# An Academic Search Engine for Personalized Rankings

Worasit Choochaiwattana*

*College of Creative Design and Entertainment Technology, Dhurakij Pundit University, Bangkok, Thailand*

**Abstract:** Rapidly increasing information on the Internet and the World Wide Web can lead to information overload. Search engines become important tools to help WWW users to discover information. Exponential increases in published research papers, academic search engines become indispensable tools to search for papers in their expertise and related fields. In order to improve the quality of search, an academic search engines' capability should be enhanced. This paper proposes a search engine for personalized rankings. In order to evaluate the performance of personalized rankings, thirty-five graduate students from the Department of Web Engineering and Mobile Application Development at Dhurakij Pundit University are participants in the research experiment. Participants are asked to use a prototype of an academic search engine to find and bookmark any research papers according to their interests, which would guarantee that each participants' list of interesting research papers could be recorded. Normalized Discounted Cumulative Gain (NDCG) is used as a metric to determine the performance of the personalized rankings. The experiments suggest that the personalized rankings outperform the original search rankings. Hence, the proposed academic search engine with personalized ranking benefits research paper discovery.

**Keywords:** Personalized Ranking, Research Paper Search Engine, Academic Search Engine.

## 1. INTRODUCTION

Although the Internet and World Wide Web (WWW) provide a new and convenient way to store and disseminate information, a rapid increase in information on WWW makes it difficult to locate pieces of information that are of interest to users. With an information overload problem, search engines have become indispensable tools to help WWW users discover the information they need. One way to improve the quality of users' search experience is to enhance search engine capabilities. Modern search engines, especially web search engines, adopt several techniques to find additional metadata to improve resource indexing and rankings of search results (Choochaiwattana 2009).

The ranking of search results then becomes a challenging task when users obtain a large number of returned search results. Research over the past decade has been concerned with the improvement of resource indexing. For enhancing the performance of rankings, additional metadata information, for instance, document title, anchor text (Brin and Page 1998; Craswell, Hawking and Robertson 2001; Eiron and McCurley 2003) , and user query log (Xue *et al.* 2004), have been used.

Given the growing number of resources returned with high similarity, various approaches to ranking the results have been examined. The similarity ranking approach measures a match between query terms and resource content. On the other hand, the static ranking measures the quality of the resource content, such as PageRank (Page *et al*. 1999) and fRank (Richarson, Prakash and Brill 2006).

Community-based research paper sharing systems, such as CiteULike, and BibSonomy, and academic search engines, such as Google Scholar, Microsoft Academic, ResearchGate, and Social Science Research network (SSRN), have become popular for researchers to discover any research paper in their fields of expertise, and related fields, according to their interests (Choochaiwattana 2010; Khabsa, Wu and Giles 2016). Retrieving academic content from this kind of system adopts a similar technology to previous research and theory in the field of information retrieval (Sanderson and Croft 2012).

With an exponentially increase in published research papers, a group of researchers examined various aspects of improving research paper searches. They have been working on proposing mechanisms for research paper indexing with social tagging (Jomsri, Sanguansintukul and Choochaiwattana 2009a; Jomsri, Sanguansintukul and Choochaiwattana 2009b; Noël and Beale 2008; Vig, Sen and Riedl 2009), developing research paper recommendation services using various techniques (Bogers, and Bosch 2008; Küçüktunç *et al.* 2013; McNee *et al.* 2002; Yin, Zhang, and Li 2007; Zhang, Wang, and Li 2008), and analyzing user behavior of an academic search engine, and gaining a better understanding of how academic search engines work (Khabsa, Wu and Giles 2016; Ishita, Agata, Ikeuchi and Yosuke 2010; Tang and Miner 2016).

*Address of correspondence to this author at the College of Creative Design and Entertainment Technology, Dhurakij Pundit University, Bangkok, Thailand; Tel: +662-954-7300 Ext. 786; Fax: +662-954-8651;
E-mail: worasit.cha@dpu.ac.th

**JEL:** D83, I23, O31.

 

This paper proposes an academic search engine for personalized rankings and investigates the contribution of personalized ranking to the task of re-ranking research paper search results. It is organized as follows: Section 2 provides details on a proposed personalized ranking. An experimental setting and evaluation are described in Section 3. Section 4 analyzes the results of the experiment, and provides a discussion. The conclusion and future research are described in Section 5.

## 2. PROPOSED PERSONALIZED RANKINGS

Typically, a search engine consists of five main components, which are a crawler, an index engine, a search application, a ranking engine, and an evaluation engine (Croft, Metzler, and Strohman, 2009). Each component performs different functions and has different responsibilities (Baeza-Yates and Ribeiro-Neto 2011). The crawler is responsible for identifying and downloading documents for the search engine. The index engine is responsible for extracting content from the downloaded documents in the document corpus and preparing document indexes, which represent the document and provide a more efficient and effective way for retrieving the documents.

In addition, the search application is a primary interface interacting with the search engine users. It is responsible for query preparing and search result displaying. The search application submits the received query to the ranking engine and then displays the search results. The ranking engine is responsible for query processing and search result ranking. It takes the user query and compares with the document index for similarity measurements. The evaluation engine is responsible for monitoring and recording interactions between the users and the search results. The search engine can use the record interaction information for tuning a search result ranking algorithm. A profile engine, laying between the evaluation engine and the ranking engine, is responsible for creating users' profile from the recorded interaction information in the log data and working with the ranking engine to perform the personalized search result ranking as illustrated in Figure **1**.

In general, the personalized ranking mechanisms consider user-paper interactions for creating each user profile. The extract keywords will be put in a set of user's keywords, which represents research interests for each user. To implement the mechanism, there are five main components – set of users, set of interactions with research paper, set of users' keywords, similarity measurement, and document corpus as illustrated in Figure **2**.

Let $N_u$ be the number of users and $N_p$ be the number of research papers in an academic search engine. Let $U$ be a set of users that contains all the users in the academic search engine; $U = \{u_1, u_2, u_3, \ldots, u_n\}$; P is a set of research papers that contains all the research papers in the document corpus, $P = \{p_1, p_2, p_3, \ldots, p_m\}$; and K is a set of keywords and contains all keywords associated with research papers, $K = \{k_1, k_2, k_3, \ldots, k_p\}$. Let $M_{up}$ be the $N_u \times N_p$ association matrix between users and research papers: $M_{up}(u_x,p_y)$ will be equal to 1 when user, $u_x$, interacts with a research paper, $p_y$. Thus, each row, or $UP_i$ in $M_{up}$, represents user interactions with research papers. In addition, for each user $u_x$, let $UKP_x$ be a set of user keywords that

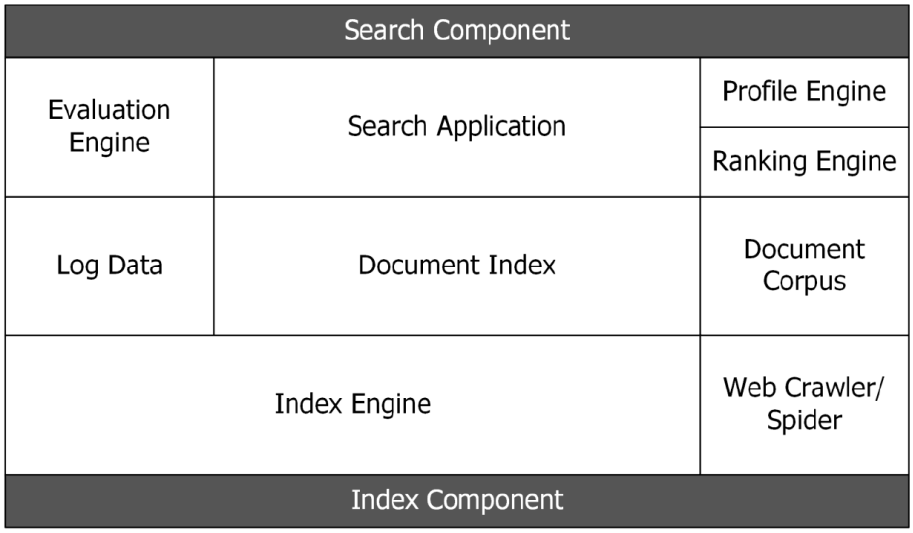


**Figure 1:** Proposed academic search engine with personalized search result ranking mechanism.

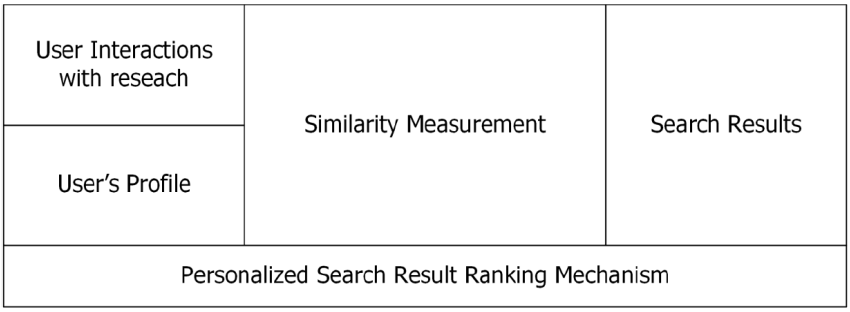


**Figure 2:** Concept of personalized search result ranking mechanism.

are derived from $M_{up}$, $UKP_x = \{< u_x, k_p > \left| u_x \in U \wedge k_p \in K \wedge M_{up}(u_x, p_y) = 1\right\}$.

The proposed rankings in this paper extends the original mechanism proposed in (Choochaiwattana 2016) by introducing user keyword discount score. Let $HLUK_x$ be a set of user keywords discount score that derive from $UKP_x$,

$$HLUK_x = \{< u_x, d_k >: \forall UP_x \in M_{up} \left| \begin{Bmatrix} \forall k_p \wedge < u_x, k_p > \in UKP_x \rightarrow AddUp(d_k) \\ \forall k_p \wedge < u_x, k_p > \notin UKP_x \rightarrow LowerDown(d_k) \end{Bmatrix} \right\}$$

When user $u_x$ submits query $q$ to the academic search engine, a similarity measurement between query $q$ and all research papers in $P$ will be computed, as show in equation (1):

$$Sim(q, p_i) = \frac{\sum_{j=1}^{t} q_j \cdot p_{ij}}{\sqrt{\sum_{j=1}^{t} q_j^{\,2} \cdot \sum_{j=1}^{t} p_{ij}^{\,2}}}. \quad (1)$$

The top 45 resources will be placed in the search result set, SRS = {$sr_1$, $sr_2$, $sr_3$,…, $sr_{45}$}. A keyword vector of user, $u_x$, extracted from $UKP_x$, $kwu_x$, will be multiplied by user keyword discount score, $HLUK_x$., $AdjustedKwu_x$. Then, a keyword vector of each search result, $ksr_y$, and $AdjustedKwu_x$ will be compared with the compute similarity score, as given in equation (2):

$$PSim(Adjusted\ Kwu_x, ksr_y) = \frac{\sum_{j=1}^{t} Adjusted\ Kwu_{xj} . ksr_{yj}}{\sqrt{\sum_{j=1}^{t} Adjusted\ Kwu_{xj}^{\,2} . \sum_{j=1}^{t} ksr_{yj}^{\,2}}} \quad (2)$$

In order to rank the search results, a personalized ranking score will be computed, as given in equation (3):

$$PRank = \alpha \cdot Sim(q, p_i) + (1 - \alpha) \cdot PSim(Adjusted\ Kwu_x, ksr_y) \quad (3)$$

The value of $\alpha$ for this particular study is 0.5. This means that the score of *Sim* and *PSim* are weight equally.

## 3. EXPERIMENT AND EVALUATION

### 3.1. Data from Microsoft Academic

Data was crawled from Microsoft Academic during June to November 2015. The crawler searched for research papers in the field of Computer Science. The final data set of research papers consisted of 71,828 records and 43,508 unique keywords. Each record of crawled research paper contained the title of the research paper, author information, keywords, references, and citation information.

### 3.2. Evaluation Metric

In order to evaluate the performance of the proposed personalized search result ranking mechanism, the Normalized Discounted Cumulative Gain (NDCG), originally proposed by Järvelin and Kekäläinen (Järvelin and Kekäläinen 2000), was used as a metric. The NDCG is devised specifically for web search evaluation, and is based on human judgments, where the human judge rates the relevance of each retrieval result on an n-point scale. For a given query, $q$, the ranked results are evaluated from the top rank down, and NDCG is calculated as in equation (4):

$$NDCG_q = M_q \sum_{j=1}^{K} \frac{(2^{r(j)} - 1)}{\log(1 + j)} \quad (4)$$

where each $r(j)$ is an integer representing the relevance rated by users, and $M_q$ is a normalization constant calculated so that a perfect ordering would obtain an NDCG value of 1.

The NDCG rewards relevant search results in the top rank more heavily than those ranked lower, and punishes irrelevant search results by reducing their contributions to NDCG.

### 3.3. Experimental Setting

Thirty-five graduate students from the Department of Web Engineering and Mobile Application

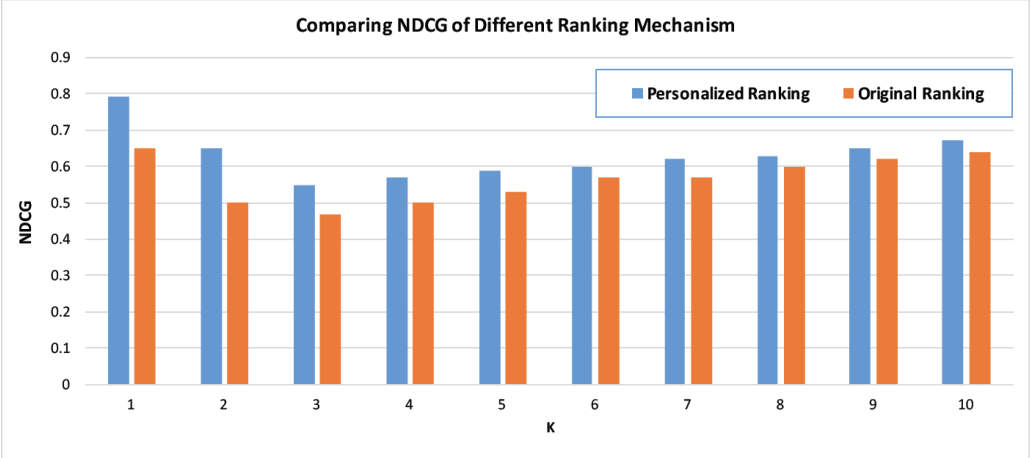


**Figure 3:** Comparison between NDCG for proposed personalized search result ranking and original ranking.

Development at Dhurakij Pundit University were invited to be participants in the research experiment. The participants were asked to use a prototype of an academic search engine to find and bookmark any research papers according to their interests, which would guarantee that each participants' list of interesting research papers could be recorded.

During the experiment, each participant was asked to search for research papers according to their interests. In order to obtain the search result list, the prototype of academic search engine was queried. Cosine similarity was used as a similarity ranking, and the top 45 resources were placed in the search result set. The personalized ranking scores of the top 45 resources were computed, and the personalized search result ranking was obtained and recorded.

The search result set was displayed in randomized order to each participant for the relevance rating. Before each participant rated relevance, they were informed that the results would be displayed in a random order. The ratings provided by each participant were then associated with the original list, and the NDCG scores were computed.

## 4. DISCUSSION

An assessment of the proposed personalized search result ranking mechanism performance was achieved by examining the NDCG values. The closer is the NDCG value to 1.0, the more effective is the search result ranking. Figure **3** shows a comparison between NDCG for the proposed personalized search result rankings and the original rankings, which suggests that the proposed personalized search result rankings provides a better set of search results as compared with the latter.

It seems that a profile derived from a list of bookmarked research papers and a set of keyword discount score can represent each individual's interest. Re-ranking the original search results according to each individual's interests and place more interesting search results in the top rank and less interesting search results in the lower rank. The proposed mechanism can be applied not only to improve research paper searches, but also to improve others resource retrievals, such as news and articles.

## 5. CONCLUSION AND FUTURE RESEARCH

This paper proposed an academic search engine for personalized ranking. It examined an issue in integrating user profiles to improve user satisfaction with the search results. The rankings based on individual preferences create more satisfied search engine users, with user profiles contributing to the task of personalized search result rankings. It can place more interesting search results in the top rank and less interesting search results in the lower rank.

The proposed personalized search result ranking mechanism was developed under the assumption that users tend to have limited fields of research areas and their research interests may change as time passes. In this respect, further analysis should be performed. It appears that user interest may change, in which case a technique to monitor user interests needs to be investigated.

## ACKNOWLEDGEMENT

The author is grateful to Chia-Lin Chang and Michael McAleer for helpful comments and suggestions.

## REFERENCES


Baeza-Yates, Ricardo and Berthier Ribeiro-Neto. 2011. *Modern Information Retrieval: The Concepts and Technology Behind Search Engine*. 2nd ed. Addison-Wesley Professional.

Bogers Toine and Antal van den Bosch. 2008. "Recommending Scientific Articles Using CiteULike." *Proceedings of the 2008 ACM Conference on Recommender Systems*. pp. 287-290.

Brin, Sergey and Lawrence Page. 1998. "The Anatomy of a Large-Scale Hypertextual Web Search Engine." *Journal of Computer Networks and ISDN Systems*. 30(1-7):107-117.

Choochaiwattana, Worasit and Michael B. Spring. 2009. "Applying Social Annotations to Retrieve and Re-rank Web Resources." *Proceedings of the International Conference on Information Management and Engineering*. pp. 215-219. https://doi.org/10.1109/icime.2009.41

Choochaiwattana Worasit. 2010. "Usage of Tagging for Research Paper Recommendation." *Proceedings of the International Conference on Advanced Computer Theory and Engineering*. pp. 439-442.

Craswell, Nick, David Hawking, and Stephen Robertson. 2001. "Effective Site Finding Using Link Anchor Information." *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 250-257. https://doi.org/10.1145/383952.383999

Croft, Bruce, Donald Metzler, and Trevor Strohman. 2009. *Search Engines: Information Retrieval in Practice*. 1st ed. Pearson Education.

Eiron, Nadav and Kevin S. McCurley. 2003. "Analysis of Anchor Text for Web Search." *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 459-460. https://doi.org/10.1145/860435.860550

Ishita, Emi, Teru Agata, Atsushi Ikeuchi, and Miyata Yosuke. 2010. "A Search Engine for Japanese Academic Papers." *Proceedings of the 10th ACM/IEEE-CS Joint Conference on Digital Libraries*. pp. 379. https://doi.org/10.1145/1816123.1816189

Järvelin, Kalervo and Jaana Kekäläinen. 2000. "IR Evaluation Methods for Retrieving Highly Relevant Documents." *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 41-48.

Jomsri, Pijitra, Siripan Sanguansintukul, and Worasit. Choochaiwattana. 2009a "A Comparison of Search Engine Using 'Tag Title and Abstract' with CiteULike - An Initial Evaluation." *Proceedings of the International Conference on Internet TechnologJy and Secured Transactions*. pp. 1-5.

Jomsri, Pijitra, Siripan Sanguansintukul, and Worasit Choochaiwattana. 2009b. "Improving Research Paper Searching with Social Tagging: A Preliminary Investigation." *Proceedings of the 8th International Symposium on Natural Language Processing*. pp. 152-156.

Khabsa, Madian, Zhaohui Wu, and C. Lee Giles. 2016. "Towards Better Understanding of Academic Search." *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*. pp. 111-114. https://doi.org/10.1145/2910896.2910922

Küçüktunç, Onur, Erik Saule, Kamer Kaya, and Ümit V. Çatalyürek. 2013. "TheAdvisor: a webservice for academic recommendation." *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*. pp. 433-434. https://doi.org/10.1145/2467696.2467752

McNee, Sean M., Istvan Albert, Dan Cosley, Prateep Gopalkrishnan, Shyong K. Lam, Al Mamunur Rashid, Joseph A. Konstan ,and John Riedl. 2002. "On the Recommending of Citations for Research Papers." *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*. pp. 116-125. https://doi.org/10.1145/587078.587096

Noël, Sylvie and Russell Beale. 2008. "Sharing vocabularies: Tag usage in CiteULike." *Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction*. pp. 71-74.

Page, Lawrence, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999 "*The Pagerank Citation Ranking: Bringing Order to the Web*." Technical Report 1999-66, Stanford University.

Richarson, Matthew, Amit Prakash, and Eric Brill. 2006. "Beyond PageRank: machine learning for static ranking." *Proceedings of the 15th International Conference on World Wide Web*. pp. 707-715.

Sanderson, Mark and W. Bruce Croft. 2012. "The History of Information Retrieval Research." *Proceedings of the IEEE 100*. pp. 1444 -1451. https://doi.org/10.1109/JPROC.2012.2189916

Tang Jie. 2016. "AMiner: Toward Understanding Big Scholar Data." *Proceedings of the 9th ACM International Conference on Web Search and Data Mining*. pp. 467.

Vig, Jesse, Shilad Sen, and John Riedl. 2009. "Tagsplanations: Explaining Recommendations Using Tags." *Proceedings of the 14th International Conference on Intelligent User Interfaces*. pp. 47-56.

Xue, Gui-Rong, Hua-Jun. Zeng, Zheng Chen, Yong Yu, Wei-Ying. Ma, WenSi Xi, and WeiGuo Fan. 2004. "Optimizing Web Search Using Web Click-through Data." *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*. pp. 118-126. https://doi.org/10.1145/1031171.1031192

Yin, Pin, Ming Zhang, and Xiaoming Li. 2007. "Recommending Scientific Literatures in a Collaborative Tagging Environment." *Proceedings of the 10th International Conference on Asian Digital Libraries: Looking Back 10 Years and Forging New Frontiers*. pp. 478-481. https://doi.org/10.1007/978-3-540-77094-7_60

Zhang, Ming, Weichun Wang, and Xiaoming Li. 2008. "A Paper Recommender for Scientific Literatures Based on Semantic Concept Similarity." *Proceedings of the 11th International Conference on Asian Digital Libraries: Universal and Ubiquitous Access to Information*. pp. 359-362. https://doi.org/10.1007/978-3-540-89533-6_44


Received on 16-02-2017 Accepted on 13-05-2017 Published on 09-06-2017

DOI: https://doi.org/10.6000/1929-7092.2017.06.36