Supporting Ranked Search in Parallel Search Cluster Networks

Fang Xiong Qiong Luo Dyce Jing Zhao
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
{xfang, luo, zhaojing}@cs.ust.hk

ABSTRACT

We investigate how to support ranked keyword search in a Parallel Search Cluster Network, which is a newly proposed peer-to-peer network overlay. In particular, we study how to efficiently acquire and distribute the global information required by ranked keyword search by taking advantage of the architectural features of PSCNs.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval–*search process*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software–*distributed systems, information networks, performance evaluation (efficiency and effectiveness)*

General Terms: Measurement, Performance

Keywords: ranked keyword search, parallel search cluster networks, peer-to-peer networks

1. INTRODUCTION

With the increasing scale and sophistication of P2P networks, keyword search techniques have been developed for data object ID search as well as for content search in various P2P network overlays. In keyword-based content search, one widely adopted ranking mechanism is TF×IDF [3]. However, this ranking process requires aggregate information such as the total number of documents that contain a specific keyword. In a dynamic, decentralized P2P network environment, special care needs to be taken to meet this requirement. For instance, PlanetP [2] approximates TF×IDF at the peer level in order to save storage and communication cost in an unstructured P2P network. As another example, Shen et al. [4] build a hierarchical summary structure that indexes at document, peer, and super-peer levels for a super-peer network.

Since the Parallel Cluster Search Network [1], or PSCN in short, is a newly identified P2P architecture, we study how to support the ranking mechanism efficiently by taking advantage of its architectural characteristics. A PSCN is composed of clusters of peers that are connected through FSLs (Forwarding Search Links) and NILs (Non-forwarding Index Links) [1]. In each cluster, peers are connected with FSLs, which transmit queries and results and allow the recipients to forward the received content. Between two clusters, there is one NIL from each peer in one cluster to one randomly selected peer in the other cluster, to transmit indexes and index updates; however, further forwarding is not allowed. Figure 1 shows an example PSCN, with the outgoing links from clusters 2 and 3 omitted for clarity.

CIKM'05, October 31–November 5, 2005, Bremen, Germany.
ACM 1-59593-140-6/05/0010.

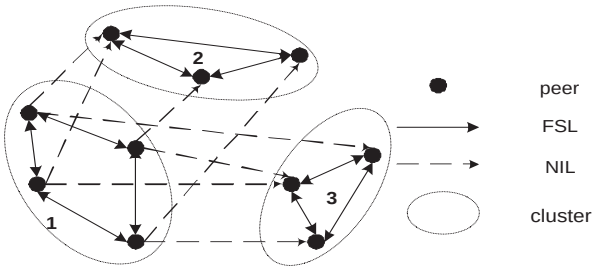

Figure 1: A Parallel Search Cluster Network (PSCN)

The process of ranked search in a PSCN is as follows:

1. The local index is built at each peer and transmitted across clusters through NILs so that each cluster has the indexes of all nodes in the network.

2. At the query time, a query is transmitted within a cluster through FSLs, and the query processing load is distributed over all peers in this cluster.

3. Finally, the querying peer merges the locally ranked query results into globally ranked ones and returns all or top-K of them to the user.

As shown in this process, there is no global index in a PSCN and query processing is distributed over peers evenly. Moreover, since each query is answered within one cluster, communication cost is saved.

2. RANKED KEYWORD SEARCH IN A PSCN

For ranking, we adopt the TF×IDF implementation suggested by Witten et al. [5]:

$$w_{D,t} = 1 + log(f_{D,t}) \quad w_{Q,t} = log(1 + N/f_t) \qquad (1)$$

where $w_{D,t}$ is the weight of term t in document D, $f_{D,t}$ the number of times that term t occurs in document D, $w_{Q,t}$ the weight of term t in query Q, N the total number of documents in the collection, and f_t the number of documents in which term t occurs.

Subsequently, the similarity between document D and query Q is calculated as follows:

$$Sim(Q, D) = \frac{\sum_{t \in Q} w_{D,t} \times w_{Q,t}}{|D|} \quad (2)$$

where $|D| = \sum_{t \in D} f_{D,t}$.

Each peer maintains the inverted index of its local documents, but it lacks the global aggregate information used in TF×IDF calculation, such as N, f_t and $|D|$. These global information will be obtained at the query time as necessary.

2.1 Search at the document level

In a PSCN, local indexes are replicated over the network through NILs. Since there is a NIL from each peer to one randomly selected peer in each of the other clusters, collectively the indexes of all peers in the network are available in each cluster.

At the query time, the querying peer forwards the query through FSLs to other peers in the cluster that it resides in. Upon receiving the query, the participating peers collect the local aggregate information related to the query based on the indexes they store locally, and return them to the querying peer. The local aggregate information are merged into the global aggregate information at the querying peer, and are distributed in the cluster. Next, each participating peer evaluates the query over the indexes it stores, ranks its query results locally, and sends the locally ranked query results back to the querying peer. Finally, the querying peer merges the locally ranked query results into globally ranked ones and returns all or top-K of them to the user.

In summary, ranked search in a PSCN has the following characteristics: (a) the most time-consuming tasks in the process of handling a ranked search query, i. e., local aggregate information collection and local ranking calculation, are distributed over all peers in one cluster; this distribution reduces search load over peers evenly, reduces the maximum requirement for the capabilities of individual peers and improves the scalability of the network; (b) there is no global index built for every existing keyword in the network; this reduces the indexing load on individual peers; (c) each query is answered within the cluster where the query is submitted without affecting the completeness of the query result, which improves the query response time and saves communication cost.

2.2 Search at the peer level

Similar to the previous work, the indexes transmitted across clusters in a PSCN can be at the peer level, instead of at the document level. At the peer level, the global ranking is done for peers to indicate which peers are most likely to possess matching documents. To obtain these documents, the querying peer needs to forward the query to the top-k_p peers in the global peer rank, each of which in turn ranks its documents locally. Finally, the querying peer merges the query results and selects the global top-K documents.

2.3 Discussion

In a P2P environment, the local index at a peer can be replicated over the network in various ways for different overlays, and ranked search is performed based on the replicated indexes. We have presented the process of ranked search in a PSCN, but the same process can be applied in other overlays with slight modification. In our experiments, we modified the algorithm for a super-peer network and an unstructured P2P network to make comparison with the PSCN.

Advanced information retrieval methods, e.g., LSI, can be applied on top of TF×IDF, to improve the quality of query results and the efficiency of query processing. Nevertheless, a simple but basic tenique as TF×IDF is sufficient for the purpose of comparing ranked search in different overlays.

3. EVALUATION RESULTS

We have experimented with ranked search in a PSCN in comparison with that in a super-peer or unstructured network. The default K value of top-K is 20.

Through the experiments, we see that compared with the processing time spent on local aggregation information collection and local ranking calculation, the time used to merge local aggregation information or to merge locally ranked results, is negligible. This suggests that it is beneficial to distribute the search workload over peers; otherwise, the bottleneck will be at the super-peers in a super-peer network or at the querying peer in an unstructured network. As a result, the processing time and the storage cost per peer in a PSCN is the lowest among the three overlays.

However, the downside of a PSCN is the flooding communication within a cluster and the index replication cost across clusters. The super-peer network wins on the network bandwidth usage and the total storage cost due to the directed query forwarding from normal peers to super-peers and the absence of index replication among normal peers.

Additionally, we find that compared with document-level indexes, peer-level indexes save 70% of the processing time, 30% of the network bandwidth usage and 30% of the storage space, with a slight decrease in precision.

4. CONCLUSION

We have presented our approach to supporting the basic TF×IDF ranking for ranked search in a PSCN and have conducted experiments to study its performance in comparison with a super-peer network and an unstructured network. We find that ranked search can be done efficiently in a PSCN by taking advantage of the architectural features. The most time-consuming tasks in the ranked search are distributed over the peers evenly, and the storage cost per peer is low. In the future work, we plan to add more advanced indexing techniques, to further reduce the index size and to improve the precision of ranked keyword search in a PSCN. A full version of this paper is available as a technical report [6].

5. REFERENCES

[1] B. F. Cooper and H. Garcia-Molina. Studying search networks with SIL. In *IPTPS*, 2003.

[2] F. M. Cuenca-Acuna, C. Peery, R. P. Martin, and T. D. Nguyen. PlanetP: Using gossiping to build content addressable peer-to-peer information sharing communities. In *HPDC-12*, 2003.

[3] G. Salton, A.Wang, and C. Yang. A vector space model for information retrieval. *Journal of the American Society for Information Science*, 18, 1975.

[4] H. T. Shen, Y. Shu, and B. Yu. Efficient semantic-based content search in p2p network. *IEEE TKDE*, 16(7), 2004.

[5] I. H. Witten, A. Moffat, and T. C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann, second edition, 1999.

[6] F. Xiong, Q. Luo and D. J. Zhao. Supporting Ranked Search in Parallel Search Cluster Networks. Technical report HKUST-CS05-14, HKUST, 2005.