Towards the Use of Mediated Knowledge-based and User-defined Views in Super-peer P2P Systems

R. Mohamed & C.D.Buckingham
School of Engineering & Applied Science
Aston University
B4 7ET, UK

M.B.Al-Mourad
School of Computing & IT
University of Wolverhampton
WV1 1SB, UK

Yaser M. A. Khalifa
Electrical & Computer Engineering,
State University of New York

Abstract - **In recent years, peer-to-peer data integration systems have attracted significant attention for their ability to communicate, collaborate and share information in a networked environment. One of the main problems that arises in such systems is how to exploit their mappings in order to answer queries posed to one peer. Our proposed framework can be used to exploit the existing mapped data together with its data location information for defining a peer's data view. This data view is expected to produce query results based on peer preferences rather than using standard query processing at the super-peer level, as practiced in current super-peer P2P systems. Our framework consists of two major components: a mediated knowledge-base at the super-peer and user-defined data views at the peer.**

I. INTRODUCTION

Recently, Peer-to-Peer (P2P) systems have become an active research area because of the opportunities for real-time communication, ad-hoc collaboration and information sharing in a large-scale environment. P2P refers to a class of systems and applications that employ distributed and autonomous nodes (peers) that co-operate in the community to share resources and services [1,2]. As the actual data is stored in various autonomous peers' data source locations, peers are 'linked' to other peers by mappings. Two basic problems arising in this architecture are: how to discover, express and compose the mappings between peers and how to exploit the mappings in order to answer queries posed to one peer [3]. The second problem is studied in this paper.

In P2P systems, peers are connected on an ad-hoc basis and the location of information is not controlled by the system. There is no guarantee that a query will be successful, even for the best query language. This is because peers only have local knowledge of the network, within which peer nodes may enter and leave frequently. The most widely known P2P network architectures are pure and super-peer networks. Our work will investigate searching issues in super-peer networks.

A search process includes aspects such as the query forwarding method, the set of nodes that receive a query-related message, the form of these messages, local processing, the stored index, and information maintenance [4]. The process depends on the system architecture used. Basically, searching can be classified as blind search, and informed search. Blind search is formally used in pure P2P systems. The original Gnutella [5] algorithm used the blind search method (also known as a flooding scheme) where each posted query is forwarded to all accessible peers within TTL ('Time-To-Leave') hops [4].

Super-peer networks use the informed search method because of the function of the super-peer node [6, 7, 8, 9], which can be seen as a hub that receives queries from connected peers (leaves). This query is forwarded to any relevant leaves and also to neighbouring hubs based on a query routing table (also known as a super-peer index). The super-peer index retains information about data stored at neighbouring peers so that the posted query will be forwarded to relevant peers only, in order to reduce the network traffic. This index would maintain either the actual location of required data as in Gnutella2 [10] or give 'directions' towards required data as in 'Routing Indices' (RIs) [11]. Hence, the super-peer node becomes the most vital component for query processing in super-peer networks. However, super-peers that provide the index for query routing are burdened by other peer nodes and by having to transform posted queries into sub-queries that become local queries for peers [23]. This project aims to reduce this burden by providing individual peers with the capability to define their own data views so that they can process posted queries locally, based on peer preferences (peer preference refers to the particular data required by a peer, where queries can be filtered and constraints imposed to return data suited to the requesting peer).

In this paper, we intend to address the aforementioned problems by proposing mediated knowledge-based and user-defined views in super-peer networks. The mediated knowledge-base aims at linking the routing index table to information about data residing at the peer locations. This knowledge base will then be used by peers to save queries made to the super-peer so that when a peer is given the same query again, it can execute the query directly without having to go through the super-peer. These saved queries are known as user-defined views and will function as a local search mechanism in order to produce query results with respect to peer preference.

The paper is organised as follows: Section 2 introduces a scenario in tourism which is used as a practical paradigm for further study. Section 3 reviews P2P data integration systems based on super-peer networks, with some motivating issues for our project. Section 4 and the

T. Sobh (ed.), Innovations and Advanced Techniques in Computer and Information Sciences and Engineering, 433–438.

subsequent sub-sections present our framework. Conclusions and a comparison of our proposed framework with centralised web-server and existing super-peer P2P systems are in Section 5.

II. SCENARIO

Consider a tourism scenario that consists of a Customer, a Basic Service (BS), and an Additional Service (AS) as shown in Figure 1. The BS and AS are information providers to the web portal. Companies who offer simple transportation services or accommodation, such as airlines and hotels, are consider as BS, while tourism and tour operator services are considered as AS. Customers require travel information from both BS and AS, while AS needs information from BS, for example, to generate travel packages. From the information provider's perspective, each of BS and AS should publicise their services on the World Wide Web (WWW). For example, hotels need to publish seasonal price information to attract more customers. This information is required by customers and also AS. At the same time, a travel agency (part of AS) may include this hotel promotion as part of its travel promotion packages that are also needed by the customer. However, this kind of information exchange that is currently available in centralized web servers among different BS and AS services is time consuming to obtain, out-of-date, and error prone, even though it is often available electronically at every level. Furthermore, different organisations may have different database schema to capture their data.

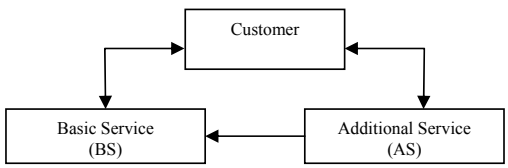

Fig.1. Relationships between three roles in a tourism scenario.

With existing WWW search engines, if customers want travel information about 'Kuala Lumpur' in Malaysia, they may type 'Kuala Lumpur travel agency' into the WWW search engine Some results containing the keywords return but may not contain any web pages for travel agency services at other areas in Malaysia because the exact keyword of other locations has not been specified.

Compared to centralized web servers, P2P systems have a distinct advantage as information providers because peers have more autonomy, both in providing their data to be shared with the community in the web portal and by maintaining their personal data. Additionally, end users can directly establish connections with other users (peers) without involving the centralized web server. Hence, we propose a framework for producing query results that considers peer preferences through user-defined data views. In this framework, the super-peer node is assisted with a knowledge-base that contains the actual location for particular information. Any peer could request these information locations to define their own data views (i.e. generate queries to other peers directly, without having to engage the super-peer). As illustrated in Figure 2, one of the peers is also a web portal and the AS information provider should be able to capture data required by the web user. An added benefit of using a knowledge-base for maintaining locations of information is that more intelligent partial keyword matching can be accomplished. The idea is that the knowledge base will generate more flexible queries that, along with the location information, can be stored as local data views within peers and executed in future without depending on super-peer query processing.

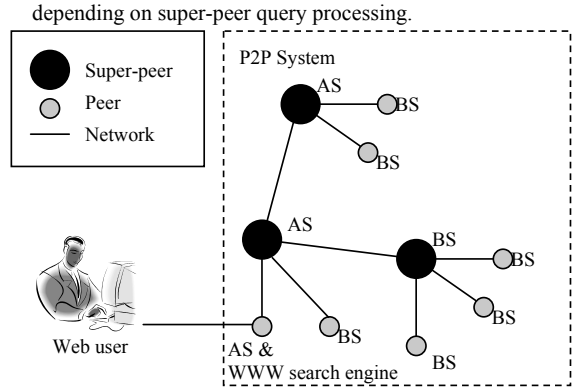

Fig.2. Illustration of a scenario in a super-peer P2P system network.

III. REVIEWS OF P2P DATA INTEGRATION SYSTEMS

P2P data integration systems are networks of autonomous peers that have recently attracted significant attention as an effective architecture for decentralized data sharing, integration and querying. Each peer shares a part or all of their resources with the community. In general, the success of such systems is achieved by increasing numbers of participants, thereby incrementing data storage and computational power of the whole system. However, as pointed out by Gribble et. al. [12], often generic P2P systems do not properly manage the semantics of data exchange. This situation leads to some drawbacks about availability and consistency of the service provided by a P2P system. It happens because of the absence of a global information schema or global knowledge for the whole peers' community. We will review some P2P data integration systems that try to avoid this by being based on a super-peer network. We are interested in how information can be shared among peers to help produce query results.

In the Edutella project [8], each participating peer has to obtain an RDF schema to be shared in the community. Additionally, each vocabulary used in the schema has to be based on a shared vocabulary dictionary. These limitations lead to decreased peer autonomy. In contrast, peers in the ORCHESTRA system [13] are free to publish and use any schema (from the same domain) for sharing in a collaborative data sharing system. There is no limit on the number of peers that can simultaneously publish and reconcile their actual schema to be shared. Peers are