Cross Language Information Retrieval Approach in Peer-to-Peer Network

M. Archana and K.A. Sumithra Devi

Department of MCA, RV College of Engineering,
Bangalore-59, India
archanams_m@yahoo.com,sumithraka@gmail.com

Abstract. Peer-to-Peer systems have emerged as popular way of sharing large volume of data. It is an application layer networks which enables network host to share resources in a distributed manner. The usability of these systems depends on effective search techniques to retrieve data. In this paper, an approach is made to list out some of the searching techniques that are applicable for the peer-to-peer network. However, most of the Peer-to-Peer information Systems is still unaware of some important features, such as cross-language information retrieval. Cross-language information retrieval is the state-of-art research area in the information retrieval research area.

Keywords: Peer-to-Peer, Cross-Language Information retrieval (CLIR), Search, Translation.

1 Introduction

Peer-to-Peer can be viewed as a communication model in which each computer has the same capabilities as the other. Any computers can initiate the communication session and it is implemented by giving each communication node both server and client capabilities, but in the recent years internet is used to exchange the information with each other directly or through an intermediate. Peer-to-Peer (P2P) networks are increasingly becoming popular because they offer opportunities for real-time communication, ad-hoc collaboration and information sharing in a large-scale distributed environment. The main advantages of the systems is its multi-dimensionality that is, they improve scalability by enabling direct and real-time sharing of services and information, enable knowledge sharing by aggregating information and resources from nodes that are located on geographically distributed and potentially heterogeneous platforms and provide high availability by eliminating the need for a single centralized component.

Peer to Peer (P2P) networks are very popular since they offer opportunities for real time communication. They also help to build adhoc network to collaborate and share information in a large scale distributed environment. Apart from this P2P is a multidimensional network where it improves scalability by enabling direct and real time sharing of services and information. It enables knowledge sharing by aggregating information and resources from nodes that are located on geographically

N. Meghanathan et al. (Eds.): CCSIT 2012, Part I, LNICST 84, pp. 394–401, 2012.

distributed and heterogeneous platforms. It also provides high availability by eliminating the need for a single centralized component.

Information retrieval is the process, where a prospective user of information is able to convert his need for information into an actual list of citations to documents in storage containing information useful to him. Information retrieval can be considered as a process where a user is able to convert the information in the form of list of citations to documents in a storage containing information useful to him. It can also be viewed as a software program that stores and manages information on documents. The system assists the users in finding the information that is needed. There are three basic processes an information retrieval system has to support. The following are the processes which an information retrieval system has to support.

- The representation of the content of the documents.
- The representation of the user's information need.
- The comparison of the two representations.
- The representation of the content of the documents and user's information need.
- The comparison of the content and user's information need.

Cross language retrieval backs the user of multilingual document collections by allowing them to submit queries in one language and retrieve documents in any of the language covered by the retrieval system. Considering the example of language1 (L1) queries on a language2 (L2) document collection. Cross-language retrieval can be achieved in three different ways [1]:

- Off-line document translation: translating L2 documents into L1 and then indexing in L1.
- Off-line index translation: indexing L2 documents in L2, then translating index into L1.
- On-line query translation: indexing L2 documents in L2 and translating L1 quires into L2.

Query translation can be applied in environment, where it would be impossible to produce translations for all available documents. In document translation, it is possible to present the user with a high quality preview of all the retrieved documents. Translating documents after they are retrieved, does not suffice because it will not help users to identify material that they wanted to be translated. Since it assumes that the user has already found the relevant document in its original foreign language, it fails to support exactly that part of a search in a multilingual environment which is the most difficult one, to formulate a query which will take the user to the foreign language document of interest.

2 Peer-to-Peer Network for CLIR

Peer-to-Peer network distinguishes itself by its distribution of power and function. They can form *ad hoc* connections between nodes for sharing all kinds of information and files. They can build connections between nodes by building adhoc connections between them. Peer-to-Peer discards hierarchical notions of clients and servers and

replaces it with equal peer nodes that function similarly as clients and servers. Different software modules communicate with each other for processing the information required for the completion of the distributed application. Each computer can access services from the software modules on another computer, as well as providing services to the other computer. The discovery process in the peer-to-peer network is much more complicated than that of the client server. Each computer should know the network addresses of the other computers running the distributed application or at least of subset. And also propagating changes to the different software modules on all the computers would also be harder.

Every computer is capable of accessing services to and fro with other computers. Each computer should know the network addresses of other computers which run the distributed application. It should be able to propagate the changes to different software modules on all the computers. However, the combined processing power of several large computers could easily surpass the processing power available from even the best single computer, and the peer-to-peer network could thus result in much more scalable applications. Peer-to-peer architecture for cross language information system can be divided into two systems

- Cooperative- information is held in the central place (description, collection index, statistics).
- Uncooperative- peer is independent and does not have information about its neighbor, but it answers to the neighbors queries.

According to network structure, peer-to-peer systems can be classified as

- Centralized network
- Decentralized network

2.1 Centralized Network

Centralized network can be treated as a combination of client-server and peer-to-peer. Alike the client-server model, some nodes in the network will act as a server, providing only the directory services. All information resources are distributed among the other peer in the networks. Centralized network also come cross the same problem as the client-server model, single point of failure and scalability. It is observed that one directory server is capable of handling only few requests from the peers and the response time will increases, if more request are placed. In this network, failure of the server will result in the total system crash down; this can be overcome by adding more servers to the networks. Most of the real world peer-to-peer systems are based on this model; Bit Torrent and Edonkey network are the best examples. Some of the benefits can be list as [2]:

- Since it is combination of client-server, transferring to this network from the existing system is very easy.
- High scalability can be achieved by improving the network usage and reducing the broadcasting.
- Finally, it is easy to manage.

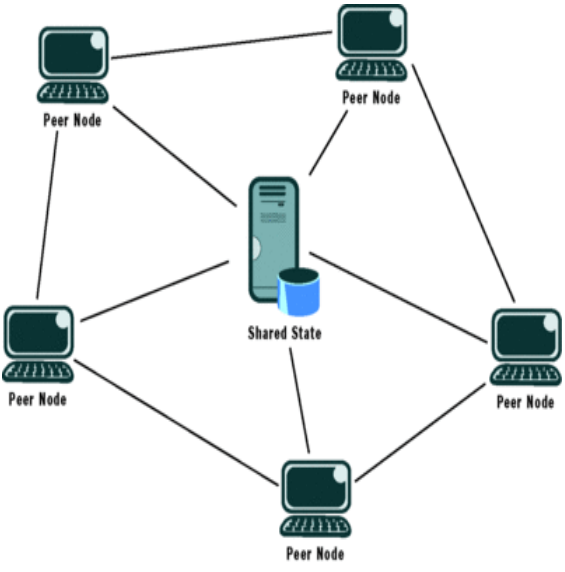

Fig. 1. Centralized peer-to-peer networks

2.2 Decentralized Network

Decentralized network is further sub divided into

1. Structured
2. Unstructured

Structured P2P also known as structured overlay networks or distributed hash tables, in which peers are grouped or clustered. In a common identifier space, each peer is responsible for a subset of identifiers and even the multiple peers are responsible for same space to acquire higher reliability. Overlay routing protocol are used by the peer to forward the messages, to carry out this process efficiently an routing table is maintained. Compared to the unstructured, the structured P2P overlay networks exhibit much lower bandwidth consumption for the search. The concept that can be used for designing the search engine for CLIR is[11]:

-Global index in structured P2P
In an unstructured system all the peers in the network are treated equally, that is they both can issue the request, respond to each other request and also route requests to other nodes. Peers flood search requests in the network, this approach is effective for the search of a popular content and performers poorly for a rare content. P2P search engine for cross language information retrieval can be designed such that:

-Local indexes in unstructured or hierarchical P2P networks
In this method documents are divided over the network and each peer maintain the list of its broadcast. To prevent the numerous documents held in the network, the queries can be answered in two different levels:

1. Peer level – which locate the group of peers with relevant document collections
2. Document level – the query is submitted to the peers and it is answered by querying the local index.

If you have more than one surname, please make sure that the Volume Editor knows how you are to be listed in the author index.

3 Search Mechanism in Peer to Peer Network

The main purpose of search mechanism in P2P network is to guide the query to the sources, so that the appropriate document can be retrieved and then translated back to the query language. The objective of this mechanism is to decrease the number of unrelated document retrieval per query and at the same time maintain a high recall rate.

In the centralized network, an index is maintained of all the documents by the participating peer. Some of the commercial information retrieval systems are web search engines and centralized P2P indexing systems. Usually in this method, all the peers in the network give the index of its entire shared document to the centralized repository, from which the required document can be retrieved. It appears as though two peers in the network are communicating directly. Any of the available translating method can be used if required.

For the decentralized network, some of the search mechanisms are [12]:

1. Breath first search

It is a widely used method in the networks. When a node has to search for a information, it generates the query message and broadcast it to the other peers in the network. If some node has a match, it responds by generating the query hit message. When the sender receives the query hit from more the one peer, it tries to download the documents from the peer with the best connectivity. One of the major sang with this method is that, each query consumes excessive network because it is propagated along all links. Therefore a node with a lower bandwidth can cause bottleneck.

2. Random Breadth-first-search (RBFS)

The drawback of the BFS is overcome in this method, here a peer which request for the information sends its search message to only a few of the peers in the network that are selected in random. In this method, a peer which request information sends messages to only a few other peers in the network which are selected randomly. It also eliminates all the disadvantages of the BFS method. Since this method randomly choices the peer, there are possibility that some of the peers in the network that contain the related information may be left unnoticed.

3. Random walkers

This method is similar to the RBFS. Here the node that needs the information forwards a query message called walker to randomly selected peer. To reduce the

time taken in getting the result, one walker is extended to n-walker, where n independent walkers are consecutively sent from the searcher. It is assumed that n-walker after T steps will reach the same number of peer as a one walker after nk steps.

4 Translation Technique

Any of the available translation methods can be used if required [9]

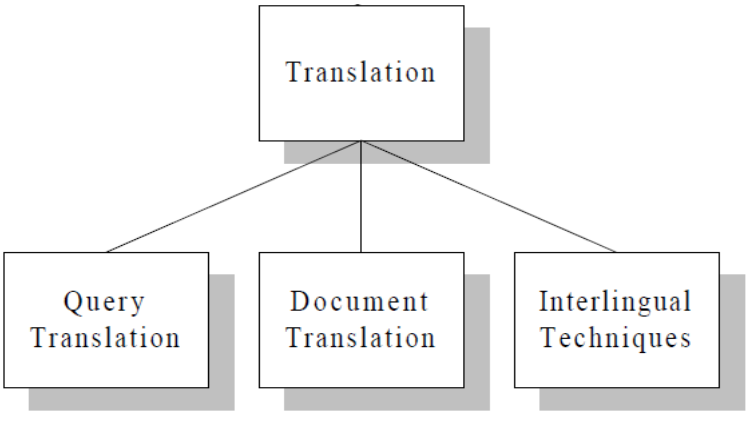

Fig. 2. Translation techniques [9]

1. Query translation

Query translation is a most general strategy in which the query is automatically converted into every supported language and is relatively efficient. The principal limitation of query translation is that queries are often short and provide little context for disambiguation. Homonymous words produce undesirable matches. Translation ambiguity causes this problem, by potentially introducing additional terms that are themselves homonymous. For this reason, controlling translation ambiguity is a central issue in the design of effective query translation techniques.

2. Document translation

Document translation is just the opposite of query translation, automatically converting all of the documents into each supported query language. It typically provides more context than queries, more effective strategies to limit the effect of translation ambiguity may be possible. Another potential advantage is that selected documents can be preseted to the user for examination without on-demand translation.

Document translation can be considered as a methodology of converting all the documents into each supported query language. The advantages of this method includes that the selected documents can be presented to the user for examination without on-demand translation. It also provides more contexts than queries, more effective strategies to limit the effect of translation etc.

3. Interlingual techniques

Interlingual techniques convert both the query and the documents into a unified language-independent representation. Controlled vocabulary techniques based on multilingual thesauri are the best examples of this approach. Controlled vocabulary techniques based on Multi lingual thesauri can be considered as one of the best example for this method. Because each controlled vocabulary term typically corresponds to exactly one concept, terms from any language may be used to index documents or to form queries. Latent semantic indexing and the generalized vector space model both use a document aligned training corpus to learn a mapping from one or more languages into a language-neutral representation. Document and query representations from either language can be mapped into this space, allowing similarity measures to be computed both within and across languages.

5 Conclusion

Peer-to-Peer is a most important part of the computer networks. Some of the methods for the centralized and decentralized network are proposed that enables the retrieval of information in the Peer-to-Peer networks with some of the translation techniques that helps the user to retrieve the required document, in the understandable languages. There are several useful CLIR techniques are known for the information retrieval. Monolingual retrieval is still more effective for free text than CLIR. Query translation, document translation, and interlingual techniques provide a range of alternatives that can be tailored to specific applications. The proposed translation techniques for CLIR do have drawback, which can be overcome by selecting the appropriate algorithms, which can be carried ad the future work of the same paper.

References

1. Hiemstra, D.: Using Language Models for Information Retrieval (2003)
2. Sankar, K.: What is peer to peer (2003), http://p2p.inetrnet2.edu/documents/what%20is%20to%20peer-5.pdf
3. Callan, J., Powell, A.L., French, J.C., Connell, M.: The effects of query based sampling on automatic database selection algorithms. Technical report IR-181, center for intelligent information retrieval. Dept. of Computer Science, University of Massachusetts
4. Crespo, A., Garcia-Molina, H.: Routing Indices for Peer-to-Peer Systems. In: Proc. of ICDCS 2002, Vienna, Austria (2002)
5. Tang, C., Xu, Z., Dwarkadas, S.: Peer-to-Peer information retrieval using self-organizing semantic overlay networks. In: Proc. of ACM SIGCOMM 2003, Karlsruhe, Germany (2003)
6. Zeinalipour-Yazti, D.: Information retrieval in Peer-Peer Systems. M.Sc Thesis. Dept of Computer Science, University of California-Riverside (June 2003)
7. Ata, B.M.A., Mohd, T., Sembok, T., Yusoff, M.: SISDOM: acmultilingual document retrieval system. Asian Libraries 4(3), 37–46 (1995)

8. Fluhr, C.: Multilingual Information Retrieval. In: Cole, R.A., Mariani, J., Uszkoreit, H., Zaenen, A., Zue, V. (eds.) Survey of the State of the Art in Human Language Technology: Center for Spoken Language Understanding, Oregon Graduate Institute, pp. 391–305 (1995)
9. Oard, D.W.: Alternative Approaches for Cross-Language Text Retrieval. In: Cross-Language Text and Speech Retrieval, AAAI Technical Report SS-97-05
10. Dorrigiv, R., Lopez-Ortiz, A., Pralat, P.: Search Algorithms for Unstructured Peer-to-Peer networks
11. Chen, H., Gong, Z., Huang, Z.: Self-Learning Routing in Unstructured P2P network. International Journal of Information Technology 11(12)
12. Zeinalipour-Yazti, D., Kalogeraki, V., Gunopulos, D.: Information retrival in Peer-to-Peer networks
13. Bawa, M., Bayardo Jr., R.J., Rajagopalan, S., Shekita, E.: Make it Fresh, Make it Quick Searching a Network of Personal Webservers. In: Proc. of WWW 2003, Budapest, Hungary (2003)
14. Archana, M., et al.: Mining The Web Information For Cross Language Information Retrieval. In: Proceeding of ICMET 2011. ASME Press (2011)
15. Archana, M., et al.: Extracting web information for CLIR. International Journal of Modeling and Optimization (IJMO) (yet to be published)