Responding to Retrieval: A Proposal to Use Retrieval Information for Better Presentation of Website Content

C. Ravindranath Chowdary(✉), Anil Kumar Singh, and Anil Nelakanti

IIT (BHU), Varanasi, India
{rchowdary.cse,aksingh.cse,anil.nelakanti.cse}@iitbhu.ac.in

Abstract. Retrieval and content management are assumed to be mutually exclusive. In this paper we suggest that they need not be so. In the usual information retrieval scenario, some information about queries leading to a website (due to 'hits' or 'visits') is available to the server administrator of the concerned website. This information can be used to better present the content on the website. Further, we suggest that some more information can be shared by the retrieval system with the content provider. This will enable the content provider (any website) to have a more dynamic presentation of the content that is in tune with the query trends, without violating the privacy of the querying user. The result will be a better synchronization between retrieval systems and content providers, with the purpose of improving the user's web search experience. This will also give the content provider a say in this process, given that the content provider is the one who knows much more about the content than the retrieval system. It also means that the content presentation may change in response to a query. In the end, the user will be able to find the relevant content more easily and quickly. All this can be made subject to the condition that user's consent is available.

Keywords: Content management systems · Personalized content · Information sharing

1 Introduction

Information retrieval (IR) systems have become integral to daily activities of millions and will retain their prominence in years to come. One of the reasons for such importance of a good IR system is the amount of data that is available on the web and the pace at which it is increasing. The number of websites reportedly increased from one in 1991 to more than one billion in September 2014[1]. Simultaneously, there was an increasing number of users availing hosted services. This increase in web usage is more than an issue of load, that was met by computationally powerful servers. The bigger challenge was to organize and

[1] http://www.internetlivestats.com/total-number-of-websites/ on 27/10/2014.

F. Daniel and O. Diaz (Eds.): ICWE 2015 Workshops, LNCS 9396, pp. 103–114, 2015.
DOI: 10.1007/978-3-319-24800-4_9

make available the huge amount of information in a readily consumable manner. This required the third entity of retrieval systems. What essentially was a two-way transaction between the host and the client has become three-way with an IR system in the middle.

Clients are served by hosts, a relation facilitated by IR systems. However, current day IR systems are more than just organizers of web links. They model user choices and preferences to serve them better. We argue that the three-entity unit of the client, IR system and the host is greater than the sum of its parts. The relation between these three entities is ignored by the current web-service architecture. We present here a proposal which will exploit this relationship to better deliver some aspects for web service usage.

Web designers write content on the pages based on the information provided by the owner of the site. Content in a website is primarily organized based on the categorization of the information and arranged appropriately by the designer. In this entire process of the current design paradigm, the *query* has no role to play during the design or presentation phases of a website. But, when a search query is given to an IR system, it retrieves links of pages that are prepared without taking query into consideration *on the host side*. This is because retrieval and content management are considered mutually exclusive, that is, the content management system does not know about the retrieval system and the retrieval system does not know how the content provider may respond to the query. Due to this shortcoming, both the content provider and IR system are under-performing. In this paper, we try to address this issue by proposing an architecture that enables the server hosting the website to present content that is based on the query posed by the user.

In sections 2 and 3, we describe some background to our work and mention some related work. In section 4, we talk about query aware content presentation, which goes beyond conventional personalization. Section 5 is about some requirements that a proposed architecture should satisfy. The architecture itself is proposed in section 6. Section 7 is about Retrieval Response Protocol, with secion 8 giving a couple of examples to show how content presentation may change in response to the query. Section 9 is the conclusion.

2 Related Work

Authors Begen et. al. [3] have studied protocols for streaming video content both on the web and non-web. They show that web streaming, in contrast to the traditional broadcasting of video, requires various techniques to ensure good user experience. It is important to note that in the context of web streaming video content, there is often a role for interaction that includes video search and retrieval.

The relevance of links (and their content) to search queries and their ranking continues to be challenging, often requiring search with multiple attempts with variations of the query. Further, different results suggested by various search engines are also explored to find the required content. This complex behavior, particularly of experienced users, was studied by Bharat [4], who suggested keeping track of the queries with all their variations and the different results that we found relevant from the search engines that were used. This indicates how complex search and retrieval of information continues to be. It is in this context that we propose content provider's involvement in delivering the most relevant content from that site for a given query.

Research shows that content organization has a significant effect on how easy or complex the document is for the viewers to understand and consume. Document annotation has been suggested by numerous authors (see for example work of Ferri et. al. [9] or of Bottoni et. al. [5]) to capture, index and present content in an easy-to-understand manner on the web. We see our work as an extension to this wherein the content is customized to match user expectations using the query that led to it.

Personalized search is a an area that is gaining increasing interest in the research community to enhance user experience and deliver better results. The idea is to use cues beyond the search query, like the user profile, the search history, the context of the search, etc. to improve results. This goal was pursued in numerous works. Matthijs and Radlinski [13] model users using past browsing activity that includes the implicit feedback from clickthrough data. This information is then used to rerank the search results for a given retrieval task. Shen et. al. [15] also employ statistical language models along with the browsing activity exploited by Matthijs et. al [13]. Vallet et. al. [19] make clusters of related contexts by dynamically modeling the content from each retrieval. This is then used to discard irrelevant contexts for a given query and retrieve on contextually relevant content. Joshi et. al [12] give an overview of various methods used for this task in [9].

3 Feedback-Aware Retrieval

A classical or bare-boned retrieval system [6] only takes into account the query for retrieval. Some modern retrieval system go further and use the information available about the user for personalizing the results [1,5,9,12,13,19]. Similarly, there has been work on Web search based on search context [4]. However, these systems do not take into account the user's activity once the user has selected and visited one of the web pages. In the proposed architecture, anonymized information about the user's activity will be made available to the retrieval system. It will, thus, be possible to design algorithms that take this activity into account. Some work in this direction was proposed by [15].

The details about this activity might include information such as the other links on the website that the user clicked on and the total time that the user spent on the website and on various pages. A retrieval system made aware of the feedback from the host server should, intuitively, perform better.

Some modern retrieval systems also provide additional links as part of the summary 'snippet' while presenting the results of retrieval to the user. Such snippets can be better prepared with the suggested feedback from the host server.

Additionally, and importantly, the host server can provide extra information about its content as part of the feedback. This extra information will be based on the query and the knowledge of the content that is available to the host server. This will allow the content provider to have a say in the presentation of the snippet for the concerned website. The retrieval system may or may not use this information, depending on the retrieval and snippet preparation algorithm.

4 Query-Aware Content Presentation

Current state of the art Web servers do not take the query into consideration while presenting the content to the user. Lot of work has been reported on improving the architecture of Web servers for various applications [3,11,20]. Many models are available to compare the architectures of the servers [10]. [22] discusses improving the performance of websites by using edge servers in Fog Computing Architecture. To the best of our knowledge there is no attempt to use the query by the host server to present the content to the user.

If a host server can present the content to the user based on the query, then it will be beneficial to both the user and the host organization. Suppose that user A gives query "popular movies in action genre + old" and that B gives "popular movies in action genre + latest". Let us assume that both the users get $Link1$ as their first link. We propose that the host server of $Link1$ should present different contents to each of them based on their query. In this case, the host server may present a list of old action movies (could be from other pages of the host server) to A and a list of new action movies to B, in both cases in addition to the content at the $Link1$.

For getting maximum benefit from this kind of architecture, the current Content Management Systems (CMS) like Drupal, Joomla, Django etc. may have to be redesigned to take user queries into account for presenting the final web page to be shown to the user. This will allow the host server (and the CMS) to play an active role in the process of content retrieval. Since the content provider knows much more about the content than the retrieval system, all that knowledge could be used to present dynamic query-aware content to the user.

While presenting the content to a user, web server can perhaps summarize the content that is available on its site and further increase the satisfaction of the user. There are many summarizers available to generate generic summaries [2,14]. Query-Specific summarization is discussed in [7,8,17,21]. A web server can go a step further and prepare a customized presentation to the user in line with [16,18]. The main point is that summarization can also be made specific to the user query on the retrieval system.

5 Privacy and Customization

Our proposal requires the retrieval system to share some information about the user and the query with the host server. It also requires the host server to provide feedback to the retrieval system based on the user's stay in the website after the user selected the website from the retrieval results. This extra sharing of information immediately raises the questions of privacy. If our proposal is implemented, its detailed version will need to include stringent requirements to address all the possible privacy concerns.

We list below some of these requirements:

- The first such requirement is that the user's identity, even if known to the retrieval system, will not be revealed to the host server. Whatever information is shared with the host server will have to be strictly anonymized so as not to reveal the user's identity.
- The second requirement is that only the relevant information will be shared. If we view this information as a list of attribute-value pairs, then only that subset of attribute-value pairs will be shared with the host server that the host server needs to know in order to better present its content.
- The third requirement is that an opt-out option will be available to both the user and the host server. The user will be made aware of the sharing of information and the user will decide whether this sharing is to be allowed or not. The information will be shared only if the user explicitly agrees to it. In the default case, there will be no sharing. Similarly, the host server will decide whether to provide feedback to the retrieval system or not and the default will be the latter.
- The fourth requirement is that both the user and the host server must be able to customize sharing of information. If they decide to share information, they will further be given the option to select the specific attributes that they are willing to share. For example, if the retrieval system knows about the user's location, age, gender and language, then the user may decide to share only location and language.
- The user will have to be informed that the activity on the visited website may be used for providing feedback to the retrieval system. And the user will then decide whether and what part of the activity on the website can be used to provide feedback to the retrieval system.

As this proposal is worked out in more detail in future work, more such requirements might be identified and will also have to be addressed.

Even after addressing these issues, one concern still remains regarding the proposed architecture. Even if the shared information is anonymized and the host server does not know the identity of the user, the retrieval system may still know the identity and be able to connect the activity of the user on the visited website with the user's identity. This raises the question whether the retrieval system will come to acquire more knowledge about the user than is warranted. This may be a problematic ethical issue and requires further investigation.

6 The Proposed Architecture

The outline of the architecture we propose is presented in Figure 1. The numbers used in this section refer to those in this figure. The scenario is that the user

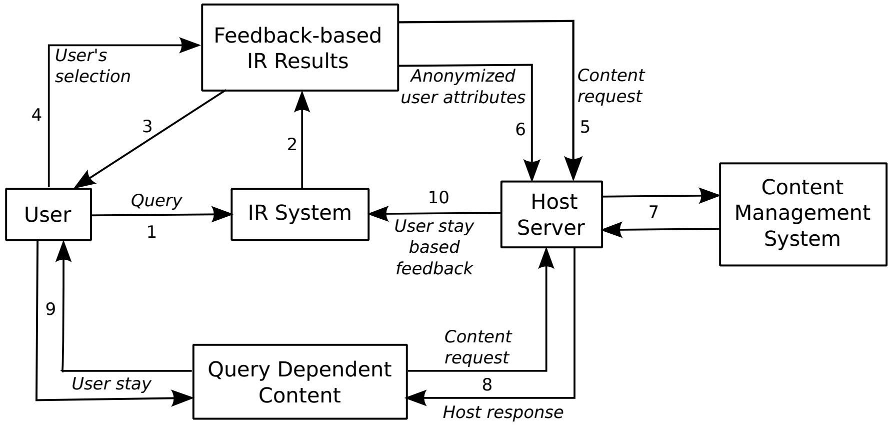


Fig. 1. The proposed architecture for more responsive IR and CMS systems

starts a retrieval system and gives a query (1). The retrieval system presents the search results to the user (2, 3). Out of them, the users selects one (4) and is taken to the destination website (5). When the user is taken to that website, the retrieval system also shares some information about the user and the query (subject to privacy requirements: see Section 5) with the server hosting the website (6). The host server uses this information to present the content (7) such that the user might have a better search experience (see Section 5). This presentation might, for example, make it easier for the user to find certain things. The host server will then provide feedback (10) to the retrieval system (again subject to privacy requirements) based on the user's stay in the website and the user's activity during the stay (8, 9). Since the information shared with the host server is anonymized, so will be the feedback given to the retrieval system. The retrieval system will now use this feedback to give better results in the future (see Section 3). The overall result will be better synchronization between the retrieval system and the host server for the purpose of presenting better results to the user. Anonymization, opt-out option and customization will be the central requirements, enforced through a protocol (see Section 7), to prevent any abuse that can result from sharing the information.

7 Retrieval Response Protocol

There are many different kinds of retrieval systems. Similarly, there are many different kinds of host servers and content management systems. If there is to be a flow of information between them as suggested in the preceding sections, then it will have to be precisely regulated so that it is possible to implement systems without any conflict. This will require a well-defined and well-designed protocol. We call this the Retrieval Response Protocol (RRP).

The Retrieval Response Protocol will regulate the flow of information between the retrieval system and the host server. The protocol will be used to initiate, maintain and close a *retrieval session*. As soon as the user selects one result from the results provided by the retrieval system in response to the user query, a retrieval session will be initiated. The ending of the session will perhaps have to be timeout based as there is no other way to know when the user has left the website.

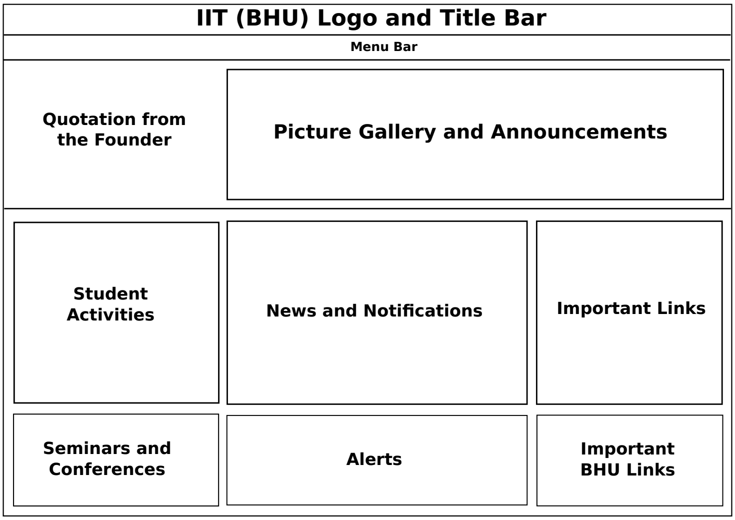


Fig. 2. Original page for the query 'IIT (BHU), Varanasi'

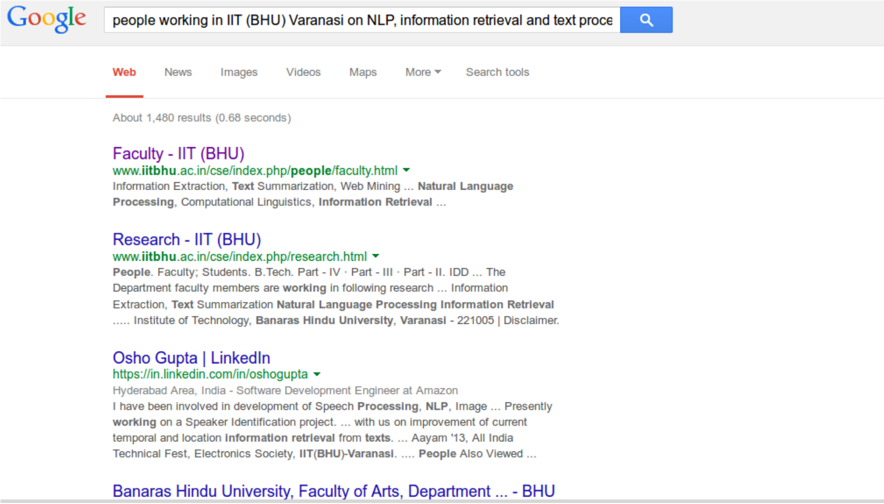
During the time the session is alive, the retrieval system will first share the information about the user and the query with the host server. After that, based on the user's activity, the host server will provide the feedback to the retrieval system. All the activity during this session will be subject to the privacy and customization requirements and the protocol design will have to take this into account.

The protocol will have to be designed to regulate this retrieval session. We leave the design of this protocol for future work.

8 Examples

In this section we show a couple of examples of how the content presentation by the host may change in response to the query in figures 2 to 6. We have taken two queries as examples:

- 'IIT (BHU) Faculty working on NLP information retrieval and text processing'
- 'Guest lectures in IIT (BHU)'


Google
people working in IIT (BHU) Varanasi on NLP, information retrieval and text proce
Web
News
Images
Videos
Maps
More
Search tools

About 1,480 results (0.68 seconds)

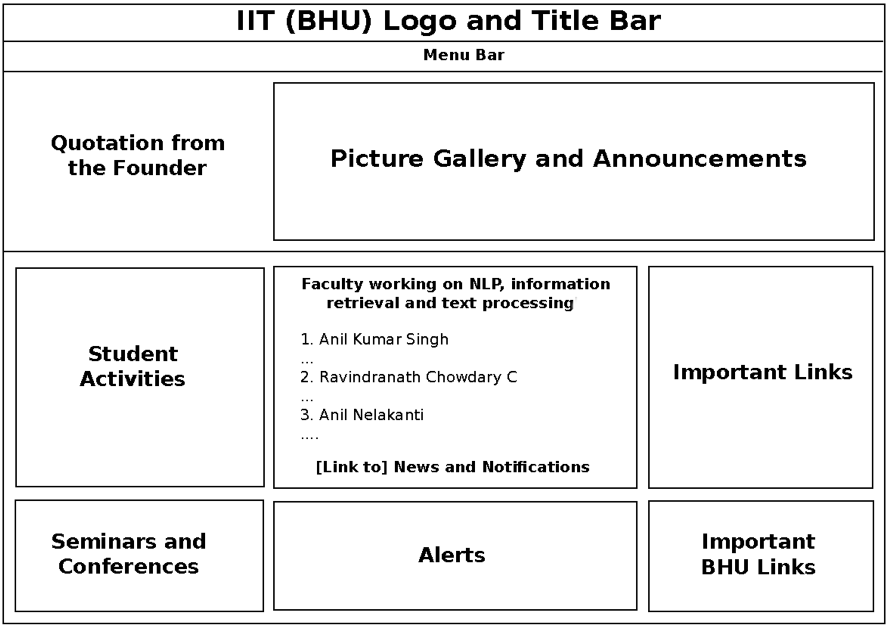
Faculty - IIT (BHU)
www.iitbhu.ac.in/cse/index.php/people/faculty.html
Information Extraction, Text Summarization, Web Mining ... Natural Language Processing, Computational Linguistics, Information Retrieval ...

Research - IIT (BHU)
www.iitbhu.ac.in/cse/index.php/research.html
People. Faculty; Students. B.Tech. Part - IV · Part - III · Part - II. IDD ... The Department faculty members are working in following research ... Information Extraction, Text Summarization Natural Language Processing Information Retrieval ..... Institute of Technology, Banaras Hindu University, Varanasi - 221005 | Disclaimer.

Osho Gupta | LinkedIn
https://in.linkedin.com/in/oshogupta
Hyderabad Area, India - Software Development Engineer at Amazon
I have been involved in development of Speech Processing, NLP, Image ... Presently working on a Speaker Identification project. ... with us on improvement of current temporal and location information retrieval from texts. ... Aayam '13, All India Technical Fest, Electronics Society, IIT(BHU)-Varanasi. .... People Also Viewed ...

Banaras Hindu University, Faculty of Arts, Department ... - BHU

Fig. 3. First Google query results


IIT (BHU) Logo and Title Bar

Menu Bar

Quotation from the Founder

Picture Gallery and Announcements

Student Activities

Faculty working on NLP, information retrieval and text processing

1. Anil Kumar Singh
...
2. Ravindranath Chowdary C
...
3. Anil Nelakanti
....

[Link to] News and Notifications

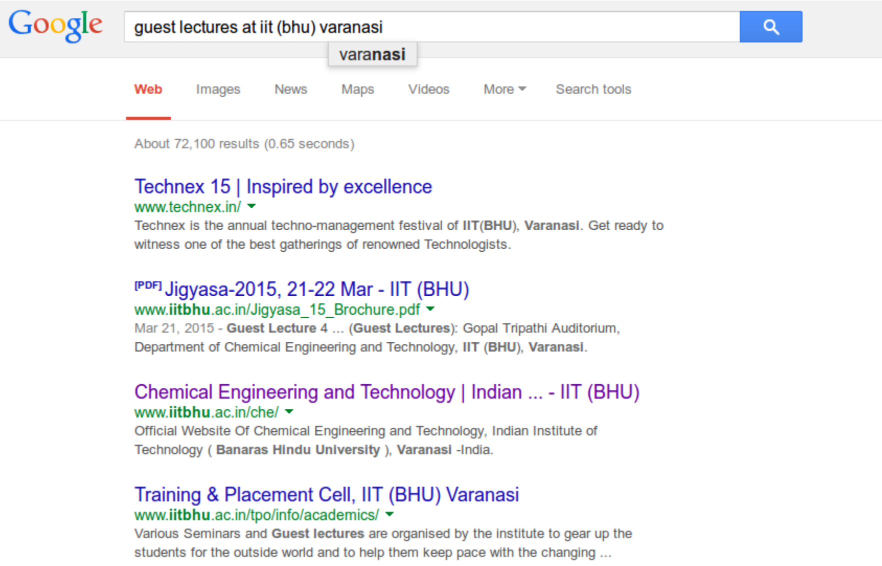
Important Links

Seminars and Conferences

Alerts

Important BHU Links

Fig. 4. Page presented in response to the first query



About 72,100 results (0.65 seconds)

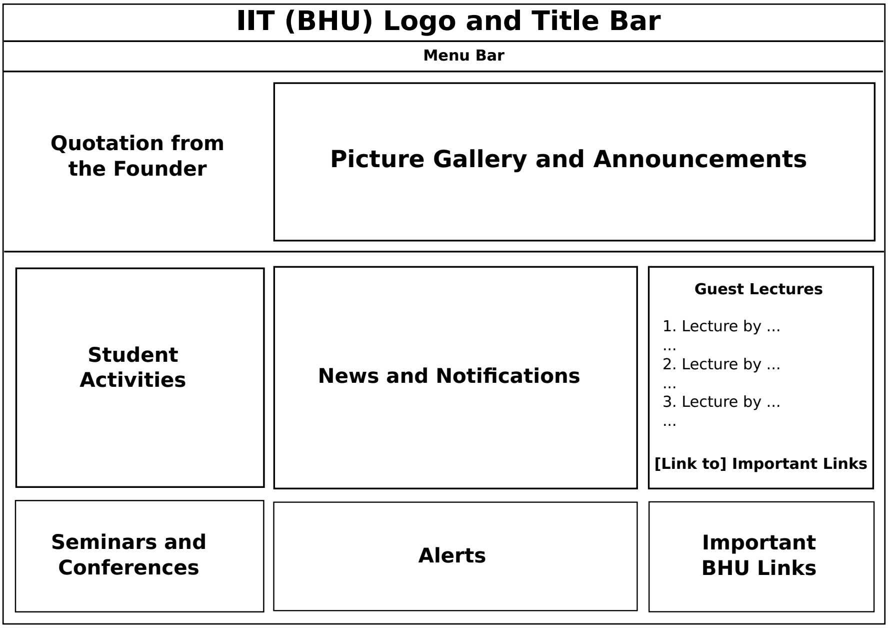
Technex 15 | Inspired by excellence
www.technex.in/
Technex is the annual techno-management festival of **IIT(BHU)**, **Varanasi**. Get ready to witness one of the best gatherings of renowned Technologists.

[PDF] Jigyasa-2015, 21-22 Mar - IIT (BHU)
www.**iitbhu**.ac.in/Jigyasa_15_Brochure.pdf
Mar 21, 2015 - **Guest Lecture** 4 ... (**Guest Lectures**): Gopal Tripathi Auditorium, Department of Chemical Engineering and Technology, **IIT** (**BHU**), **Varanasi**.

Chemical Engineering and Technology | Indian ... - IIT (BHU)
www.**iitbhu**.ac.in/che/
Official Website Of Chemical Engineering and Technology, Indian Institute of Technology (**Banaras Hindu University**), **Varanasi** -India.

Training & Placement Cell, IIT (BHU) Varanasi
www.**iitbhu**.ac.in/tpo/info/academics/
Various Seminars and **Guest lectures** are organised by the institute to gear up the students for the outside world and to help them keep pace with the changing ...

Fig. 5. Second Google query results



IIT (BHU) Logo and Title Bar

Menu Bar

Quotation from the Founder

Picture Gallery and Announcements

Student Activities

News and Notifications

Guest Lectures

1. Lecture by ...
...
2. Lecture by ...
...
3. Lecture by ...
...

[Link to] Important Links

Seminars and Conferences

Alerts

Important BHU Links

Fig. 6. Page presented in response to the second query

9 Conclusion

In the current information retrieval paradigm, the host does not use the query information for content presentation. The retrieval system does not know what happens after the user selects a retrieval result. And the host also does not have access to the information which is available to the retrieval system. We presented the outline of an architecture that addresses these issues. The aim is to provide a better search experience to the user through better presentation of the content based on the query and better retrieval results based on the feedback to the retrieval system from the host server. The retrieval system will share some information with the host server and the host server in turn will provide relevant feedback to the retrieval system based on the user's stay in the website. The host uses all the query related information for dynamic content presentation.

Most of this is not done as part of conventional personalization reported in the literature, where the host (with its content management system) does not really come into the retrieval picture except in just returning the pages requested. In conventional personalization, things are looked at only from the point of view of the retrieval system. One main idea here is to have varying presentation of the content based on the query (with the user's consent).

This revised paradigm for information retrieval also introduces the issues of privacy which will have to be addressed stringently. It also needs a new protocol for content retrieval response, which we briefly described. This protocol will regulate the flow of information between the retrieval system and the host server subject to the privacy and customization requirements. If some of what we propose is not feasible, say, due to privacy concerns, the rest could still be implemented.

We realize that the architecture proposed here is a bit sketchy. A direction for future research will be to flesh out the details of this architecture, including the retrieval response protocol.

References

1. Badros, G., Lawrence, S.: Methods and systems for personalised network searching (2005). US Patent Application 20050131866
2. Balahur, A., Kabadjov, M.A., Steinberger, J., Steinberger, R., Montoyo, A.: Challenges and solutions in the opinion summarization of user-generated content. J. Intell. Inf. Syst. **39**(2), 375–398 (2012)
3. Begen, A., Akgul, T., Baugher, M.: Watching video over the web: Part 1: Streaming protocols. IEEE Internet Computing **15**(2), 54–63 (2011). doi:10.1109/MIC.2010.155
4. Bharat, K.: Searchpad: explicit capture of search context to support web search. In: Proceedings of the 9th International World Wide Web Conference, pp. 493–501 (2000)
5. Bottoni, P., Ferri, F., Grifoni, P., Marcante, A., Mussio, P., Padula, M., Reggiori, A.: E-document management in situated interactivity: the wil approach. Universal Access in the Information Society **8**, 137–153

6. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. In: Proceedings of the Seventh International Conference on World Wide Web 7. WWW7, pp. 107–117. Elsevier Science Publishers B. V., Amsterdam (1998). http://dl.acm.org/citation.cfm?id=297805.297827
7. Chowdary, C.R., Kumar, P.S.: ESUM: an efficient system for query-specific multi-document summarization. In: Boughanem, M., Berrut, C., Mothe, J., Soule-Dupuy, C. (eds.) ECIR 2009. LNCS, vol. 5478, pp. 724–728. Springer, Heidelberg (2009)
8. Chowdary, C.R., Sravanthi, M., Kumar, P.S.: A system for query specific coherent text multi-document summarization. International Journal on Artificial Intelligence Tools **19**(5), 597–626 (2010)
9. Ferri, F., Grifoni, P., Padula, M.: Using shape to index and query web document contents. Journal of Visual Languages and Computing **13**, 355–373
10. Harji, A.S., Buhr, P.A., Brecht, T.: Comparing high-performance multi-core webserver architectures. In: Proceedings of the 5th Annual International Systems and Storage Conference, SYSTOR 2012, pp. 1:1–1:12. ACM, New York (2012). http://doi.acm.org/10.1145/2367589.2367591
11. Hashemian, R., Krishnamurthy, D., Arlitt, M., Carlsson, N.: Improving the scalability of a multi-core web server. In: Proceedings of the 4th ACM/SPEC International Conference on Performance Engineering. ICPE 2013, pp. 161–172. ACM, New York (2013). http://doi.acm.org/10.1145/2479871.2479894
12. Joshi, C., Jaiswal, T., Gaur, H.: An overview study of personalized web search. International Journal of Scientific and Research Publications **3** (2013)
13. Matthijs, N., Radlinski, F.: Personalizing web search using long term browsing history. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, pp. 25–34. ACM, February 2011
14. Radev, D.R., Jing, H., Budzikowska, M.: Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In: NAACL-ANLP 2000 Workshop on Automatic Summarization, pp. 21–30. Association for Computational Linguistics, Seattle (2000)
15. Shen, X., Tan, B., Zhai, C.: Context-sensitive information retrieval using implicit feedback. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR 2005, pp. 43–50. ACM, New York (2005). http://doi.acm.org/10.1145/1076034.1076045
16. Shibata, T., Kurohashi, S.: Automatic slide generation based on discourse structure analysis. In: Dale, R., Wong, K.-F., Su, J., Kwong, O.Y. (eds.) IJCNLP 2005. LNCS (LNAI), vol. 3651, pp. 754–766. Springer, Heidelberg (2005)
17. Sravanthi, M., Chowdary, C.R., Kumar, P.S.: QueSTS: a query specific text summarization system. In: Proceedings of the 21st International FLAIRS Conference, pp. 219–224. AAAI Press, Florida, May 2008
18. Sravanthi, M., Chowdary, C.R., Kumar, P.S.: Slidesgen: automatic generation of presentation slides for a technical paper using summarization. In: Proceedings of the Twenty-Second International Florida Artificial Intelligence Research Society Conference, Sanibel Island, Florida, USA, May 19–21, pp. 284–289 (2009). http://aaai.org/ocs/index.php/FLAIRS/2009/paper/view/22
19. Vallet, D., Fernandez, M., Castells, P., Mylonas, P., Avrithis, Y.: Personalized information retrieval in context. In: Proceedings of 3rd Int. Workshop Modeling Retrieval Context 21st Nat. Conf. Artif. Intell. (2007)

20. Veal, B., Foong, A.: Performance scalability of a multi-core web server. In: Proceedings of the 3rd ACM/IEEE Symposium on Architecture for Networking and Communications Systems. ANCS 2007, pp. 57–66. ACM, New York (2007). http://doi.acm.org/10.1145/1323548.1323562
21. Yin, W., Pei, Y., Zhang, F., Huang, L.: Query-focused multi-document summarization based on query-sensitive feature space. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management. CIKM 2012, pp. 1652–1656. ACM, New York (2012). http://doi.acm.org/10.1145/2396761.2398491
22. Zhu, J., Chan, D., Prabhu, M., Natarajan, P., Hu, H., Bonomi, F.: Improving web sites performance using edge servers in fog computing architecture. In: IEEE 7th International Symposium on Service Oriented System Engineering (SOSE), pp. 320–323, March 2013