# Research on Personalized Search Engine Technology

**Authors:** GU Liping

**Date:** 2016-08-01T00:00:00+00:00

## Abstract

Personalized search engines constitute a user-driven optimization approach for web page ranking results. Based on ontology and the semantic web, user modeling can produce accurate query results through three mechanisms: limiting search modalities, filtering search results, and participating in the search process. Therefore, the user model for personalized search engines can be regarded as a model for user-driven personalized search services. The research conclusion integrates previous studies and proposes#2; a new model of user behavior (user interests, user preferences, user query records) - user profile (user behavior and keyword groups) - user modeling (relevance algorithms and ranking algorithms) - personalized service, which can serve as a guideline for digital libraries to develop personalized search engines.

## Full Text

## Preamble

**Research on Personalized Search Engine Technology**
(Department of Library and Information Science, National Taiwan University, Taipei 100671, Taiwan)

**Abstract:** Personalized search engines represent a user-driven approach to optimizing web page ranking results. Based on ontologies and the semantic web, user modeling can produce accurate query results through three approaches: constraining search methods, filtering search results, and becoming part of the search process itself. Therefore, personalized search engine user models can be viewed as models for user-driven personalized search services. This study integrates previous research and proposes a new model of "user behavior (user interests, user preferences, user query history) → user profile (user behavior and keyword sets) → user modeling (relevance algorithms and ranking algorithms)

→ personalized services," which can serve as a guideline for digital libraries developing personalized search engines.

**Keywords:** information retrieval; information searching; information seeking behavior; user engagement; personalized digital libraries

**Author Biography:** Gu Liping (1978- ), male, from Taipei, Taiwan, postdoctoral research assistant in the Department of Library and Information Science, National Taiwan University. Research direction: decision support systems.

## 1. Technology: Methods for Optimizing Search Engines

### 1.1 User Modeling to Constrain Search Methods

A simple (or direct) approach to implementing personalized search engines involves presetting user interests or preferences before the search begins. When users log into the system, retrieval is conducted within predefined scopes such as subject domains, document types, or publication dates specified by the user. This represents the personalized system model commonly adopted by digital library information retrieval systems. Currently, this approach is not widely used in personalized search engine systems, but it exhibits two important trends worth considering for digital libraries.

The first trend involves integrating user interest forms, preference settings, and web page ranking algorithms to provide personalized search services. The technical approach combines classic flat ranking lists with search engines, allowing users to interactively query through hierarchical folder tags (subjects), enabling knowledge extraction, query optimization, and search result personalization during the browsing process. This service model is similar to personalized digital libraries but places greater emphasis on secondary queries during browsing, further queries based on results, and auxiliary queries integrated with other intelligence analysis systems. It can be considered an evolved version of personalized digital libraries.

As previously mentioned, search engines and database retrieval systems have different inherent conditions and problem-solving models, and current personalized digital library systems differ from personalized search engines accordingly. However, the technology of using user modeling to constrain search methods in personalized search engines is not complex, as its underlying technique simply involves adding system-preset queries before the user' s retrieval query. Since search engine queries typically do not require users to input search formulas but only keywords, users perceive this as personalized search. In reality, this technology used by most digital libraries merely hides portions of database retrieval system conditional statements. Nevertheless, in personalized search engines, while the underlying technology is the same, the overlay techniques vary widely, offering valuable lessons for personalized digital libraries.

### 1.2 User Modeling to Filter Search Results

If user modeling that constrains search results weaves user interests and preferences into a fishing net, then user modeling that filters search results functions as a double-layered funnel. The principle is the same: to screen or filter search results, with the former occurring before searching and the latter after searching. However, the underlying technology for the latter is relatively more complex. Currently, this approach is more widely applied in personalized search engine systems and demonstrates two important trends for digital libraries to consider.

The first trend involves establishing user profiles from user behavior and combining them with domain ontologies (associations of keyword sets) to provide personalized search services. The technical approach analyzes user click records, estimates user interests to build ontologies, and uses ontologies to replace vocabulary in users' current queries. When calculating user interests to optimize the query process, the system must effectively identify user preferences and establish a profile for each user. Once such profiles are available, the system must determine user interest sets among numerous query matching schemes. Therefore, in this model, "user behavior" refers to user interests and preferences. Based on this model, another type of personalized digital library can be developed.

The second trend involves extracting keywords from document content and combining them with user search records to establish user profiles for personalized services. The technical approach identifies relevant query terms from web snippets in search results while using agglomerative clustering algorithms to generate personalized query clusters, enhancing the clustering effectiveness of personalized search engines. Alternatively, using self-organizing map algorithms (SOM) to establish user interest databases after user retrieval, employing text mining methods to optimize differentiated results in personalized search. In this model, where search engines suggest semantically related query terms, users can select search vocabulary that reflects their information needs.

A simple comparison reveals that in user modeling that constrains search results, users preset search formulas whose front-end portions are hidden by the information system. If personalized search fails to provide needed information, users must either acknowledge that their original settings were imperfect or select all "user interests" and cancel all "user preferences" (effectively abandoning personalized search) to obtain relevant information. This model in personalized digital libraries leaves information-seeking users "silent victims" who cannot voice their complaints. However, in user modeling that filters search results, user modeling sets the back-end search formulas, and the system automatically performs secondary retrieval after the user' s search, hiding this back-end portion. Consequently, users do not face a choice of "whether to personalize" but enter a process of "already personalized for you." In a sense, this represents a "Don' t be evil" approach, where the personalized search system assumes responsibility for users' inability to find information rather than shifting blame to end users.

### 1.3 User Modeling as the Search Process

User modeling can serve as both fishing net and funnel in search engines, functioning to preset queries before user retrieval and automatically perform secondary retrieval (with relevance recommendations) afterward. User modeling can also become a Rubik' s cube, optimizing the matching of multiple search results during user retrieval. Its underlying technology is more complex than the previous two approaches, building upon their search results and technical methods while following fundamentally different technical routes. It exhibits two important trends that represent essential references for next-generation personalized digital libraries.

The first trend involves user modeling techniques derived from artificial intelligence applications. The technical approach uses genetic programming (GP) learning machine technology based on evolutionary theory to optimize document weights in vector space, achieving personalized web search ranking from individual queries to different ranking results. Alternatively, using fuzzy sets and fuzzy logic to score user satisfaction and optimize search. Whether genetic algorithms or fuzzy logic, the underlying data comes from user interests, preferences, and queries. Developing user modeling based on user behavior and transforming it into user profiles to establish personalized services represents a developing trend.

The second trend involves applying user profiles in information retrieval systems and web search engines. The technical approach dynamically structures user profiles (establishing relevant phrase sets for user interests) based on observed user behavior and actions, for use in extended query functions of information retrieval systems to alter search engine ranking order. The focus of this technical route is not to have user modeling screen and filter search results but to change search results themselves. In this model, user interests, preferences, query histories, and related phrases in user profiles constantly change, with user profiles participating in web page ranking and document relevance ranking.

User modeling as the search process offers many possibilities and represents the primary trend in future research on personalized search engines and even search engines in general. Its enormous potential lies in non-traditional user engagement, which has not yet fully manifested in search engine services or personalized digital libraries.

## 2. Application: Optimizing Digital Library Retrieval Systems

Researchers used 500 terms to query four search engines—Google, Yahoo, Live, and Ask—and analyzed the results among 42,758 hits, finding that Google and Yahoo prefer to cite their own services (such as YouTube and Yahoo Answers). Digital libraries do not have similar problems. However, traditional personalized digital libraries employ only one of the three personalized search engine

technologies and tend to approach personalized services from the perspective of database retrieval systems rather than web search engines.

Adopting the second perspective can enrich digital libraries' information organization and retrieval. For example, can blogs and microblogs in the medical field be considered medical resources and information resources for digital libraries? Research shows that patients and nurses describe their lives while doctors publish healthcare-related information on blogs, and these content differences can be leveraged by search engines for ranking improvements to enable user models to search for appropriate knowledge sources. Thus, information services supporting medical teams require personalized search engines in digital libraries.

E-Service includes four modes: collaboration, customization, integration, and adaptation. The spirit of personalized service is that individuals can contribute and receive customized or personalized information recommendations in a collaborative environment, obtaining timely or just-in-time support through an integrated system or process. This requires digital library personalized search engines to provide precise search results to save end users time in information seeking behavior, allowing them to use that time for other work. Personalized services have never been limited to the information provision aspect of personalized digital libraries but extend throughout end users' entire workflow. Research shows that basic science researchers typically search databases or web engines using keywords without integrating library resources or services into their work, suggesting that library resources should be accessible through their professional websites, personnel relationships with key administrative departments should be cultivated, and campus academic information should be centralized and managed in institutional repositories.

Currently, various approaches have been used to build new digital library systems. For example, manually editing user interests into text classification trainers, personalized catalog systems combining user interests and classification catalogs are faster and easier for discovering relevant information than categorization systems (CAT) and list interface systems (LIST). Another example involves building Arabic and English product catalog retrieval systems using ontologies (requiring bilingual ontologies to optimize search engines due to different natural languages). Yet another example involves establishing fuzzy concept network archival retrieval systems based on user profiles to provide personalized web pages and related documents according to user preferences. These studies demonstrate the importance of user models for digital libraries.

Users' inconsistent relevance judgments, rankings, and relevance criteria change the evaluation of personalized search systems, particularly in measuring and estimating the randomness of ranking similarity and relevance criteria. Based on this theory, developing a new model of "user behavior → user profile → user modeling → personalized service" becomes necessary.

When digital libraries develop personalized search engines, first, the search engine must effectively identify user interests and establish a profile for each in-

dividual user. Second, once such profiles are available, the search engine must match results with a ranking approach that aligns with a given user' s interests. However, users will not actively express personal preferences, so it is essential to leverage users' historical behavior records to mine possible patterns in user behavior and establish user profiles. Third, ontologies capable of semantic approximate reasoning must be built based on their past query records, i.e., keyword terms.

In this process, user profiling is the foundational element of personalized applications. Many user profiles are built on what users are interested in rather than what they are not interested in. Through personalized query clustering methods, testing optimization strategies for users' positive and negative preferences can utilize agglomerative clustering algorithms to optimize personalized query results.

Agglomerative clustering algorithms are essentially grouping algorithms that initially treat each point as a cluster, then repeatedly merge points through some distance measurement until all points are aggregated. Further extensions of this technology combine with creating and using personas to determine user groups from user behavior and recommend information based on user groups and current user query terms.

In other words, the establishment of all next-generation digital library systems revolves around the model of "user behavior → user profile → user modeling → personalized services."

## 3. User-Driven Personalized Search Services

Following the emergence of search engines such as Yahoo, Google, and Bing, people have always been delighted initially but soon become dissatisfied with service models like subject content (categorical) retrieval, popular statistical ranking retrieval, and similar content (clustering) retrieval. People want these technologies to serve them while also wanting their search results to belong to themselves rather than the masses. This desire has given rise to the personalized search engine service model. It integrates all users' search processes, including queries, browsing, clicks, and dwell time, as the basis for analyzing user behavior models. Simultaneously, it derives relevant documents, excludes non-relevant documents, and performs web page or document ranking from these user models, providing personalized search content according to differences in user interests and preferences within user profiles. In short, it represents a model of "from integrated information to differentiated services."

This model is not difficult to find in commercial activities. In the past, most B2C e-commerce system search engines only allowed users to search for product numbers, types, and prices, neglecting the use of agent technology to participate in transactions between buyers and sellers. Researchers suggest that employing agent technology can comprehensively consider factors such as price, quantity, brand, packaging, and delivery time to optimize search engines, achiev-

ing personalized recommendations of the most suitable products to current users through repeated transactions and retrieval processes. This technology can serve as a reference for digital libraries.

This paper systematically reviews three user modeling approaches for optimizing search engines: constraining search methods, filtering search results, and becoming the search process itself. For optimizing digital library retrieval systems, it proposes a model of "user behavior (user interests, user preferences, user query history) → user profile (user behavior and keyword sets) → user modeling (relevance algorithms and ranking algorithms) → personalized services."

Among these, user modeling as the search process still has much research potential. If user intentions can be better utilized, text snippet extraction can be generalized, such as using statistical language models to capture commonalities between documents and user intentions. Meanwhile, employing instance algorithms similar to PageRank (InstanceRank) can reduce instance set size by selecting the most representative instances from learning libraries. Regarding future research directions for extending personalized search engines, studies show that social media websites constitute a significant portion of search results, indicating that social media should be part of personalized search engines. Additionally, using continuous principal component analysis (CPCA) for Fourier series calculations to obtain translation, rotation, flipping, and scale of three-dimensional models enables search engines to search 3D model databases using shape similarity. This suggests that virtual societies can be part of personalized search engines. These research directions warrant subsequent tracking and exploration.

## References

[1] PAGE L, BRIN S, MOTWANI R, WINOGRAD T. The pagerank citation ranking: bringing order to the Web (1999). [ED/OL] [2010-10-27] http://ilpubs.stanford.edu:8090/422/.

[2] KIM H, CHAN PK. Personalized search results with user interest hierarchies learnt from bookmarks [J]. Advances in Web mining and Web usage analysis, 2006.

[3] JIANG X, TAN AH. Learning and inferencing in user ontology for personalized Semantic Web search [J]. Information sciences, 2009(16).

[4] KEY P, DENG L, NG W, LEE DL. Web dynamics and their ramifications for the development of Web search engines [J]. Computer networks, 2006(10).

[5] CAMBAZOGLU BB, KARACA E, KUCUKYILMAZ T, TURK A, AYKANAT C. Architecture of a grid-enabled Web search engine [J]. Information processing and management, 2007(3).

[6] STEGERS R, FEKKES P, STUCKENSCHMIDT H. MusiDB—A personalized search engine for music [J]. Journal of Web semantics, 2006(4).

[7] BAR-LLAN J, KEENOY K, YAARI E, LEVENE M. User rankings of search engine results [J]. Journal of the American society for information science and technology, 2007(9).

[8] FERRAGINA P, GULLI A. A personalized search engine based on Web-snippet hierarchical clustering [J]. Software—practice & Experience, 2008(2).

[9] STAMOU S, KOZANIDIS L, TZEKOU P, ZOTOS N. Ontology-driven personalized query refinement [J]. Journal of Web engineering, 2009(2).

[10] HONG JL, SIEW EG, EGERTON S. Information extraction for search engines using fast heuristic techniques [J]. DATA & KNOWLEDGE ENGINEERING, 2010(2).

[11] LEUNG KW T, NG W, LEE DL. Personalized concept-based clustering of search engine queries [J]. IEEE transactions on knowledge and data engineering, 2008(11).

[12] HUNG CL, CHI YL, CHEN TY. attentive self-organizing neural model for text mining [J]. Expert systems with applications, 2009(3).

[13] FAN WG, PATHAK P, WALLACE L. Nonlinear ranking function representations in genetic programming-based ranking discovery for personalized search [J]. Decision support systems, 2006(3).

[14] GURSKY P, HORVATH T, JIRASEK J, KRAJCI S, NOVOTNY R, PRIBOLOVA J, VANEKOVA V, VOJTAS P. User preference Web search –experiments with a system connecting Web and user [J]. Computing and informatics, 2009(4).

[15] KUMAR H, KANG S. Exclusively Yours: Dynamic Individuate Search by Extending User Profile [J]. New generation computing, 2010(1).

(Editor: Zhou Xiaohui) 2011

*Note: Figure translations are in progress. See original paper for figures.*