File Correspondences Dictionary Construction in Multilingual P2P File Sharing Systems

Hongding Wang[1,2], Shaohua Tan[1,2], Shiwei Tang[1,2],
Dongqing Yang[1] , Yunhai Tong[1,2]

[1]School of Electronics Engineering and Computer Science,
[2]National Laboratory on Machine Perception
Peking University, 100871, China
{hdwang,tsh,tsw,dqyang,yhtong}@pku.edu.cn

Abstract. Sharing files discovery is a fundamental problem in P2P networking. This paper presents a name-based approach for identifying sharing file correspondences in multilingual P2P systems. The problem is first analyzed through comparing the names of the sharing files in different nodes of a real P2P community, which name those files in different languages. Then based on the relationships of those files names, a computer-aided method is proposed to solve the problem. Furthermore, the framework and identifying procedure of this method have been discussed in the paper.

1 Introduction

Nowadays, Peer-to-Peer (P2P) systems, which are distributed systems consisting of interconnected nodes able to self-organize into network topologies with the purpose of sharing resources such as content, CPU cycles, storage and bandwidth, are very popular in resource sharing field. They are capable of adapting to failures and accommodating transient populations of nodes while maintaining acceptable connectivity and performance without requiring the intermediation or support of a global centralized server or authority [1]. Due to such characteristics, users are free to join and leave at any time in P2P system [26].

Dynamic resource discovery in different nodes is a central problem in P2P environment, which means the capability of finding the existing resources in the network that best match the requirements of a given resource request [2][26]. Sharing files are one kind of the most important resources in P2P systems. So for file exchanges and retrieval with file name is a simple and convenient way in P2P systems, however, semantic heterogeneity often makes name-based query be difficult. Due to sharing files named in different languages, for instance, a node names them in English, while another in Chinese and one in Chinese Pinyin, one popular problem in P2P systems called naming conflicts arises. . In this paper, we propose a name-based approach for identifying sharing file correspondences in P2P networking.

The rest of this paper is organized as follows. In Section 2, we review existing work in this field. In Section 3, we discuss the details of motivating examples in multilingual P2P systems. Section 4 presents the method of identification of sharing file correspondences in multilingual P2P systems. Section 5 provides the experimental results. Finally, we conclude this paper in Section 6.

2 Related Work

The P2P model is made popular by file-sharing applications, for these applications, such as Kazaa, Overnet, BitTorrent and Maze, provide improved scalability and performance [3-9,18].

Given task requirements and resource policies, the resources discovery problem arises in P2P and Grids [10, 11]. With heterogeneous ontology descriptions, paper [2] gave a semantic matching-based approach to cope with the dynamic resource discovery problem in distributed contexts. Furthermore, in order to simplify the systematic building of Grid applications, paper [10] suggested how to describe Grid resources with ontology. A matchmaker approach is presented for resource discovery in the Grid with a set of rules in [11] to improve the effectiveness of the matching process. The EDUTELLA project developed a P2P infrastructure for metadata sharing in RDF format [12]. Besides, a metadata model of RDF for encoding semantic information was introduced allowing peers to handle heterogeneous views on the domain of interest [13]. Based on epistemic logic, authors of [14] gave an approach to P2P data integration, which was more suitable than the commonly adopted semantics based on FOL. With P2P technology, Lu Yan et al provided a framework of global storage system- SkyMin, the name sever of the system manages the indexes of the peers' resources, being used to find sharing resources [19]. In order to connect heterogeneous information providers and share those heterogeneous information resources, paper [20] proposed a super-peer topology for schema-based P2P networks and discussed how the schema information can be used for routing and clustering in such a network with existing information integration concepts of mediator-based information systems. Furthermore, paper [26] studied some problems of information retrieval in a P2P file sharing system, and paper [30] gave a survey of anonymous P2P file sharing. As is a fact, in many P2P file sharing systems, querying was usually performed by using file names [3,4,8,9,27], so file name correspondences identification is very useful, and researchers have studied the genomic resources table mapping problems in P2P systems [28].

As is well known, the dual problem of synonymy and homonym is a major topic addressed in information retrieval research [17]. The dual problem exists in monolingual situation, as well as in multilingual situation, so Cross-Language Information Retrieval (CLIR) attracts more and more interests of computer scientists [21-25]. As translating documents is very expensive, most researchers in this field opt to take the query translation approach [24].

Though resource discovery have been studied extensively in P2P systems, to our knowledge, we are not aware of any previous work that has considered sharing files discovery involving multiple languages in P2P systems. This paper intends to propose

a name-based method to solve the problem of sharing file correspondences identification in multilingual P2P systems.

3 Motivating Examples

Here we first give a scenario of sharing files discovery in multilingual P2P systems, which comes from a real hybrid P2P system-Maze system [8,9,18]. Maze is one of the first large-scale deployments of an academic research project, with over 210,000 registered users and more than 10,000 users online at any time, sharing over 140 million files [8].

(a) files named in English (b) files named in Chinese (c) files named in Pinyin

Fig.1. Examples in multilingual P2P systems

Figure 1 illustrates three working interfaces drawn from Maze community. We mainly discuss the sharing files at each node. Without name standards, the sharing files are named in different language by different nodes in Maze system. More often, in order to conveniently use, the files are usually classified in different categories, and put into different folders corresponding to those categories, but we take files and folders as sharing files. From Figure 1, we find that a node names its sharing files in English (as Figure 1-a shows, hereafter node a); the second names its sharing files in Chinese (as Figure 1-b shows, hereafter node b); and last one names its sharing files in Pinyin (as Figure 1-c shows, hereafter node c). Furthermore, even in a node, its sharing files may be named in different languages because it doesn't rename those files' names when the node replicates them from other nodes. In fact, the same category sharing resources are named in different languages, as Table 1 shows.

Table 1. Sharing file names of different nodes in Figure 1

Resource	Node a	Node b	Node c
learning	learning		xuexi
videos	Movie	电影	dianying
music	Music		
software	software	软件	ruanjian
picture		图片	tupian

From Table 1, we can see that the naming standards of files of the node a, node b and node c are different, i.e., for the same information resource, node a names it in English, while node b in Chinese, and node c in Pinyin, it is the relationship among the file names in those peers.

Moreover, we have analyzed the amount of file names in different languages in Maze system with statistic technique during Jan of 2006. Each time, we searched the sharing files (focusing on four main categories resources) with randomly selecting keys in Chinese, English, and Pinyin respectively on three computers based on the existing terms in Maze system, and recorded the amount of returned results. Because of the high churn rate, the statistics are interrelated closely with the experimental period. We have got the approximate amount of the four kinds of resources named in different languages after many trials in Jan of 2006, as Table 2 shows.

Table 2. Percent of file names in different languages in Maze system (Jan., 2006)

Resource	In Chinese	In English	In Pinyin
Video	51.6%	43.5%	4.9%
Music	42.9%	46.4%	10.7%
Software	59.2%	36.5	4.3%
Games	55.8%	41%	3.2%

Such phenomena prevail in P2P networking of China, because P2P paradigm dictates a fully distributed, cooperative network design, where nodes collectively form a system without any supervision [15]. Due to lack of standards and different naming manners, the problem of the files named in different languages in Maze system arises. Though different node names files in different languages, the file names are as much as descriptions of content [26]. In fact, in many file sharing systems, querying is accomplished by using simple value searches with file names [3,4,27]. Therefore, file names correspondences are very useful for file retrieval in P2P systems, and [28] have studied maintenance and management of the mapping tables in P2P systems, however without mentioning multilingual problems. Though comparing file names to identify correspondences has limitations, it is a convenient and effective method for finding file correspondences in P2P systems [28].

As far as we know, Maze system hasn't resolved this problem, when searching a file, it merely returns the relevant files named in the same language the keys use.

4 Identification Method

We first assume that files represented with synonyms in different nodes in a P2P community are the same sharing file. Therefore, if we find files named with synonyms in different nodes in a P2P community, they are file correspondences.

4.1 Preliminaries

Definition 1. *Synonyms in multilingual context* are terms representing the same information resources, which are independent of the languages they use. For example, terms such as 'movie' and '电影 (dianying)' are synonyms in multilingual context.

Definition 2. *Name-based file correspondences* are those files if and only if their names are synonyms in multilingual context. So 'movie' and '电影 (dianying)' are

name-based file correspondences. Name-based file correspondences are denoted as $\Leftrightarrow_{name}$ in this paper, which have properties illustrated as following:

a) Reflexivity: a $\Leftrightarrow_{name}$ a;

b) Symmetry: if a $\Leftrightarrow_{name}$ b, then b $\Leftrightarrow_{name}$ a;

c) Transitivity: if a $\Leftrightarrow_{name}$ b, b $\Leftrightarrow_{name}$ c, then a $\Leftrightarrow_{name}$ c.

With the relationships of the file names, file correspondences can be identified in multilingual P2P systems. In the three nodes illustrated in Figure 1, the file correspondences are shown as following:

(1)learing $\Leftrightarrow_{name}$ xuexi;

(2)movie $\Leftrightarrow_{name}$ 电影 $\Leftrightarrow_{name}$ dianying;

(3)software $\Leftrightarrow_{name}$ 软件 $\Leftrightarrow_{name}$ ruanjian;

(4) 图片 $\Leftrightarrow_{name}$ tupian.

4.2 File Correspondences Identification Method

With the relationships of those file names, we propose a computer-aided system to identify name-based file correspondences in multilingual P2P systems. Figure 2 shows the framework of the system, which consists of three components--Chinese-English translator, Pinyin translator and matching module.

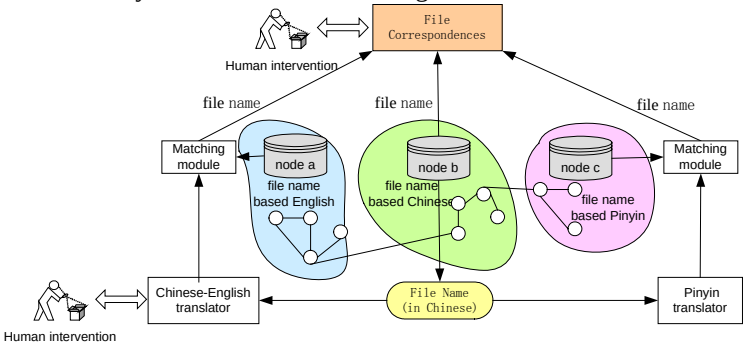

Fig. 2. Framework of file correspondences identification system

As Figure 2 illustrates, given a file named in Chinese of node b (on behalf of peers naming files in Chinese), in order to find its corresponding file named in English in node a (on behalf of peers naming files in English), the Chinese-English translator translates the file name into English, which is a CLIR problem [21-25]. Usually, two ways are able to solve the problem, one way is by means of online bilingual dictionary (i.e., Chinese-English) or machine translation [16], which is a straightforward approach, but human interventions needs to be involved for word sense disambiguation, the other uses domain-dependent thesaurus [2]. In this paper, we adopt to use online bilingual dictionary, as well as domain-dependent thesaurus in the system.

Here we neglect the routing path and DHT (distributed hash table) of P2P systems, merely study the problem of file correspondences.

At the beginning of correspondence identification, the system makes use of online bilingual dictionary to accomplish Chinese-English translation with expert's guidance and modification. For example, given a file which is named '电影' (dianying) in node b, the online bilingual dictionary translates it into English, several choices may be given by the dictionary, such as 'movie', 'film', 'cinema', 'cine', and so on. Obviously, 'movie' and 'film' are related closely, so the terms such as 'movie' and 'film' are chosen, while others are deleted. At the same time, a specific, domain-dependent bilingual thesaurus is constructed with the results of online bilingual dictionary. Then the domain thesaurus grows with accumulated knowledge of both bilingual dictionary and domain experts. Along with the growth of the thesaurus, the system can make use of it to Chinese-English translation later.

Based on what the Chinese-English translator having done, we retrieve files with those terms given by the Chinese-English translator respectively in node a. For one of the terms, if there are files returned from node a, it is a candidate correspondence of '电影', otherwise, it is not. Hence we get the candidate correspondences of '电影' in node a. For a candidate correspondence file name, the matching module begins to work with its returned files from node a. Two ways can accomplish the matching task the given file and the returned files. One way is by means of the structural characteristics (for example, using the extend name of a file, and so on), and the other is by means of contents of the files. Here we only make use of the structural characteristics to match the given file and the returned files. With one candidate name, if there are some of its returned files from node a having the same structural characteristics of the given file of node b, it is a name-based file correspondence of the given file of node b, therefore, name-based file correspondences are identified between node a and node b.

At the same time, in order to find name-based file correspondences in node c (on behalf of peers naming files in Pinyin), the Pinyin translator gives the Pinyin of the given file name of node b with existing tools. Due to just one Pinyin name given by the Pinyin translator for a given file name of node b, human interventions are unnecessary here. The Pinyin name is a candidate file name in node c corresponding to the given file in node b. Then the matching module works as mentioned above, name-based file correspondences are identified between node c and node b.

Due to terms are words or phrases in P2P file sharing systems, human modification is not as difficult as that of sentences translation in machine translation. Moreover, the mature domain-dependent thesaurus can be reused in the same domain in the future, so it is worth constructing such a domain-dependent thesaurus with human interventions.

Finally, with the transitive property of name-based file correspondence, file correspondences among different nodes in multilingual P2P systems are identified. By means of those file names correspondences in multilingual P2P systems, the file correspondences dictionary can be constructed. Certainly, the domain experts need to modify the file correspondence dictionary manually at the end of the construction of such a dictionary, deleting the wrong ones and adding the ones that the system fails to report. For example, a term 'video' in node a obviously is a file correspondence of

‘电影’ in node b, so the domain experts have to insert it into the file correspondences dictionary.

5 Experimental Results

With the method introduced in Section 4, we have constructed a preliminary domain-dependent file correspondences dictionary. Of course, as mentioned above, during the procedure of the dictionary construction, domain experts’ intervention is necessary. Furthermore, a prototype has been implemented to evaluate the merit of the file correspondences dictionary in multilingual P2P systems. We have downloaded randomly about 10,000 files from Maze system. Due to many duplicate names in those files, so we add prefix as new identifiers to each name when it is saved in the prototype. For example, we rename a file named ‘music.wma’ to ‘00001_music.wma’. With the similar name matching mechanism as file search engine of Maze system, we retrieve those files with its original names, not taking care of its prefix that we add when they are stored in the prototype. Two measures used in information retrieval field [29] are defined to evaluate the prototype in this paper. Precision is the ratio of the number of relevant file names retrieved to the total number of file names retrieved. Recall is the ratio of the number of relevant file names retrieved to the total number of relevant file names in the P2P systems. Figure 3 illustrated the experimental results.

$$precision = \frac{\left|\{relevant\ \ file\ names\} \bigcap \{retrieved\ \ file\ names\}\right|}{\left|\{retrieved\ \ file\ names\}\right|}$$

$$recall = \frac{\left|\{relevant\ \ file\ names\} \bigcap \{retrieved\ \ file\ names\}\right|}{\left|\{relevant\ \ file\ names\}\right|}$$

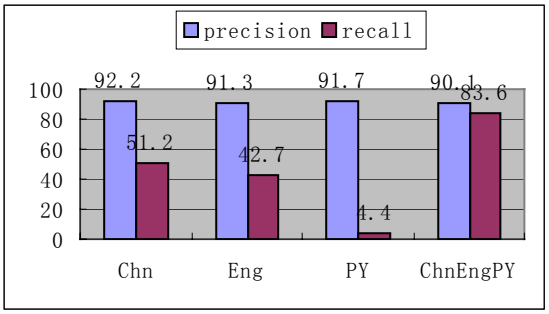

Fig. 3. Performance of the prototype

Figure 3 shows that the precision of the prototype just decreases a little, because a few errors are inevitable in the file correspondences dictionary, however, the recall has been improved notably. In fact, with the file correspondences identification procedure, it uses keys in three languages to retrieve sharing files for a query, so the number of the files retrieved almost includes all the relevant files named in the three languages. If nodes of the P2P community are aided with the file correspondences

dictionary, they will return most of the relevant files for a given query. As a result, more relevant sharing files are discovered in multilingual P2P systems.

It is an effective approach to improve the abilities of files share and download in P2P systems. According to the mechanism of files share and download in Maze system [9], when a peer accepts a file download request, the request will be transmitted to her friends to find more downloadable locations, then the requester can download the file from several sites at the same time. With the file correspondences dictionary, it can find more mirror images in an interest group for a file query, which can speed up the files download and serve for more requesters at the same time.

There are some problems related to the method. First of all, in order to improve the qualities of the file correspondences dictionary, our future work will make full use of the content of the files. Secondly, different from query translation approach in CLIR, we just translating queries into relevant languages of the target resources, the reason is that we merely find file correspondences with their names, which are heuristic clews to the end users. Moreover, the method introduced in section 4 includes two identification procedures, users can decide to choose either one or two of them when retrieval resources in the multilingual P2P systems.

6 Conclusion

In this paper, we have addressed a problem about sharing files discovery in P2P networks, and presented a name-based approach for identifying sharing file correspondences in multilingual P2P systems. While sharing files discovery has been studied extensively in P2P field, we are not aware of any previous work that has considered identifying sharing file correspondences in multilingual P2P systems. We analyzed the problem in detail, and based on the relationships of the files' names in multilingual P2P systems, we proposed a computer-aided method to solve the problem. The components and procedure of constructing the file correspondences dictionary among different P2P nodes have been discussed in the paper. The experimental results show that such a dictionary is helpful to retrieval sharing files in P2P systems.

Acknowledgments. We are grateful to anonymous reviewers for their insightful comments on the paper. This work is supported partly by the National Natural Science Foundation of China under Grant No.60403041.

References

1. Stephanos Androutsellis-Theotokis and Diomidis Spinellis. A Survey of Peer-to-peer Content Distribution Technologies. ACM Computing Surveys, 36(4):335–371, 2004.
2. S. Castano, A. Ferrara, S. Montanelli, G. Racca. Matching Techniques for Resource Discovery in Distributed Systems Using Heterogeneous Ontology Descriptions. In Proc. of ITCC'04, 2004.
3. FANNING, S. Napster. http://www.napster.com.
4. The Gnutella web site: http://gnutella.wego.com.

5. The Kazaa web site. http://www.kazaa.com.
6. Overnet. http://www.overnet.com.
7. Cohen, B. Incentives Build Robustness in Bittorrent. In Proc. of 1st Workshop on Economics of Peer-to-Peer Systems, 2003.
8. Mao Yang, Hua Chen, Ben Y. Zhao, Yafei Dai, Zheng Zhang. Deployment of a Large Scale Peer-to-peer Social Network. In Proc. of WORLDS'04, 2004
9. Hua Chen, Xiaoming Li, Jinqiang Han. Maze: a Social Peer-to-peer Networking. In Proc. of IEEE International Conference on E-Commerce Technology for Dynamic E-Business, 2004.
10. M. Cannataro and C. Comito. A Data Mining Ontology for Grid Programming. In Proc. of SemPGRID '03, 2003.
11. H. Tangmunarunkit, S. Decker, and C. Kesselman. Ontology-based Resource Matching in the Grid – the Grid Meets the Semantic web. In Proc. of SemPGRID '03, 2003.
12. Nejdl et al. EDUTELLA: a P2P Networking Infrastructure Based on RDF. In Proc. of WWW'02, 2002.
13. J. Broekstra et al. A Metadata Model for Semantics-based Peer-to-peer Systems. In Proc. of SemPGRID '03, 2003.
14. D. Calvanese, G. De Giacomo, M. Lenzerini, and R. Rosati. Logical Foundations of Peer-To-Peer Data Integration. In Proc. of PODS'04, 2004
15. R. Schollmeier. A Definition of Peer-to-Peer Networking for the Classification of Peer-to-Peer Architectures and Applications. In Proc. of P2P'01, 2001.
16. R. F. Simmons, Technologies for machine translation, FGCS, 2(2):83--94, 1986.
17. Charles H. Heenan. A Review of Academic Research on Information Retrieval. http://eil.stanford.edu/publications/charles_heenan/Academic Info Retrieval Research. pdf, 2002.
18. The Maze web site. http://maze.tianwang.com.
19. Lu Yan, Moisés Ferrer Serra, Guangcheng Niu, Xinrong Zhou, Kaisa Sere. SkyMin: A Massive Peer-to-Peer Storage System. In Proc. of GCC'04, 2004.
20. Alexander Löser, Wolf Siberski, Martin Wolpers, Wolfgang Nejdl. Information Integration in Schema-based Peer-To-Peer Networks. In Proc. of CaiSE'03, 2003.
21. Adriani, Mirna and Croft, W. Bruce. The Effectiveness of a Dictionary-Based Technique for Indonesian-English Cross-Language Text Retrieval. CLIR Technical Report IR-170, University of Massachusetts, Amherst, 1997.
22. Adriani, Mirna. Using Statistical Term Similarity for Sense Disambiguation in Cross-language Information Retrieval. Information Retrieval, 2(1): 67-78, 2000
23. Ballesteros, L., and Croft, W. Bruce. Resolving Ambiguity for Cross-language Retrieval. In Proc. of ACM SIGIR'98, 1998.
24. Adriani and C.J. van Rijsbergen, Term Similarity-Based Query Expansion for Cross-Language Information Retrieval. In Proc. of ECDL'99, 1999.
25. Adriani and C. J. van Rijsbergen. Improving Cross-Language Information Retrieval Performance Using Automatic Phrase Translation Technique. In Proc. of RIAO'00, 2000.
26. Wai Gen Yee, Ophir Frieder. On search in peer-to-peer file sharing systems. In Proc. of ACM SAC'05, 2005.
27. M. Harren, J. M. Hellerstein, R. Huebsch, B. T. Loo, S. Shenker, and I. Stoica. Complex Queries in DHT-Based Peer-to-Peer Networks. In Proc. of IPTPS'02, 2002.
28. Kementsietsidis, M. Arenas, and R. Miller. Mapping Data in Peer-to-Peer Systems: Semantics and Algorithmic Issues. In Proc. of SIGMOD'03, 2003.
29. R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval, Reading, MA. Addison-Wesley, 1999.
30. Tom Chothia, Konstantinos Chatzikokolakis. A Survey of Anonymous Peer-to-Peer File-Sharing. In Proc. of NCUS'05, 2005.