情報処理学会
グループウェアとネットワークサービス研究会
2006年11月16日, 17日

Towards Better Search and Sharing: the Growing Role of Relevance Feedback for Personalized Web Search

Roman Y. Shtykh and Qun Jin
{roman@akane., jin@}waseda.jp
Networked Information Systems Laboratory, Faculty of Human Sciences,
Waseda University

Abstract

We are living in the era of fast information growth when we are hardly able to self-orientate in the volume of information surrounding us and manage it. In the situation of huge information overflow that we are experiencing already now and will be experiencing sharply in the future we are trying to better understand how to treat information around us looking for new methods of organizing new information pieces or retrieving slices of information best matching our current needs.

In this paper we are examining the growing role of relevance feedback to improve information retrieval in Web search systems and representing a shift to its active use in BESS (BEtter Search and Sharing) collaborative search system.

Keywords: Relevance feedback, information retrieval, collaborative search and sharing.

パーソナライズドWeb検索における適合性フィードバックの増大する役割 ――Better Search & Sharingの試み――

シュティフ　ロマン　　金　群
{roman@akane., jin@}waseda.jp
早稲田大学大学院人間科学研究科

情報の急増する時代に生きる我々にとって、周囲の莫大な量の情報に対して自らを順応させたり、その情報量を管理したりすることはますます困難になってきている。我々が現在実感している情報氾濫は将来も進行すると予想されるので、周囲の情報の扱い方をより深く理解することは重要であろう。そこで、我々は新しい情報を整理したり、個人のニーズに最もマッチする情報を検索する新たな方法を模索している。

本論文では、情報検索の性能を一層改善させるために、ウェブ検索においてますます大きな役割を果たすようになってきている適合性フィードバックについて論じ、それを積極的に活用するようにした、「より良い情報検索・共有」を提供する仕組みBESS（BEtter Search & Sharing）の進化を示す。

キーワード：適合性フィードバック、情報検索、協同検索・共有

1 Introduction

Judging from the latest technical articles in the Internet and proceedings of conferences related to information retrieval (IR) more and more attention seems to be paid to an individual person as an active evaluator of the shared Web information. For an individual such a role has been typical in groupware systems, but not so big in IR field, particularly in search engine systems and in business, with the exception of some studies on standard feedback techniques like document

rating, reading time measurement, page scrolling, interactions with numerous applications, etc. [1-3].

What is the reason for such a boost of research activities and general public attention to it? Do IR systems face the limits of retrieval capabilities?

Presumably, not. But with the growth of popularity of social networks, that shew again that there is no algorithms or the whole system capable to make decisions better than humans, the role of an individual person grew much stronger. Many got back to consider it seriously.

In the situation of information overflow when decision quality decreases due to overwhelming amount of information [4] there became important new methods of more precise information search – search of the information the users sets as a goal of their search. Therefore many scientists and enterprises turn again to information filtering and relevance feedback in particular. Relevance feedback changes its forms, becomes more complicated and sophisticated - nobody now considers explicit feedback only when talking about relevance feedback, and new attempts to make feedback more natural and precise through combining its explicit and implicit forms are taken.

2 Information filtering for information retrieval

Till recent past most common approaches to rank search results focused on similarity of a query and a page, as well as the overall page quality. For a long time relevance feedback was considered mostly as "the process of modifying the query to improve the effectiveness of the remainder of the search, based upon partial relevance judgments." [5] In iterative retrieval every new query Q_{i+1} was supposed to be closer to "optimal" query than a sequence of queries $<Q_i>$ executed till the current moment.

During last several years the situation has been changing. Recently scientists in IR retrieval field started showing keen interest in feedback for predicting user intentions to rank search results and produce better ranked lists [6-8]. Similar research was conducted in the past as well and was even employed in several production systems. For example, DirectHit algorithm [9] that takes into account the number of clicks on links and time a user spent on the page used to be employed in a number of search systems like DirectHit, Lycos and MSN. Recently we can observe a new increase of activities in this area producing more elaborate algorithms for feedback and becoming closer to real world systems. This can be seen from recent Microsoft Research reports mentioned above and businesses of other Internet business vendors employing such techniques. As an example, Russian Internet advertisement company Begun [10] stated that behavioral advertisement based on watching user actions on certain web sites and search engines brings about the same amount of profit as semantic advertisement. They have been already researching this new direction in advertisement for three years and possess information about the behavior of about 20 million users. To say more, "Begun" system creates user profiles that are dynamically updated according to user actions on specified web sites. The main source of information for profile formation is user search queries leading to advertised sites, user routes from one site/page to another, history of interactions with certain advertisements. Moreover, users are organized into groups and a part of prediction algorithm uses profile similarities to fill the gaps (grey spots) in other users' profiles and do correct predictions. Similar research is being done by Microsoft AdCenter Labs [11]. Using probabilistic approaches they are building user profiles based on page views, searches, and other online behaviors for targeted advertisement. Further, such profiles are clustered and segments of customers with similar interests are created.

Recently Google has introduced a beta-version of personalized search that produces search results ranked with attention to personal search histories [12]. This was a long-planned move preceding by acquirements of such personalization expert companies as Outride Inc. [13] and Kaltix Corp. [14]. Another search engine giant Yahoo! also collects personal histories [15], but we are not aware if they are integrated with the search engine to change result ranks somehow.

As to Microsoft recent researches [6-7] we mentioned above, it is the first research done on real datasets used in Web search systems trying to interpret collective user behavior to predict user preferences for search results. Implicit feedback interpretation is enhanced by modeling query-dependent deviations from the expected "noisy" user behaviors. Currently it is not clear where the results will be used, but probably the algorithms are already used with Windows Live ™ [16] in its Live Search.

3 Evolution of BESS project: Shift from ontologies to active use of relevance feedback

3.1 Better targeting user information needs through user activity analysis

The rapid growth of attention to social networks and their diversity growth, advances in information filtering research made influence upon our research project's original plan [17][1] that sets its goals as improvement of Web search result retrieval through using user feedback and static profiling, and ontologies to determine the scope of user search intentions, the domain of the query. The project initially puts the main emphasis on use of ontologies for improving relevance of search results. Although the ontologies can improve the search significantly by allowing for the search to look for pages on a concrete concept instead of being based on ambiguous and badly specified keywords, they still cannot express all user search intentions and scope of the current search fully taking into account the current information needs (session needs), short-term and long-term information needs (user profile, background, etc.) of a user. To get better idea about these needs and improve personalized search we use such implicit feedback information as clickthrough history, dwelling time for every document, group membership and activities inside the group and explicit feedback through the Web browser toolbar[2] (Figure 1). Though explicit feedback can disrupt search user activities, explicit measures are found to be much more accurate than implicit ones [18].

Figure 1. Explicit feedback toolbar

Tentatively we consider the elements of user-system interactions (which include elements taken at the proxy (Figure 2) and elements derived from those data) to enhance search as shown in Table 1 and 2.

Taking above-mentioned user interactions with the search engine as implicit feedback and combining it with user explicit feedback we are forming dynamic user profiles which reflect user information needs and help us understand current search intentions to better evaluate user contributions (done by explicit feedback) to the system and get better search results. In other words, user profiles are used as the main element of search personalization – both query augmentation and re-ranking of the documents submitted by the user to BESS.

Table 1. Elements of user-system interaction captured at the server

1. Queries	
Query	Search query string
Url	Retrieved url (final url in case of redirection)
Time	Query timestamp
Is_click	Click or not
2. Feedback	
Query	Search query string related to current feedback
Url	Retrieved url
Time	Feedback timestamp

Table 2. Derived data

Document dwell time
Domain dwell time
Number of clicks per query
Number of feedbacks per query
Number of feedbacks by a user
Overall feedback value of a user
Feedback history and their helpfulness
Number of feedbacks by a group
Overall feedback value of a group
Number of feedbacks inside a group
Similarity of feedbacks of other group users

1 The tentative project code is BESS that stands for "BEtter Search and Sharing"

2 Explicit feedback mechanism can be implemented in a different, more application-independent, way

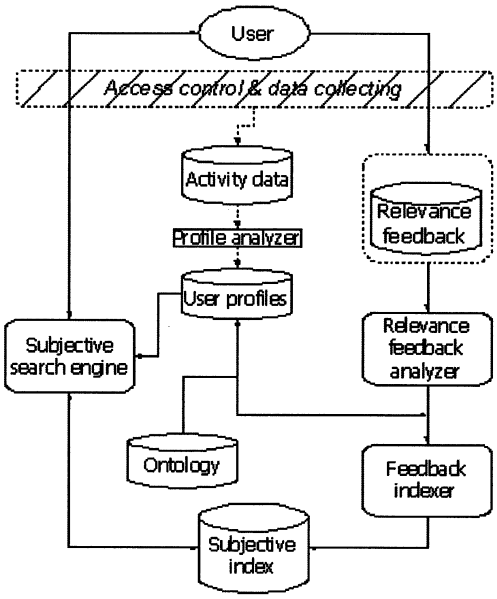

Figure 2. Improving search through use of relevance feedback in BESS

3.2 Collaborative user model

By forming user profiles and grouping dynamically we remove the necessity to create models for groups – group model is constructed from individual user model. Furthermore, the need for user role disappears together with motivation which was presented as an upgrading of user privileges inside the group. Users' weight inside the system is decided judging from their search and after-search activities, cooperation of members are thought to be minimal and we believe that "diversity and independence are important because the best collective decisions are the product of a disagreement and contest, not consensus and compromise....Paradoxically, the best way for a group to be smart is for each person in it to think and act as independently as possible." [19]

Collaborative user model is shown in Figure 3. Generally, profile is an automatically generated instance that is a part of any User and Group but it is presented as a separate instance here to emphasize User-Group relationships - users and groups are not tightly coupled, they have loose relationships only through their profiles. Domain is extracted from Profile with use of ontologies to remove ambiguities or set better the search scope for more precise definition of current user information needs.

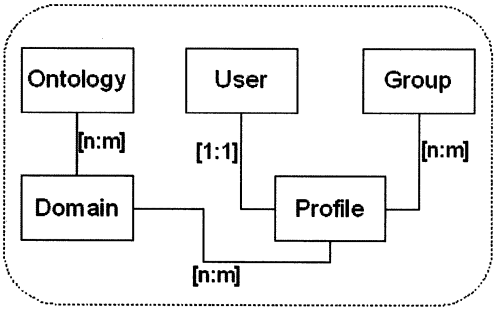

Figure 3. Collaborative user model

4 BESS and main approaches to improve information retrieval

There can be two main approaches to work with data in general, in our case to improve document retrieval in particular, – Data First and Structure First [20]. In our research we don't structure any data in a system-specific way from the beginning to make them easily searchable but we try to enhance search results any other search engine provides by analyzing user activities. Hence, we follow Data First – Structure Later approach.

Semantic Web [21] can be a good example of an alternative approach – Structure First. This initiative intends to wrap Web information with meaning, in this way creating a self-descriptive environment and enhancing the precision of search. Semantic Web is supported by a number of projects like IEML (Information Economy Meta Language) [22], for instance, that formulates a new alphabet describing the subtleties of all natural languages in order to create a more intuitive information index, enabling searching by concept instead of vague keywords. Another example is RDF (Resource Description Framework) [23] – a W3C specification and a widely used knowledge modeling approach describing data in collections as subject-object-predicate triples – and RDF query languages. Such initiatives makes information handling very easy and efficient, but quite complicated for an ordinary Web user.

5 Summary

In the paper we pointed at the growing interest to implicit feedback techniques in Information

Retrieval, increasing sophistication of filtering algorithms and keen interest of Internet businesses to this area. Also we briefly explained about the shift of our main approach in BESS collaborative search project from ontologies to captured and analyzed user activities to increase the precision of document retrieval in the Web search system.

At the current stage of design and development we don't integrate subjective index data (the data that is based on user contributions and posessed by the system) and objective index data of conventional search engine (This can be a topic for further research though). Currently we search only on the data submitted by registered users of the system.

References

[1] M. Morita and Y. Shinoda. Information filtering based on user behavior analysis and best match text retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '94)*, 1994, pp. 272-281

[2] R. W. White, I. Ruthven, and J. M. Jose. Finding relevant documents using top ranking sentences: An evaluation of two alternative schemes. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '02)*, 2002, pp. 57-64

[3] J. Budzik and K. Hammond. Watson: Anticipating and contextualizing information needs. In *Proceedings of the 62nd Meeting of the American Society for Information Science*, 1999, pp. 727-740

[4] Morville, P. Ambient Findability. O'Reilly Media, 1st Ed., 2005, pp. 165-166

[5] I. A. Witten, A. Moffat, and T. C. Bell. Managing gigabytes: compressing and indexing documents and images, 2nd Ed., p.214

[6] E. Agichtein, E. Brill, and S. Dumais. Improving Web Search Ranking by Incorporating User Behavior. In *Proceedings of the ACM Conference on Research and Development on Information Retrieval* (SIGIR), 2006

[7] E. Agichtein, E. Brill, S. Dumais, R. Ragno. Learning user interaction models for predicting web search result preferences. In *Proceedings of the ACM Conference on Research and Development on Information Retrieval* (SIGIR), 2006

[8] X. Shen, B. Tan, and C. Zhai. Context-sensitive information retrieval with implicit feedback. In *SIGIR '05*, pp. 43–50.

[9] DirectHit, http://www.searchengines.com/directhit.html

[10] Begun, http://begun.ru/

[11] Microsoft AdCenter Labs, http://adlab.microsoft.com/

[12] Google Personalized Search, http://labs.google.com/personalized/

[13] Google Press Center, Google Acquires Technology Assets of Outride Inc., http://www.google.com/press/pressrel/outride.html

[14] Google Press Center, Google Acquires Kaltix Corp., http://www.google.com/press/pressrel/kaltix.html

[15] Yahoo! MyWeb, http://myweb.yahoo.com/

[16] Windows Live ™, http://www.live.com/

[17] S. Chan and Q. Jin. Collaboratively Shared Information Retrieval Model for e-Learning. In *Proceedings of ICWL2006*, 2006

[18] D. M. Nichols. Implicit ratings and filtering. In *Proceedings of the 5th DELOS Workshop on Filtering and Collaborative Filtering*, pp.31-36, 1997

[19] J. Surowiecki. The Wisdom of Crowds, Anchor; Reprint edition (August 16, 2005), pp. xix-xx.

[20] Stephano's Linotype, http://www.betaversion.org/~stefano/linotype/news/93/

[21] Semantic Web, http://www.w3.org/2001/sw/

[22] IEML, http://www.ieml.org

[23] Resource Description Framework (RDF), http://www.w3.org/RDF/