The SOWES Approach to P2P Web Search Using Semantic Overlays

Christos Doulkeridis
Dept. of Informatics
AUEB
Athens, Greece
cdoulk@aueb.gr

Kjetil Nørvåg
Dept. of Computer Science
NTNU
Trondheim, Norway
Kjetil.Norvag@idi.ntnu.no

Michalis Vazirgiannis
Dept. of Informatics
AUEB
Athens, Greece
mvazirg@aueb.gr

ABSTRACT

Peer-to-peer (P2P) Web search has gained a lot of interest lately, due to the salient characteristics of P2P systems, namely scalability, fault-tolerance and load-balancing. However, the lack of global knowledge in a vast and dynamically evolving environment like the Web presents a grand challenge for organizing content and providing efficient searching. Semantic overlay networks (SONs) have been proposed as an approach to reduce cost and increase quality of results, and in this paper we present an unsupervised approach for distributed and decentralized SON construction, aiming to support efficient search mechanisms in unstructured P2P systems.

Categories and Subject Descriptors

H.3.3 [**Information storage and retrieval**]: Information search and retrieval; H.3.1 [**Information storage and retrieval**]: Content Analysis and Indexing; I.7.m [**Document and text processing**]: Miscellaneous

General Terms

Algorithms, Design, Performance

Keywords

Distributed and peer-to-peer search, semantic overlay networks

1. INTRODUCTION

The advent of the World Wide Web in combination with efficient search engines like Google and Yahoo! has made an enormous amount of information easily available for everybody. It is generally accepted nowadays that current web search technologies work well and have made access to huge information corpora feasible. They also provide efficient ranking, thus easing separation of important results. Nevertheless, this shining image of centralized web indexing and searching is starting to blur and a number of issues question their future applicability: 1) current search engines only cover a small fraction of the documents on the Web, for example recent studies have shown that Google currently indexes less than 1 % of the total (static and dynamic) Web [3], 2) many web pages are rarely accessed by search engines, resulting in outdated search index contents, 3) web search providers possess the control over information, what should be indexed, how it is presented, the ranking of indexed pages and the ability to filter out material that is deemed controversial (censoring), and 4) web search is still based on exact matching techniques, calling for large scale and credible use of semantics.

These aforementioned issues can be solved by employing a P2P network of web servers, where participants are willing to share data and computing cycles, and where the control of information is completely distributed among the participating peers. In this paper, we present SOWES, a scalable approach to P2P web searching. Previous approaches to P2P web search have relied on the use of structured P2P networks, storing indexing information in DHTs [2, 4]. However, scalability and support for semantics can be difficult in such systems, so instead we base our approach on unstructured P2P networks. Such systems in their basic form suffer very high search costs in terms of both consumed bandwidth and latency, so in order to be useful for real applications, more sophisticated search mechanisms are required. We solve this problem by employing *semantic overlay networks* (SONs) [1], where peers containing related information are connected together in separate overlay networks. If SONs have been created, queries can be forwarded to only those peers containing documents that satisfy the constraints of the query context, for example based on topic, user bookmarks/profiles or features extracted from previous queries.

We will in the following describe the SOWES approach to web search, based on a scalable approach for SONs creation.

2. SOWES

A main feature of SOWES is the unsupervised, distributed and decentralized generation of SONs. Peers that contain semantically related information become part of a logical cluster (i.e., SON). Connections are created between the peers within a cluster. In this way, a query relevant to a cluster can be forwarded only to peers that belong to this cluster. Each cluster is represented by one or more special peers, henceforth named *cluster representatives*, which maintain connections to other cluster representatives and they are responsible for forwarding queries to the most relevant clusters, avoiding the excessive cost of flooding the entire network. Two important challenges in a P2P system employing SONs are 1) creation of SONs and 2) SON-based searching.

2.1 SOWES Construction

Creation of semantic overlays in SOWES is a multi-phase distributed process. We assume that peers store documents and have already organized themselves in an unstructured P2P network.

In the first phase, clustering takes place on local data on each peer. Clustering is performed on *feature vectors* F_i (a vector of tuples where each tuple contains a feature f_i and a weight w_i) extracted from documents using a feature extraction process that includes tokenization, stemming, stop-word removal and finally keeping the top-k features based on their local TF/IDF values. Feature vectors are also used to represent clusters. In the subsequent phases, first 1) local zones of peers are created, by randomly selected peers (*initiators*), then 2) semantic clusters are formed, based

WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-332-9/06/0005.

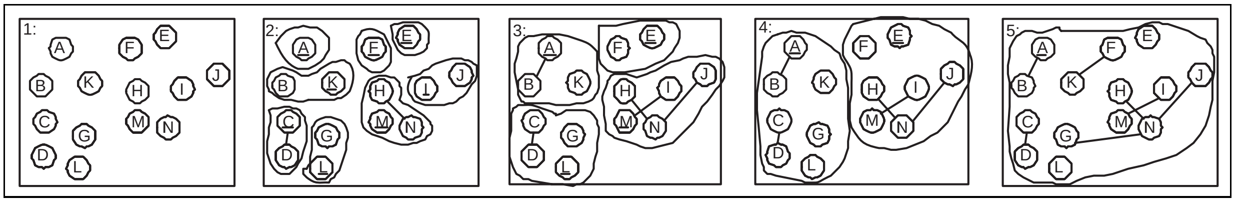

Figure 1: Example of SOWES creation.

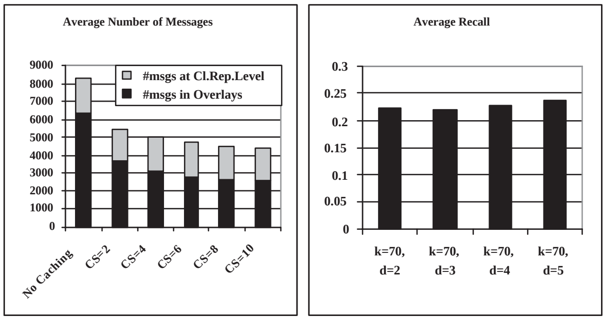

Figure 2: Average number of messages and average recall.

on data stored on the peers in a zone. In the next steps, zones and clusters are merged recursively until global clusters are obtained.

An example of SOWES creation is illustrated in Figure 1. In the first step, local clustering is performed, for simplicity we assume in the example that only one cluster is formed at each peer and use a zone size of $S_Z = 2$ peers. The zone initiators in each step are identified by underlined identifiers, and the zones at each step are illustrated by lines around the peers in the zone. In the second step, a certain percentage of the peers are selected to be initiators and each creates a zone consisting of an average of S_Z peers. At this point, clusters that are similar are merged and links created between the clusters. Note that in the figure only one link is used between cluster elements (peers or existing clusters), but in general d ($d > 1$) such links will be created, to ensure fault-tolerance. Subsequently, the merged clusters will be treated as one cluster, i.e., represented by one feature vector and a list of representative peers. In the third step, a percentage of the existing zone initiators are selected to be zone initiators for the next level, and create new zones where each new zone consists of an average of S_Z zones. Again similar clusters are merged, in this example the clusters containing A and B are merged into a new cluster, the H/N-cluster is merged with the J cluster, and M with the I cluster. In the fourth step, new zones are created, but none of the clusters are similar enough to be merged. In the fifth and final step, the zone covers the entire network, F is merged with K, and G is merged with the H/N/J cluster. Thus the final result is the following clusters: A/B, C/D, K/F, E, L, G/N/J/H, and M/I. The connections between the clusters, created at the final step, are not shown in the figure.

2.2 Searching in SOWES

A query for web documents originates from a querying peer Q_P. Routing is performed by directing and evaluating the query to appropriate peers and then returning matching results. In our context, query processing is performed by first determining which clusters might contain relevant data (*inter-cluster routing*), followed by searching one or more of these clusters (*intra-cluster routing*).

Inter-cluster routing refers to query routing at cluster representative level, which aims to identify similar cluster descriptions to the query. In order to limit the costly intra-cluster searching, the inter-cluster routing is performed in two steps. In the *first step*, a search for appropriate clusters is performed, using the links created among cluster representatives to route queries. In the *second step*, Q_P determines the most appropriate clusters (based on the results of step 1), and selectively forwards the query for intra-cluster searching. In order to reduce the cost of subsequent searches, the cluster representatives cache the result of queries, in order to reduce the cost of intra-cluster searching for frequent queries.

In order to demonstrate the feasibility and efficiency of SOWES, we have developed a simulation environment that covers all phases of SOWES generation and searching. In the experiments a number of peers (up to 20,000) were interconnected using a topology created by the GT-ITM topology generator. We used the Reuters text collection and we generated random queries from the terms. In Figure 2, the reduction of the average number of (intra-cluster routing) messages is depicted, when caching of query results is employed for various cache sizes. Also the associated average recall is presented for different values of d and top-k, for a network of 8,000 peers. These premature results exhibit small recall, however we believe that better distributed clustering algorithms will improve the search results and we intend to focus on this issue in our future work. For more details on the experiments we refer to [5].

The main contributions of our approach are 1) an architecture for P2P web search based on an unstructured P2P network, 2) algorithms for unsupervised, distributed and decentralized construction of semantic overlays, and 3) appropriate algorithms for web search based on the use of SONs. Our work differs from super-peer approaches, since cluster representatives do not have a constant role and any peer can take their place. Furthermore, when compared to gossiping protocols, our approach guarantees that a query will reach a thematically relevant peer, no matter how many hops are required, since the created overlays cover the entire network, grouping similar peers together. In our future work, we intend to improve the quality and efficiency of SOWES.

3. REFERENCES

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

[2] F. Cuenca-Acuna, C. Peery, R. Martin, and T. Nguyen. Planetp: Using gossiping to build content addressable peer-to-peer information sharing communities. In *Proceedings of the 12th IEEE International Symposium on High Performance Distributed Computing*, 2003.

[3] J. Li, B. Loo, J.M.Hellerstein, M. Kaashoek, D. Karger, and R. Morris. On the feasibility of peer-to-peer web indexing and search. In *Proceedings of the 2nd IPTPS'03*, 2003.

[4] S. Michel, P. Triantafillou, and G. Weikum. MINERVA Infinity: A Scalable Efficient Peer-to-Peer Search Engine. In *Proceedings of Middleware'05*, 2005.

[5] K. Nørvåg, C. Doulkeridis, and M. Vazirgiannis. Semantic overlays for P2P web searching. Technical report, AUEB, 2005 `http://www.db-net.aueb.gr/index.php/publications/technical_reports/`.