P-Download: A New Personalization Approach for a Content-Based Search System

Keejun Han, Juneyoung Park, Donghee Hong, Jinsup Shin and Mun Y. Yi*

Department of Knowledge Service Engineering, Korea Advanced Institute of Science and Technology
373-1, Guseong-dong, Yuseong-gu, Daejeon 305-701, South Korea
{keejun.han, j.park89, lucy.hong, js.shin, munyi}@kaist.ac.kr

With the explosive growth of information provided on the Web, personalization of search continues to be an important issue, particularly in the context of content-based search systems as the Internet started to evolve from being a simple information provider to a rich content provider. Building upon the recent findings in personalization strategies, the present research proposes a new search personalization algorithm that creates a synergetic effect by combining the download information with the current state of the art click-based algorithm. By assessing the log data of a user's personal click-history in relation to the download information, the proposed method offers substantial advantage in creating a more specific user profile for personalization. A large dataset from a real-life content-based search system has been analyzed and tested for the evaluation of the proposed personalization method. The results largely support the significance of the proposed approach, highlighting the importance of downloading information in content-based search systems as a key ingredient for effective personalization. The findings have practical implications for content search service providers.

Key Words: Personalization, Content-based search, Big data, Log data, Click-based algorithm

1. INTRODUCTION

The exploding growth of online information during the past decade has made search engines an indispensable part of the Internet experience, effectively demonstrated by the rapid growth of many search engine providers. It has become clear that the Internet has reached a point where simple navigation cannot suffice to allow users to retrieve the information that they require. The sheer number of pages and contents available on the Web became too enormous for a singular search query to be able to find the exact information that the user wants to retrieve. As the basis of the Internet users has become more diverse with ever-increasing adoption of the Internet technologies and applications, a single query can be sent with different expectations. A Personalized search system offers a potential solution to this problem of the current search systems.

In this study, our goal is to propose a new personalized search algorithm and validate its effectiveness by comparing it with P-Click [1], which is known as a state of art personalization algorithm. P-Click is a personalization algorithm that derives an individual preference profile from the user's click-history. Because the user's personal click-history is a log dataset that is automatically and constantly recorded, the dataset easily becomes significantly large and thus brings stability in performance and match between the user and the data. Such implicit approach to extracting user information is known to be much robust compared to the explicit approach in which the user profile is specified by the user themselves who are often

* Mun Y. Yi is the corresponding author.

reluctant on providing search preference and interests. However, P-Click is not totally bullet proof in its application.

Despite the high performance and significance as a personalization algorithm, the effectiveness of P-Click in a content-based search system remains questionable. A content-based search system is an upcoming search environment where various multimedia materials are downloadable, whereas a traditional Web-based search system allows access to the websites only. Since the rise of Web 2.0, contents have been not only generated and shared among users but also delivered to them in a fashion to meet what the user needs, as Web 2.0 'works for the user'[2]. Providing the availability and accessibility to the specific contents that the user needs significantly alter the user's experience in this form of new Web, and a simple webpage with text information does not suffice anymore. The personalization of a content-based search system has the distinct purpose of satisfying the deliverables required in the new phenomena in the Web.

This study uses the central aspect of a content-based search system with which a user downloads the content that he or she finds value in, and proposes a new algorithm, called P-Download, that augments P-Click by exploiting the content downloading information. We used data set retrieved from a real-life content-based search system known as Korean Traditional Knowledge Portal[1]. Using the dataset, we tested and compared the effectiveness of both P-Click and P-Download to examine how the new perspective used in P-Download can provide a relatively superior performance for content-based search systems. In summary, this paper's main contributions are as follows:

- We develop a new personalization algorithm that is specialized for a content-based search system.
- We examine the validity of a new algorithm by comparing it with the current state of the art algorithm, which is P-Click.
- The real-life data explicitly demonstrates the stability of the performance of the new perspective used in P-Download algorithm.

The rest of the paper is proceeded as follows. Section 2 introduces related works and section 3 proposes the algorithm P-Download. Section 4 describes the dataset used. Section 5 describes the experiment and the results obtained from the experiment. Section 6 concludes our study with a summary of findings and future research implications.

2. RELATED WORK

The goal of personalization is to provide right contents to right users in accordance with their search needs and interests [3]. Identifying the user's goals and needs is accomplished through the creation of a *user profile* that consists of a set of *things* (e.g., values, terms, twits, tags) that represent the user. There are largely two ways to utilize the user profiles: query expansion and re-ranking. While a traditional query expansion, which selects additional terms normally to improve *recall*, heavily focus on how to construct a list of candidate terms to be added for expansion, a

[1] http://www.koreantk.com

personalized query expansion approach considers both aspects of *precision* and recall by adding different terms for different users on the same query *q* [4]. Re-ranking, on the other hand, given a query *q,* more focuses on re-ordering the initial search results by re-weighting each document in the list [5]. In this paper, because our aim for this study is to augment the existing algorithm with downloading information, we chose the re-ranking approach to observe the direct effects of downloading information on the final search ranks. Compared to the re-ranking by *pseudo-relevance feedback,* which utilize the top-N initially retrieved document [6], using search log data is more powerful, directly giving more weights to the terms existing on the clicked pages. If a page is clicked and downloaded, the terms in the page can be more positively weighted during the re-ranking process.

User profiles can be made from users' direct inputs [7]. This approach is to ask users to provide their general interests. Those interests are then used to filter search results by computing similarities between the retrieved pages and interests. However, using direct inputs from users suffers from a large number of missing and malicious inputs from users because users are reluctant to provide explicit feedback about their search results or interests [8]. Thus, many of later works on personalized search focused on building user profiles automatically from the past search history of users. In this case, search-log data, which is essentially a large data about users' search activities, can be efficiently utilized to construct user profiles implicitly. It records search related activities of all of the users, making it possible to predict the preferences of users based upon their past activities recorded in the data. Because the approach that uses the search log data is the primary means to construct the user profile [9][10], we focus on eliciting a robust personalized algorithm from the real search log data.

There are three ways of personalization using a search log data: historical click-based algorithm, user-topical-interest-based algorithm, and group-based algorithm denoted as P-Click, S-Topic (or LS-Topic), and G-Click in this paper [1]. Among those, the most efficient algorithm in a real dataset was proven to be P-Click. The underlying assumption for P-Click is that for a query q submitted by a user u, the Web pages frequently clicked by u in the past are more relevant to u than those hardly clicked by u. However, in P-Click, many of noise clicks are abused to compute the personalization score and they adversely affect the overall accuracy of personalization. In this paper, we propose a new algorithm using the download information of retrieved contents to minimize the effect of those noise clicks on the personalization score.

In spite of several advantages of using search log data for personalization, those data suffer from *data sparsity*. Analysis on the sparsed data should largely rely on approximation and prediction to create user profiles, thus, there have been attempts to increase the number of available data in the dataset. Most popular method is *collaborative filtering (CF)* that discovers a similar group of users and incorporate the preferences of the group of users to secure the performance of personalized search [10]. A unique approach that applies *singular vector decomposition (SVD)* in the 3-dimensional data of query, user, and page discovers the latent relationships among those contained in click-through data [11]. In this paper, we propose a flexible algorithm that can potentially apply those existing methods to solve the data sparsity problem.

In content-based search, most of personalization works focused on developing a recommendation system. It aims at recommending items that had not yet been considered by users, but, might be preferred. Although there are three techniques [12] for recommendation: Collaborative filtering, content-based filtering, and hybrid approach, CF still shows the best performance among those. However, there is still a large room left for personalized search to perform sufficiently high on the content-based search services. To the best of our knowledge, this is the first work to utilize the downloading information to strengthen the content-based personalized search techniques on those services.

3. PROPOSED METHOD

Search logs record the activities of users, which reflect their interests while performing search. In the traditional Web search, search logs are generally consisted of queries, the URLs that users clicked, and the number of times that they clicked. In contrast, content-based search data has the following information: user queries, the URLs of contents, actions performed on the URL (click or download) and the time that they performed the corresponding actions. The logs are then separated by sessions that consist of a single query and all of the clicked Web pages after issuing the query. Note that downloading actions do not always appear in every session because users may not download any content if they are not able to find relevant contents to their needs, implying that the total number of downloading actions is less than that of the clicking actions. A partial sample of search log data is shown in Table 1. Based upon those logs, our approach forms a 3-tuple of <q, p, u> that consists of query (q), document (p) based on a data set that shows the user's (u) past clicking and downloading activities.

Table 1. Sample entries of search logs. C denotes a click action and D denotes a download action

Session ID	Query	Contents ID	Action	Time
1	Immune	BOGHBE_2010_v23n4_10	C	Xxxx
1	Immune	BOGHBE_2010_v23n4_20	C	Xxxx
1	Immune	BOGHBE_2010_v23n4_20	D	Xxxx
2	ADHD	HBSKB9-2004-v15n1-239	C	Xxxx
…	…	…	…	…

The underlying assumption for P-Click is that for a query q submitted by a user u, the Web pages frequently clicked by u in the past are more relevant to u than those hardly clicked by u. Equation 1 shows the calculation done to gain the P-Click score from the tuple <q,p,u>. While $|Clicks(q,p,u)|$ of equation 1 represents the number of times the user 'u' has clicked the document 'p' for the query 'q'. $|Clicks(q,\blacksquare,u)|$ represents the total number of documents that the user 'u' clicked for the query 'q'. The β score represents the smoothing value for the equation, and is defined to have the value of 0.5 in this study. Dou [1] compares the performance of P-Click with 4 other algorithms, denoted as L-Topic, S-Topic, LS-Topic and G-

Click in his study of personalization algorithm. The performance of P-Click was the most stable in multiple test conditions and it outperformed other algorithms.

$$S^{P-Click} = \frac{|Clicks(q,p,u)|}{|Clicks(q,\blacksquare,u)| + \beta} \cdots \text{Equation 1}$$

In comparison, the P-Download algorithm takes an entirely different perspective for the tuples. Instead of the number of clicks that is used in P-Click, P-Download takes into account the nature of the content-based search system and uses the number of downloads the user generated. In a content-based search system, the user not only clicks on the webpage to acquire the information in need, but also clicks on the content given that the information provided shows that the content is what he or she was looking for. This aspect separates the algorithm from P-Click as it more specifically matches the user behavior pattern in a content-based search system.

P-Download algorithm is also constructed by the same three tuple <q,p,u>. However, the tuples are used to calculate the P-Download score by Equation 2. $|Downloads(q,p,u)|$ represents the number of time the user 'u' downloaded the document 'p' for the query 'q', and $|Downloads(q,\blacksquare,u)|$ represents the total number of documents downloaded by user 'u' from the query 'q'. Because the number of download is much smaller than the number of clicks, we adjusted the smoothing value γ to be at 0 for it to have a more impact to the final value.

$$S^{p-download\_only} = \frac{|Downloads(q,p,u)|}{|Downloads(q,\blacksquare,u)| + \gamma} \cdots \text{Equation 2}$$

The P-Click algorithm suffers from a reduced performance from the noise caused by the user clicking on documents that does not match their needs. Thus, in order to create a synergetic effect and potentially maximize the performance of the algorithms, we combine the two algorithms above. The P-Download algorithm's consideration of user's final selection of the content could significantly reduce the noise from the P-Click algorithm. Equation 3 calculates the combined score $S^{pd-click}$ where α score represents the impact factor between 0 and 1 which determines the ratio for implementing the score from P-Click and P-Download. In this paper, α is empirically set as 0.

$$S^{p-download} = \alpha \cdot S^{p-click} + (1-\alpha) \cdot S^{p-download\_only} \cdots \text{Equation 3}$$

4. DATASET

The purposes of this section are to verify if the characteristics of our dataset are consistent with those found in [14][16] and find user behaviors newly shown in content-based searches. It is necessary to show that our dataset is consistent with other previous search log datasets, in order to secure the reliability of the results obtained based upon our dataset. By doing so, our proposed algorithm can be potentially applied into not only our dataset but also other similar datasets.

For this study, we collected search query logs from Korean Traditional Knowledge Portal (KTKP) which is operated by Korea Patent Office for a comprehensive service of Korea's traditional knowledge, providing various types of contents such as scientific papers, patents, and prescriptions about Korean medicine. Unlike other search services that provide web-based documents, the KTKP provides contents that are relevant to a given query.

4.1 Statistics about Dataset

For our study, we collected 5 years of search log data from KTKP. In the collected dataset, the queries without any clicks were removed because they did not contain any meaningful information. We also removed the records of users accessing the service from external web portals because their user IDs are anonymous. User IDs are necessary to identify each individual user. Table 2 summarizes the basic statistics of the dataset. The number of clicks/queries indicates that users normally click less than two pages per queries. This tendency is similarly shown in [13][14]. On the other hand, the number of downloads/clicks implies that users rarely download contents although they click a few pages. It sounds plausible that users click the candidate items and only download the most relevant items among them. Thus, downloading information can be a powerful indicator for the relevance of the items. Furthermore, 43% of queries in the dataset are repeated at least once while 69% of those queries are repeated by the same user. These results are mostly consistent with those given in [15] and support the assumption that personalized search is useful with regard to this dataset.

For the experiment, we chose the data from January 2012 and April 2013 to form a sample dataset because the current ranking provided by KTKP is not consistent with old data. The sample dataset was split into two parts: a training dataset and a test dataset. The training set consisted of the log data of the first 11 months and the testing data consisted of the log data of the last 5 months. Again, note that all of the data without false clicks and anonymous users were used for the overall analysis of the dataset (reported in this section) and the sample dataset were used for the evaluation of the algorithms (reported in the next section).

4.2 Statistics about Queries

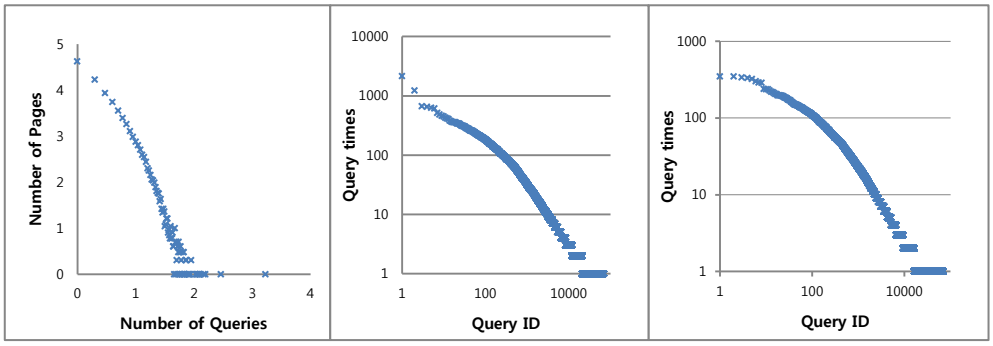
The analysis for the queries is required to verify if the query behaviors of the dataset are similar with that given in [16]. Figure 1 (a) plots the distributions of queries and pages. In the figure, the large portion of pages are only associated with few queries, while few pages are associated with a large number of queries. In other words, the rule of power law is exhibited in the graph, implying that there exist queries that are largely affected by personalization.

Figure 1 (b) plots the distributions of query frequency. In this figure, the first query is the most frequent one and the last is the least popular one. Figure 1 (c) plots the distribution of number of users with each query. Both figures also conform to the rule of power law as shown in [1]. The power-law distribution is commonly observed in the analysis of search log dataset in previous studies [13][14][16], strongly supporting that our dataset is closely consistent with other datasets used in

those studies. It demonstrates that our algorithm run on this dataset can be also successfully applied in other datasets.

Table 2. Basic statistics of dataset

Item	ALL (trimmed)	Sample
# users	9,084	3,196
# queries	238,149	47,536
# distinct queries	70,404	19,708
# clicks	398,331	67,833
# downloads	105,112	26,989
# clicks / queries	1.6726	1.4269
# downloads / clicks	0.2639	0.3978



(a) (b) (c)

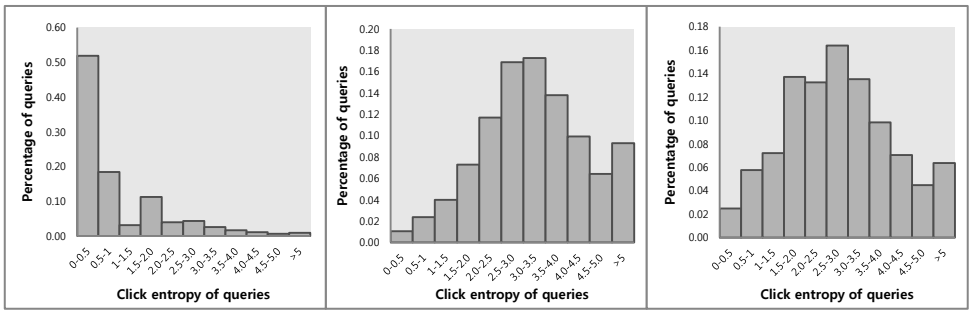
Figure 1.Query Popularity Distribution (a) The distribution of pages and queries (with logarithm on X and Y), (b) Distribution of query frequency (log scale), and (c) Distribution of user number of queries (log scale)

4.3 Distribution of Query Click Entropies

The performance of personalization may be unsatisfactory if queries have less variation [17]. Query click entropy suggested in [1] is a good indicator for click variation. If page p is only clicked by query q, the entropy is 0. Smaller entropy indicates that most users agree to click few pages on the same query. Meanwhile, higher entropy means that query is either informational or ambiguous, promising the higher effectiveness of personalization in this case [18].

Figure 2 (a) shows the click entropy distribution. Approximately, 50% of queries have low click entropy between 0 and 0.5 which is consistent with that in [1]. However, Figure 2 (b) and Figure 2 (c) have different behaviors from those in [1]. Unlike previous studies, the click entropies are heavily skewed to the right on the repeated queries in our dataset. It means that there exist a large number of queries that differentiate desired items, invoking a noticeable click variation. A possible

explanation for this is that users seek for different types of contents although their query is identical. In KTKP, various types of contents are provided such as papers, prescriptions, and patents. Depending on the needs of users, they click and download the different types of items. As is the case in KTKP, content-based search services begin to provide various types of contents rather than focusing on a single type of content (e.g., Amazon, e-bay). Due to the result of click entropies, personalization seems to be more useful for the various types of content-based services in improving the effectiveness of the search compared to the traditional services.



(a) All queries (b) Queries with query times >5 (c)Queries with user number >2

Figure 2. Distribution of query click entropy

5. EXPERIMENT

Our experiments aim at answering the following research questions:

- Does personalization perform well in content-based search?
- What is the effect of using download information in the real dataset?
- What is the best weight balance between click and download information in our algorithm?

5.1 Experiment Measure

To evaluate the performance of our algorithm, we used the *Mean Average Precision (MAP)* and *Normalized Discounted Cumulative Gain (NDCG)* measures, considering that we more focus on whether or not our algorithm improves *Precision* rather than *Recall,* because users are only likely to look at a few items that are highly ranked in the search list. Average Precision (AP) for query *s* is defined as follows:

$$\mathrm{AP} = \frac{1}{R}\sum_{i=1}^{l}\frac{R_i}{i}\delta_i \cdots \text{Equation 4}$$

where R is the number of relevant contents, R_i is the number of relevant contents up to *i*th position in the sequence of retrieved contents. δ_i is 1 if the *i*th content is relevant to *s*, otherwise 0. *l* denotes the number of contents in the list. MAP is then calculated as follows:

$$\text{MAP} = \frac{1}{Q}\sum_{q=1}^{Q} AP(q) \cdots \text{Equation 5}$$

where Q is the number of queries. Compared with AP, DCG is a somewhat more sophisticated measure because it gives more weights on the items that are highly ranked in the search list. It is computed as:

$$\text{DCG(p)}\begin{cases} G(1) & ,if\ p = 1 \\ DCG(p-1) + \dfrac{G(p)}{\log(p)}, & otherwise \end{cases} \cdots \text{Equation 6}$$

where p is a particular rank position, DCG(p) denotes the DCG value accumulated at a particular rank position p and G(p) denotes gain value and its value is fixed at 1 if the content is relevant at *p*. Finally, DCG is normalized from 0 to 1 by *IDCG* (Best possible DCG value) as follows:

$$\text{NDCG(p)} = \frac{DCG(p)}{IDCG(p)} \cdots \text{Equation 7}$$

In general, MAP and NDCG have similar effects on evaluating personalization performance, and our experimental results confirm that those two measures are in fact consistent. In the two measures, *l* and *p* are equally set as 5 for the experiment.

5.2 Experimental Setup

In the experiment, we defined U_1 to be the top 50 downloaded query results from the query in the KTKP. Afterwards, for the documents that are $x_i \in U$, we used the suggested personalization algorithm to calculate the personalization score. Afterwards, we defined U_2 to be the re-ranked query results that has been sorted in a descending order according to their personalization scores. Finally, we calculated U_d as the final ranking by combining U_1 and U_2 through Borda's method [19]. In our experiment, we set U_1 as the *baseline*. Notice that this baseline is the original Web search method without any personalization. We also similarly calculate final ranking based on P-Click algorithm, which is used for the comparison with the proposed personalization method.

Furthermore, we found, for many of queries, users selected only the top results, suggesting that the baseline has done the best on those query. Except for those queries, users selected more than the top results. Thus, we denote those queries as *not-optimal queries* and we examine the search performance in two different query types.

5.3 Results

5.3.1 Overall Performance

Table 3 shows the overall effectiveness of the personalization strategies on the test queries. We find:

(1) Both the click-based personalization method P-Click and the download-based personalization P-Download consistently outperform the baseline method overall. For instance, on all test queries, P-Click has a 13.82% improvement over the baseline method and P-Download has an 18.52% improvement over the baseline method (using MAP@5). P-Click and P-Download also show significant improvements (6.55% and 13.77%) over the baseline for the not-optimal queries. These results show that personalization does improve content-based search performance.

(2) Our proposed method P-Download outperforms P-Click. Again, P-Download has significant improvements (5.09% and 5.45% using NDCG@5 and MAP@5 respectively) over P-Click on all queries. For not-optimal queries, it also shows better performance (4.77% and 7.72% using NDCG@5 and MAP@5 respectively) than P-Click. These results provide empirical evidence that utilizing download information can augment the click-based search strategy by identifying the contents that are clicked and downloaded.

Table 3. Overall performance of personalization strategies

Method	All		Not-optimal	
	NDCG@5	MAP@5	NDCG@5	MAP@5
Baseline	0.3885	0.3810	0.3627	0.4120
P-Click	0.4020	0.4421	0.3853	0.4409
P-Download	**0.4236**	**0.4676**	**0.4046**	**0.4778**

5.3.2 Impact of Parameter

Recall the parameter α in Equation 3 that balances the impact between click and download information. The smaller α is, the bigger the impact for download is. We chose MAP because it has a larger gap between the lowest and highest value than NDCG, in order to clearly observe the performance varies as α changes. Figure 3 shows the MAP value against varying α from zero to one. It shows the best performance when we only consider download information only (α=0). In other words, personalization performs the best when we only use download information. Actually, the result that α is not optimized somewhere between 0 and 1, but optimized at 0 is unexpected because the number of download information is much less than that of click information in our dataset, thus, we initially expected that using only downloading information would suffer from the insufficient number of available sources for personalization. The possible reason for this unexpected result is because we only look at top 5 items in the sequence

of retrieved contents. For those top 5 items, the number of download information is sufficiently enough to perform effective personalization. This assumption is reliable because MAP is the highest (MAP = 0.3) when α reaches 0.4 if we consider top 10 items in our additional experiment. However, our approach to concentrate on measuring precision for top 5 items is still reasonable because, as previously shown in Table 2, users mostly look at less than 2 items (# clicks / queries = 1.6726) in average for each query. To sum up, these results show that download information empowers the personalization on especially highly ranked contents in the search list.

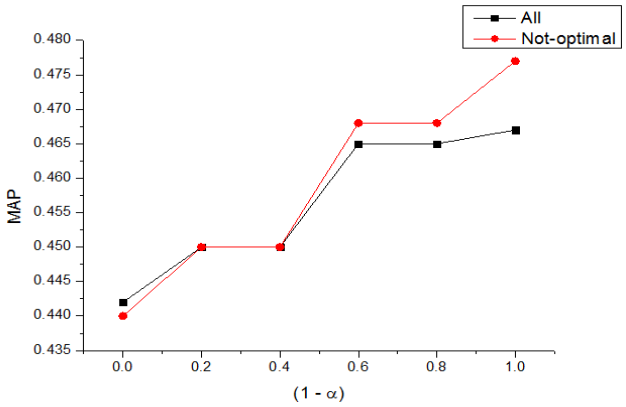


Figure 3. Impact of the parameter α

6. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a new personalization algorithm, P-Download, utilizing download information of contents by assigning more weights to contents that are clicked and downloaded. The assumption for using download information is that, download action can be considered as the final confirmation that the chosen content highly fits needs of a user. We used a large real search log data from KTKP that is the most popular content-based search engine in providing various contents about Korean traditional knowledge. Through the analysis of the dataset, we confirmed that personalization can perform well for not only web-based search but also content-based search.

Experimental results also show that the proposed personalization approach consistently outperforms the baseline condition without personalization and the click-based approach with personalization. Although the download information is not as plentiful as the click information, the algorithm has been found to still work well on top-ranked items in the search results. Although our algorithm provides definite performance improvements, it only can work on repeated queries. It is also affected by the availability of the data. Our future work, thus, should include incorporating other group-based personalization techniques into the proposed algorithm in order to overcome those limitations, in addition to utilizing textual information from the retrieved contents.

ACKNOWLEDGEMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2011-0029185).

REFERENCES

[1] Dou, Z., Song, R., and Wen, J., "A Large-scale Evaluation and Analysis of Personalized Search Strategies," In *Proc. the Int'l Conf. on World Wide Web*, Banff, Alberta, Canada, pp. 581-590, May, 2007.
[2] Miller, P., Web 2.0: Building the New Library, Ariadne, 2005.
[3] Shahabi, C., and Chen, Y., C., "Web Information Personalization: Challenges and Approaches," In *Proc. the 3rd Int'l Workshop on Databases in Networked Information Systems*, 2003.
[4] Chirita, P., A., Firan C., S., and Nejdl., W., "Personalized Query Expansion for The Web," In *Proc. the Int'l Conf on ACM SIGIR*, pp. 7-14, 2007.
[5] Shen, X., Tan, B., and Zhai, C., "Implicit User Modeling for Personalized Search," In *Proc. the Int'l Conf. on ACM CIKM*, pp 824-831, 2005.
[6] Yu, S., Cai, D., Wen, J., and Ma, W., "Improving Pseudo-relevance Feedback in Web Information Retrieval Using Web Page Segmentation," In *Proc. the 12th Int'l Conf. on WWW*, 2003.
[7] Pitkow, J., Schutze, H., Cass, T., Cooley, R., Turnbull, D., Edmonds, A., Adar, E., and Breuel, T., Personalized Search, *Commun*, ACM, 45(9), pp. 50-55, 2002.
[8] Carrol, J., M., and Rosson, M., B., "Paradox of the Active User," *Interfacing thought: Cognitive Aspects of Human-Computer Interaction,* pp. 80-111, 1987.
[9] Shen, X., Tan, B., and Zhai, C., "Implicit User Modeling for Personalized Search, In *Proc. the Int'l Conf. on ACM CIKM*, pp. 824-831, 2005.
[10] Sugiyama, K., Hatano, K., and Yoshikawa., "Adaptive Web Search Based on User Profile Constructed without Any Effort from Users," In *Proc. the Int'l Conf. on WWW*, pp. 675-684, 2004.
[11] Sun, J., T., Zeng, H., J., Liu, H., Lu, Y., Chen, Z., "CubeSVD: A Novel Approach to Personalized Web Search," *In Proc. the Int'l Conf. on WWW*, pp. 382-390, 2005.
[12] Ben, S., J., Konstan, J., A., and Riedl, J., "E-commerce Recommendation Applications," *Applications of Data Mining to Electronic Commerc*e, pp. 115-153. 2001.
[13] Silverstein, C., Marais, H., Henzinger, M., and Moricz, M., "Analysis of a Very Large Web Search Engine Query Log", *SIGIR Forum*, 33(1), pp. 6-12, 1999.
[14] Jansen, B., J., Spink, A., Bateman J., and Saracevic, T., "Real Life Information Retrieval: A Study of User Queries on the Web, *SIGIR Forum*, 32(1), pp. 5-17, 1998.
[15] Teevan, J., Adar, E., Jones, R., and Potts, M., "History Repeats Itself: Repeat Queries in Yahoo's logs," In *Proc. the Int'l Conf. on SIGIR*, pp. 703-704, 2006.
[16] Xie, Y., and O'Hallaron, D., R., "Locality in Search Engine Queries and Its Implications for Caching," In *Proc. the Int'l Conf. on INFOCOM*, 2002.
[17] Teevan, J., Dumais, S., T., and Horvitz, E., "Beyond the Commons: Investigating the Value of Personalizing Web Search, " In *Proc. the Int'l Conf. on PIA*, 2005.
[18] Lee, U., Liu, Z., and Cho, J., "Automatic Identification of User Goals in Web Search," In *Proc. the Int'l Conf. on WWW*, pp. 391-400, 2005.
[19] Dwork, C., Kumar, R., Naor, M., and Sivakumar, D., "Rank Aggregation Methods for the Web," In *Proc. the Int'l Conf. on SIGIR*, pp. 613-622, 2001.