An Active Query Routing Methodology for P2P Search Networks

Srikanth Kallurkar and R. Scott Cost

University of Maryland Baltimore County, USA
skallu1@umbc.edu

Abstract. Peer-to-Peer computing paradigm may provide a solution to the retrieval problem in an ever burgeoning volume of online and digital information. While research has focused on the means of collaboration as a tool for query routing, we feel that there is a disconnect in the way P2P networks are handled and the expectations of performance in the real world. In this work, we discuss the motivations for functional peer connectivity and suggest Metadata placement as a methodology for an effective and efficient dual for query routing in Peer-to-Peer networks. We present an abstract architecture for a semi-structured overlay P2P search network..

1 Introduction

Peer-to-Peer (P2P) computing has emerged as a possible solution to almost every problem that is, or can be composed of several subproblems. Inherently, the idea of solving a problem by decomposing it into smaller subproblems to produce a final solution lends itself to efficiency. While this is true, retrieval applications in a P2P environments require not only be the search process efficient but also be effective. Such consideration, in our view, promote a need for controlable operations in the P2P networks. Retrieval in P2P network is much like a Distributed Information Retrieval (DIR) process. Traditional research in DIR focuses on retrieval problems such as reducing false positives and false negatives among potential sources of information for a query spread across a network. A P2P search (or retrieval) network model is a built atop a DIR model [4] but also has the following issues that affect retrieval effectiveness and efficiency.

1. Dynamism of peers: Peers can join and leave a network at will. Such dynamics of the network mean that query results will depend upon content availability.
2. Restrictions on search: Peers may not cooperate for certain types of queries or may restrict access to certain types of users.
3. Heterogeneous Information sources: Peers may host different types of content thus use different types of indexing and retrieval models.

Daswani and Garica-Molina [7] suggested the following aspects that define data sharing in P2P networks

1. Topology: The dynamics of peer interconnections.
2. Routing: The protocols for query message routing.

3. Placement: The positioning of data and associated metadata in relation to the topology and the routing protocols.

Query routing deals with querying distributed information sources. The discussion of query routing strategies provides for the general impression that query routing and metadata are inseparable entities, The query routing process assumes enormous significance for effective query results. Without loss of generality, we can say that the best query routing strategy is one where a query is routed to every source that has relevant information and to none of those that have no relevant information (given that we can distinguish informational relevancy from non-relevancy). For such a strategy, a query should be evaluated against metadata about all the sources in a network, which leads to the question of the how is metadata propagated in a P2P network. For example, in a broadcast network, metadata may be broadcast to all peers in the network. While this is suitable for relatively small, and static networks, such a deployment may not be feasible where in larger and more dynamic environments. The research so far has focused on optimal routing strategies under the common assumption that the metadata will be in place for the routing process, i.e. metadata about a peer distributed to all the members of the network. The effectiveness of a query routing strategy then, is only as good as consistency provided by the metadata a query can be evaluated against and, its efficiency is bounded by the distance a query travels to reach the peers hosting relevant metadata. Thus, there is a need for an optimum metadata placement strategy that can provide for optimum locations of metadata in a P2P network such that updates to metadata is quick and the query travel distance is minimized.

Furthermore, the actual routing process is dependent upon the ability of the router to evaluate a query against the metadata about various sources. We suggest the use of a layer of abstraction to ease the implementation of a router. Abstraction, as described in the Object Oriented Paradigm, provides data encapsulation by defining the modes of data access, a layer of transparency to the data operations and the freedom to change the design and implementation of such operations behind a standard API. We observe that the P2P query routing problem contains a similar need for data encapsulation. A layer of abstraction applied over metadata passes the responsibility of providing the “means of evaluation” from the router to the info-peer. However, we note that abstraction does not absolve the router of performing the actual routing decisions. The intuitively observable advantage of abstraction is the enforced conformance requirement on the info-peers to provide a standard interface. This interface provides a grounding for Metadata usage for routing purposes while allowing freedom of metadata representation.

In this work, we present a high level architecture of functionally structured overlay P2P network (Section 3). We describe a Metadata Placement methodology that builds on top of the overlay network and provides for effective query routing strategies (Section 4). Section 2.1 provides a background for query routing for content search in P2P networks. Finally, we summarize our contributions and comment on the future work in P2P retrieval (Section 5).

2 Background

In this section, we provide a discussion of query routing strategies and related metadata technologies.

2.1 Query Routing

In Gnutella [12], queries are routed by a peer broadcasting the query to all its neighbors, which recursively do the same, thus effecting a network-wide distribution of the query. The process is usually performed until a certain hops (edge between a pair of peers) has be reached. In super-peers networks like Kazaa [16], super-peers represent a set of clients and process queries on behalf of the clients by indexing the clients data, i.e routed among super-peers. Structured networks on the other hand provide point-routing by determinstic data locations by using DHTs [11, 13]. Crespo and Garica-Molina [5] describes the use of routing indices as hints about the best paths a query should take to reach documents of interest. Three types of routing indicesmethods were described. The first one, Compound Routing Indices, annotated paths between peers with expected results based on known information about immediate neighbors such as number of documents on a topic. The second method, called Hop-count, guessed the expected results up to a certain number of peers beyond the neighbors. A cost model provided the number of messages needed to find all the results upto the given number of hops. The third method approximated the cost model to include information about peers beyond the hop-count with some loss in accuracy. Cuenca-Acuna et al [6] described a randomized mechanism, called a Gossiping layer, for propagating content updates in a P2P search network. Peers were randomly chosen to inform as well inquire about changes. The peers knowledge base consists of a local searchable index and bloom filters about all peers in the network. They used Bloom filters [2] for extremely succinct document collection descriptions. The assumption is that the gossiping layer covers the entire network. Subsequently however, the authors cite a provision for subsections of peers covered by the gossiping layer for faster updates among the members of each subsection. We can thus observe that the problem of metadata propagation relates simply to the scale of propagation which in turn is dependent on the degree and scope of metadata consistency needed. Bawa et al [1] used term vectors as metadata about peers, peers were clustered by topic and queries are routed to clusters and within clusters. Nejdl et al [10] described super-peers for routing queries in the P2P network. Super-peers were repositories for RDF descriptions of peer content as well as those of neighboring super-peers. Peers registered with a chosen super-peer and did not take part in query routing. The super-peer could route the query to its registered peer andor forward it to the neighboring super-peers. Clusters of peers are formed as similar content peers join the same super-peer. Tzitkzikas et al [14] used Taxonomies as content descriptors for peer content in a P2P network. Peers could maintain their own taxonomies and also mappings to taxonomies of other peers. Queries were evaluated over the taxonomies and mappings.

2.2 Metadata

Metadata can be defined as content description used as representation of a resource. A resource can be a single document, a collection of documents, an online service etc. Metadata is then broadly classified as structured and unstructured. Structured metadata is usually associated with describing single resource, such as a single document and is usually of a prescribed format, for example, RDF [15], Dublin Core [8]. On the other hand, unstructured metadata is usually associated with routing queries to collections of documents, say in a DIR system and are normally content derived such as term histograms [17, 9, 3]. In the following sections, we describe an architecture for functional overlay network that allows for a need-based metadata placement strategy.

3 A High Level P2P Architecture

The commonly observed architecture of P2P network has peers that publish their content on the P2P network. Such peers have obligation to forward queries that they receive according to the routing protocol of that network. We present a slightly varied version of such a network where a peer can choose the function it will serve in the network. The functions are those of routing queries and serving information. Peers pertaining the routing function are named routers while those serving information are named info-peers. In this architecture, a peer serves at least one of the functions. A router offers the facility of routing to info-peers by accepting metadata about them. The motivation is that a peer can be just an info-peer and not entail the costs associated with P2P overheads such as query traffic. The introduction of such relaxation however raises the issue of motivations for peers to be just routers. For a peer to assume a routing function, it should be able to allocate enough computing and storage power through which it can make routing decisions.

In this work, we refer to the info-peers are "peers" as a flexibility criterion. An information source may chose to serve the P2P community only by sharing its content through a peer interface while not sharing the burden of query distribution, however, retaining the choice of becoming a router. We note that such an architecture has some similarities to the super-peer architecture [16], in terms of represent ion of clients by the super-peers. Our suggested architecture divides the membership by the routing or information serving functions, thus the P2P paradigm is still maintained in that data is still distributed and query can be processed at the info-peers. Also, the nature of the alliances is determined by the pair of interested peers and is negotiable.

The architecture consists of a logical P2P overlay network of routers The routers form routing alliances with the info-peers such that routers are representatives of the info-peers in the router network. The routers form ad-hoc connections among themselves to form the P2P overlay network. In addition, they also form inter-router alliances. Thus, the architecture is a semi-structured network, in that the structure is not rigid and is congigurable by the participants. The router overlay provides for super

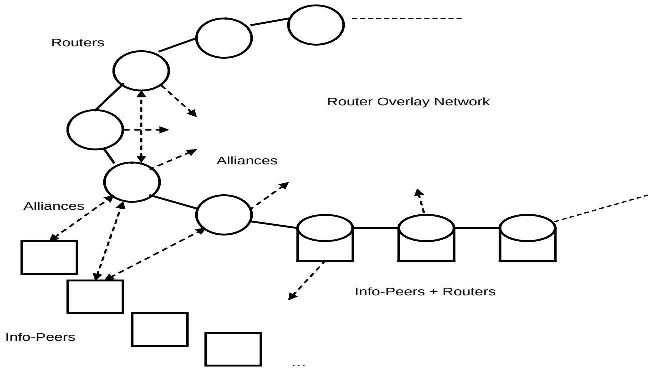

Fig. 1. P2P Architecture - Router Overlay Network

alliances between routers beyond the simple neighborhood information carried by super-peers. Figure 1 shows a schematic of the router overlay, which consists of routers as well as info-peers with a routing function. The alliances (Section 4.1) are shown as dotted line to indicate that the alliances are temporary.

4 Metadata Placement

A routing process for a query can be described as a 3-tuple representation $< Router >< Metadata >< Info - Peer >$.
Such a representation indicates that a query at a router is evaluated against metadata stationed at that router. The Metadata Placement (MP) methodology involves all the three query routing entities. There are two components to Metadata Placement methodology, routing alliances and active metadata, described below.

4.1 Routing Alliances

Routing alliances are a bipartite agreement for exchange of metadata that can exist between a router and an info-peer as well as between a pair of routers. An alliance between any pair of peers carries the explicit agreement by the router to serve the routing function on behalf the other member of the alliance. An alliance can be initiated by either of the members.

1. Router's Perspective: When a router joins the network it broadcasts an advertisement message that describes its capabilities. A router advertisement consists of
   - Processing Capability: CPU power, storage, memory, networking capability
   - Distance from the Information Peer: Number of hops
   - Monetary Cost of services: Dollar value for subscription

A peer that receives such an advertisement can then solicit the router for alliance formation. Router alliances that exist between a pair of routers allow for inter-router cooperation. An inter-router alliance can be observed as Business-to-Business (B2B) model of subcontracting jobs and assembling the final product. For example, an inter-router alliance can be of the nature that a query about a particular topic can be subcontracted to a member of the alliance which specialized in the topic.

2. Info-Peer's Perspective: When an info-peer joins the network, it broadcasts bids for routers. Routers can respond to the bids with their capabilities which are then used by the info-peer to establish alliances with the peers it selects. It also receives unsolicited advertisements by routers which are also evaluated by their capabilities. After the initial alliances are formed with the routers, the info-peers alliances are managed by metadata objects, described in the next section.

4.2 Active Metadata

We use the term Active Metadata to describe an encapsulation of metadata that can perform various actions on behalf of the information source to effect maximal exposure to the source. In the P2P context, an Info-Peer creates a metadata description most appropriate for its content and encapsulates it within an active metadata object. The mode of data access is set by the Info-Peer. The various actions include, but are not restricted to,

1. ability to return a similarity value of the query to the metadata
2. ability to interpret the results of the previous queries to either relocate or replicate to other peers and,
3. ability to interact with the host router to gather network related information such as peer connectivity
4. ability to migrate
5. ability to replicate

The active metadata objects can be implemented as remote objects to restore privacy and security of the metadata.

4.3 Routing Methodology

The routing methodology allows for a query to spread in the router overlay network along the edges of the network and along the alliances. The decision as to which edges should the query be routed to is decided by the router by interacting with its alliances, those with info-peers through the active metadata and with the allied routers. A router queries the active metadata objects through the standard interface, which return similarity value and checks whether query falls under an alliance category with any of the routers. If the selection process is below a satisfactory threshold, i.e. none of the active metadata return satisfactory similarity values and none of the router alliances are suitable for the query, the query is routed along the normal P2P edges in the router overlay network.

4.4 Placement Methodology

The routers evaluate the alliances with current Info-Peers they host metadata for, while the active metadata acts on behalf of the info-peers to evaluate the routers it is placed with. The router has a utility function with which it can evaluate the alliances, based on

1. The amount of resources used by the router to route the queries against the successful results from the info-peers,
2. The effectiveness of query results from the selected peers. Such an evaluation may involve an user feedback loop.

The utility function determines the value of each alliance over time period to allow for evaluating the need for either continuing the same alliances or looking for more. Such a function can be viewed as an economic barometer for the router to judge the value of alliances. For example, if only 10% of metadata stationed at the router provided usefully results for the queries, then the router can reform alliances with existing info-peers or form new alliances altogether. On the other hand, active metadata uses the collected statistics from the router and the queries processed over a period of time to evaluate the value of current alliance. For example, if a high percentage of queries were successfully evaluated at a router or the router is about to go offline, then the metadata object solicit network information from the router and solicit replication permissions from the router's neighbors. The object may also provide the info-peer with its evaluations of the router to determine the priority of updates, highly successful routers may be updated ahead of less successful ones.

5 Conclusion and Future Work

We have described a novel placement methodology that can be implemented to suit the dynamism of the peers and the changes in the network. We also suggested a functionally structured P2P network where such a routing methodology can be implemented. We suggest the study of information needs that can provide for a generalized model for metadata placement in P2P networks.

References

1. Mayank Bawa, Gurmeet Manku, and Prabhakar Raghavan. Sets: Search enhanced by topic segmentation. In *Proceedings of the Twenty Sixth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003.
2. B. H. Bloom. Spacetime tradeoffs in hash coding with allowable error. *Communications of the ACM*, 13(7):422–426, 1970.
3. J. P. Callan, Z. Lu, and W. Bruce Croft. Searching distributed collections with inference networks. In E. A. Fox, P. Ingwersen, and R. Fidel, editors, *Proceedings of the Eightteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 21–28, Seattle, Washington, 1995. ACM Press.

4. Jamie Callan. *Advances in Information Retrieval*, chapter Distributed Information Retrieval, pages 127–150. Kluwer Academic Publishers, 2000.
5. Arturo Crespo and Hector Garica-Molina. Routing indices for peer-to-peer systems. In *Proceedings of the Twenty Second International Conference on Distributed Computing Systems*, Vienna, Austria, 2002.
6. Francisco Matias Cuenca-Acuna, Christopher Peery, Richard P. Martin, and Thu D. Nguyen. Planetp: Using gossiping to build content addressable peer-to-peer information sharing communities. In *Proceedings of the Twelfth IEEE International Symposium on High Performance Distributed Computing*, 2003.
7. Neil Daswani, Hector Garcia-Molina, and Beverly Yang. Open problems in data-sharing peer-to-peer systems. In *Proceedings of Ninth International Conference on Database Theory*, 2003.
8. http://dublincore.org.
9. L. Gravano H. Garcia-Molina and A. Tomasic. Gloss: Text-source discovery over the internet. *ACM Transactions on Database Systems*, 24(2):229–264, 1999.
10. Wolfgang Nejdl, Martin Wolpers, Wolf Siberski, Christoph Schmitz, Mario Schlosser, Ingo Brunkhorst, and Alexander Lser. Super-peer-based routing and clustering strategies for rdf-based peer-to-peer networks. In *Proceedings of the Twelfth International ACM World Wide Web Conference*, 2003.
11. Sylvia Ratnasamy, Paul Francis, Mark Handley, Richard Karp, and Scott Shenker. A scalable content-addressable network. In *Proceedings of ACM SIGCOMM*, 2001.
12. Gnutella Specification. www9.limewire.comdevelopergnutella_protocol_40̇ṗdf.
13. Ion Stoica, Robert Morris, David Karger, M. Frans Kaashoek, and Hari Balakrishnan. Chord: A scalable peer-to-peer lookup service for internet applications. In *Proceedings of ACM SIGCOMM*, 2001.
14. Yannis Tzitkzikas, Carlo Meghini, and Nicolas Spyratos. Taxonomy-based conceptual modeling for peer-to-peer networks. In *Proceedings of Twenty Second International Conference on Conceptual Modelling*, pages 446–460, 2003.
15. http://www.w3c.org/RDF/.
16. Kazaa Website. http://www.kazaa.com.
17. Jinxi Xu and W. Bruce Croft. Cluster-based language models for distributed retrieval. In *Proceedings of the Twenty Second Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 254–261, 1999.