World Applied Sciences Journal 29 (Computer Sciences, Engineering and Its Applications): 14-18, 2014
ISSN 1818-4952

DOI: 10.5829/idosi.wasj.2014.29.csea.2207

Fresh Information Reterival Using P2P Web Search

R.udayakumar, K.P.Thooyamani and Khanaa

Prof. School of Computing Science, Bharath University, Chennai

Abstract: The fast development of the World Wide Web and Dynamic nature makes it a challenge for searching and retrieving of information that is more recent.The WWW is a rapidly growing and changing information source. Its growth and change rates make the task of finding recent information harder. it has more communication delay. In this paper, in order to reduce the delay we proposed a P2P Web search that connects an a-priori unlimited number of peers, each of which maintains a personal local database and a local search facility. Each peer posts a small amount of metadata to a physically distributed directory layered on top of a DHT-based overlay network that is used to efficiently select promising peers from across the peer population that can best locally execute a query. We also uses Textual Entailment approach for searching a relevant document for a given keyword ie) to retrieve not only textual documents that have specified keywords, but also to discover semantically equivalent or entailed documents from given keywords.

Key words: Information retrieval · Distributed hash table · Textual Entailment · Peer to peer

INTRODUCTION

Internet is explosively expanding and IT evolution caused by Internet is changing our society. World Wide Web (WWW) is the most popular service in Internet. WWW is a hypertext system, which is regarded as a large-scale database. Search engines are used for Web page retrieval. It was difficult to search Web pages because of no index in WWW. However, search engines make it possible to search Web pages by indexing. Conventional search engines employ centralized architecture, in which a robot traverses Web pages by following links. However, such a robot wastes a long time to traverse. Waseda University's Senrigan wasted a week to collect all of Japanese documents. Today, it wastes a longer time because we have more pages. We call the complete time of this travel the update interval, which is the longest time until all updated pages are collected. Internet is also used as business tool. However, the update interval is too long to use Internet for business. There are many distributed search engine to reduce the update interval e.g) Cooperative Search Engine (CSE). Assume that there is a local search engine, which searches only local documents, in each Web server. CSE integrates these local search engines by using Meta search engines, which communicate with each other. In this way, CSE realizes a global search engine. CSE can reduce the update interval very much but it has performance issues because communication delay occurs at search time. In this paper, we describe the P2P framework to reduce the delay and retrieve the fresh information.

Peer-to-peer (P2P) overlay network has become a substantial research topic in recent years in network and web applications. The central strength of such kind of systems is the capability of sharing resources, information and services. Typically P2P systems are highly distributed and self-organizing systems that contain huge amounts of heterogeneous data, for example, textual documents, audio/video files, multimedia and so on. As the amount of data is continuously increasing and the number of available peers is growing too, more efficient searching algorithms will be needed to accomplish tasks. For this reason, P2P networks have appeared as an attractive architectural paradigm for the information retrieval (IR) area, especially to build new search engines to deal with huge amounts of heterogeneous and continuously changing data. However, P2P retrieval methods still pose a lot of research challenges. Search methods are typically limited to simple keyword queries and only provided

Corresponding Author: R. udayakumar, Prof. School of Computing Science, Bharath University, Chennai,

title-base search facility, which means the end user cannot retrieve the content unless he knows a unique keyword. These methods lack support for the equivalent semantically content search. Our goal here is to automatically discover not only such documents, but also other documents in which the given search keys can be inferred from their context.

The remainder of this paper is organized as follows: In section 2, we describe about overview of the P2P searching and Textual Entailment. P2P information retrieval process and architecture, in section 3 and its behaviors in section 4. Finally, we summarize our conclusions.

P2P Searching: Many P2P search algorithms have already been deployed that critically depend on peer topology and query routing. There are several different architectures for P2P networks: *Centralized* approach maintains a constantly updated directory at a central location. That contains the object location, ID assignment, etc. *Decentralized* approaches can either follow the *pure* model, with all peers equally making, routing and answering requests, or a *hybrid* one, where peers are divided into leaf-nodes and super-peers[2]. Taxonomy classified the decentralized approaches into *structured* and *unstructured*. The structured P2P network is tightly controlled and files are placed at specified locations that make subsequent queries easier to satisfy. There are two types in this regard: *loosely structured* systems, where file placement is based on hints and highly structured systems. Search methods in an unstructured P2P network can be categorized as either *blind* or *informed*. In a *blind* search, nodes do not store any information regarding locations[6]. In *informed* approaches, nodes locally store metadata that assists in the search for the queried objects. In our paper we go for informed approach.

Textual Entailment: The concept "textual entailment" is a new approach that is applied in the natural language processing field. It is used to indicate the state in which the semantics of one natural language written text can be inferred from the semantics of another text. Specifically, if the truth of a text segments entails the truth of another text segment. For example, given the texts:

- For their discovery of ulcer-causing bacteria, Australian doctors Robin Warren and Barry Marshall have received the 2005 Nobel Prize in Physiology or Medicine.
- Robin Warren was awarded a Nobel Prize.

It is clear that the semantics of the second one can be inferred from the semantics of the first one; then, it is said that textual entailment exists between both texts. The first articulation is called text statement T but the second one is called hypothesis statement H[10]. It is clear from the above example that the textual entailment is a one-way relation, i.e. the hypothesis statement is entailed from the text statement but the reverse is not usually true. This is given by T=>H, which means "H is entailed from T". The recognition of textual entailment requires processing at the lexical level (as in the above example, the recognizing of synonymy between "received" and "was awarded"), as well as at the syntactic level and the sentence semantics level. This informal definition is somewhat based on (and assumes) common human understanding of language, as well as common background knowledge.

Textual Entailment Search Algorithm

The Outline of Algorithm Is as Follows:

```
Algorithm TE_SEARCH is
  ForwardQuery (Peer p, Query q) is
    selectedPeers = selectKMostRelevantPeers(p,q,k);
    expandSize = selectedPeers.size();
    setVisited(q,p);
    if (expandSize > 0) then
        splitTTL(q, expandSize);
        for each r in selectedPeers do
            receiveQuery(r,q);
    else
      querySource = get neighbor who submitted
                    the query;
      receiveQuery(querySource, q);
  End;
  ReceiveQuery (Peer r, Query q) is
    if (not visited(r,q)) then
          addVisited(r,q);
          addSourceNeighbour(q,r.queriesMap);
          decreaseTTL(q);
          setLocalResult(r,q);
          For each document d in D(r) do
              For each keyword k in K(d) do
                  if k => q then
                      addtoLocalResult(d,r,q);
                      exit;
          sendLocalResults(r,q);
          if (moreHopsExists(q)) forwardQuery(r,q);
    else if (moreHopsExists(q)) then
            joinedQuery = joinQueryEntries(r,q);
            updateLocalMap(r,q,joinedQuery);
            forwardQuery(r,joinedQuery);
  End;
Begin
  setInitialTTL(q);
  addVisited(initial,q);
  forwardQuery(initial,q);
End;                                                    □
```

P2P Information Retrieval: This section briefly discusses the P2P Information Retrieval architecture. Our proposed searching algorithm based on this architecture is illustrated in the next section. We have adopted P2P-Information Retrieval (P2P-IR) architecture[15]. However, here this architecture will deal with an unstructured P2P

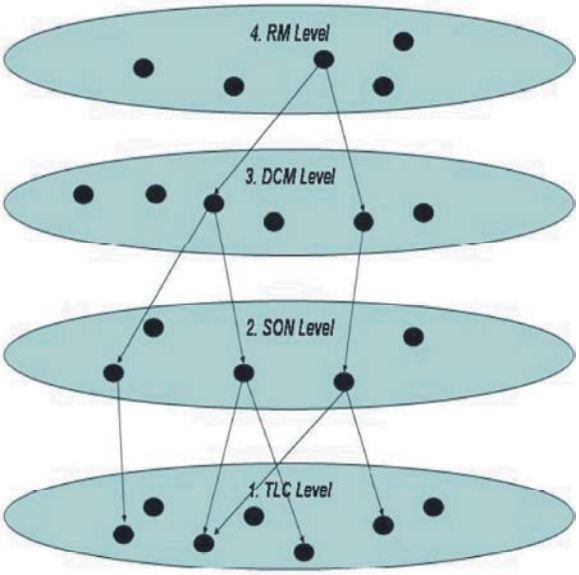

Fig. 1: Outline of P2P information retrieval

peer network. This P2P-IR architecture consists of four levels, as shown in Figure 1. Each level can be conceived of as a space of objects with a specific topology and with certain functions to access and manipulate these objects. Each level mainly uses functions of the level underneath it to implement its functions. Briefly we can illustrate each level as follows:

Transport Layer Communications Level (TLC): This level is responsible for communicating the peers with each others. Each peer runs software that supports communication with other peers through TCP/IP and UDP/IP protocols. This software identifies associated peers via physical address.

Structured Overlay Networks Level (SON): Deals with various P2P network structures; structured or unstructured P2P.

Document and Content Managements Level (DCM): Manages how we retrieval the desired document. Although the specified P2P model plays a role in answering this question, there are common tasks of document management to all models, which will serve as building blocks for the definition of retrieval models (Level 4). For example, maintaining a distributed document repository, associating term sets with documents, maintaining vocabularies and cluster hierarchies. Most of these tasks come from basic operations associated with text-based information retrieval.

Retrieval Models Level (RM): Manipulates the same previous DCM functions (Level 3). Therefore, it must have a common framework with proposed DCM network. Such a framework characterizes notions like representation of user queries, the provision of ranking and clustering functions and interfaces for basic information retrieval tasks. It provides functions for constructing vocabularies and document indexing, such as extracting keys from documents.

System Design: We assume architecture of a P2P Web search as follows. Each peer is fully autonomous and has its own local search engine and has a local index that can be built from the peer's own crawls or imported from external sources and tailored to the user's thematic interest profile. Peers are willing to share metadata about their local indexes (or specific fragments of local indexes) by publishing it into a P2P network. This conceptually global but physically distributed directory, which is layered on top of a Chord [10] style Distributed Hash Table (DHT), contains compact statistics and quality-of-service information. For failure resilience and availability, the responsibility for a term is shared and replicated across multiple peers. Notice that, unlike [6], we use the DHT to partition the term space, such that every peer is responsible for the metadata of a randomized subset of terms within the global directory. We do not distribute documents or index lists across the directory.

Query processing works as follows. In a preliminary step, every peer publishes statistical metadata (Posts) about a subset of terms in its local index to the directory. A hash function is applied to the term in order to determine the peer currently responsible for this term. This peer stores all Posts for this term from across the directory in a PeerList. Posts contain contact information about the publishing peer together with statistics to calculate IR-style relevance measures for a term (e.g., the size of the inverted list for the term, the maximum average score among the term's inverted list entries, or some other statistical measure or Textual Entailment results) and other information, e.g., regarding quality-of-service.

The query initiator collects the PeerLists for all query terms from the distributed directory and combines this information to find the most promising peers for the current query. This step is referred to as query routing. Query routing has been a research issue for many years [4, 7], but typically focuses on disjoint data sets. A number of these strategies have been evaluated in previous work [2]. However, naturally, the peers' data

collections often highly overlap, as popular documents are highly crawled. We have developed strategies to combine overlap estimation with the available score/ranking information into an overall quality-novelty measure that can boost the effectiveness of query routing [1] in such an environment.

Demonstrations: Our demonstration aims at illustrating the whole functionality of our system as well as its ease of use in a live demo using five PCs. To make this a true P2P demo, we additionally invite all visitors to join our network instantly with their notebooks. After physically connecting to our ad-hoc LAN (cable and wireless; using network equipment we bring along) and automatically obtaining an IP address, visitors can browse a local web page, hosted on one of our PCs. To ease live deployment, we have prepared thematically focused collections using the BINGO [3] crawler and stored them in a local database. Alternatively, provided an outbound network connection, visitors can perform a live crawl originating from arbitrary starting points using BINGO. Next, peers publish statistical metadata about their local indexes to the directory. This metadata is subsequently used to identify promising peers for particular queries. Peers can inspect the metadata they received from remote peers, as every peer maintains a random subset of the directory[16]. Arbitrary keyword queries can be entered into a form field, just like in one of today's popular web search engines.

The metadata is used to identify a tune able number of promising remote peers for a query using query routing strategies such as CORI [4]. Users can instantly inspect the resulting peer ranking. The query is sent to these selected peers who indicate this fact in real-time. The user can also interactively validate this decision by inspecting the peers' metadata. The results obtained from the remote peers are merged into one global result list by the peer initiating the query and is presented to the user in form of a result list indicating the URL, the page title, the origin peer and the document score. Cached copies of the documents in the prepared collections have also been stored to the database, so that the user can instantly validate the relevance of a document to a query.

CONCLUSION

In this paper, we propose a search engine that is capable of indexing and searching frequently updated web information. The approach utilizes the P2P concept with DHT to achieve scalability and content coverage. Experiment results show that it has good load scalability as the contents increase. The recall and precision rate of the result are satisfactory as well.

REFERENCES

1. Bender, M.S., P. Michel, G. Triantafillou, Weikum and C. Zimmer, 2005. Improving collection selection with overlap awareness in p2p search engines. In sigir.
2. Saravanan T. and R.Udayakumar, 2013. Optimization of Machining Hybrid Metal matrix Composites using desirability analysis, Middle-East Journal of Scientific Research, ISSN:1990-9233, 15(12): 1691-1697.
3. Bender, M.S., G. Michel, Weikum and C. Zimmer, 2005. The minerva project: database selection in the context of p2p search. In BTW, pp: 125-144.
4. Bookmark-nduced Gathering of Information with Adaptive Classification into Personalized Ontologies. http://www.mpi-sb.mpg.de/units/ag5/software/bingo/
5. Callan, J., 2000. Distributed information retrieval. Advances in information retrieval, Kluwer Academic Publishers., pp: 127-150.
6. Saravanan, T., G. Saritha and R. Udayakumar, 2013, A Robust H-Infinity Two Degree of Freedom Control for Electro Magnetic Suspension System, Middle-East Journal of Scientific Research, ISSN:1990-9233, 18(12): 1827-1831.
7. Huebsch, R., J.M. Hellerstein, N. Lanham, B.T. Loo, S. Shenker and I. Stoica, 2003. Querying the internet with pier. In VLDB, pp: 321-332.
8. Li, J., B. Loo, J. Hellerstein, F. Kaashoek, D. Karger and R. Morris, 2003. On the feasibility of peer-to-peer web indexing and search. In In 2nd International Workshop on Peer-to-Peer Systems (IPTPS).
9. Nottelmann, H. and N. Fuhr, 2003. Evaluating different methods of estimating retrieval quality for resource selection. In SIGIR, pp: 290-297. ACM Press.
10. Udayakumar, R., V. Khanna, T. Saravanan and G. saritha, 2013. Cross Layer Optimization For Wireless Network (Wimax),Middle-East Journal of ScientificResearch,ISSN:1990-9233,16(12):1786-1789.
11. Ratnasamy, S., P. Francis, M. Handley, R. Karp and S. Schenker, 2001. A scalable content-addressable network. In SIGCOMM, pp: 161-172. ACM Press,
12. Rowstron, A. and P. Druschel, 2001. Pastry: Scalable, decentralized object location and routing for large-scale peer-to-peer systems. In IFIP/ACM Middleware, pp: 329-350.

13. Stoica, I., R. Morris, D. Karger, M.F. Kaashoek and H. Balakrishnan, 2001. Chord: A scalable peer-to-peer lookup service for internet applications. In SIGCOMM, pp: 149-160. ACM Press.
14. Suel, T., C. Mathur, J. Wu, J. Zhang, A. Delis, M. Kharrazi, X. Long and K. Shanmugasunderam 2003. Odissea: A peer-to-peer architecture for scalable web search and information retrieval. Technical report, Polytechnic Univ.
15. Thooyamani, K.P., V. Khanaa and R.Udayakumar, 2013. Blue tooth broad casting server, Middle-East Journal of Scientific Research, ISSN:1990-9233, 15(12): 1707-1712.
16. Saravanan, T. and R. Udayakumar, 2013. Optimization of Machining Hybrid Metal matrix Composites using desirability analysis, Middle-East Journal of Scientific Research, ISSN:1990-9233, 15(12): 1691-1697.