A Survey on Search and Retrieval of Information in Peer to Peer Networks

Amrita Yadava
Department of Computer Science Engineering
Rajasthan Technical university , Kota
Jaipur, India
amrita.yadava@yahoo.com

***Abstract*- The approach builds on work in unstructured P2P systems and uses only local knowledge. This paper gives applications of content based source selection and document retrieval in peer to peer networks.**

***Index Terms*- peer to peer, retrieval, search, index, network**

I. INTRODUCTION

As Internet has become a part of our daily lives, it is very important to get all the desired things which we want. Nowadays people use internet to watch videos, their favourite tv shows, news and listen music, also with the help of various search engines like google, yahoo, bing etc. they get what they need. These search engines help us finding information in fractions of second. This project aims in doing the same task like these search engines but by using a different approach , that is peer to peer approach. This approach uses computer in our homes to do so.

II. PEER TO PEER NETWORKS

These networks uses nodes, which is a computer connected to a network. This network facilitates communication between the connected nodes through various protocols enabling many distributed applications. The Internet is the largest contemporary computer network with a prolific ecosystem of network applications. Communication occurs at various levels called layers. The lowest layers are close to the physical hardware, whereas the highest layers are close to the software. The top layer is the application layer in which communication commonly takes place according to the client-server paradigm: server nodes provide a resource, while client nodes use this resource. An extension to this is the peer-to-peer paradigm: here each node is equal and therefore called a peer. Each peer could be said to be a client and a server at the same time and thus can both supply and consume resources. In this paradigm, peers need to cooperate with each other, balancing their mutual resources in order to complete application-specific tasks. For communication with each other, during task execution, the peers temporarily form overlay networks: smaller networks within the much larger network they are part of. Each peer is connected to a limited number of other peers: its neighbours. Peers conventionally transmit data by forwarding from one peer to the next or by directly contacting other, non-neighbouring, peers using routing tables. The architecture of a peer-to-peer network is determined by the shape of its overlay network(s), the placement and scope of indices, local or global, and the protocols used for communication. The choice of architecture influences how the network can be utilised for various tasks such as searching and downloading. In practice the machines that participate in peer-to-peer networks are predominantly found at the edge of the network, meaning they are not machines in big server farms, but computers in people's homes. Because of this, a peer-to-peer network typically consists of thousands of low-cost machines, all with different processing and storage capacities as well as different link speeds. Such a network can provide many useful applications, like: file sharing, streaming media and distributed search. Peer-to-peer networks have several properties that make them attractive for these tasks. They usually have no centralised directory or control point and thus also no central point of failure.

III. ARCHITECTURES

There ismuch architecture for a peer-to-peer network. The choice for one of these affects how the network can be searched. To enable search, one requires an index and a way to match queries against entries in this index. It is important to realize that what the index is used for is application-specific.This could be mapping filenames to concrete locations in the case of file sharing, user identifiers to machine addresses for instant messaging networks, or terms to documents in the case of information retrieval. The challenge in all the above stated cases is keeping the latency low while maintaining the properties like load balancing and self organization. Indexing, query routing and query processing are the three tasks for searching the efficiency of latency. Indexing helps in finding where the index is stored and the cost at which it is accumulated. Query routing indicates the path of query from which it is sent or received and whether that path is efficient or not to answer the query. Query processing tells that which peer performs the query since if more peers are involved in query processing then latency increases. Thus these are the three subtasks which are needed in retrieving information. Below are stated the four architectures which are used in file sharing and information retrieval.

A. Centralised Global Index

Early file sharing systems used a centralised global index located at a dedicated party, usually a server farm, that kept track of what file was located at which peer in the network. When peers joined the network they sent a list of metadata on files they wanted to share containing, for example, filenames, to the central party that would then include them in its central index.. The most famous example of this type of network is Napster. This approach avoids many problems of other peer-to-peer systems regarding query routing and index placement. However, it has at least two significant

drawbacks. Firstly, a central party limits the scalability of the system. Secondly, and more importantly, a central party forms a single point of technical, and legal, failure.

B. Distributed Global Index

In this system, both the index and the data are distributed in such networks. These indices conventionally take the form of a large key-value store: a distributed hash table . When a peer joins the network it puts the names of the files it wants to share as keys in the global index and adds its own address as value for these filenames. Other peers looking for a specific file can then obtain a list of peers that offer that file by consulting the global distributed index. Each peer stores some part of this index. There are many ways in which a hash table can topologically be distributed over the peers. However, all of these approaches have a similar complexity for lookups: typically O (log n), where n is the total number of peers in the network.

C. Strict Local Indices

Peers join the network by contacting bootstrap peers and connecting directly to them or to peers suggested by those bootstrap peers until reaching some neighbour connectivity threshold. A peer simply indexes its local files and waits for queries to arrive from neighbouring peers. An example of this type of network is the first version of Gnutella. This network performs search by propagating a query from its originating peer via the neighbours until reaching a fixed number of hops, a fixed time to- live, or after obtaining a minimum number of search results: query. Unfortunately, this approach scales poorly as a single query generates massive amounts of traffic even in a moderate size peer-to-peer. Thus, there have been many attempts to improve this basic flooding approach.

D. Aggregated Local Indices

Networks that use this approach have at least two, and sometimes more, classes of peers: those with high bandwidth and processing capacity are designated as super peers, the remaining 'leaf' peers are each assigned to one or more super peers when they join the net-work. A super peer holds the index of both its own content as well as an aggregation of the indices of all its leafs. This architecture introduces a hierarchy among peers and by doing so takes advantage of their inherent heterogeneity. Searching proceeds in the same way as when using strict local indices. However, only the super peers participate in routing queries. Since these peers are faster and well connected, this yields better performance compared to local indices, lower susceptibility to bottlenecks, and similar resilience to churn.

IV. CHANGES MADE IN THE SYSTEM

An unstructured P2P system is used where each user stores locally its own data and performs the search and retrieval functions. It helps in reducing response time and have good load balancing properties. Thus, to optimize the overlay by establishing connections between peers based on the criterion of network proximity. In particular, peers minimize the network distance from their neighboring nodes by establishing connections to nodes that belong to the same network. Content based search and retrieval can be used in a variety of contexts. Since keyword based approach is already used earlier, therefore multi-term processing is used in this methodology. Also privacy protection and optimal search results are been focused.

V. CONCLUSION

Peer-to-peer technology is a robust solution to the ethical and technical problems and deserves more attention. This paper gives a clear definition of peer-to-peer information retrieval and what distinguishes it from related and overlapping fields, like file sharing and federated information retrieval. These concrete contributions may aid in the design and construction of a large-scale peer-to-peer web search engine, which is the primary goal.

ACKNOWLEDGMENT

I thank all staff members of my college and friends for extending their cooperation during my seminar. I would like to thank my parents without whose blessings; I would not have been able to accomplish my goal. Above all I thank the almighty God for His blessings, without which any of this would not have been possible.

REFERENCES

[1] McGraw-Hill. 2002. McGraw-Hill Dictionary of Scientific and Technical Terms 6th Ed. McGraw-Hill Professional.

[2] Zeinalipour-Yazti, D., Kalogeraki, V., and Gunopulos, D. 2004. Information Retrieval Techniques for Peer-to-Peer Networks. Computing in Science & Engineering 6,

[3] Kulathuramaiyer, N. and Balke, W.-T. 2006. Restricting the View and Connecting the Dots - Dangers of a Web Search Engine Monopoly. Journal of Universal Computer Science.

[4] White, A. 2009. Search Engines: Left Side Quality versus Right Side Profits. Working paper, Toulouse School of Economics. Dec. http:// dx.doi.org/10.2139/ssrn.1694869 (Retrieved June 25th 2012).

[5] Tene, O. 2008. What Google Knows: Privacy and Internet Search Engines. Utah Law Review 2008, 4, 1434–1490.

[6] Lewandowski, D., Wahlig, H., and Meyer-Bautor, G. 2006. The Freshness of Web Search Engine Databases. Journal of Information Science 32, 2,131–148.

[7] Bergman, M. K. 2001. The Deep Web: Surfacing Hidden Value. Journal of Electronic Publishing 7.

[8] Galanis, L., Wang, Y., Jeffery, S., and DeWitt, D. 2003. Processing Queries in a Large Peer-to-Peer System. In Proceedings of the Conference on Advanced information Systems Engineering (CAiSE'03), Klagenfurt, AT.

[9] Oram, A. 2001. Peer-to-Peer: Harnessing the Power of Disruptive Technologies 1st Ed. O'Reilly Media.