SemSearch: A Scalable Semantic Searching Algorithm for Unstructured P2P Network

Wei Song[1], Ruixuan Li[1,*], Zhengding Lu[1], and Mudar Sarem[2]

[1] College of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, Hubei, P.R. China
[2] School of Software Engineering, Huazhong University of Science and Technology, Wuhan 430074, Hubei, P.R. China
weisong@smail.hust.edu.cn,{rxli,zdlu}@hust.edu.cn, mudar66@hotmail.com

Abstract. Resource searching in the current peer-to-peer (P2P) applications is mainly based on the keyword match. However, more and more P2P applications require an efficient semantic searching based on the contents. In this paper, we propose a novel scalable semantic searching algorithm named SemSearch for the unstructured P2P networks. For the consistency and flexibility of semantic analysis, we integrate global and local semantic information to do the semantic analysis. Moreover, SemSearch transfers the searching requests to the peers whose shared resources are more semantic similar to implement semantic searching. We further evaluate the performance of SemSearch through the simulation experiments.

Keywords: SemSearch, Peer-to-peer, Semantic analyzing, Semantic searching.

1 Introduction

The resource searching in most current P2P systems is based on the keyword match. However, some P2P applications need an efficient resource searching over the contents of resources. There is not a scalable semantic searching algorithm in the real P2P applications. So, improving the semantic analysis and implementing a scalable P2P semantic searching is the main motivations of this paper. In this paper, we integrate global and peer's local semantic information to analyze the resource. Moreover, we propose SemSearch, a novel scalable semantic searching algorithm for the unstructured P2P networks.

The main contributions of this paper include: 1) Proposing a scalable semantic analysis method to achieve the consistency and scalability of the semantic analysis; 2)

* This work is supported by National Natural Science Foundation of China under Grant 60403027, 60773191, 70771043, National High Technology Research and Development Program of China under Grant 2007AA01Z403, China Postdoctoral Science Foundation under Grant 20060400846, Natural Science Foundation of Hubei Province under Grant 2005ABA258.

Y. Zhang et al. (Eds.): APWeb 2008, LNCS 4976, pp. 631–636, 2008.

Proposing an efficient P2P semantic searching algorithm, called SemSearch; 3) Carrying out the simulation experiments to evaluate the performance of SemSearch.

The rest of the paper is organized as follows. In the next Section, we present some related work in this background. Section 3 presents the details of the SemSearch protocols. Section 4 evaluates the SemSearch performance through the simulation experiments. Finally, the paper is concluded in Section 5.

2 Related Work

The current P2P resource searching can be classified into two categories. One is based on the resource description (keyword, file name, etc.) match such as Gnutella, Bitcomet, and other popular file sharing P2P applications. Some searching algorithms in the structured P2P networks include: Chord [1], Tapestry [2], and CAN [3] also belong to this category. In this category, the keyword match is used to judge whether resources satisfy a searching request or not. Many P2P applications have employed this method. However, the keywords are difficult to represent the resource contents and users' interests. So, the searching results can not satisfy the users.

The second category is based on the resource contents. Semantic Overlay Network (SON) [4] first introduces the concept of the semantic searching. We can classify the P2P semantic searching into two sub-categories: 1) Semantic searching based on the static semantic analysis, in which peer extracts the resource semantic information from a static semantic model. pSearch [5] used a semantic vector to describe the resources. SemreX [6] was a semantic P2P overlay based on a concept tree. The static semantic analyses are easy implemented. However, this method is difficult to be extended. Hence, such method only adapts to the special applications. 2) Semantic searching based on the scalable semantic analysis. In this sub-category, the semantic analysis is extendable by the self-study. H.T. Shen et. al. proposed a semantic P2P framework [7] based on Hierarchial Summary Structure. RS2D [8] was a risk driven semantic P2P research search service. FCAN [9] implemented the content query over structured P2P overlay. And Semantic Small World [10] researched the semantic searching from the views of small world theory and semantic clustering. These semantic searches are scalable and flexible. Nevertheless, with the network size growing, it is difficult to keep the consistency of semantic analyzing and search.

3 SemSearch Protocols

3.1 Semantic Description in SemSearch

Resnik has proposed that the key to the similarity of two concepts is the extent to which they share information[11]. Therefore, in SemSearch, we measure the word frequency in the resources to describe the resources. SemSearch has a global semantic knowledge base which is a keyword matrix, named Global Semantic Keyword Matrix (GSKM). A GSKM row describes a semantic cluster, for example, GSKM shown in (1) describes four semantic clusters.

$$\begin{bmatrix} \text{operation system} & \text{process} & \text{priority} & \text{scheduling} & ... \\ \text{database} & \text{view} & \text{oracle} & \text{pattern} & ... \\ \text{semantic web} & \text{ontology} & \text{OWL} & \text{meta - model} & ... \\ \text{distributed system} & \text{replication} & \text{P2P} & \text{Grid} & ... \end{bmatrix} \quad (1)$$

While a peer P joins SemSearch, it first loads the GSKM $A^{m\times n}$. Afterwards, P measures the frequency of A's elements in each resource r_i to build r_i's keyword frequency matrix $B_i^{m\times n}$. B_i describes the keyword's distribution in r_i's contents. We further normalize B_i to build a relative keyword frequency matrix $P_i^{m\times n}$, which describes the keyword frequency proportion in the resource r_i.

The sum of P_i's elements in a row describes r_i's keyword frequency proportion in a semantic cluster. So, we introduce the resource semantic vector $C=(1,1,\ldots,1)^{1\times n}P^T$ to represent the resource contents' matching degree with each semantic cluster. Usually, the shared resources in a peer are similar. Consequently, we describe the semantic information of a peer by its local shared resources. Hence, we introduce the peer's semantic vector $R=\frac{1}{t}\sum_{i=1}^{t}C_i$ to represent the peer's semantic information.

3.2 Extended Semantic Analysis in SemSearch

The global GSKM can not reflect the comprehensive resource contents. Moreover, an exhaustive GSKM is quite difficult for a P2P system. Therefore, we build peer's Local Semantic Keyword Matrix (LSKM) to make semantic analysis scalable.

A peer P with t shared resources constructs a word frequency vector $B(b_1,b_2,\ldots,b_t)$ for GSKM $A^{m\times n}$'s each keyword. Each component in B is a keyword's frequency locally. The similar meaning words usually have the similar distributions, so P looks for the new similar meaning words to extend LSKM. While P extends its LSKM $L^{m\times N}(N>n)$, it first copies A's elements into L and set the elements $l_{ij}(n<j\leq N)$ as null. Moreover, we use the average word frequency vector to judge whether a new word is a semantic cluster's synonyms. This paper uses the cosine distance of two vectors to measure their similar degree. The cosine distance of two vectors x and y is defined as follows: $\theta(x, y)=\cos^{-1}\frac{x\cdot y}{|x||y|}$.

P full-text retrieves its local resources and collects a keyword candidates list. Afterwards, it computes the similar degree of every candidate and semantic cluster. Each semantic cluster in L has a cosine distance threshold $\hat{\theta}$. If the cosine distance of a candidate keyword and a semantic cluster's average word frequency vector is less than the threshold, then the keyword joins this semantic cluster.

When a new keyword S attempts to join the i^{th} semantic cluster, if the i^{th} row of L has null values, then S replaces a null value. Otherwise, S replaces the keyword which has the maximal cosine distance to the semantic cluster. And we set $\hat{\theta}$ as the elements' maximal cosine distance to the average word frequency vector. SemSearch peer periodically calls the extension algorithm to extend LSKM.

3.3 Semantic Search of SemSearch

A SemSearch searching request $q(g,\beta)$ is made up of a query vector g and a cosine distance threshold β. The searching request $q(g,\beta)$ searches the semantically similar resources whose cosine distance to g is less than β. The searching source peer P selects a set of keywords from its interests to construct the query vector. P constructs a matrix $V^{m\times t}$ in which the element v_{ij} is the j^{th} resources' frequency in the i^{th} semantic cluster. Moreover, for a set of inquired keywords U, P constructs an inquired keyword frequency vector $W(w_1,w_2,...,w_t)$ in which w_i represents U's total frequency at the i^{th} resources. The vector W represents the distributions of the inquired keywords at local resources. So, we use the matrix V and vector W to build query vector g.

$$g=\frac{WV^T}{\left|WV^T((1,1,..,1)^{1\times m})^T\right|} \tag{2}$$

The peers return local semantically similar resources whose cosine distance to g is less than β. Furthermore, peers select no more than r neighbors whose peer semantic vectors are near to g to transfer the searching request.

The searching results of SemSearch depend on the threshold β. Therefore, we use an adaptive method to determine β value. When a peer first launches a searching request, it uses β_{init} as the initial threshold. Afterward, the peer adjusts the β value from β_{min} to β_{max} based on the previous return responses.

4 Simulation Experiments

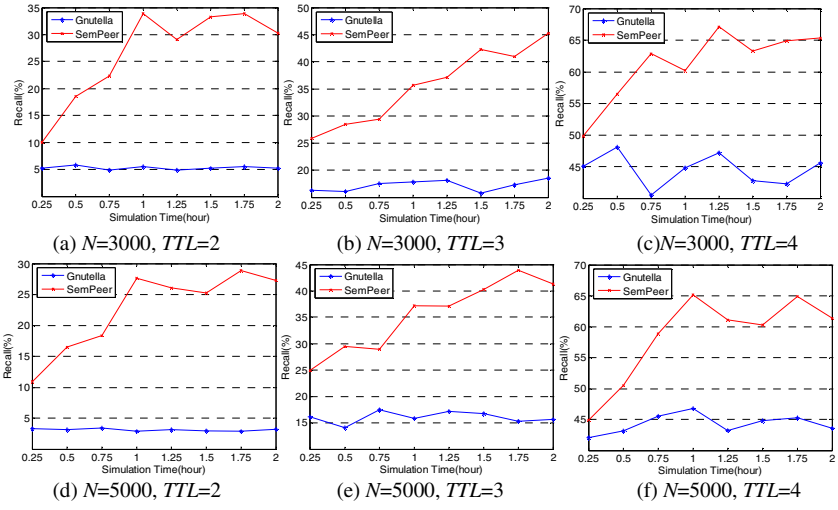
In this Section, we evaluate the performance of SemSearch by simulations. The network size is 3000-5000, and the peer's connection degree follows the power law distribution of exponent A=3.0 and $Degree_{average}$=4.0. The shared resources are the documents in ACM database and classified based on ACM Computing Classification System 98 [12]. We select 3000 documents from each top catalog except General Literature and Computing Milieux. Moreover, the GSKM is a 9×5 matrix and LSKM is a 9×10 matrix. In the simulations, a peer launches two searching requests for each minute. More experimental parameters are show in Table 1.

Table 1. Parameter and settings in the simulation experiments

Parameter	Parameter Meaning	Default Value
β	cosine distance threshold	0.7
TTL	search radius	2-4
r	Max peers that searching request is transferred to	3

4.1 SemSearch Recall Rate

We do the simulation experiments to compare the average recall rate of SemSearch and Gnutella in various searching *TTL* and network size. The experimental results are shown in Figure 1.



(a) *N*=3000, *TTL*=2 (b) *N*=3000, *TTL*=3 (c)*N*=3000, *TTL*=4

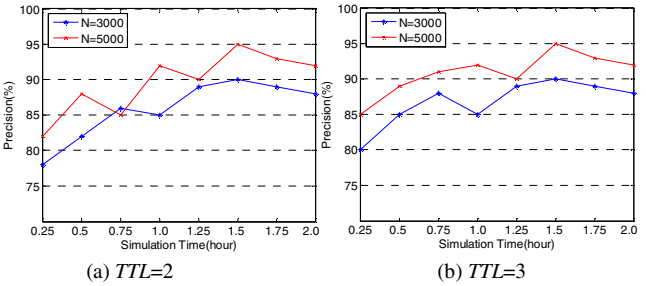
(d) *N*=5000, *TTL*=2 (e) *N*=5000, *TTL*=3 (f) *N*=5000, *TTL*=4

Fig. 1. Recall Comparison between Gnutella and SemSearch

As shown in the experimental results, SemSearch gets a higher recall rate. Furthermore, we can easily discover that the SemSearch recall rate rise over time. Analyze this phenomenon, we have found that the peer's semantic analysis is more accurate with the LSKM extending. So, the recall rate of semantic searching rises.

4.2 SemSearch Precision

We have measured the precision to evaluate the consistency of semantic searching. Based on the ACM CSS classification, a peer launches a searching request for the resources in one catalog. If the return resources are in the same catalog, we consider the returned resources are semantically correct. The experimental results are shown in Figure 2.



(a) *TTL*=2 (b) *TTL*=3

Fig. 2. Precision ratio of SemSearch over time

The experimental results show that SemSearch keeps a high precision (about 90%) in the experiments. Further, analyzed the experimental results, we have found that the 40%-50% semantically inapprehensive returned results still are similar to the initial

searching requests. So, we can draw a conclusion that the semantic analysis of SemSearch is available, and the returned documents are the peer's required ones.

5 Conclusion and Future Work

In this paper, we propose a novel scalable semantic searching algorithm for the unstructured P2P networks, named SemSearch. Comparing with the existing P2P semantic searching algorithms, the semantic analysis of SemSearch is more scalable and easy to be implemented. The experimental results show that SemSearch is a scalable and efficient P2P semantic searching algorithm. In the future work, we will further research using the relationship of various semantic clusters to improve the semantic analysis and resource searching. Also, we plan to implement SemSearch over an actual P2P application to further evaluate and improve its performance.

References

1. Stoica, I., Morris, R., Karger, D., Kaashoek, M.F., Balakrishnan, H.: Chord: a scalable peer-to-peer lookup service for internet applications. In: ACM SIGCOMM Computer Communication Review, vol. 31, pp. 149–160. ACM Press, New York (2001)
2. Zhao, B.Y., Huang, L., Stribling, J., Rhea, S.C., Joseph, A.D., Kubiatowicz, J.: Tapestry: A Resilient Global-Scale Overlay for Service Deployment. J. IEEE Journal on Selected Areas in Communications 22, 41–53 (2004)
3. Ratnasamy, S., Francis, P., Handley, M., Karp, R., Shenker, S.: A Scalable Content-Addressable Network. In: The ACM SIGCOMM 2001, pp. 161–172. ACM Press, New York (2001)
4. Crespo, A., Garcia-Molina, H.: Semantic overlay networks for P2P systems. Technical report, Stanford University (2002)
5. Tang, C., Xu, Z., Mahalingam, M.: pSearch: Information retrieval in structured overlays. ACM SIGCOMM Computer Communications Review 33(1), 89–94 (2003)
6. Chen, H.H., Jin, H., Ning, X.M., Yuan, P.P., Wu, H., Guo, Z.X.: SemreX: A semantic similarity based P2P overlay network. J. of Software 17(5), 1170–1181 (2006)
7. Shen, H.T., Shu, Y.F., Yu, B.: Efficient semantic-based content search in P2P network. IEEE Transaction on Knowledge and Data Engineering 16(7), 813–826 (2004)
8. Klusch, M., Basters, U.: Risk Driven Semantic P2P Service Retrieval. In: The 6th IEEE International Conference on Peer-to-Peer Computing (P2P 2006), pp. 161–170. IEEE Computer Society, Los Alamitos (2004)
9. Wang, J., Yang, S., Gao, Y., Guo, L.: FCAN: A Structured P2P System Based on Content Query. In: The 5th International Conference on Grid and Cooperative Computing (GCC 2006), pp. 113–120 (2006)
10. Li, M., Lee, W., Sivasubramanian, A.: Semantic small world: an overlay network for Peer-to-Peer search. In: The 12th IEEE International conference on Network Protocols (ICNP 2004), pp. 228–238. IEEE Computer Society, Los Alamitos (2004)
11. Resnik, P.: Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. J. of Artificial Intelligence Research 11, 95–130 (1999)
12. ACM CCS, `http://www.acm.org/class/1998/`