Systematic Support for Content Bundling in BitTorrent Swarming

Jinyoung Han, Taejoong Chung, Hyunchul Kim, Ted "Taekyoung" Kwon, Yanghee Choi
School of Computer Science and Engineering
Seoul National University, Seoul, Korea
Email: {jyhan, tjchung, hkim}@mmlab.snu.ac.kr, {tkkwon, yhchoi}@snu.ac.kr

I. Introduction

BitTorrent [3] has achieved a huge success, being estimated to account for 27-55% of today's Internet traffic [1]. The ever increasing usage of BitTorrent is due to some attractive properties of its swarming systems. First, cooperation among peers in a swarm results from the tit-for-tat based incentive mechanism to improve the overall system performance in terms of throughput. Second, the tit-for-tat strategy also prevents the free-riding problem. Third, the swarming technique scales well even in the presence of massive flash crowds for popular contents.

Despite the strengths of BitTorrent, its swarming system suffers from a fundamental limitation: little or no availability of contents often. That is, peers arriving after the initial flash crowd may end up with finding the content unavailable, not to mention unpopular contents [5]. Recently, Menasche *et al.* [5] showed that *bundling* [1] is one solution to mitigate this availability problem; it improves the availability and reduces download times for unpopular contents by combining multiple files into a single swarm.

This paper is motivated by our conjecture: "So far, content bundling in BitTorrent has been done manually and in an ad hoc manner decided by publishers. If bundling is supported systematically in an automatic and efficient fashion, we believe that the system may enhance the availability and download speed."

As a first step towards developing such a systematic content bundling scheme in swarming systems, we make the following contributions: (1) this is the first work that raises and explores the possibilities and benefits of systematic content bundling in swarming systems. (2) we evaluate three systematic bundling algorithms (*cosine*, *Levenshtein distance*, and *matching coefficient distance* [4]) based on the content similarity and find that the *cosine clustering* algorithm outperforms the other two in terms of bundling accuracy and efficiency. (3) we find that the cosine bundling algorithm clusters 60% of all the contents to generate bundles without any manual intervention of publishers, with $> 98\%$ accuracy. (4) we find that the systematic support for bundling increases the number of content files in a swarm, which improves content availability. (5) we observe that movies and TV shows (i.e., video files) are more "clusterable" than other content type, which implies more performance gain.

[1] Bundling is a common strategy adopted by BitTorrent publishers by which a publisher packages a number of related files (e.g., episodes of a sitcom) and disseminates them via a single larger swarm [5], instead of disseminating individual files via separate swarms.

II. Similarity-based Content Bundling

A. Methodologies

To bundle the same or highly correlated contents, we propose to use a few criteria to estimate the content similarity to cluster [2] contents. Once the similarity of contents is estimated by one of the criteria, multiple clusters (or bundles) are formed based on the degree of the content similarity. Among multiple candidate criteria to estimate the content similarity such as content size, category (e.g., movie, TV show or game), and content hash, we choose the torrent title because the torrent title is exactly what BitTorrent users use to search, download, and upload contents. To calculate the similarity of torrent titles, we adopt three popular text classification algorithms: 1) cosine, 2) Levenshtein distance, and 3) matching coefficient distance. [3]

We collected the data log of BitTorrent swarms from Torrentz [2] starting at 10 PM (GMT+9, Korea Standard Time), Nov 24, 2009 for 15 minutes. Each data of the 54,115 torrents includes torrent title, category, size, torrent creation time, and the number of seeds and leechers.

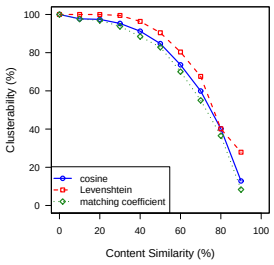
B. Results

We first measure the number of clusters and the clusterability, which indicates the ratio of the number of clustered files to the total number of content files, when we apply the three algorithms. Figures 1a and 1b show the clusterability and the number of clusters as we vary the threshold of clustering (or content similarity) from 10% to 90% for each algorithm. Though 60% or 70% of content similarity is a tight condition, the clusterability indices of the three algorithms are plotted around 60%, which suggests that contents are highly clusterable into bundles.

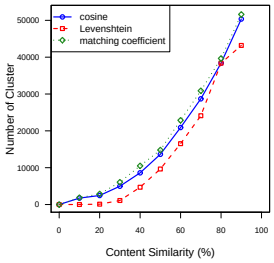
To measure the clustering accuracy, we investigate top 20 clusters in terms of the number of files in a bundle for each algorithm. We manually check these clusters whether individual content files are bundled into appropriate clusters. Figure 1c shows the cluster accuracy for each algorithm is perfect when content similarity is over 80%. The cosine

[2] In this paper, cluster/clustering and bundle/bundling are used interchangeably.

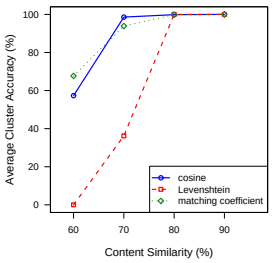
[3] Given space limit, readers refer to [4] for details.



(a) Clusterability



(b) Number of cluster



(c) Average cluster accuracy

Fig. 1: Clusterability and number of clusters with accuracy check. When we apply similarity-based clustering algorithms, content files are highly clusterable with high accuracy.

algorithm outperforms the other two in terms of clustering accuracy and efficiency as shown in Figure 1c and Table I.

TABLE I: Computation time of each algorithm

Content Similarity	cosine	Levenshtein	matching coefficient
70%	60 min	203 min	132 min
80%	115 min	256 min	204 min
90%	166 min	510 min	268 min

We also measure the average number of content holders (or seeds) those who keep at least a full copy of a content file in a swarm when one or more content files are bundled in a swarm. As shown in Figure 2, compared to the original (non-clustered) BitTorrent, there are more content holders in a swarm, which implies that the systematic bundling scheme can improve content availability. Besides, those content holders in a swarm may contain highly correlated contents possibly in different swarms, and thus the proposed systematic support can further improve availability. For example, even if there are no seeds for full episodes of the popular movie "Star Wars", there may be seeds for each episode of "Star Wars". In this case, if a bundle for "Star Wars" is systematically formed (not manually), the availability will be enhanced.

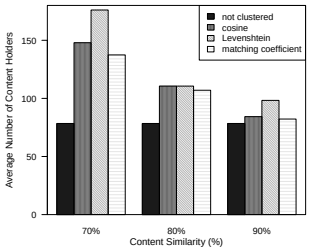
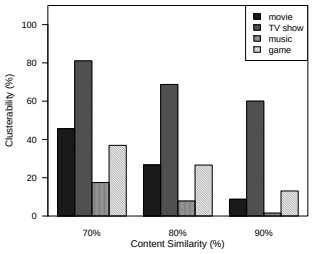
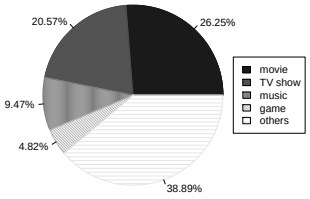


Fig. 2: Average number of content holders in a swarm

Finally, we investigate the clusterability and the number of clusters based on the content type. We categorize all the contents into movie, TV show, music, game, and others. We apply the cosine algorithm and vary the content similarity from 70% to 90% in Figure 3. As shown in Figures 3a and 3b, contents of movies, TV shows and games are more clusterable than music. This is because most of song titles in the same album are different.



(a) Clusterability based on content type



(b) Pie of the number of cluster for similarity 80%

Fig. 3: Clusterability and the number of clusters of the 5 types of contents.

III. Concluding Remarks

Content availability is a serious problem in today's peer-to-peer swarming systems. The proposed systematic support for bundling contents can enhance the availability and download performance with little overhead. To the best of our knowledge, this is the first work that explores the possibilities and benefits of title-based content bundling. We will study more efficient way of content delivery to enhance the bundling performance by constructing partial multicast trees among peers.

Acknowledgment

This work was supported by NAP of Korea Research Council of Fundamental Science and Technology and the ITRC support program [NIPA-2010-C1090-1011-0004] of MKE/NIPA. The ICT at Seoul National University provided research facilities for this study.

References

[1] The impact of p2p file sharing, voice over ip, instant messaging, one-click hosting and media streaming on the internet. http://www.ipoque.com/resources/internet-studies/internet-study-2008_2009.

[2] Torrent search engine. http://www.torrentz.com/.

[3] B. Cohen. Incentives build robustness in bittorrent. In *1st Workshop on Economics of Peer-to-Peer Systems*, 2003.

[4] M. Kabir. Similarity matching techniques for fault diagnosis in automotive infotainment electronics. *CoRR*, abs/0909.2375, 2009.

[5] D. S. Menasche, A. A. Rocha, B. Li, D. Towsley, and A. Venkataramani. Content availability and bundling in swarming systems. In *ACM CoNEXT*, 2009.