Developing an Extensible Framework for Content Based Searching in Super Peer P2P Network

Muhammad Nazrul Islam*, Md. Ashiqul Islam* , Imam Jafar Shadaque* , and Md. Razib Hayat Khan†
* Khulna University of Engineering & Technology, CSE Department, Khulna, Bangladesh,
e-mail: nazrul_bd80@*yahoo.com*
† Royal Institute of Technology (KTH), Department of Computer and System Sciences, Sweden,
e-mail: raz_cit@yahoo.com

***Abstract*—Searching is an important factor in p2p network for content retrieval. Most of the searches in p2p system are title-based with their limited functionality. Without knowing the unique filename we can't retrieve the content of the file in title based search. Here super peer p2p network is designed that supports content-based search for relevant documents. At the beginning, a general and extensible framework is proposed which is based on hierarchical summary structure for searching similar documents in p2p network. The summary structure is formed by Vector Space Model (VSM), Latent Semantic Indexing (LSI) and Singular Value Decomposition (SVD) techniques. Than an effective document searching is developed by summarizing and maintaining all documents within the network with different factors. Finally at the end, the experimental result is verified on a real p2p prototype and large scale network. The results show the effectiveness, efficiency and scalability of the proposed system.**

***Keywords*—Peer-to-peer, Content search, Hierarchical summary, Indexing, Title-based search.**

I. INTRODUCTION

Peer-to-Peer (P2P) computing has recently attracted a great deal of research attention. In a P2P system, a large number of nodes (e.g., PCs connected to the Internet) can potentially be pooled together to share their resources, information and services. Our system has several key features: unit level, peer level and super peer level. Summaries are first represented as vectors, which are further optimized by LSI [1] techniques and represented as high-dimensional points. In this paper, we address the problem of semantic-based content search in the context of document retrieval. Given a query, which may be a phrase, a statement or even a paragraph, we look for documents that are semantically close to the query. We propose a general and extensible framework for semantic-based content search in P2P network. The super-peer P2P architecture [18] which is more efficient for contents look-up is employed as the underlying architecture. To facilitate semantic-based content search in such a setting, a novel indexing structure called Hierarchical Summary Indexing Structure, is proposed. With such an organization, all information within the network can be summarized with different granularity, and then efficiently indexed.

II. RELATED WORKS

We will first review previous work on P2P architecture. [2] Extends [3]'s hybrid architecture to design *super*-peer network. Summary techniques are crucial in P2P systems. Effective summarization of peer information is absolutely needed in P2P network. In this paper, we propose a hierarchical summary indexing structure for efficient semantic-based content search in super-peer P2P network, which can support complex semantic-based queries. Another related area is high-dimensional indexing. In the literature, many high-dimensional indexing methods have been proposed [4]. However, existing methods are typically not efficient for more than 30-dimensions and are not scalable [5] due to the "dimensionality curse" phenomenon when the dimensionality reaches higher.

III. GENERAL FRAMEWORK FOR SUPER PEER P2P BASED SEARCH

In this section, we present a novel Hierarchical Summary Indexing framework for P2P-based document search system. We shall first discuss the super-peer P2P architecture, and then look at how such a structure can facilitate the design of the proposed framework.

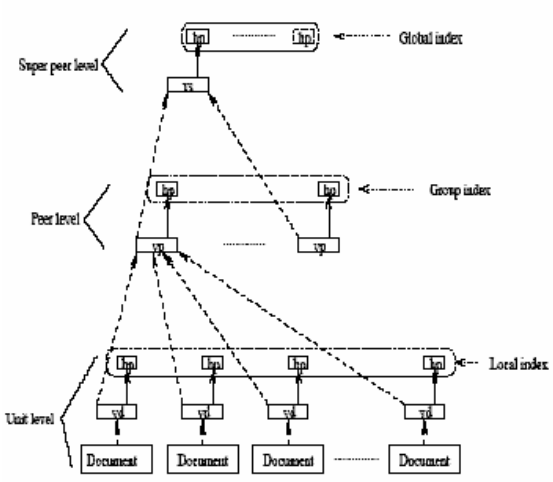

Fig. 1. Hierarchical Summary Indexing Structure.

A. Super- peer P2P Network

A *super-peer* is a node in a peer-to-peer network that operates both as a server to a set of clients, and as an equal in a network of super-peers. A straightforward query processing mechanism in super peer network works as follows. A peer (client) submits its query to the super peer of its group. The super peer will then broadcast the query to other peers within the group. At the same time, the

super peer will also broadcast the query to its neighboring super peers.

B. Hierarchical Summary Indexing Structure

Summarization is a necessary step for efficient searching, especially when the amount of information is very large. A summary is a very compact representation. In our framework, we introduce a new interesting concept, Hierarchical Summary Indexing Structure (Summary and Indexing), which is closely related to the super-peer P2P architecture we employed. Our scheme essentially summarizes information at different levels. We have employed three levels of summarization in our framework. The lowest level, named as *unit level*, an information unit, such as a document is summarized. In the second level, named as *peer level*, all information owned by a peer is summarized. Finally, in the third level, named as *super level*, all information contained by a peer group is summarized. Figure 1 depicts such a structure for document summary. Figure 1 shows each level of summary has a corresponding index built on top of it. Figure 2 shows the hierarchical summary indexes in a peer group.

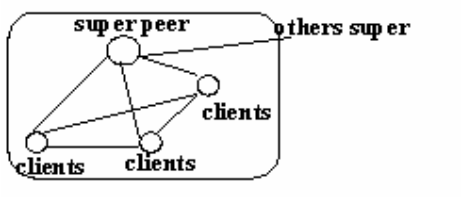

Fig. 2. Summary indices in a peer group.

Figure 1 depicts such a structure for document summary. Figure 1 shows each level of summary has a corresponding index built on top of it. Figure 2 shows the hierarchical summary indexes in a peer group.

IV. SEMANTIC BASED CONTENT SEARCH SYSTEM

Suppose there are a large number of peers in the network, and each peer contains a large number of documents, what we want to achieve is to find the most relevant documents as quickly as possible.

A. Building Summary

For each level, our summarization process consists of two steps by techniques of Vector Space Model (VSM) [6] and Latent Semantic Indexing (LSI) [1] respectively. Briefly, in VSM, documents and queries are represented by vectors of weighted term frequencies. Latent Semantic Indexing (LSI) has been proposed to optimize the vector prepared by VSM. A technique known as Singular Value Decomposition (SVD) is used to reduce this concept space into a much lower dimensionality.

Algorithm 1 indicates the main routine of building summary in the hierarchical structure as shown in figure 3. We illustrate the algorithm in Example 1.

1. foreach peer
2. foreach document
3. generate its vector vd by VSM
4. generate peer weighted term dictionary vp
5. foreach document vector vd
6. transform it into D(vp) dimensionality
7. generate high dimensional point for vd by SVD
8. pass vp to its super peer
9. foreach super peer
10. generate group weighted term dictionary vs
11. foreach vp
12. transform it into D(vs) dimensionality
13. generate high dimensional point for vp by SVD
14. pass vs to other super peers
15. generate global weighted term dictionary vn
16. foreach vs
17. transform it into D(vn) dimensionality
18. generate high dimensional point for vs

Fig. 3. Algorithm 1 for building hierarchical structure summaries.

EXAMPLE 1 (AN EXAMPLE OF HIERARCHICAL SUMMARY BUILDING) *Table 1 provides a small P2P network with eight documents;* di_m^n *represents the* i^{th} *document which is in* m^{th} *peer of* n^{th} *group. The process of summary building is depicted in Figure 4, where the*

TABLE I.
A TABLE OF DOCUMENTS

Id	Document
$d1_1^1$	Monitoring XML data on the web
$d2_1^1$	Approximate XML joins
$d3_1^2$	Outlier detection for high dimensional data
$d4_1^2$	High dimensional indexing using sampling
$d5_2^1$	Document clustering with committees
$d6_2^1$	Document clustering with cluster refinement
$d7_2^2$	Title language model for information retrieval
$d8_2^2$	Document summarization in information retrieval

summary dimensionality for each level is reduced to 2. For simplicity, the weight of a term is represented by its frequency only. As we can see, vectors of documents vds within a peer form the vector of peer vp. Based on vp, each vd is transformed into D(vp)-dimensional vector which is then reduced into a 2-dimensional document summary by SVD. Take a look at the first peer which

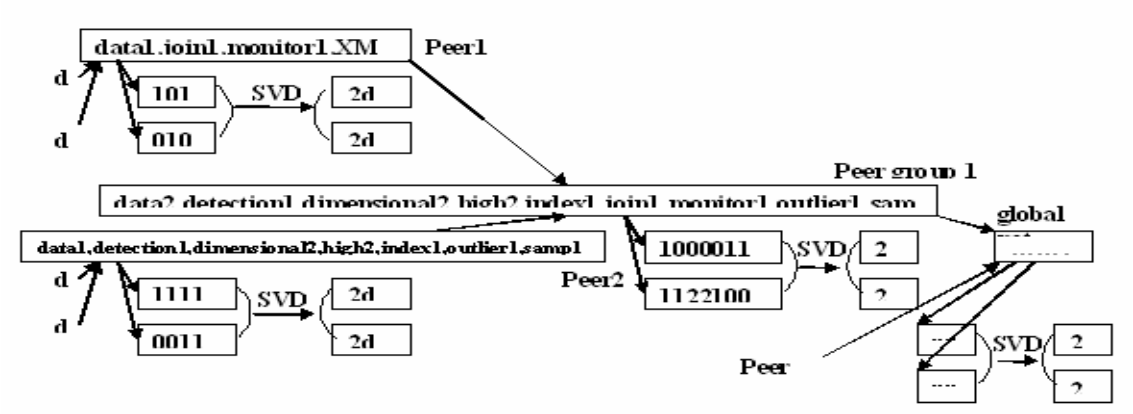

Fig. 4. An example of hierarchical summary building.

contains documents $d1_1^1$ *and* $d2_1^1$. *Both documents are merged to form its vp of together with term weights, where D(vp) is 5. Based on vp, both documents are mapped into 5-dimensional vectors of (1,0,1,1,1) and (0,1,0,1,0) respectively, which are in turn reduced into a much lower 2-dimensional points by SVD. Same process is applied to generate a peer's and supper peer's summary.*

B. Query Processing

Figure 4 depicts how a query is being processed in a P2P network. In the figure, dark arrow indicates the direction of a query being transmitted and blank arrow indicates the route of results being returned. When a peer issues a query Q, Q is first passed to its super peer, followed by the hierarchical indexing search in order of global index, group index and peer index, which is the reverse order of the summary construction.

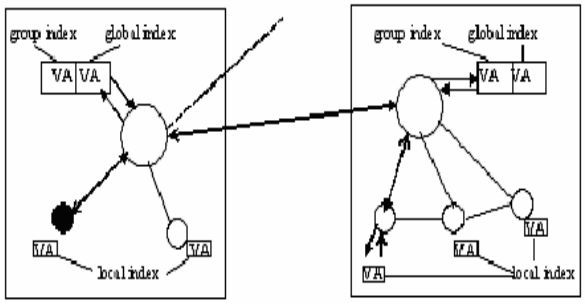

Fig. 5. The routing of query processing initialized by dark peer.

V. Semantic Based Content Search System

One crucial difference between P2P and traditional information retrieval is that P2P network is dynamic in nature. A peer can join and leave the network at any time. Hence the summarization and indexing techniques have to be able to handle dynamic operation efficiently. We propose the following peer insertion algorithm 2 in our hierarchical indexing structure as shown in figure 6.

$$AIR(dic, dic_{future}) = \frac{\sum_{i=0}^{D[dic_{future}]} \left| dic[i] - dic_{future}[i] \right|}{\sum_{i=0}^{D[dic_{future}]} dic[i]}$$

.Where *dic* and dic_{future} are the current and future weighted term dictionary at group or global level (*vs* or *vn*). $D[dic_{future}]$ is the dimensionality of the future dictionary.

1. Build peers local index
2. Pass peers vp to its super peer
3. if $AIR_{group} > \theta_{group}$
4. Rebuild and index group peers summary
5. Update super peers
6. Broadcast vs to other super peers
7. for each super peer
8. if $AIR_{global} > \theta_{global}$
9. Rebuild and index super peers summary
10. else
11. Generate peers high - dimensional point
12. Insert the point into groups index

Fig. 6. Algorithm 2 for peer insertion.

VI. Cost Models

In this section, we will evaluate our hierarchical summary indexing structure in a super-peer network based on the following types of metrics: *Storage Overhead, Query Response Time and Loads*.

A. Storage Overhead

The storage overhead in our structure includes peer overhead and super peer overhead. For peer overhead, each peer contains its documents' summary; local index built on the document summary, local current and future

dictionaries, together with the SVD's Singular Vectors. Hence the total peer Storage Overhead (SO) is:

$$SO_{peer}=D_{doc}*N_{doc}+VA_{local}+2D(vp)+D_{doc}*D(vp)$$

Where D_{doc} is the dimensionality of summarized high-dimensional points of documents, N_{doc} is the number of documents in the peer , VA_{local} is the size of local VA-file on points of documents, and D(vp) is the dimensionality of peer's term dictionary.

B. Query Response Time and Load

The client peer first forwards its query to its group's super peer. Its super peer then searches its global index and selects the K_{group} most relevant groups to which it forwards the query. In each selected group, the super peer searches its group index and forwards the query to the K_{peer} most relevant individual peers. So the times of a query being forwarded is:

$$Times_{query} = 1 + K_{group} + K_{group} * K_{peer}$$

At each level of index, KNN search is performed. Correspondingly, the total processing time is computed as:

$$Time=Time_{global}+K_{group}*Time_{group}+K_{group}*K_{peer}*Time_{local}$$

Where $Time_{global}$ refers to the processing time of KNN search in global VA-file.

VII. Experimental Result And Discussion

We have evaluated the proposed hierarchical summary and indexing scheme in a real P2P setting as well as via simulation. In this section, we report the results of the performance study.

A. Experimental Setup

Table 2 gives some experiment parameters and their default settings for both the real system and the simulator respectively.

TABLE II.
PARAMETERS AND SETTENGINS

Name	Default Value	Description
Network Type	Power-Law	Topology of the network, with out degree 3.2
Max User Wait Time	60s	Time for a user to wait an answer
Query Rate	8e-3	The expected number of queries per user per second
TTL	5	Time-To-Live of an message
Network Size		Number of peers in the network
Peer Group Size		Number of peers in each peer group
K_{group}		Number of super peers to return
K_{peer}		Number of peers for a super peer to return
K_{doc}		Number of documents for a peer to return

B. Retrieval Precision

In this experiment, we examine the effectiveness of our summary technique. We first implement a relatively small real network to show that our proposals are very practical and applicable to P2P systems. Our real network has 30 nodes. We use 4 benchmark collections of documents which were used by Smart [7], together with their queries and human ranking. Table 3 presents the characteristics of the datasets.

TABLE III.
CHARACTERISTICS OF REAL DATASETS

	MED	CISI	CACM	TIMES
Number of documents	1033	1460	3204	425
Number of queries	30	76	64	83
Number of terms occurring in more than one document	5831	5743	4867	10337

i) Effect of Dimensionality: The precision is measured by the ratio of the number of relevant documents over the number of returned documents. Figure 7 shows the changes of the average precision when the summary for the documents is reduced to different dimensions with SVD technique.

Next, we study the retrieval precision at the group level with the peer level summaries. Figure 8 illustrates the variation of the average precision as the number of dimension increases. Lastly, we repeat the experiment on the highest level of hierarchy to test if the correct groups that contain the relevant documents can be returned.

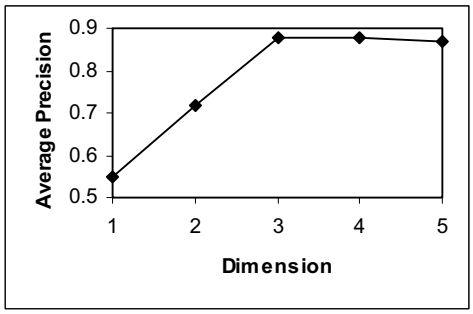

Fig. 7. Document Level Summary Precision

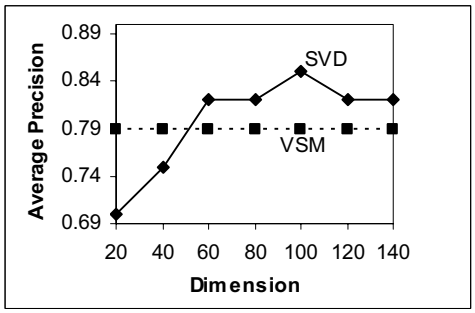

Fig. 8. Peer Level Summary Precision.

At this level, its precision is measured by the ratio of the number of relevant peer groups over the number of returned peer groups. The result is shown in Figure 9. From Figure 8 and Figure 9, we can see that different dimensionality of summary may achieve different precision. The smallest value with highest precision is

always chosen as the final dimensionality of summary at each level.

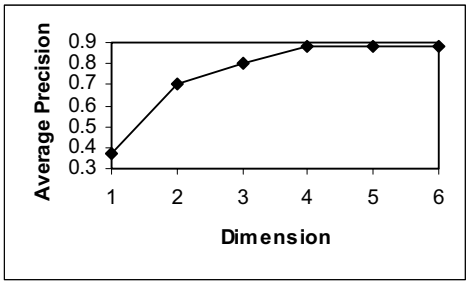

Fig. 9. Super Peer Level Summary Precision.

ii) Precision of the Whole System: In the above subsection, we have seen how the dimensionality of summary affects the precision at each individual level. In this experiment, we integrate the three levels and test the overall precision of the whole system. The precision is measured by the ratio of the number of relevant documents over the number of returned documents after the whole network has been searched. Obviously, the precision of the whole system is expected to be lower than the precision at documentation level since the precision is further reduced at higher levels.

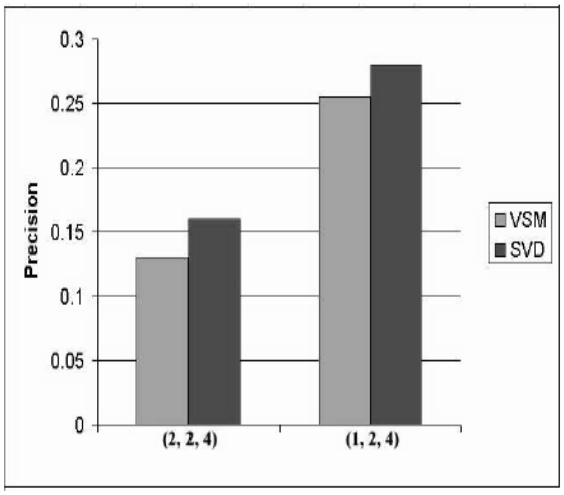

Fig. 10. Overall Hierarchical System Summary Precision.

VIII. CONCLUSION

We have examined the issues to support content-based searching in a distributed peer-to-peer information sharing system. We have proposed the first general and extensible hierarchical framework for summary building and indexing in P2P network. Based on this framework, we have presented an effective two-step summarization technique to transform large size representations of documents, peers, and super peers into small high-dimensional points. A prototype and a simulated large scale network have been designed to evaluate the system performance. Our experiments showed that such a hierarchical summary indexing structures can be easily adopted and our prototype system achieves remarkable achievements.

REFERENCES

[1] C.H. Papadimitriou, H. Tamaki, P. Raghavan, and S. Vempala, "Latent semantic indexing: A probabilistic analysis", In *PODS*, 1998.

[2] B. Yang and H. Garcia-Molina, "Designing a super-peer network", In *ICDE*, 2003

[3] B. Yang and H. Garcia-Molina, "Comparing hybrid peer-to-peer systems", In *VLDB'2001*, 2001.

[4] S. Berchtold and D. A. Keim, "Indexing high-dimensional spaces: Database support for next decade's applications", *ACM Computing Surveys*, 33(3):322–373, 2001.

[5] R. Weber, H. Schek, and S. Blott, "A quantitative analysis and performance study for similarity search methods in high dimensional spaces", In *VLDB*, pages 194–205, 1998.

[6] S. K.M. Wong, W. Ziarko, V. V. Raghavan, and P. C.N. Wong, "On modeling of information retrieval concepts in vector spaces", In *TODS*, 1987.

[7] C. Buckley, A. Singhal, M. Mitra, and G. Salton, "New retrieval approaches using smart". In *TREC 4.*, pages 25–48., 1995.

[8] Heng Tao Shen, Yanfeng Shu and Bei Yu, "Efficient Semantic-Based Content Search in P2P Network", IEEE transactions on knowledge and data engineering, vol. 16, no. 7, July 2004.

[9] F. M. Cuenca-Acuna and T. D. Nguyen. Text-based content search and retrieval in ad hoc p2p communities. In *International Workshop on Peer-to-Peer Computin*g, 2002.

[10] I. Stoica, R. Morris, D. Karger, F. Kaashoek, and H. Balakrishnan. Chord: A scalable peer-to-peer lookup service for internet applications.In *SIGCOMM*, 2001.

[11] Freenet. *http://freenet.sourceforge.com/*.

[12] P. Triantafillou, C. Xiruhaki, M. Koubarakis, and N. Ntarmos. Towards high performance peer-to-peer content and resource sharing systems. In *CIDR*, 2003.