Organizational Approaches for Peer-to-Peer based Information Retrieval System

Haizheng Zhang
Computer Science Department
University of Massachusetts, Amherst

In recent years, Peer-to-Peer based information retrieval systems have received significant interest among computer scientists. A peer-to-peer based information retrieval system consists of a set of nodes connected in a peer-to-peer fashion. Each node hosts a document collection to share with other nodes and these nodes work collectively to provide information retrieval service to users. In such a system, an information retrieval task can be considered as a search session during which nodes forward the query to their neighbors, perform local searches and/or return search results. While the promise of this type of applications is attractive, the underlying technology is challenging. The lack of complete, up-to-date information of states of other nodes in the network requires sophisticated strategies for effective distributed search. In addition, the presence of concurrent search sessions adds an another level of complication: nodes may not be able to complete forwarding and perform local searches for all queries they have received in a timely fashion due to bandwidth and processing capacity limitations.

This thesis frames a P2P IR problem into a multi-agent framework and attacks it from an organizational perspective by exploring various adaptive, self-organizing topological organizations and designing appropriate coordination strategies for large-scale agent organizations. Specifically, two protocols have been designed to create semantic based implicitly clustered agent organizations and explicit multi-level hierarchical agent organizations respectively in the context of single-query peer-to-peer based information retrieval systems, i.e, each external query is processed until completion before another external query is allowed to enter the system.

In forming implicitly semantically-close clusters, agents exchange their resource descriptions to expand their local information about the content distribution over the network. Agents then prune the topology based on predefined rules. Two search strategies were evaluated on the reorganized topology. The experimental results demonstrated that the topology reorganization process combined with a context-aware search algorithm can improve considerably the information retrieval performance.

In forming an explicit, multi-level topical hierarchical structure to facilitate locating relevant documents, agents join different groups in the hierarchy based largely on their content similarity. The group formation is achieved by organizing the agent-view structures properly so as to place semantically similar agents together to form explicit groups in an incremental and distributed manner. A context-aware search algorithm is also designed to take advantage of the hierarchical organization. During the search process, agents in the network follow various cooperation strategies to forward queries and return results in the network. This approach further improved performance over the implicit approach..

These studies on single-query systems shed light on how to conduct a distributed search on content sharing systems. However, they did not take into account important issues in real content sharing systems where there may be significant loading and scheduling issues in the networks when there are multiple external queries being processed in the network. In order to handle multiple, concurrent search sessions in the system, an agent control mechanism is proposed to engineer the query flow in the entire network based only on agents' local observations of network traffic and agent loading so as to improve the mean effective propagation speed of search queries. The elements of such a control mechanism include resource selection, local search scheduling and feedback-based load control. In particular, with the feedback based load control unit, an agent not only considers the capacity of its own communication channels, but also takes into account its neighboring agents' service rate, which is acquired dynamically from its neighboring agents.

In addition to the local agent control mechanisms, the algorithm proposed in[2] does not perform well in concurrent distributed information retrieval systems since it tends to increase dramatically the burden of top-level mediators, thereby creating hot spots in the network. In order to reduce the potential hot spots in the network, an improved two-phase search algorithm[3] has been designed to address this problem based on this novel agent control mechanism. The first phase of this query routing algorithm occurs when the agent who initiates the query, forwards the query along the lateral links to the agents at the same level to locate relevant clusters. After comparing the similarity measures returned from these agents, the initiator selects the most similar agents to proceed to the second phase of the search. The second

search phase is primarily conducted inside each group. In particular, agents only forward the query along upward links or downward links during the second phase search. This way, the traffic load is more evenly distributed among the various levels of the hierarchy given that the entry points of the queries are randomly distributed in the network. Hence, the new search strategy mitigates the hot spots problem introduced in the previous algorithm.

Anticipated Intellectual Contributions

This work is one of the first attempts to address this issue from an organizational perspective. The agent organization of a peer-to-peer information retrieval system includes both the underlying topological network and the coordination mechanism among agents in the process of query routing. Efforts from both aspects have been made to achieve better information retrieval performance and improve system throughput. The organization formation and reorganization algorithms are conducted in a bottom-up fashion. The advantage of these organizational approaches partly lies in the fact that only partial information is needed and no global information is required. While this protocol is designed for the peer-to-peer based information retrieval domain, it has broad implications on distributed multi-agent community formation in general and sheds light on forming complex agent organizations in a parallel, distributed and incremental fashion.

In addition to the contributions to topological agent organization formation, I have also explored the coordination mechanisms to the distributed search process. Based on the implicit clustered agent organizations and hierarchical agent organizations, search algorithms have been designed accordingly. Most previous studies oversimplify agent network environments by implicitly assuming that network bandwidth and the agent computational power is unlimited[7]. They usually overlook the profound impact of concurrent interactions among agents in the network. Hence, although the algorithms designed under such assumptions may achieve good performance in some cases, they do not adapt well to the changes in the network environment because the performance of resource selection and query routing algorithms depends on factors including the network traffic, agent computational capacity, and underlying topological organizations. Additionally, even if the initial network situation is given as expected, the deployment of these algorithms can bring changes to the initial environment and result in unpredictable network situations. These changes may invalidate the assumptions of these algorithms and lead to chaos situations. In contrast, I study distributed content sharing systems by taking an adaptive, self-organizing organizational approach and exploiting feedback information to dynamically adjust agents' query routing and resource selection behaviors. Therefore, agents take different strategies under varying network situations. Instead of forming agent organizations based on the static properties including content distribution, I plan to explore learning techniques to form and reorganize agent organizations based on run-time characteristics. The anticipated contribution of such learning-based organizational approaches is to provide a mechanism to adapt agent organizations to the ever-changing network situations. Overtime, agent organizations can monitor their structural properties and the run-time environmental factors and perform self-diagnosis. Based on learning algorithms and their local observations about the organization, agents can then reorganize the organization in a collective fashion in order to achieve a better performance.

In conclusion, distributed information retrieval systems can be naturally modeled as multi-agent systems. Studying distributed information retrieval applications from a multi-agent perspective leads to the design of many novel coordinated search strategies for various underlying topological organizations. On the other hand, the distributed information retrieval domain provides an arena for various multi-agent techniques. The large-scale nature provides multi-agent research with a concrete testbed to design, implement and test many technologies in multi-agent research. I expect my results can contribute to both areas and provide insights for other large-scale applications

.

References

[1] Haizheng Zhang, W. Bruce Croft, Brian Levine, and Victor Lesser A Multi-Agent Approach for Peer-to-Peer Information Retrieval; The Third International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2004) .

[2] Haizheng Zhang, Victor Lesser; A Dynamically Formed Hierarchical Agent Organization for a Distributed Content Sharing System; IEEE/WIC/ACM International Conference on Intelligent Agent Technology, 2004.

[3] Haizheng Zhang, Victor Lesser; A Queueing Theory Based Analysis on Agent Control Mechanisms in Peer-to-Peer based Information Retrieval Systems; In CIKM 2005 Workshop on Peer to Peer based Information Retrieval (P2PIR 2005)

[4] S. Zilberstein and A. I. Mouaddib. Reactive Control of Dyanmic Progressive Processing. Proceedings of the 16th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 1999

[5] A. Arnt, S. Zilberstein, J. Allan and A. I. Mouaddib. Dynamic Composition of Information Retrieval Techniques. Journal of Intelligent Information Systems. 23(1): 67-97, 2004

[7] Jie Lu and Jamie Callan; Content Based retrieval in hybrid peer-to-peer networks; Proceedings of the twelfth international conference on Information and Knowledge management, 2003