KB large and hence it will be no problem to store it on any peer.

In the future, some more experiments are planned using more test collections in order to better separate domain and size effects.

3.2 A new P2PIR testbed

For further experiments, besides using an existing P2PIR testbed such as the ones proposed in [17, 11], I would like to develop a new testbed which has a more realistic distribution of content and queries and to propose a new evaluation measure.

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 guarantee this is to identify peers with authors of documents (as done in [2]). I propose to use the CiteSeer\(^2\) database, treating each author as a peer and assigning each document that an author has written to the corresponding peer.

The use of scientific papers gives rise to a simple query generation and distribution algorithm: each peer will ask for the titles (or some randomly selected keywords) of papers that are referenced in its own papers. This makes sense because, in my 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).

Finally, since there are no relevance judgments for these queries, the performance of distributed retrieval algorithms will be measured by comparing it to a centralised setting. For this, I would like to propose a new measure that reflects the capability to rank, and does so better than naïve approaches that only use simple precision and recall on some defined sets of returned documents [17, 19]. The measure is closely related to mean average precision and can be computed as follows:

\[ \text{We assume that a query in the P2PIR system returns a ranked list } A \text{ of the best } k \text{ documents it can find (after merging results, that is)} \]

\[ \text{This will be compared to the ranked list } C \text{ of all documents returned by a centralised search engine.} \]

\[ \text{We now mark the positions of all documents in } A \text{ within } C. \text{ As an example, consider } A = \{ L, M, O \} \text{ and } C = \{ K, L, M, N, O, P \}. \]

\[ \text{Now we compute} \]

\[ \sum_{i=1}^{k} \frac{m(A_i) \text{prec}(A_i, ..., A_1)}{\text{min}(k, |C|)} \] \hspace{1cm} (1)

where \(m(D) = 1\) 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 retrieval functions also use other variables that require global knowledge of the collection (e.g. average document length). It will have to be investigated in the future if this is a serious problem.

\(^1\)http://citeseer.ist.psu.edu
the $k$ documents found (or $|C|$, if there are less to be found anyway). In our example, this yields $\frac{1}{2}(\frac{1}{4} + \frac{1}{2} + \frac{1}{4})$.

Note that if we use a reference corpus for weight estimation, then rankings are global and hence the scores of documents within $A$ 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.

3.3 Further proposed experiments

Now, using this experimental setup, I would like to address the second main question of my thesis: is there a way to keep peer profiles very compact – so that they can be sent around and stored in routing tables – and still guarantee good recall when matching queries against compressed profiles? And: can world knowledge be helpful for expanding queries?

In order to define this more precisely, imagine profiles to consist of the $n$ highest weighted terms from a peer’s document collection – where $n$ is dictated by the available storage and bandwidth capacities of peers. Further, the weighting of a term within a peer is assumed to be some function of its frequency within the peer’s local collection and some global reference corpus – e.g., a BM25 weight where the $t f$ part is calculated using the local frequency and the $idf$ part is estimated via the reference corpus.

Then, it may well occur that some term $t$ has been pruned from all profiles so that a query for $t$ cannot be routed properly, even if there are documents satisfying it. To overcome this, the query could be expanded with related terms, preferably ones that are likely to be found in peers’ profiles and that lead to the right peers in order to answer $t$.

The experiments that I propose basically aim at evaluating various query expansion strategies, similar in spirit to what has already been done in [29], but using the new evaluation framework and extended by some more sources of information. This idea is also somewhat similar to [6], where performance of ordinary (i.e., centralised) retrieval is improved using large external corpora for query expansion. The sources of information to be investigated in my thesis include the following:

- The web: This variant is cheap in terms of storage, provided that peers have access to a search engine. Queries can be sent to a web search engine and the results (either full-text documents or just snippets) can be used to expand the query.
- Other large collections: expanding queries with terms that co-occur with the query terms in a large collection is expensive in terms of storage because an index of that collection must be held at each peer. However, the results in [6] indicate that it might be more effective than using the web if one chooses a collection from an appropriate domain.
- Thesauri: some general-purpose or – if available – domain-specific thesaurus may be used to expand queries. This involves the high cost of creating the thesaurus, but consumes little storage on peers.

Expansion using these global knowledge sources will then have to be compared to other expansion strategies:

- Local feedback: as a query gets routed through the P2P network, each peer that it encounters consults its local database for answers. If any matching documents are found, these may be assumed to be relevant and used for expansion (pseudo relevance feedback). As far as storage cost is concerned, this is very cheap, but it was also found to be of little use in many cases in [29] because it only starts to work if something has been found already.
- Profile expansion by query caching: Since there are usually only a few rather popular queries, it might boost recall if peers add terms to their profile, which are taken from recent queries that they have answered successfully. Of course, this means that other terms have to be evicted from the profile to keep it at size $n$. In order to evaluate this, one needs to have a query log where queries are listed in order of arrival, e.g., from a large search engine. Optimally, the log should fit the collection, such that all queries have non-empty result sets. For a start, however, I propose to use any publicly available query log together with the WT10g collection distributed on peers as proposed in [17].
- Profile expansion by collaborative tagging: if persons are allowed to see the profiles of peers that are neighbours of their own peer in the overlay network, then they might tag these profiles. These tags may be propagated to the corresponding peers and included in a “tag profile”. This is, however, extremely difficult to simulate and hence to evaluate.

More strategies and expansion methods are thinkable and may be added to this preliminary list later. However, testing the above strategies should give a good impression of the extent to which world knowledge may be helpful or to which it might be replaced by something else.

4. CONCLUSION AND QUESTIONS FOR DISCUSSION

The use of unstructured peer-to-peer systems for text retrieval has a number of interesting applications and shows great promise for the future. However, many problems remain unsolved, some of which are to be tackled in this dissertation. Preliminary experiments on using reference corpora for estimating global term weights showed promising results. Applying more world knowledge in the task of expanding queries for query routing in the face of compressed peer resource descriptions has to be examined more closely, but is expected to outperform alternative choices such as local feedback (see the preliminary results in [29]).

At the doctoral consortium, it would be highly helpful for me to discuss the proposed evaluation framework and the experiments planned for the second part with the committee and other doctoral students. I will greatly welcome all comments and recommendations regarding the evaluation of query routing in unstructured P2P systems.

5. REFERENCES


APPENDIX

A. STUDENT’S STATEMENT

The SIGIR doctoral consortium is an excellent opportunity for me to discuss the work presented here with experienced IR workers and other doctoral students. Since my Ph.D. project is situated within an evolving and rather young discipline, namely peer-to-peer information retrieval, where there are not yet standard ways for evaluation, I feel the need to discuss my proposed evaluation methodologies with experienced IR researchers before putting them into practice.

I am sure that I can also greatly benefit from the experience, comments and recommendations of senior IR researchers as far as my overall conception of work and the particular experiments planned are concerned.

Finally, I hope that some of the results that I have obtained so far will be interesting to others, in which case the doctoral consortium is an excellent opportunity to present them.

– Hans Friedrich Witschel

B. SUPERVISOR’S STATEMENT

Hans Friedrich Witschel has been working in my department on information retrieval since three years, first funded by the German Research Foundation on a project on content based P2PIR, and since one year as a scholar in our department’s graduate school. He is presently writing his thesis on foundations and algorithms for distributed IR, considering in particular content based P2PIR. He is making good progress and will most likely be finished early next year, so the timing for him to attend this year’s SIGIR07 Doctoral Consortium would be perfect.

Due to the department’s focus on natural language processing algorithms for IR, he would very much benefit from presenting and discussing his ideas with senior IR researchers and fellow students. Although he already has achieved a number of interesting results in detail, he still needs to define the overall structure of his dissertation and to conduct experimental evaluations.

Hans Friedrich Witschel is one of our most talented Ph.D. students and capable to pursue an academic career. The opportunity to discuss his ideas and results with experienced IR researchers will be most beneficial for him. I strongly recommend Hans Friedrich Witschel to the SIGIR07 Doctoral Consortium.

– Prof. Gerhard Heyer