Semantic Information Retrieval over P2P Network

Yulian Yang

Semantic Information Retrieval over P2P Network

Yulian YANG

Université de Lyon, CNRS
INSA-Lyon, LIRIS, UMR5205, F-69621, France
yulian.yang @ liris.cnrs.fr

RÉSUMÉ. *Porteurs de nombreux avantages en termes d'évolutivité, de décentralisation et d'auto-organisation, les réseaux P2P se présentent comme une alternative intéressante lorsqu'il s'agit de publier et d'accéder à de l'information. Plusieurs travaux de recherche ont traité du problème du routage des requêtes au sein de systèmes d'information soutenus par des réseaux P2P. Cependant, peu d'entre eux semblent intéressés pour savoir si les mécanismes de routage sémantique de requêtes permettent effectivement de fournir aux utilisateurs des informations sémantiquement pertinentes en fonction de leurs besoins. Ainsi, dans ce travail nous proposons d'étudier l'application de la sémantique à la Recherche d'Information traditionnelle ainsi qu'aux systèmes d'information basés sur des réseaux P2P. Nous démontrons l'écart qui subsiste entre la recherche d'information sémantique et le routage sémantique des requêtes. Finalement, nous présenterons les défis offerts par la recherche d'information sémantique lorsqu'elle est appliquée aux réseaux P2P.*

ABSTRACT. *As an alternative to publish or access information, P2P network is attracting more and more attention because of its advantages in scalability, decentralization and self-organization. While a lot of research has been focused on semantic query routing in information system over P2P network, little work concerns if semantic query routing can satisfy the requirement of semantic information retrieval, and to what extent the query routing mechanism can provide the information semantically related to users' query. This paper studies the application of semantics in traditional information retrieval and information system over P2P network, demonstrates the gap between semantic information retrieval and semantic query routing, and finally present the challenge of semantic information retrieval over P2P network.*

MOTS-CLÉS : *Recherche d'information, réseau P2P, Sémantique, Echelle Sémantique, Routage Sémantique*

KEYWORDS: *Information retrieval, P2P network, Semantics, Semantic layer, Semantic routing*

1. Introduction

Peer-to-peer (P2P) network refers to the system or application which performs a task in a decentralized manner by employing distributed resources such as computing power, storage, data and network bandwidth (MIL 02). It has the character of scalability, decentralization and self-organization. One of the major applications over P2P network is file/content sharing. Various file sharing systems have been designed and used in real application. For example, the first decentralized file sharing system Gnutella and a number of BitTorrent based services. Although some of these systems have been undergoing a series of the lawsuits, P2P network remains to be an efficient mechanisms for information sharing and acquisition.

As the technology and application of P2P network is developing, the resources in P2P network will become diverse, and the information demand of users will not be satisfied just via key-word matching. For example, the users would not just like to find a movie named 'Inception', which can be implemented by matching the query 'Inception' and the documents' titles. They might want some information expressed by more complicated query, such as 'a file about the history of the pop music' or 'a book written by a female author who was born in Beijing' stored in a distant peer. In this case, information retrieval (IR) over P2P network is supposed to answer the query by exploring the semantics of resources in peers.

Recently, the research on centralized semantic IR has developed rapidly with the technology for semantic web (SET 07). When it comes to IR over P2P network, a lot of work is concentrated on the scalability and the traffic as well as semantic query routing (LUU 08, ROS 08). Semantic query routing refers to forwarding the query to target resources by evaluate their semantic relationship on topic level. However, the queries in semantic IR is often specific, and little work has been done so far to study the gap between semantic query routing and achieve semantic IR.

In this paper, we study the issues of performing semantic IR over P2P network. Specifically, we study the application of semantics in centralized IR system in Section 2, and demonstrate current IR mechanisms over P2P network in Section 3. In Section 4, the gap between semantic IR and semantic query routing is analyzed. The challenges in semantic IR over P2P network are presented in Section 5.

2. Semantic information retrieval

Semantic IR aims to identify the relevant information to a query via evaluating their semantic relationship rather than key-word matching. We divide it into two levels, according to the semantic relationship between users' queries and the retrieved information. We assume the information to be retrieved is a collection of text documents.

In one level, a query can be answered by evaluating its semantic relevance to the content of each document. For example, the user intends to find the documents about the marriage of Eva Longoria and Tony Parker, which is very popular on the inter-

net recently. The semantic IR system is supposed to return the documents about their wedding and divorce, because wedding and divorce are both related to marriage. A lot of approaches are proposed for this, such as the probabilistic approach (JON 00), language model (CRO 03), LSA (DEE 90) and approach employing the theory in machine learning (YUE 07). These approaches assume the terms concurrent in the same context have a common concept. However, these approaches can't represent the semantic relationship between the terms. The specification of the semantic relationship is essential for performing more efficient IR, especially when the query submitted includes multiple keywords (TIA 06).

In the other level, the result of a query not only concerns the content in individual document, but also its relations with the outside information (e.g. the content in the other documents). For instance, the query aims to find documents about 'mobility' whose author once published a paper titled as 'peer to peer computing'. In order to implement such a query, we need to find the author of the paper 'peer to peer computing' in the author list of this paper, then continue to find the other documents written by this author. Thanks to the technology in semantic web, ontology and information extraction, semantic IR in this case is becoming promising (FEN 03), and relevant work has been done in (GRA 05).

3. Current information retrieval over P2P network

Information retrieval over P2P (P2P-IR) network aims to forward a query to target resources (query routing), perform relevance evaluation on the target resources, and then return the top-k most relevant documents to the user who submits the query. Most of the work in this field focuses on how to achieve semantic query routing, which refers to forward the query to the peers which hold the semantically related resources to it. To a certain extent, it often refers to query routing when people mention P2P-IR.

DHT is the most common type of structured P2P network so far, which indexes the resources in the network via a hash table and distributively stores the table in peers. The indexing granularity of the resources often is the document (LUU 08) or peer (MAT 05). Structured P2P network can response the queries quickly, but it can only support key matching (IVA 06) which present the employment of semantics to a great extent. Relevant work is done in CAN (a semantic P2P layer), which represent each document with a vector and map the documents into a semantic space. Close documents are semantic related in the semantic space. The query is also mapped to the semantic space in the same way. The documents close to it in the semantic space are identified as relevant ones.

In IR over unstructured P2P network, the query has to be flooded through the network to find the peers sharing the relevant data, e.g. Gnutella. Therefore, a lot of research in this subfield mainly concentrate on forwarding the query semantically. Except random walking algorithms (GKA 06), (COH 07) proposed a guided search over associative overlays. (SAR 04) introduced the percolation search algorithm for locating and retrieval content in random network. (ZHU 05) presented an approach which build semantic links between peers and then perform query routing on it.

There also exists some work which employs super-peer mechanism or takes advantage of both structured and unstructured P2P network to perform IR. For example, (PAP 10) proposed to combine DHTs and peer clusters for efficient full-text P2P indexing, (DOU 10) implemented a super-peer based P2P network for organizing similar content in the network. A semantic overlay network was implemented in (CRE 05).

4. Semantic information retrieval over P2P network

Although a lof of work has been done both in semantic query routing in P2P network and the semantic IR in centralized systems, no existing work concerns to what extent the semantic query routing mechanism can satisfy semantic IR. We will study it in this section.

4.1. *Semantic information retrieval and semantic query routing*

Query routing and IR both aim to find relevant resources, but they have intrinsical difference when it comes to semantic IR. Generally, semantic query routing is often performed on a topic level while semantic IR is more specific. Let's see how semantic IR would fail during the semantic query routing.

In structured P2P network, take semantic layer over CAN for example. The system maps the documents in the network into the semantic space, where each document is represented as a vector and similar documents would be close to each other. The similarity between documents is evaluated by the distance between vectors, and it has no semantic implication. Moreover, the vector for each document is calculated via SVM or LSI. It doesn't include information about the specific relationship between terms or keywords. Therefore, It can not handle very well with the queries with more than one key words. For the queries which concern the documents and its relations with outside information, it would be more difficult for the system to handle.

For semantic query routing in unstructured P2P network, the query is forwarded to the peers which have relevant resources. Here, the problem appears again. How to evaluate the relevance between a query and a collection of resource ? If the relevance is too specific, the query takes the risk that it has no relevance to all the next-hop peers. If the relevance is general, the query would be forwarded to too many peers which no doubt is a waste of traffic. In addition, it still cannot deal with the query like 'documents about mobility whose author once published a paper titled as peer to peer computing', especially when the concerned documents are stored in different peers.

4.2. *Challenges*

In conclusion, the challenges in semantic P2P-IR lies in the following aspects :

1) the user's queries are specific, while query routing over the semantic layer of P2P network is general, often on the topic level. So there is a risk that the query might be forwarded to peers which has no answers to it, but something only related on the topic level, where the topic of the query often can be evaluated via Ontology. Take the web search over P2P network for example, the user would like to request some

specific query like ‘the possible cause of headache’. The query might be forwarded to the peers which have the resources about ‘health’, but we have no idea about the probability that the peers has the information of ‘the cause of a headache’. Apparently, this will cause a waste of traffic and computing cost if the peers actually have no such information.

2) the ‘relevance’ and ‘related’ defined in the semantic layer of structured P2P network is vague. This results in a system which can not provide accurate results for a query. They can not deal with the queries which include several keywords and have explicate relationship among them. Let alone more complicated queries, which might be concerned with multiple documents located in different peers.

3) It’s hard to determine the generalization level of the outline for each peer. To make P2P network be a real self-organized one, peers are assumed to manage their resources by themselves, and only share an outline to make sure relevant queries can be forwarded to them. This can also decrease the traffic cost of the network. However, what kind of outline can precisely and concisely represent the resource in a peer, and also facilitate the query routing of the P2P network ?

5. Conclusion

This paper studied semantic P2P-IR. We analyzed the semantic IR in centralized systems, presented the research achievements and issues in current information systems over P2P network. The challenges of semantic P2P-IR are demonstrated.

6. Bibliographie

[COH 07] COHEN E., FIAT A., KAPLAN H., « Associative search in peer to peer networks : Harnessing latent semantics », *Computer Networks*, vol. 51, n° 8, 2007, p. 1861-1881.

[CRE 05] CRESPO A., GARCIA-MOLINA H., « Semantic Overlay Networks for P2P Systems », *Agents and Peer-to-Peer Computing*, vol. 3601, p. 1-13, Springer Berlin / Heidelberg, 2005.

[CRO 03] CROFT W. B., JOHN L., *Language Modeling for Information Retrieval*, vol. 13 de *The Information Retrieval Series*, Springer, 2003.

[DEE 90] DEERWESTER S., DUMAIS S. T., FURNAS G. W., ET AL., « Indexing by Latent Semantic Analysis », *Journal of the American Society for Information Science*, vol. 41, 1990, p. 391-407.

[DOU 10] DOULKERIDIS C., VLACHOU A., NORVAG K., ET AL., « Efficient search based on content similarity over self-organizing P2P networks », *PEER-TO-PEER NETWORKING AND APPLICATIONS*, vol. 3, n° 1, 2010, p. 67-79.

[FEN 03] FENSEL D., WAHLSTER W., LIEBERMAN H., ET AL., *Spinning the Semantic Web : Bringing the World Wide Web to Its Full Potential*, MIT Press, 2003.

[GKA 06] GKANTSIDIS C., MIHAIL M., SABERI A., « Random walks in peer-to-peer networks : Algorithms and evaluation », *Performance Evaluation*, vol. 63, n° 3, 2006, p. 241-263.

[GRA 05] GRAUPMANN J., SCHENKEL R., « The Light-Weight Semantic Web : Integrating Information Extraction and Information Retrieval for Heterogeneous Environments », *SIGIR-HDIR*, 2005.

[IVA 06] IVANA P., MARTIN R., TOAN L., FABIUS K., ET AL., « Beyond term indexing : A P2P framework for web information retrieval », *Informatica*, vol. 2, 2006, page 2006.

[JON 00] JONES K. S., WALKER S., ROBERTSON S. E., « A probabilistic model of information retrieval : development and comparative experiments », *Information Processing and Management*, 2000, p. 779-840.

[LUU 08] LUU T., SKOBELTSYN G., KLEMM F., ET AL., « AlvisP2P : scalable peer-to-peer text retrieval in a structured p2p network », *VLDB*, 2008.

[MAT 05] MATTHIAS B., SEBASTIAN M., PETER T., ET AL., « Improving collection selection with overlap awareness in P2P search engines », *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, 2005.

[MIL 02] MILOJICIC D. S., KALOGERAKI V., LUKOSE R., ET AL., « Peer-to-peer computing », rapport, 2002, HP Labs, Palo Alto.

[PAP 10] PAPAPETROU O., SIBERSKI W., NEJDL W., « PCIR : Combining DHTs and peer clusters for efficient full-text P2P indexing », *COMPUTER NETWORKS*, vol. 54, n° 12, 2010, p. 2019-2040.

[ROS 08] ROSTAMI H., HABIBI J., LIVANI E., « Semantic routing of search queries in P2P networks », *Journal of Parallel and Distributed Computing*, vol. 68, n° 12, 2008, p. 1590-1602.

[SAR 04] SARSHAR N., BOYKIN P. O., ROYCHOWDHURY V. P., « Percolation Search in Power Law Networks : Making Unstructured Peer-to-Peer Networks Scalable », *Proceedings of the Fourth International Conference on Peer-to-Peer Computing*, P2P'04, 2004, p. 2-9.

[SET 07] SETCHI R., TANG Q., « Concept Indexing Using Ontology and Supervised Machine Learning », *Transactions on Engineering, Computing and Technology*, vol. 19, 2007, p. 221-226.

[TIA 06] TIAN C., TEZUKA T., OYAMA S., ET AL., « Improving Web Retrieval Precision Based on Semantic Relationships and Proximity of Query Keywords », *DEXA'06*, 2006, p. 54-63.

[YUE 07] YUE Y., FINLEY T., RADLINSKI F., ET AL., « A support vector method for optimizing average precision », *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, 2007, p. 271-278.

[ZHU 05] ZHUGE H., LIU J., FENG L., ET AL., « Query routing in a peer-to-peer semantic link network », *Computational Intelligence*, vol. 21, 2005, p. 197-216.