Knowledge Discovery in Peer-to-Peer Application Using a Data Warehousing Approach

Min-Feng Wang, Jui-Hwa Feng, and Meng-Feng Tsai
Department of Computer Science and Information Engineering
National Central University, Jhongli, Taiwan 32001
Tel: (886-3)422-7151 Ext 34500, Fax: (886-3)422-2681
E-mail: 945402023@cc.ncu.edu.tw

Abstract— **Recently, BitTorrent has emerged as a very popular and scalable peer-to-peer file distribution mechanism. It has been successful at distributing large file quickly and efficiently. With large number users of BitTorrent, there should be much information about the user data flowing through network. The rich information may imply the users' habitual behavior, data access pattern, interested file, and so on. The information is useful for general users, network managers, and designer, etc. Hence, we propose a framework on BitTorrent which combined with data warehouse and data mining techniques to offer an efficient and systematic analysis for users.**

Keywords: P2P, Data Mining, Data Warehouse

I. INTRODUCTION

BitTorrent is a protocol that enables fast downloading of large files using minimum Internet bandwidth. Unlike other download methods, BitTorrent maximizes transfer speed by gathering pieces of the file you want and downloading these pieces simultaneously from people who already have them [18]. For the high scalability, fault tolerance, and the load balance, BitTorrent become the most popular application based on the P2P paradigm. It has proved extremely popular according to the CacheLogic estimates that BitTorrent generated about 30% of all US Internet traffic and 53% of all P2P traffic in June 2004 [1] which grows on persistently.

Because of the popularity of BitTorrent, there are many analyses on it [13]. Previous works of analysis on BitTorrent are usually focused on the performance of the algorithms used in BitTorrent [5], [6]. Our study is focus on the BitTorrent users' characteristics; such as client's interesting files, network utilization in different time, etc. We collect and analyze users' data log which includes IP address, time, and some features to predict user's habitual behavior. Because of the log data size is large and much useless data needed to be filtered, it is impossible to investigate each record in details. We build the data warehouse and access data from the data cube can efficiently derive the needed information. Then we can use the data mining techniques to generalize association rules to predict the result for unknown data.

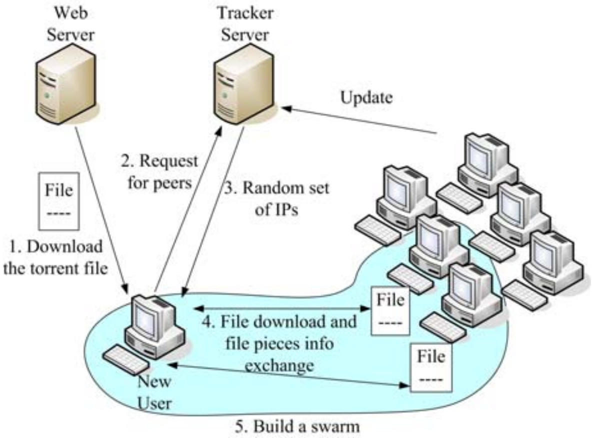

Fig. 1. Steps to join the BitTorrent network

We propose a user-based data analysis on BitTorrent. Exploited the data cube structure to access data efficiently and view the data from various aspects and to employ multidimensional analysis. Then, use the data mining techniques to mine some association rules for decision making.

II. BACKGROUND AND RELATED WORK

A. BitTorrent Overview

Bittorrent is a P2P application for sharing files that capitalizes the resources (access bandwidth and disk storage) of peer nodes to efficiently distribute large contents [5]. Figure 1 shows steps to join the BT network. The basic idea is to divide the file into equal-sized blocks (typically 32-256KB) and have nodes download the blocks from multiple peers concurrently. For avoiding a delay between blocks being sent, BitTorrent breaks blocks further into sub-blocks in the wire to enable pipelining of requests so as to mask the request-response latency [4]. A user that has not downloaded the entire file may have completely several blocks, which it can upload to other clients. This allows clients to share the workload even as they are still downloading.

B. Data Warehouse and OLAP

As coined by W. H. Inmon, the term “Data warehouse” refers to a “subject-oriented, integrated, time-variant and nonvolatile collection of data in support of management’s decision-making process” [7]. A data warehouse is a repository of integrated information, available for queries and analysis. Data and information are extracted from different sources. If users want to execute queries, the system only needs to search the data warehouse instead of the source databases. For this reason, it can save much more query processing time for users.

On-Line Analytical Processing (OLAP) technique has been proven to be one of the most popular tools for on-line, fast, and effective multidimensional data analysis [8], it can save much time because of massive amount of data stored in data warehouse. To provide fast and multidimensional analysis of data in a data warehouse, the OLAP tool precomputes aggregation over data and organizes the result as a *data cube* [14] composed of several dimensions, each representing one of the user analysis perspectives. The typical operations provided by OLAP include *roll-up, drill-down, slice,* and *dice* and *pivot* [9].

C. Data Model

Star schema is the most common and popular dimension model used in data warehouse proposed by Kimball [10]. A star schema consists of a fact table and several dimension tables. The fact table stores the bulk of the data which is a list of foreign keys corresponding to dimension table with no redundancy, and numeric measure of user interests. The dimension table contains a set of attributes. Furthermore, the attributes in a dimension table may form either a hierarchy or a lattice.

D. Association Rule Mining

The concept of association rule mining is to search interesting relationships among items in a given data set. For example, the supermarket information that customers who purchase diapers also tend to buy beers at the same time is represented in association rule below:

Diaper → Beer [sup = 2%, conf = 60%].

Rule *support* and *confidence* are two measures of rule interestingness. The support of 2% means that 2% of all customers purchase transactions contain both diaper and beer. The confidence of 60% means that 60% of the purchase transactions that contain diaper also contain beer. Typically, an association rule is considered interesting if it satisfies a minimum support threshold and a minimum confidence threshold that are set by users or domain experts. The most popular and influential mining algorithm is Apriori [9], and the concept of multi-dimensional association rules is first proposed by H. Zhu [12].

III. System Framework

Our system framework is shown in Figure 2, including the data collection from the Internet, generalize the data to a certain level of abstraction, integrate the collection data into data warehouse, and mining the association rules for user, systematically.

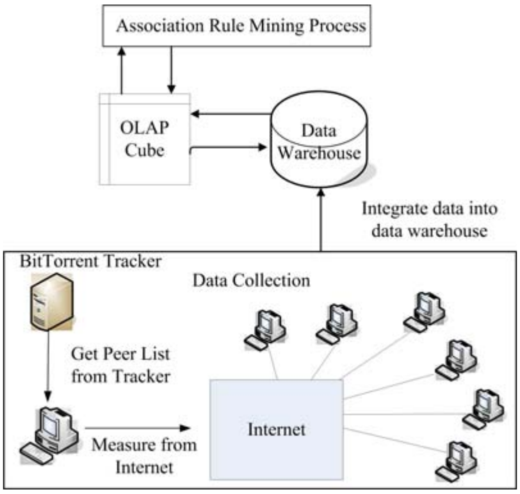
Association Rule Mining Process
OLAP Cube
Data Warehouse
Integrate data into data warehouse
BitTorrent Tracker
Data Collection
Get Peer List from Tracker
Measure from Internet
Internet

Fig. 2. System framework

A. IP Addresses and Locations

In our experiments, we choose *torrentportal.com* (a large websites which contains various torrent files) as our BiTorrent community. From torrentportal, we select 12 different and popular files then classified according to the file size and file type.

After deciding which files to measure, we conduct a *PeerGet* program to contact the tracker for each specific file and gathering the IP addresses of peers downloading the same file. The *PeerGet* program returns the IP list from the tracker can guarantee the hosts in the list are all BitTorrent users.

Since the IP address consists of 32 bits, the distinct IP addresses exceed hundreds of millions in the Internet. As well known, the IP addresses have the hierarchy structure which have the same prefix will classify into the same subnetwork. IP addresses in the same subnetwork usually have similar characteristic, e.g. round-trip time, location, and bandwidth. Same as the concept hierarchy of IP address, data warehouse also supports the hierarchical level. Therefore, we distribute the IP address with a hierarchical structure which group the IP addresses with identical first 24 bits for providing a more general and locality analysis of the data. For a more general view, we observe that there are almost 90% of IP addresses with the same first 16 bits come from the same country in our dataset. Thus, we further group them into a higher level.

We provide approximate information about geographical location of each peer. The geographical location can help to analysis data in a specific area or country. It’s helpful for system provider or designer to make some business decision or adjust for better performance. Since querying the web-based WHOIS clients, we can obtain the location information of each peer. We just collect the country of each peer. Here, we conduct a program called *LocationGet* to contact the

TABLE I
RAW DATASET

File Type	File Size	Time	Country	IP	Band-width	RTT
apps	1GB	3/19/22	DE	87.78.99.166	0.42	523
video	2GB	3/30/04	US	69.128.23.1	0.30	302
apps	3GB	4/11/17	CA	69.157.24.26	15.7	343
audio	5MB	3/23/10	JP	218.217.22.9	1.95	188
…	…	…	…	…	…	…

web-based WHOIS [3] clients and grab the information of country of each peer.

B. Round-trip time measurement

Round-trip time (RTT) is the time takes to send a packet to a remote host and receive a response; used to measure delay on a network at a given time. Besides that, RTT can provide distance information in terms of latency. Latency is the easiest distance metric to provide, and the most generally useful. Francis et al [15], propose a virtual topology use the Round-trip propagation and transmission delay as the distance between two hosts for server or peer selection. Thus, for the multi-function of RTT, we use it as the measure in our framework of data warehouse. For the users view point, we use RTT to offer a multidimensional analysis for peer selection. On the contrary, according to the change of RTT, a network manager can analyze a region or a node traffic trend.

Fping [2] is a ping like program which use the Internet Control Message Protocol (ICMP) echo request to determine if a host is up and it's round-trip time. Fping can ping many hosts in parallel, and for getting the trend of each peer's network load, we conducted Fping of each peer we get from *PeerGet* program, and last around one month and execute one time an hour.

Note that, Fping is use to determine if a host is up and it's RTT. So there are several situations we can't pingable the hosts:

1. Hosts are down
2. The host's network traffic is huge, the packet we sand may loss
3. Hosts behind the firewall or the measure packets are blocked by some anti-virus software.

We filter out the situations of above, and just record the pingable measurements as our data sources.

C. Bandwidth Measurement

The bandwidth between two hosts is the smallest capacity in the path. Knowledge of the capacity of a path can be put to good use in various scenarios. Using such information, designer can devise a more load balance system. Further, same as RTT, bandwidth can offer to users for peer selection [17]. But different from RTT, bandwidth does not interfere by the network traffic or change on the go. It has small variation, so we just need an accurate tool for bandwidth measurement. Thus, we use *CapProbe* as our capacity estimation tool which presented by Kapoor et al. [16].

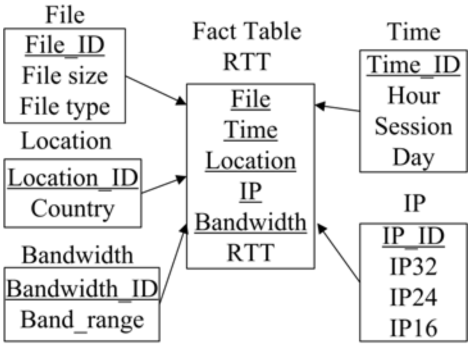

Fig. 3. Star schema of our system

We summarize the steps of data collection and the raw dataset is depicted in Table I:

1. We conducted the PeerGet program on 12 torrent files and last for more than one month to gather IP addresses interesting in the given file.
2. To use the LocationGet program to find the apply country of each IP address.
3. To execute the Fping and CapProbe periodically to measure the RTT and bandwidth of each IP address respectively. And filter out the data that cannot pingable.

D. Data Generalization and Data Transformation

After collecting data, for provide users more meaningful and understand information, we generalized the numerical attributes (i.e. RTT, Bandwidth) to a certain level of abstraction. There are two reasons for that:

1. Users do not understand the meaning behind the numerical data
2. For the mining step, it could not accumulate sufficient support to pass the minimum support threshold

For example, we generalize the numerical attribute "RTT" as following:

***good** (1 ~ 220), **average** (221 ~ 350), **mediocre** (351 ~ 750),*
***poor** (higher than 750)*

And the association rule between IP address and RTT as following:

IP (x, "140.115.50.1") => RTT (x, "good")

We transform the collected data into the data warehouse, the star schema shows in Figure 3. There are five dimension tables and a fact table with RTT as the measures. The IP, Time, and File dimensions have the hierarchy structure. For the Time dimension, we generalize higher levels of abstraction. We divided 24 hours into 4 Sessions: morning, noon, evening, and midnight, and 4 sessions a day for a general view of data.

E. Association Rule Mining Process

This is the final step in our framework to offer to users some general association rules for decision making.

TABLE II
NUMBER OF RULES OF EACH DATASET

	Distinct values of each dimension	**Support**	**Number of rules**
A (raw dataset)	(5, 149, 169479, 4)	550	50130
		600	39632
		650	18230
B (data cube)	(3, 149, 107026, 4)	1200	9481
		1400	5987
		1600	4430
C (data cube)	(3, 149, 7102, 4)	3000	7701
		4000	6076
		7000	3570
D (raw dataset)	(169479, 11, 4)	550	20458
		600	13939
		650	7788
E (data cube)	(107026, 11, 4)	1200	4448
		1400	2771
		1600	1977

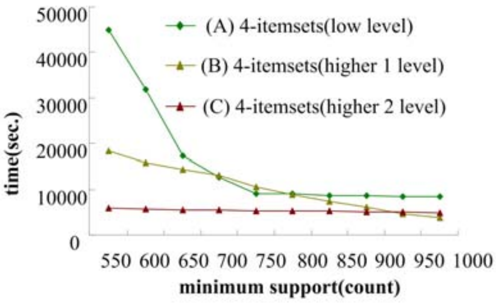

Fig. 4. Association rules mining time with 4 dimensions

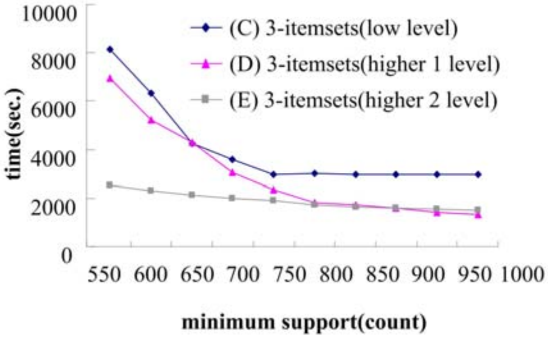

Fig. 5. Association rules mining time with 3 dimensions

Generally, discover association rules can be decomposed into two steps:

1. Find all frequent itemsets with support above minimum support.
2. Use the frequent itemsets to generate the desired rules.

We use the Apriori algorithm to mine association rules. The Apriori algorithm, first proposed by Agrawal and Srikant [11], is an interesting and influential algorithm. The main idea of this algorithm is based on the prior knowledge of frequent *k*-itemsets to generate candidate (*k*+1)-itemsets.

IV. EXPERIMENT RESULTS

We describe our experiment results and analyze the outcomes of the association rule mining. Through the clean and integrated of data, we still have more than 46 millions rows of data. Following we compare the execution time with different situation, firstly. Then analyze the results of association rule mining.

A. *Execution Time Comparison*

We compare the execution time that mining association rules with select from raw dataset and select from cube. We show that with different dimensions, minimum support, and level.

Figure 4 and Figure 5 show the execution time of association rules mining with different dimensions and minimum support. In Figure 4, the (A) dataset is extract from raw dataset that contains 4 dimensions and without any hierarchy structure. On the contrary, dataset (B) and (C) are same as dataset (A) which extract from data cube, and roll-up with 1 and 2 level respectively. We can obtain that the mining execution time of dataset from data cube (contains the build time of cube) is lower than dataset from raw data significantly. From Figure 5, significantly, we can easier obvious that mining from data cube can decrease execution time. Table II shows the rules number of each dataset in above figures.

Not surprised, the association rules from raw dataset are many times than rules from data cube. But these rules are too detail and complexity, it is hard for users to make use of these

TABLE III
TWO DIMENSIONS ASSOCIATION RULES FOR PEER SELECTION

Rules	Confidence (%)	Support (count)
IP ("210.24.99") => **RTT** ("good")	99.91%	1221
IP ("80.36.121") => **RTT** ("mediocre")	1.0	1307
Country ("TW") => **RTT** ("good")	91%	94458
Country ("KE") => **RTT** ("poor")	1.0	7831
• • •	• • •	• • •

TABLE IV
THREE DIMENSIONS ASSOCIATION RULES FOR PEER SELECTION

Rules	Conf.	Sup.
Country ("US"), **IP** ("206.169.170") => **RTT** ("good")	99.9%	2573
IP ("166.111.249") => **Bandwidth** ("8.0-10.0"), **RTT**("good")	99.6%	1912
IP ("202.33.24"), **Day** ("Saturday") => **RTT** ("good")	100%	819
• • •	• • •	• • •

rules. Generalize a certain level of abstraction can help users to analyze and use them easily. Following shows the difference between low level and high level rules and the components of high level rules from low level rules.

B. Analyze the Association Rules

In this section, we analyze the results of association rules the mining from the data cube. First, for the peer selection strategy, as previous chapter mentioned [15] use round-trip delay as the criterions to help peer or server selection. Following, we also take some rules that related to the round-trip time as examples:

Table III shows some rules that can help to select peers. We generalize the RTT to a certain level of abstraction with four distinct values. Therefore, when peer selection, we also classify peers into four levels that apply to the four distinct values. First, we choose the RTT with "good" level, and then choose the "average" and so on. Not only actively choose but also passively filter out the peers with RTT "poor". Further, we can consider multi-dimensional to help peer selection more accurately. Table IV displays some multi-dimensional rules:

For above rules, we present rules with more dimensions for peer selection or analysis. We consider more than one factor that increases the accuracy of peer selection strategy. Further, the IP dimension can roll-up higher level for more generalize information.

V. CONCLUSIONS AND FUTURE WORKS

We devise a data warehouse framework for collecting BitTorrent users' information. Because of multidimensional structure, we provide versatile point of views for various analysis tasks. Through data warehouse, the result of rules presents a more generalized abstraction. Compared to raw dataset mining, it is also more efficient to do association rules or other data mining tasks.

We collect data from Internet firstly, then transfer the collected data into a data warehouse, and exploited the OLAP and data mining techniques to mine some association rules for users. For example of peer selection, compare to the previous works on peer selection, we can provide a more generalized, efficient, and consider more than one factor of rules for users.

In our framework, we use the Apriori algorithm to implement the association rule mining task. If the dimensions or the distinct values of each dimension become complicated, the execution time of association rule mining still waste much time on scan the dataset. One of our future works is to conduct innovate algorithm to speed up the mining steps. Further, we can expand more dimensions to the star schema for analysis diversity. Like available bandwidth, it is variable at any time of each peer. We can add it to the measure attribute of data warehouse for peer select more accurately.

For the scalability, we hope to build our framework on tracker servers in the future. Because tracker servers are the centralized components, each host needs to connect them to join the BitTorrent network. Thus, from the view of trackers, we can get more meaningful rules that adapt to the whole BitTorrent users.

ACKNOWLEDGMENT

This work is sponsored by National Science Council, Taiwan, under grant NSC 96-EC-A-02-S1-029.

REFERENCES

[1] CacheLogic. The true picture of peer-to-peer filesharing. http://www.cachelogic.com/research, July 2004.

[2] Fping – a program to ping hosts in parallel. http://www.fping.com/.

[3] WHOIS - a TCP-based query/response protocol in order to determine the owner of a domain name, an IP address. http://en.wikiwedia.org/wiki/Whois

[4] B. Cohen, “Incentives build robustness in bittorrent,” in *Proceedings of the first Workshop on the Economics of Peer-to-Peer Systems,* Berkeley, USA, 2003.

[5] M. Izal, G. Urvoy-Keller, E. W. Biersack, P. Felber, A. Al Hamra, and L. Garc′es-Erice, “Dissecting bittorrent: Five months in a torrent’s lifetime,” in *Passive and Active Measurements,* Antibes Juan-les-Pins, France, April 2004.

[6] D. Qiu and R. Srikant, “Modeling and performance analysis of bittorrent-like peer-to-peer networks,” in *ACM SIGCOMM*, Portland, OR, USA, August 2004.

[7] W. H. Inmon and C. Kelly, *Rdb/VMS: Developing the Data Warehouse*, QED Publishing Group, Boston, Massachussetts, 1993.

[8] S. Chaudhuri and U. Dayal, “An overview of data warehouse and OLAP technology,” in *ACM SIGMOD Record*, Vol. 26, pp. 3 59-370, 1997.

[9] J. Han and M. Kamber, *Data mining: Concepts and Techniques*, MORGAN KAUFMANN PUBLISHERS, 2000.

[10] R. Kimball, *The Data Warehouse Toolkit Practical For Building Dimensional Data Warehouses*, JOHN WILEY & SONS, INC. 1996.

[11] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules,” in *Proceedings of the 20th VLDB Conference*, pp. 487-499, 1994.

[12] H. Zhu, *On-Line Analytical Mining of Association Rules*, SIMON FRASER UNIVERSITY, December, 1998.

[13] J. Pouwelse, P. Garbacki, D. Epema, and H. Sips, “The Bittorrent P2P File-sharing System: Measurements and Analysis,” in *4th International Workshop on Peer-to-Peer Systems (IPTPS’05)*, Feb. 2005.

[14] J. Gray, A. Bosworth, A. Layman, and H. Pirahesh, “Data cube: a relation aggregation operator generalizing group-by, cross-tabs and subtotals,” in *Proceedings of International Conference on Data Engineering*, pp. 152-159, 1996.

[15] P. Francis, S. Jamin, V. Paxson, L. Zhang, D. F. Gryniewicz, and Y. Jin, “An architecture for a global Internet host distance estimation service,” in *Proceedings of IEEE INFOCOM*, New York, NY, Mar. 1999.

[16] R. Kapoor, L. Chen, L. Lao, M. Gerla, and M. Sanadidi, “CapProbe: a simple and accurate capacity estimation technique,” in *Proceedings of the 2004 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications,* pp. 67-78, 2004.

[17] T. S. E. Ng, Y. hua Chu, S. G. Rao, K. Sripanidkulchai, and H. Zhang, “Measurement-based optimization techniques for bandwidth-demanding peer-to-peer systems,” in *Proceedings of IEEE INFOCOM*, April 2003.

[18] http://www.wsbtv.com/technology/4390621/detail.html