Analysis of Time-Dependent Query Trends in P2P File Sharing Systems

Masato Doi, Shingo Ata, and Ikuo Oka

Graduate School of Engineering, Osaka City University
3–3–138 Sugimoto, Sumiyoshi-ku, Osaka 558–8585, Japan
{doi@n.,ata@,oka@}info.eng.osaka-cu.ac.jp

Abstract. In P2P file sharing systems, time dependent characteristics of query (query trends) for a file become much important to forecast the demand of the file in future. Prediction of future demand would be effective for the efficient use of the caching mechanism, however, the accurate prediction of query trend is difficult because patterns of query trends may differ significantly according to a nature of keyword used in the query. Identification of query pattern is one of important roles for accurate forecast of future query demand. In this paper, we propose a new method to classify measured query trends into some typical query patterns. We first measure query trend for each keyword in the most famous P2P file sharing system in Japan, and analyze the pattern of query trends by using clustering technique with Discrete Fourier Transform. We then apply our method to the measurement results and show that most of keywords can be categorized into one of four typical trend patterns.

1 Introduction

In recent years, Peer-to-Peer (P2P) file sharing system attracts attention and is used by many users. In P2P file sharing system, since each node works as the server and client, they can share and exchange information directly and equally. Therefore P2P file-sharing system is superior to client/server model in scalability and fault tolerance.

Now, there are many P2P file sharing systems (e.g., Napster, Gnutella, KaZaa, Cabos, Limewire, BitTorrent), and Winny and Share are the most popular file sharing systems in Japan.

In these P2P file sharing system, many of performance improvements such as search efficiency, load balancing, and anonymization is achieved by the use of caching mechanism. Since disk spaces for caching are limited, cache replacement algorithm is important for the efficient use of cached resource. As typical cache replacement algorithms are Least Recently Used (LRU), Least Frequently Used (LFU) and First In First Out (FIFO), and LRU is adopted in many file sharing system.

These cache replacement algorithms decide caches duration by referred time and the referred frequency, therefore caches of keywords that are accessed same

C.S. Hong et al. (Eds.): APNOMS 2009, LNCS 5787, pp. 181–190, 2009.

time or same frequency are processed as same condition regardless of the kind of contents.

However, cache replacement by checking the last accessed time is not always the best way to achieve the efficiency because the time dependent variance of query demand (we refer as *query trend* in this paper) may differ completely according to the type of keyword used in query. For example, a keyword related to an event is requested significantly around the time of the event, while in other periods there are few requests for the keyword. On the other hand, there are some requests at all time for an FAQ-like keyword. If we apply the LRU to event-related keywords, many of cache files has been created according to the recent requests of the keyword. However, after the time that the event has passed, the request for the keyword will decrease rapidly, and many of cached files will not be used which waste disk resources of peers. In order to solve this problem, it is necessary to introduce a new cache replacement algorithm which takes query trend of keyword into consideration.

Time-dependent trends of queries can be classified into some patterns. If we know which trend a keyword belongs to, we can predict how it will be requested. In this paper, we analyze time-dependent trends of queries in P2P file sharing system and classify trend patterns by clustering. Specifically, we collect keyword (filename) queries by crawling the P2P file sharing system. We then classify trend patterns by cluster analysis and analyze the nature of each cluster.

This paper is organized as follows: Section 2 introduces related works of popularity distribution and transition in some network applications. Section 3 describes the method of collecting search queries in P2P file sharing system which is most used in Japan now. In Section 4, we describe our proposed method of classifying query trends into typical patterns based on similarities of trends. We then present the experimental results of our analysis by applying measured query trends in section 5. Finally conclude our paper with future research topics in Section 6.

2 Related Work

In this section, we describe related works about popularity in web server and video-sharing site.

In [1], Padmanabhan et al investigated popularity distribution of contents and time-series transition of popularity based on data measured in web servers. As the result, [1] showed that the distribution of popularity of accessed contents roughly follows power-law as well as shown in other studies. [1] also analyzed time-series transition of popularity of top 100 contents and showed that 60-70% of contents still exists in top 100 contents after 5 days. However, [1] only considered popularity in web servers as a whole, and how changes popularity of individual content changes was not analyzed.

[2] studied User Generated Content (UGC) posted on Video-on-Demand (VoD) system such as YouTube. As a result of investigating popularity distribution of video, [2] showed popularity distribution almost follows power-law and

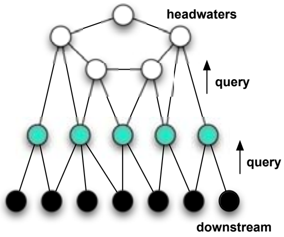


Fig. 1. Typical network topology of measured P2P file sharing system

fits best by power-law with an exponential cutoff. However, although [2] only analyzed the relation between days after video is posted and the number of accesses, [2] did not analyze the access patterns of individual content.

[3] analyzed transition of contents popularity in VoD system. In [3], patterns of transition of contents popularity are only two type (outbreak type, persistent type), however, they are insufficient to represent actual patterns of transition of contents popularity.

There are many studies about a cache replacement algorithm in P2P file-sharing system [4, 5, 6]. However, they did not take into consideration trend patterns.

As above, there are few studies about patterns of time-series transition of contents popularity, even if it exists, it performed only observation and speculation. Also there are no studies about trend patterns in P2P file sharing system. If we can classify patterns of query trends and identify which pattern a query trend for a keyword belongs to, we can predict demand of queries for the keyword in future. Predicting future requests enables designing new cache replacement algorithm that takes trend patterns into account for achieving more efficient cache utilization.

3 Measurement of Search Queries

In this section, we describe the method to measure search queries in P2P file sharing system which is most used in Japan. In Pure P2P network, since there is no management server, it is impossible to collect all search queries. We therefore consider the method of collecting queries as much as possible. In this paper, we develop a crawler program for collecting search queries.

3.1 Collecting Search Queries by Winny Crawler

In this paper, we focus on a P2P file sharing system which is mostly used in Japan. The topology is formed by kind of hierarchy according to line speeds of peers that are specified by users. Fig. 1 shows a typical network topology in the P2P system. Search queries are transmitted from peers with slow line to peers with high-speed line. Therefore, it has a nature that queries are gathered in peers with high-speed line. If we connect our peer by specifying higher link speed, our peer would be connected to the upper layer of the hierarchy and the peer can collect more keywords. Furthermore, since keyword of a query is encrypted by RC4, we develop a crawler which can decode encrypted queries.

Our crawler has node list and query list. Node list is the list of nodes available by monitoring advertisement messages from other nodes. When the crawler is needed to connect with another nodes, the crawler selects a node randomly from the node list. Query list handles information of search queries transmitted by other nodes. It includes identifier of query message, IP addresses of the node that generated the query and keyword exist. If the crawler receives same queries multiple times, the crawler does not discard queries with the same ID. At the analyzing phase, however we consider queries with the same ID to be single query. We describe the outline procedure of operation of procedure crawler as follows.

Outline of operation of Winny crawler

1. Create an initial node list which includes a set of nodes that the crawler firstly connects.
2. Connect to nodes listed in the node list by specifying that the node has high-speed link so that the node can be placed at the higher layer of the hierarchy.
3. - If the crawler receives the search query message, the crawler adds the query message into the query list.
   - If the crawler receives the advertisement node from connected nodes, the crawler adds information of the node in advertisement message into the node list.
   - If the crawler receives the request of connection termination or unknown message from the connected node, the crawler terminates the connection to the node.

We corrected search queries by our crawler from April 1, 2009 to April 31, 2009. In the period, the crawler always connects about 500 nodes.

4 Method of Classifying Trend Patterns

In this section, we describe the method of classifying trend patterns of search queries. We define trend patterns as time-series of number of queries per day. We classify trend patterns of each keyword by cluster analysis.

4.1 Similarity of Time-Series Trends

To apply the clustering method, it is necessary to define distance between time-series transit of query. There are two methods which calculate in time domain in frequency domain respectively. In time domain calculation, distance between time-series transition is defined as Euclidean distance between vectors by considering time-series of transition as a vector.

However, Euclidean distance between vectors is not applicable directly because two time transitions are not completely synchronized, i.e., we have to slide the sequence of the trend to adjust the start of transitions to the same time.

In [7,8,9], DFT/DWT is used to calculate distance between time-series transition from similarity in frequency domain. However, if we use DWT the synchronization problem is still remained, and we therefore use the DFT for calculation of the similarity.

4.2 Similarity Derivation Based on Discrete Fourier Transform

DFT is Fourier Transform in discrete group and transformation from time domain to frequency domain. DFT of the time-series $x = [\mathbf{x_t}] = [x_0, x_1, \ldots, x_{n-1}]$ is represented by in the form of $f = [\mathbf{f_j}] = [f_0, f_1, \ldots, f_{n-1}]$, and is defined by

$$f_j = \sum_{k=0}^{n-1} x_k e^{-\frac{2\pi i}{n} jk} \quad j = 0, 1, \ldots, n-1. \tag{1}$$

Note that i is the imaginary number ($i^2 = -1$) and DFT has symmetrical property like $f_j = f_{n-j}$. When the demention n of x_t is power-of-two, Fast Fourier Transform (FFT) can be applied to calculate in $O(nlogn)$.

Since our purpose is to categorize the *trend patterns*, absolute values on query trends (e.g., number of queries/hour) is not our focus. We therefore apply DFT to values normalized by maximum number of queries/hour appeared in the trend.

4.3 Distance between Time-Series Transition

After DFT/FFT, two time-series transitions p and q are transformed to $f(p)$ and $f(q)$ respectively. The distance between two time-series transitions can be considered as Euclidean distance of the coefficient sequence of $f(p)$ and $f(q)$ which is calculated from

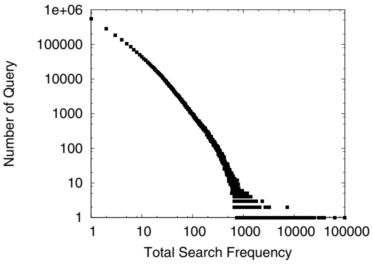
$$Distance(p, q) = \sqrt{\sum_{j=1}^{\frac{2}{n}} (|f_j(p)| - |f_j(q)|)^2}. \tag{2}$$

4.4 Clustering Based on Similarity

We perform hierarchical clustering by distance obtained by Eqn. (2). We employ Ward's clustering method [10]. Ward's clustering method achieves to minimize the sum of squares of two arbitrary clusters. Distance between cluster t and cluster r is calculated by

$$S_{tr} = \frac{n_p + n_r}{n_t + n_r} S_{pr} + \frac{n_q + n_r}{n_t + n_r} S_{qr} - \frac{n_r}{n_t + n_r} S_{pq}, \tag{3}$$

t is the cluser obtained by combining p and q, and n_t, n_p, n_q, n_r are the element count in t, p, q, r respectively.



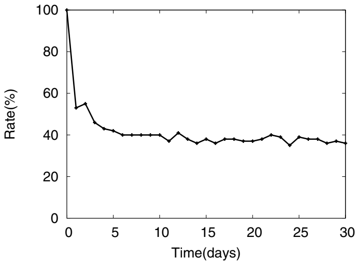


Fig. 2. Distribution of search frequency per keyword in Winny

Fig. 3. Change of keyword ranking

5 Measurement and Analytic Results

5.1 Statistical Characteristics of Measured Query Messages

We run our developed crawler for one month (from 2009/4/1 to 2009/4/31). After the measurement, we obtain 60,531,074 query messages having 3,899,980 keywords in total. The total number of unique nodes is 156,708. Fig. 2 shows total frequency of search distribution of each query in measurement period. In this figure, the horizontal axis represents the total search frequency, and the vertical axis represents the number of keyword corresponding the total search frequency. [11] presented that popularity distribution of contents follows power law. Distribution of search frequency per keyword in file sharing system also almost follows the nature of power law.

We also investigate how popularity changes in file-sharing system. We analyze how most frequent keywords vary by days. For this purpose, we first obtain top 100 most frequent keywords at the first day. We then obtain the ratio of keywords that remain top 100 keywords at the second day or more. Fig. 3 shows rate of remaining keywords in top 100 with progress of time. In this figure, the horizontal axis represents time (day), and vertical axis represents the rate of remaining keywords in top 100. We can observe that about 30% of keywords are still remained in top 100 after 30 days. Additionally, since about a half of keywords replaced rapidly (within 5 days).

5.2 Result of Classifying Trend Patterns Using Clustering

We perform clustering with 753 keywords with that have over 1,000 queries. We investigate transition of the number of queries every one hour. Then, in order to investigate rough transitions, we smooth transitions of the number of queries in 24 hours. We apply the moving average of the number of queries instead of the actual number of queries. The moving average can be obtained by

$$q_i = \frac{m_i + m_{i-1} + \ldots + m_{i-23}}{24}, \tag{4}$$

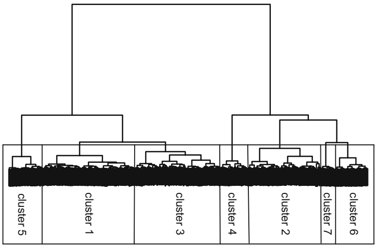


Fig. 4. Dendrogram of clustering

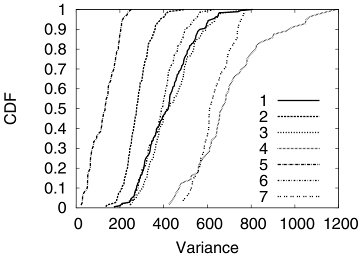


Fig. 5. Variance of the number of searches for each keyword

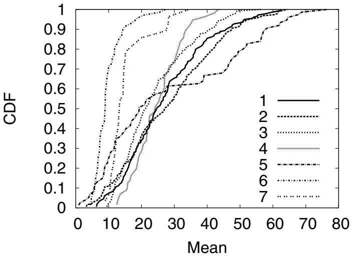


Fig. 6. Mean of the number of searches for each keyword

Where q_i is the i-th average of queries and m_i is the number of queries measured at time i. Additionally, we normalize transitions of the number of queries by the maximum number of queries.

We show dendrogram of clustering analysis in Fig. 4. We obtain 7 clusters in total shown in Fig. 4. In followings we analyze nature of each clusters.

Fig. 5 shows variance of the number of queries of each keyword by the cumulative distribution of each cluster. We can observe that the variance of cluster 5 is smaller than other clusters from Fig. 5. Therefore, the number of queries of keywords in cluster 5 keeps almost average value and it is found keywords in cluster 5 are searched constantly.

Fig. 6 shows mean of the number of queries of each keyword by the cumulative distribution of each cluster. This figure shows that mean for Cluster 6 is smaller than other clusters from fig. 6. In this paper, since we normalize transitions of the number of queries at maximum, small mean is caused by the keywords that are searched intensively in short time. Therefore, the smaller the mean is, the more the pattern of transition becomes sharpen. Hence, keywords in cluster 6 has the trend pattern that both increase and decrease of the number of queries are quite rapid.

Fig. 7 shows the duration from the time when the query rate exceeds 5% of the maximum and the time when the query rate reaches the maximum. The cumulative distributions of the duration are shown in these figures. Fig. 8 shows

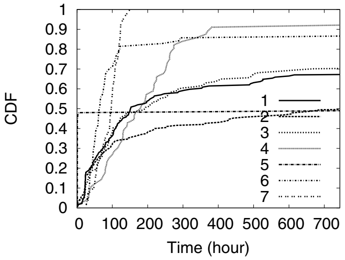


Fig. 7. Time concerning increasing to maximum, after exceeding 5% of maximum

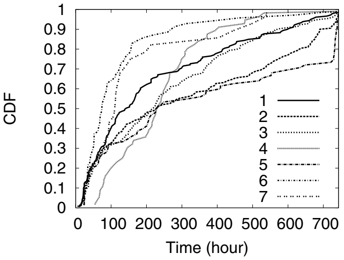


Fig. 8. Time concerning decreasing to maximum to 5% of maximum

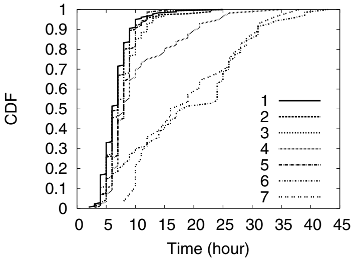


Fig. 9. Maximum continual increase time

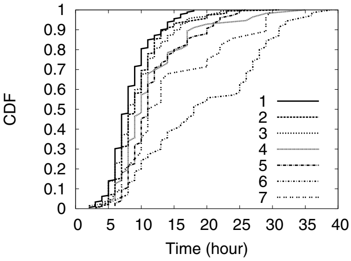


Fig. 10. Maximum continual decrease time

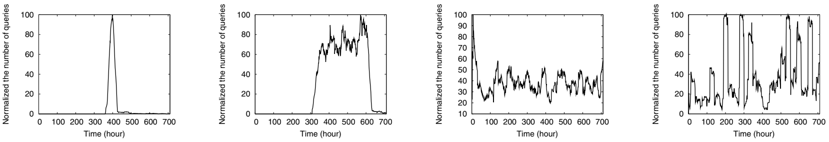
the results of opposite side, i.e., the distribution of duration from the maximum to the 5% of the maximum. From these figures, Since Cluster 6 takes small value, Cluster 6 has the trend pattern that the number of searches increases rapidly and decreases rapidly. Additionally, Cluster 7 takes small value next to Cluster 6. Cluster 7 also have the trend pattern that the number of searches increases rapidly and decreases rapidly. On the other hand, Cluster 4 takes decent small mean and time, Cluster 4 has the trend pattern that the number of searches increases gradually and decreases gradually.

Fig. 9 and Fig. 10 show CDFs of the maximum duration time of increase of the number of queries for each keyword and that of decrease, respectively. From these figure, we can observe that the duration time of increase and decrease in Cluster 1, Cluster 2 and Cluster 3 have small values. Therefore, we consider that such clusters have the trend pattern that the number of queries varies drastically.

Thereby, it is found that trend patterns can be classified into four types. *Instant* and *temporal* have the nature that accesses concentrate on a specific period and *assurance* and *fickleness* have the nature that accesses happen constantly. Furthermore, in the trend pattern that accesses concentrate on a specific period, both time taken by increasing to maximum and decreasing 5% of maximum are 22 hours at the shortest. While the trend pattern that the

Table 1. Typical trend pattern

trend pattern	cluster number
instant	6, 7
temporal	4
assurance	5
fickleness	1, 2, 3



(a) instant (b) temporal (c) assurance (d) fickleness

Fig. 11. Typical Trend Patterns

number of searches increases and decreases rapidly exits, the trend pattern that both time taken by increasing to maximum and time taken by decreasing to 5% of maximum take over 500 hours also exists. Additionally, in the trend pattern that the number of searches keeps constantly, while the trend pattern that has small variance of the number of searches and that is constantly searched exists, the trend pattern that has large variance of the number of searches and that the number of searches intensely changes also exists. Table. 1 shows aspect of each cluster and Fig. 11 shows typical trend patterns of each aspect.

6 Conclusion and Future Works

In this paper, we have considered search queries as index of popularity in P2P file sharing system and have classified trend patterns of search keywords from the numerical feature. First, we have collected search queries in P2P file sharing system and have collected data concerning time-series transitions of search keyword. We have then analyzed total number of queries distribution of each keyword. In the result, we have shown it has almost follows the nature of power law as reported in earlier studies. Then we have classified trend patterns by hierarchical clustering and have analyzed each cluster. In the result, we have shown trend patterns can be categorized into one of four typical types.

As future research topics, we need to analyze trend patterns of search keywords in other P2P file sharing systems. Moreover, for suggestion of new cache replacement algorithm, we need to consider the method to identify which trend a keyword belongs to and the optimal cache replacement algorithms for each pattern.

Acknowledgement

This work was partially supported by a Grant-in-Aid for Young Scientists (A) (No. 19680004) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan.

References

1. Padmanabhan, V.N., Qiu, L.: The content and access dynamics of a busy web site: Finding and implications. In: Proceedings of ACM SIGCOMM 2000, Stockholm, Sweden, August 2000, pp. 111–123 (2000)
2. Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.-Y., Moon, S.: I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system. In: Proceedings of ACM IMC 2007, San Diego, CA, October 2007, pp. 1–14 (2007)
3. Mori, T., Asaka, T., Takahashi, T.: Analysis of popularity dynamics in youtube-like vod. In: IEICE Technical Report (NS2008-103), Japan, November 2008, vol. 108, pp. 87–92 (2008)
4. Cohen, E., Shenker, S.: Replication strategies in unstructured peer-to-peer networks. In: Proceedings of ACM SIGCOMM 2002, Pittsburgh, PA, August 2002, pp. 177–190 (2002)
5. Dunn, R.J., Gribble, S.D., Levy, H.M., Zahorjan, J.: The importance of history in a media delivery system. In: Proceedings of IPTPS 2007, Bellevue, WA (February 2007)
6. Obayashi, N., Asaka, T., Takahashi, T., Sakaki, J., Shinagawa, N.: Load balancing with content classification in P2P networks. IEICE Transactions on Communications J90-B, 720–733 (2007)
7. Wu, Y.-L., Agrawal, D., Abbadi, A.E.: A comparison of DFTand DWT based similarity search in time-series databases. In: Proceedings of ICKM 2000, New York, pp. 488–495 (2000)
8. Vlachos, M., Meek, C., Vagena, Z.: Identifying similarities, periodicities and bursts for online search queries. In: Proceedings of SIGMOD, June 2004, pp. 131–142 (2004)
9. Kontaki, M., Papadopoulos, A.: Efficient similarity search in streaming time seqences. In: Proceedings of SSDBM, Santorini Island, GR, June 2004, pp. 63–72 (2004)
10. Ward, J.H.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58(301), 236–244 (1963)
11. Gummadi, K., Dunn, R., Saroiu, S., Gribble, S., Levy, H., Zahorjan, J.: Measurement, modeling, and analysis of a peer-to-peer file-sharing workload. In: Proceedings of SOSP 2003, Bolton Landing, NY (October 2003)