Reliability Verification of Search Engines' Hit Counts: How to Select a Reliable Hit Count for a Query

Takuya Funahashi and Hayato Yamana

Computer Science and Engineering Div., Waseda University,
3-4-1 Okubo, Shinjuku-ku, Tokyo, 169-8555 Japan
{takya,yamana}@yama.info.waseda.ac.jp

Abstract. In this paper, we investigate the trustworthiness of search engines' hit counts, numbers returned as search result counts. Since many studies adopt search engines' hit counts to estimate the popularity of input queries, the reliability of hit counts is indispensable for archiving trustworthy studies. However, hit counts are unreliable because they change, when a user clicks the "Search" button more than once or clicks the "Next" button on the search results page, or when a user queries the same term on separate days. In this paper, we analyze the characteristics of hit count transition by gathering various types of hit counts over two months by using 10,000 queries. The results of our study show that the hit counts with the largest search offset just before search engines adjust their hit counts are the most reliable. Moreover, hit counts are the most reliable when they are consistent over approximately a week.

Keywords: search engine, information retrieval, hit count, reliability, trustworthiness.

1 Introduction

Nowadays, many studies adopt search engines' hit counts to estimate the popularity of queried terms. Here, search engine hit count is defined as the number returned by a search engine as the search result count. Hit counts are widely used in fields such as machine translation research [1], word similarity measurements [2], and word clustering research [3]. For example, "Honto? Search" system [4] was proposed to help the user to determine trustworthiness of a statement that he or she was unconfident about, such as whether "the Japanese Prime Minister was Junichiro Koizumi" is true or false. The system compares the hit count of a statement with the hit count of its alternative statement to determine it. Thus, unreliable hit counts result in unreliable knowledge extraction. The use of hit counts in such a diverse set of fields claims that they be accurate and reliable.

Although hit counts are used for many studies, hit counts are unreliable because they "dance," i.e., change, in some situations. To the best of our knowledge, hit counts dance when a user clicks the "Search" button multiple times or clicks the "Next" button on the search results page, or when a user queries the same term on separate days. Previous researches have attempted to demonstrate the characteristics of hit count transition [5][6][7]; however, these researches do not clarify the reliability

F. Daniel and F.M. Facca (Eds.): ICWE 2010 Workshops, LNCS 6385, pp. 114–125, 2010.

of the hit counts, i.e., how to select the most reliable hit count for a query from among its variation.

The purpose of this paper is to investigate the trustworthiness of search engines' hit counts to adopt them with various types of studies, while we do not know the real hit counts. First, we demonstrate how hit counts dance by classifying the "hit count dance" into three cases: 1) when we click the "Search" button multiple times in a short time interval, 2) when we click the "Next" button multiple times in a short time interval, and 3) when we click the "Search" button on separate days. Fig. 1 shows these three cases. Second, we propose a new scheme to provide a basis to adopt hit counts on the basis of the observed hit count transition. In the experiment, we verify hit counts from three major search engines, Google, Yahoo!, and Bing, by using their "search APIs." Though we investigate three major search engines, the difference of hit counts among search engines is out of scope of this paper.

The rest of this paper is organized as follows: In Section 2, we review related work. Section 3 describes a process to verify hit counts from the viewpoint of providing a basis to adopt hit counts. Finally, the paper is concluded in Section 4.

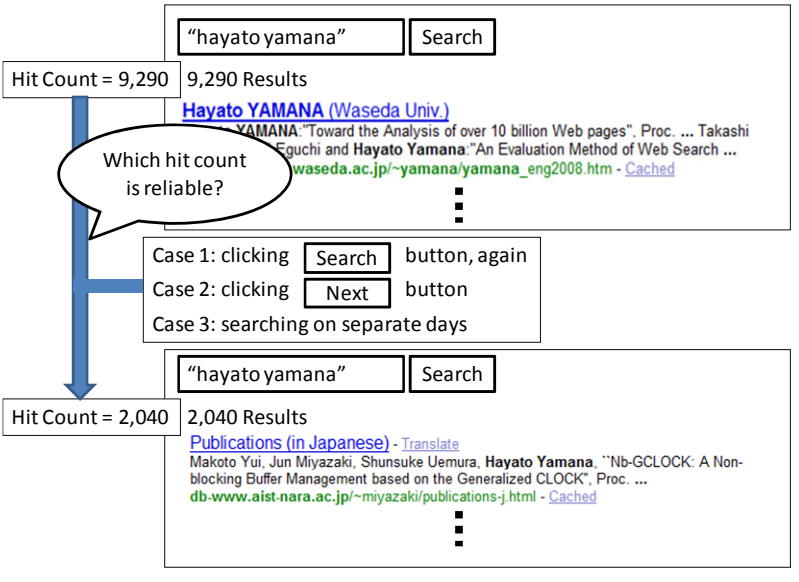

Fig. 1. Hit Count Dance

2 Related Work

Kilgarriff [5] reported that queries repeated the following day gave hit counts that varied by more than 10% as compared to the counts obtained the day before with 9 times in 30 days. He assumed that the reason was because queries were sent to different computers, at different points in the update cycle of their index, and with different data in their cache. He mentioned that the arbitrariness of search engine counts should be taken into account.

Thelwall [6] compared three major search engines with respect to four viewpoints—hit counts, the number of search result URLs, the number of search result sites, and the number of search-result top-level domains obtained by using 2,000 single word queries. Here, the three search engines considered were Google, Yahoo!, and Live Search. He compared these three search engines to clarify the difference among them on the basis of the above-mentioned four viewpoints; however, he was not concerned about the transition of the hit counts day by day, nor concerned about the transition when clicking "Next" button.

Uyar [7] proposed "Percentage of Error," how differ between the hit count and the number of returned documents, using queries whose hit counts were less than 1,000. Uyar assumed that the actual number of returned documents was reliable. Here, we are able to retrieve 1,000 documents at maximum for each query from search engines. "Percentage of Error" is defined as the difference between a hit count of the first result page and the number of returned documents, normalized by the number of returned documents. Uyar used 1,000 queries that were generated as random numbers with seven and eight digits. Uyar compared Google, Yahoo! and Live Search to conclude that Google provided the most accurate hit counts for both single and multiple term queries. Google provided less than 10% error for 78% of single queries. His main goal was to compare three search engines to find out the search engine returning accurate hit counts; however, he did not concerned about the hit count transition within the search engine. In addition, Uyar investigated the average daily hit count change ratio. However, he was not concerned about the transition of the hit counts day by day. Moreover, these results cannot be directly generalized to queries that return a higher number of documents because his research was based on queries that returned less than 1,000 documents.

As shown above, previous researches showed the hit count dance characteristics to find out the search engine to provide the most accurate hit counts among three major search engines. However, they did not clarify the basis to adopt the accurate hit counts within a search engine, i.e., how to select the most reliable hit count for a query from among the variation of hit counts when clicking the "Search" button multiple times in a short time interval or when clicking the "Next" button multiple times in a short time interval, or when clicking the "Search" button on separate days.

In this paper, we show the characteristics of the hit count dance in more detail in Section 3. Then, we propose a new scheme for selecting accurate hit counts to enhance their reliability.

3 Experiment and Trustworthiness of Hit Count

Our research goal is to find out the most reliable hit count for a query from among the variation of hit counts within a search engine, when clicking the "Search" button multiple times in a short time interval or when clicking the "Next" button multiple times in a short time interval, or when clicking the "Search" button on separate days. Here, the difference of hit counts among search engines is out of scope of this paper.

3.1 Experiment

Three types of experiments were performed on the basis of the three "hit count dance" cases described in Section 1:

1) Clicking the "Search" button many times
2) Clicking the "Next" button step by step to reach the last search result page
3) Searching with the same query on different days

We used 10,000 queries provided by Yahoo! JAPAN as the top 10,000 frequent queries in December 2007. The dataset is provided for the Japanese national "info-plosion" project [8]. Yahoo! JAPAN provides the top 10,000 frequent queries as they are. This implies that the written language of the queries is mainly Japanese but not limited to Japanese. The distribution of the queries' length, number of space-separated words, is shown in Table 1.

Table 1. Distribution of queries' length

Length of query (word)	Frequency
1	9,522
2	424
3	42
4	1
5	1
Total	10,000

In the experiment, the queries were submitted to Google, Yahoo!, and Bing via their search APIs, with these APIs' default setting; non-phrase search, normal-safe filter and the number of search engine returned is 10. Although we submit 10,000 queries to each of these search engines, sometimes, they return errors such as "no results" or "connection timed out." We omitted such erroneous results from our experimental result. These experiments were performed from October 2009 to December 2009.

Case 1: Clicking the "Search" button many times. To demonstrate the characteristics of hit count transition in case 1, we use the coefficient of variation, called CV. CV is a normalized measure of dispersion as shown below:

$$CV = \frac{standard\ deviation}{average} = \frac{\sqrt{variance}}{average} \tag{1}$$

Each query out of the 10,000 queries was submitted to all the search engines, 100 times in 5 min, i.e., each search engine performed the same search 100 times in 5 min. Here, we submitted the same query many times in a short time span to avoid the other effects such as search engines' index update cycle. After gathering 100 variations of hit counts with the same query, CV is calculated both for each query and for each search engine.

Table 2 shows the result of case 1, where the numbers in total are not equal to 10,000 because search engines sometimes return errors as described. As shown in Table 2, Google returns almost the same hit counts, i.e., consistent hit counts. For only 9 queries out of 10,000, the hit count dances with a CV less than 0.1%. Yahoo! also returns consistent hit counts because 99.4 % queries have a CV of less than 0.1%. Additionally, the maximum CV is less than 5%. Bing's hit counts dances a little as compared to those of the other search engines; however, 97.5% queries have a CV of less than 0.5%. Only one query has a CV of more than 20%; this deviation is related to the "porn words" that resulted in some effects of Bing's search filter. As a result, we are able to conclude that hit counts rarely dance in case 1. Even when hit counts dance, the CV is less than 5% except Bing's rare case.

Table 2. Distribution of hit counts when the "Search" button is clicked many times

Range	Frequency		
	Google	Bing	Yahoo!
$CV = 0.0\%$	9,977	699	9,096
$0.0\% < CV <= 0.1\%$	9	2,555	730
$0.1\% < CV <= 0.5\%$	0	6,191	46
$0.5\% < CV <= 1\%$	0	171	4
$1\% < CV <= 5\%$	0	56	1
$5\% < CV <= 10\%$	0	12	0
$10\% < CV <= 20\%$	0	4	0
$20\% < CV <= 100\%$	0	1	0
$100\% < CV$	0	0	0
Total	9,986	9,689	9,877

Case 2: Clicking the "Next" button continuously to reach the last page of the search results. In order to find out the characteristics of the hit count transition for a given query when clicking the "Next" button continuously, we gathered HitCount(1,10), HitCount(11, 20), ..., and HitCount(991, 1000) for each query. Here, HitCount(N, N+9) is defined as the hit count when we set the number of results per page to 10 and the "search offset" to N, i.e., retrieving from the N-th ranked result. For example, HitCount(1,10) shows the hit count when the top 10 search results are retrieved. HitCount(11,20) shows the hit count when the search results from 11th to 20th are retrieved, i.e., after clicking "Next" button once. HitCount(21,30) shows the hit count after clicking "Next" button twice.

In order to find out the transition patterns of the hit count with clicking the "Next" button, we define "Deep hit count Vector," in short DV, for each query. DV is defined as below based on HitCount(1, 10), HitCount(11, 20), ..., and HitCount(991, 1000). Each element in DV indicates the difference ratio of the hit count compared to the HitCount(1, 10).

$$DV = \left\{\frac{HitCount(1,10)}{HitCount(1,10)}, \frac{HitCount(11,20)}{HitCount(1,10)}, \ldots, \frac{HitCount(991,1000)}{HitCount(1,10)}\right\} \quad (2)$$

In this experiment, we submitted the queries only to Bing and Yahoo! because their APIs are able to return the top 1,000 documents at maximum as search results, while Google's API returns only the top 64 documents at maximum.

After calculating the set of DVs for 10,000 queries, we apply k-means clustering technique with cosine similarity to all pairs of DVs to extract transition patterns of the hit counts. Here, we omit the DVs that lack some elements because of errors such as "no results" or "connection timed out." We vary its clustering size k from 1 to 6; then, selected the best size by manual. Here, we choose the best size based on the following two points;

1. Start offsets when hit count dances, i.e., changes, begin are clearly different among clusters.
2. Curves of change ratio are clearly different among clusters.

Finally, we chose the best clustering size—1 on Bing and 2 on Yahoo!. Fig. 2 and Fig. 3 show the clustered hit count transition of DVs, each of whose curves represents the mean of the DVs clustered into the same cluster.

On Bing, the best clustering size is 1 where 100% of queries are clustered in. All DVs show the same characteristics such that the hit count is almost consistent till the search offset becomes more than 900. Nevertheless, after a search offset exceeds 900, the hit count falls down to 1.3% compared to HitCount(1, 10). The final hit count, HitCount(991, 1000), is the same number that a user is able to actually retrieve from Bing. This means that Bing adjusts the hit count to be the same as the number a user is able to actually retrieve such as 1,000. Hence, we have to omit the last hit count number on Bing.

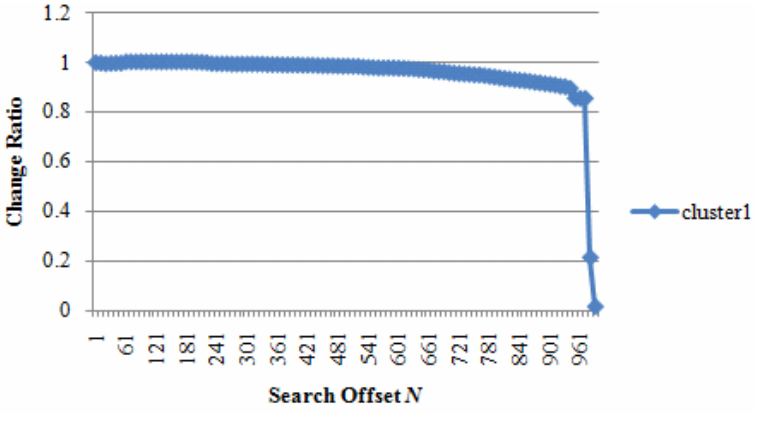

Fig. 2. Clustering result of hit count transition when clicking the "Next" button continuously to reach the last page of the search results (Bing)

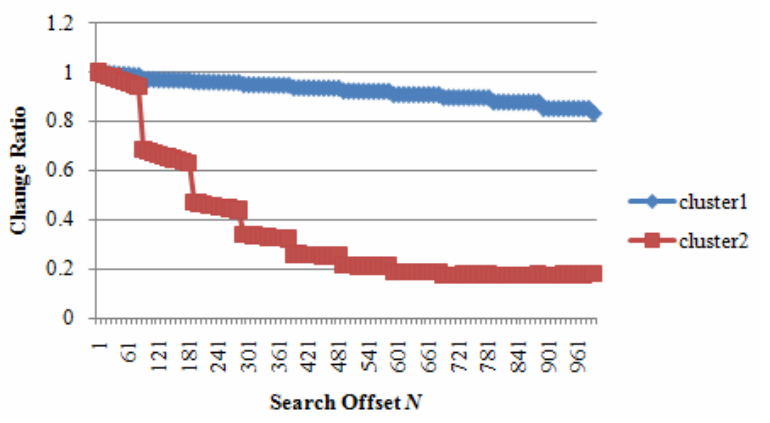

Fig. 3. Clustering result of hit count transition when clicking the "Next" button continuous to reach the last page of the search results (Yahoo!)

On Yahoo!, we have two clusters. Cluster 1 includes 86% of queries whose hit counts decrease slowly. Cluster 2 includes 14% of queries whose hit counts decrease faster than those of cluster 1 and saturate at around 20% of HitCount(1, 10).

From the above observation, we can assume that search engines return roughly estimated hit counts HitCount(1, 10) at first. With an increase in the start offset, hit counts decrease to re-calculate the hit counts by using more matched results as shown in Fig.4.

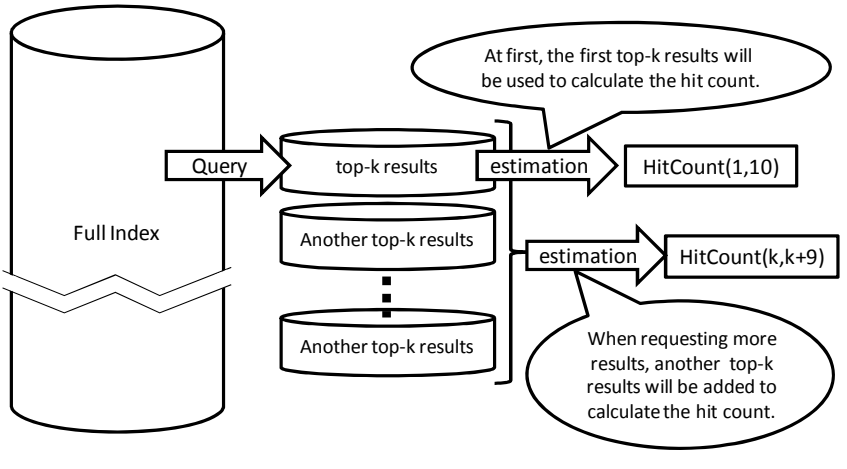

Fig. 4. How to estimate hit counts (assumption)

Here, we should consider many speed-up techniques adopted by search engines such as distributed processing [9], early termination [10], and index pruning [11]. These techniques enable fast searching; however, they might have some side-effects while calculating hit counts. We assume that these side-effects result in a rough

estimation of HitCount(1, 10). From this viewpoint, when we search with a larger search offset, we will be able to get a more reliable hit count except in the case when search engines adjust the hit count to the number that a user actually retrieves. Since hit counts are estimated by using sampled Web pages that are matched with the query through the searching process, hit counts become more reliable when a larger search offset is used. The larger the search offset is, the larger is the number of sampled Web pages that will be used for estimating the hit count. Therefore, we can conclude that the most reliable hit count is HitCount(k, k+9) where k is the largest number, usually 991 for a search having more than 1,000 matched pages. If a search engine adjusts the last hit count, we should use the hit count just before the adjusted hit count.

Case 3: Searching on different days. In order to demonstrate the characteristics of the hit count transition in case 3 and to find out what the reliable hit count during the period is, we gathered hit counts during two months, from October 11th, 2009, to December 12th, 2009, by using the same 10,000 queries. In order to find out the transition patterns of the hit counts when searching on different days, we define "Variational ratio Vector," in short VV, for each query as shown below. Each element in VV indicates the change ratio of the hit count compared to the hit count on October 11.

$$VV = \left\{ \frac{HitCount(\text{Oct}, 11)}{HitCount(\text{Oct}. 11)}, \frac{HitCount(\text{Oct}. 12)}{HitCount(\text{Oct}. 11)}, \cdots, \frac{HitCount(\text{Dec}. 12)}{HitCount(\text{Oct}. 11)} \right\} \quad (3)$$

where HitCount(*date*) is the hit count searched on *date*

where HitCount(*date*) is the hit count searched on *date*

In order to clarify the hit count transition, we apply k-means clustering technique with a cosine similarity to all pairs of VVs. Here, we omit the VVs that lack some elements because of errors such as "no results" or "connection timed out." We varied its clustering size from 1 to 6 and selected the best size by manual on the same criterion used in case 2. Finally, we chose the best clustering size—4 on Google, 5 on Bing, and 3 on Yahoo!.

Fig. 5–Fig. 7 show the clustering result of VVs, each of whose curves represents the means of VVs clustered into it. Table 3 shows the size of each cluster, i.e., the percentage of queries clustered into the same cluster.

Table 3. Percentage of queries clustered into each cluster

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Google	80.3%	11.0%	7.2%	1.5%	-
Bing	59.1%	19.4%	15.1%	5.6%	0.8%
Yahoo!	67.2%	29.6%	3.2%	-	-

On Google, the largest cluster is cluster 1 whose hit count is almost consistent during the period. The second largest cluster is cluster 2 whose hit count is also consistent but with pulse noise between Oct. 20th and Oct. 22nd. The next largest cluster is cluster 3 whose hit count increases slowly. Finally, the smallest cluster is cluster 4 whose hit count increases eight-fold times around Dec. 2nd.

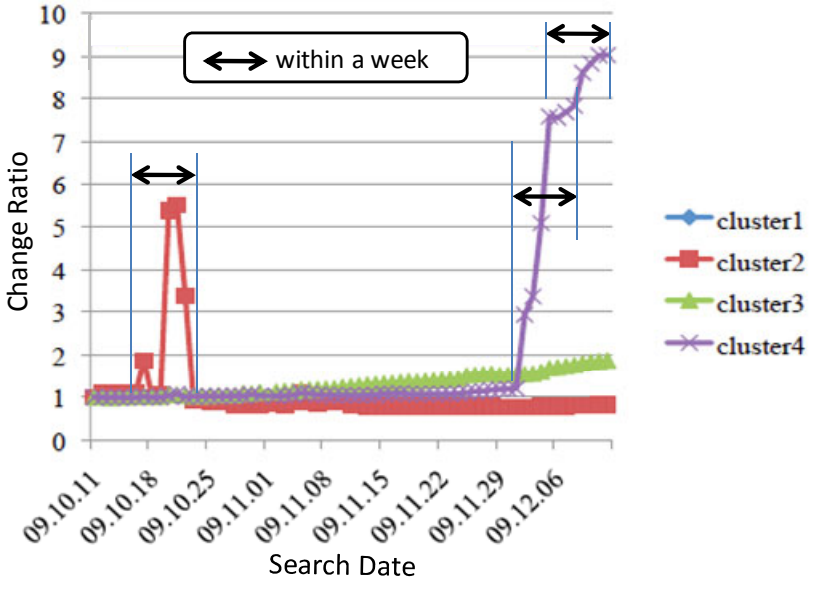

Fig. 5. Clustering result of daily hit count transition (Google)

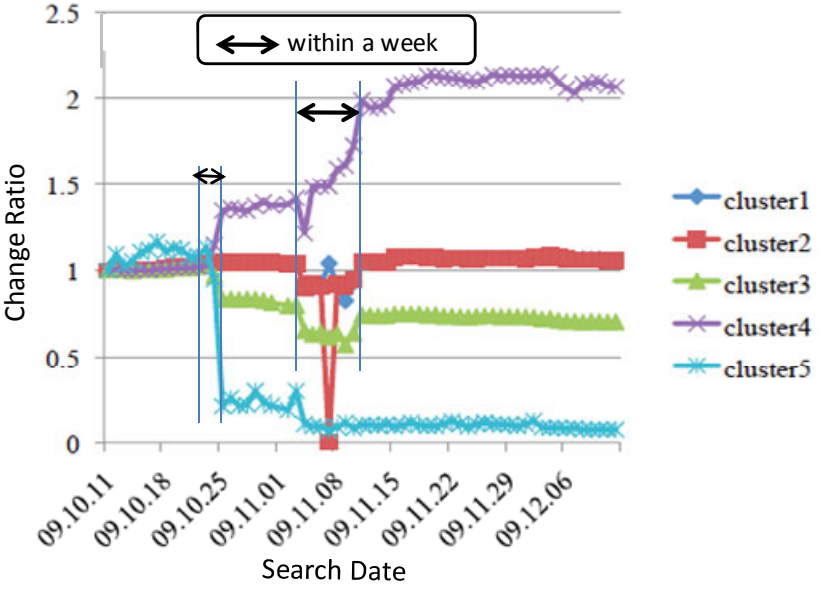

Fig. 6. Clustering result of daily hit count transition (Bing)

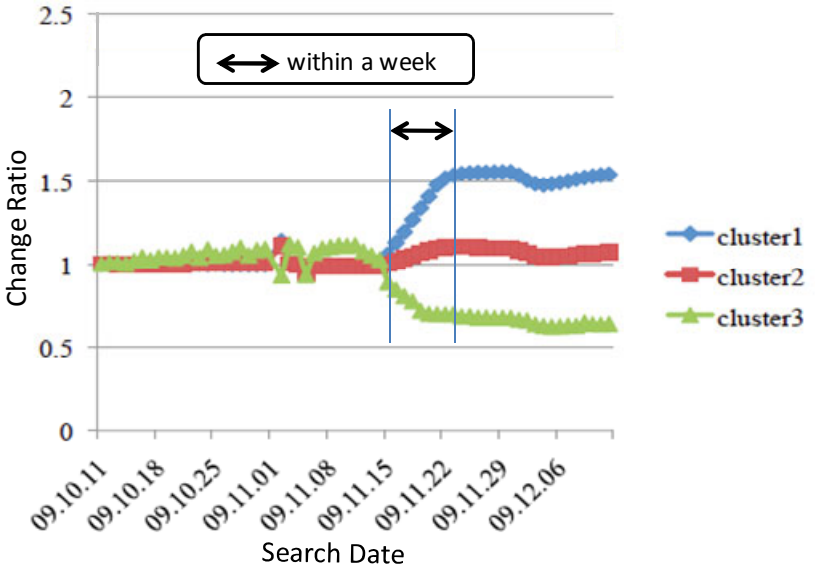

Fig. 7. Clustering result of daily hit count transition (Yahoo!)

On Bing, hit counts vary on Oct. 25th. The phenomenon may be resulted from Bing's search index update on that day. From Nov. 3rd to Nov. 11th, the hit counts dance very hard. Finally, hit counts calm down again from Nov. 11th. These transitions seem to repeat "stay" and "dance."

On Yahoo!, hit counts are almost consistent except between Nov. 13 and Nov. 20. Yahoo! seems to update its index during this week. After Nov. 20th, we can observe increased, decreased, and stable clusters.

As a result, we summarize that hit counts repeat "stay" and "dance" as their temporal transition whose increase or decrease speed depends on the clustered queries. The "dance" phase continues up to one week as shown in Fig.5 to Fig.7. During the "dance" phase, hit counts usually change over 30% in comparison with those at the beginning of the phase. Hence, we are able to conclude that hit counts are more reliable when they keep stable numbers during a week in comparison with when they change, because dancing hard implies an index updating phase.

3.2 Basis to Select a Reliable Hit Count for a Query

Based on 3.1, we are able to provide the most reliable hit count for a query within a search engine, i.e., the most reliable hit count among its variation when clicking the "Search" button multiple times in a short time interval or when clicking the "Next" button multiple times in a short time interval, or when clicking the "Search" button on separate days.

From the experiment in Case 1, we do not have to consider the hit count dance caused by clicking the "Search Next" button because it does not have a large effect. Instead, we should consider the hit count dances in Case 2 and Case 3. From the

experiment in Case 2, we are able to obtain the most reliable hit count, HitCount(k, k+9), where k is defined as the largest number, usually 991 for the search having more than 1,000 matched pages. If a search engine adjusts the last hit count, we should use the hit count just before the adjusted hit count. Moreover, based on the experiment in Case 3, hit counts seem to repeat "stay" and "dance" phenomena. We should use the hit counts that keep almost the same number during a week in order to avoid the search engines' index updating phase or some other changing phase. Here, the "stay" phase is defined as the phase where the changing ratio of hit counts is less than 30%.

4 Conclusion

In this paper, we provide a basis to adopt hit counts with various types of studies. Previous researches investigated the hit count dance characteristics to find out the search engine to provide the most accurate hit counts among three major search engines. However, they did not clarify the basis to adopt the accurate hit count for a query within a search engine, i.e., how to select the most reliable hit count for a query from among its variation when clicking the "Search" button multiple times or when clicking the "Next" button, or when clicking the "Search" button on separate days. In this paper, we analyzed the characteristics of hit count transition by gathering various types of hit counts over two months by using 10,000 queries. We have concluded that the hit counts with the largest search offset just before search engines adjust their hit counts are the most reliable. Moreover, hit counts are the most reliable when they are consistent, i.e., less than 30% change, over approximately a week.

There remain two aspects after this work, which have not yet been explored fully. The first aspect is the consideration of the queries' famousness. In this paper, we selected well-used queries. Considering the architecture of search engines such as result cache, the query selection may have some effect on "hit count dance." Hence, we should also use non-famous queries. The second aspect is the verification time period. We performed our experiment from October 2009 to December 2009. Two months might be insufficient for the conclusion of the hit count dance. Therefore, we should continue our experiment and make hit count dance more obvious.

Acknowledgement

The authors are grateful for the financial support by the Grant-in-Aid for Scientific Research from the Ministry of Education, Science, Sports and Culture (No. 21300038). We would like to thank our anonymous reviewers who have provided helpful comments on the refinement.

References

1. Kilgarriff, A., Gefenstette, G.: Introduction to the Special Issue on the Web as Corpus. J. of Computational Linguistics 29(3), 333–347 (2003)
2. Cilibrasi, R.L., Vitanyi, P.M.B.: The Google Similarity Distance. IEEE Trans. on Knowledge and Data Engineering 19(3), 370–383 (2007)

3. Matsuo, Y., Sakai, T., Uchiyama, K., Ishizuka, M.: Graph-based Word Clustering using Web Search Engine. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing, pp. 542–550 (2006)
4. Yamamoto, Y., Tezuka, T., Jatowt, A., Tanaka, K.: Honto? Search: Estimating Trustworthiness of Web Information by Search Results Aggregation and Temporal Analysis. In: Dong, G., Lin, X., Wang, W., Yang, Y., Yu, J.X. (eds.) APWeb/WAIM 2007. LNCS, vol. 4505, pp. 253–264. Springer, Heidelberg (2007)
5. Kilgarriff, A.: Googleology is Bad Science. J. of Computational Linguistics 33(1), 147–151 (2007)
6. Thelwall, M.: Quantitative Comparisons of Search Engine Results. J. of the American Society for Information Science and Technology 59(11), 1702–1710 (2008)
7. Uyar, A.: Investigation of the Accuracy of Search Engine Hit Counts. J. of Information Science 35(4), 469–480 (2009)
8. info-plosion, http://www.infoplosion.nii.ac.jp/info-plosion/ctr.php/m/IndexEng/a/Index/ (accessed 28/4/2010)
9. Manning, C.D., Raghavan, P., Schutze, H.: Introduction to Information Retrieval. Cambridge University Press, New York (2008)
10. Anh, V.N., de Kretser, O., Moffat, A.: Vector-Space Ranking with Effective Early Termination. In: Proc. of the 24th Ann. Int'l ACM SIGIR Conf., pp. 35–42 (2001)
11. Carmel, D., Cohen, D., Fagin, R., Farchi, E., Herscovici, M., Maarek, Y.S., Soffer, A.: Static Index Pruning for Information Retrieval Systems. In: Proc. of the 24th Ann. Int'l ACM SIGIR Conf., pp. 43–50 (2001)