The Impact of Modes of Mediation on the Web Retrieval Process

Mandeep Pannu, Rachid Anane, Anne James
Faculty of Engineering and Computing
Coventry University,UK
{m.pannu, r.anane, a.james}@coventry.ac.uk

Abstract. This paper is concerned with the investigation of mediation between users and Web search engines and the impact of different modes of mediation on the Web search effectiveness. This involves the integration of explicit, implicit and hybrid modes of mediation within a content-based framework, facilitated by the adoption of the Vector Space Model. The work is supported by an experimental evaluation of the impact of different mediation modes on documents retrieval process in terms of recall and precision. The results of the experiments indicate that the mediation framework improves the quality of the retrieval process, and that the difference in the quality of the results is statistically significant.

Keywords: User profiling, Personalisation, Implicit profile, Explicit profile, content-based

1 Introduction

Most current Web search engines are designed to serve a generic user irrespective of individual needs and interests. This raises the fundamental issue of how to identify and select the information that is relevant to a specific user. The retrieval process can be improved through personalisation of the search according to the specific needs and interests of the users. Implicit and explicit approaches can be used for user profiling. In the implicit approach, the behaviour of the users and their activities are observed and information is collected without the direct involvement of the user. On the other hand, explicit profile generation requires the users to directly provide specific information in order to create an individual user profile.

This research is an integral part of the effort aimed at overcoming the limitations of classic search engines. A mediation framework which is proposed allows the filtering of the results generated by classical search engines with the use of information contained in a user profile. The framework incorporates content-based information retrieval techniques, and it is facilitated by the adoption of the Vector Space Model (VSM**).** The proposed framework has been used to create a critical evaluation of variants of explicit, implicit and hybrid profiling techniques.

This paper is structured as follows. Section 2 is concerned with the background of this research. Section 3 presents the architecture of the proposed framework. Section 4 deals with the experimental evaluation of the framework in relation to the classical search engines, and offers pointers for further work. Section 5 concludes the paper.

2 Research Scope

The relation between a user query and Web pages is problematical and is driving the research in the field of information retrieval. Users have a variety of needs and the retrieval systems are often unable to satisfy adequately the requirements fulfil of an individual user.

2.1 Related Work

Personalised systems are designed to help users overcome the limitations of Web search by extracting keywords based on individual preferences. Personalisation can be implicitly or explicitly generated.

Explicit profile creation involves asking users for specific information in order to create an individual user profile. Salton et al. [1] considered user involvement as a powerful way of improving the relevance of the search results, and systems based on the information explicitly provided by users are constantly being developed [2-3-4]. In the explicit user profile generation users can build their own profile according to their specific interest and needs. Methods for generating explicit profile include asking the user to approve/disapprove a document [4], to give rating from a scale of values [5] or to engage in a dialogue in a natural language [3].

One of the first personalisation systems designed by Lieberman [1] was implicitly assuming an interest in a document if it was bookmarked, and a lack of interest if the document was left without saving or following hyperlinks inside it. Other approaches use techniques such as capturing mouse clicking [2] or tracking user mouse movements, as the mouse pointer can be used for reading [6]. A more recent approach used by Hussein and Elsayed [7] involves capturing the users' facial expression to estimate their interest in a document. Implicit feedback can be as effective as explicit feedback [8].

Although implicit methods are the focus of many research programmes, their reliability is still an issue [2]. Moreover, Paulson and Tzanavari [9] have pointed out that implicitly generated profiles are often not useful once users change their area of interest.

3 Proposed Framework

The primary goal of this research is to introduce a mediation framework, which act as an interface between a user and a classical search engine to provide personalised search results, without violating the privacy of the user. The framework can also be used as a vehicle for the investigation of different modes of user profiling. The major issue in evaluating an information retrieval approach is the amount of documents available and the quality of the results. Instead of developing a search engine, techniques can be evaluated by filtering only a subset of Web documents, where this subset would be retrieved from a base Web search engine API. The proposed framework is part of this endeavour.

3.1 Design Requirements

For the framework to be useful, it has to meet several objectives. First, it has to allow the implementation of custom methods for building user profiles. The framework should provide a programming interface that supports the tracking of actions detected in a Web browser, like navigating or clicking. A programmer modifying the framework in order to evaluate different filtering techniques should be able to do so by only modifying the filtering method, and by handling events from the browser to gather implicit or explicit information. Gathering explicit information may require some modification to the graphical user interface (GUI). The framework should support transparency for all other operations like retrieving search results from a base Web search engine or maintaining a database.

3.2 Mediation Framework Architecture

The framework combines a content-based approach and Vector Space Model for the filtering of results. The content-based approach was adopted because of its focus on the interaction between the profile of a single user and the content of a document. The VSM was used for the determination of the similarity between personalised profiles and documents. It offers formal clarity, efficiency in documents representation and consistent use of weights for the terms in documents, profiles, and query representations.

The overall architecture of the mediation framework is presented on the Figure 1. The user can use the provided interface to access documents on the Web. Every action in the browser is handled by the framework, and the information about it stored in the database, together with a VSM term vector containing keywords for visited document, the URL of the document and the time of the event. Similarly, the explicit feedback provided by the user is stored in the database.

When the user starts the search process by selecting one of the currently implemented profiling methods (explicit, implicit or hybrid), the application will generate a VSM representation of a selected profile, based on the information stored previously in the database. If the hybrid profile is selected, then both implicit and explicit vector representations are created separately and are then merged together.

Queries are used to retrieve a number of documents from a base Web search API. All the returned documents are indexed and represented in a vector form. The vector contains a list of terms extracted from the document, together with a weight for each of the term.

The system attempts to sort the documents by calculating the cosine similarity between the vector representing the user profile and the vectors representing each document. The cosine similarity is calculated by multiplying values for corresponding terms, and dividing the sum by the length of both vectors. The system then returns documents with the highest value of similarity. Depending on the configuration, the system can either return a constant number of documents, or only documents for which some similarity threshold is satisfied.

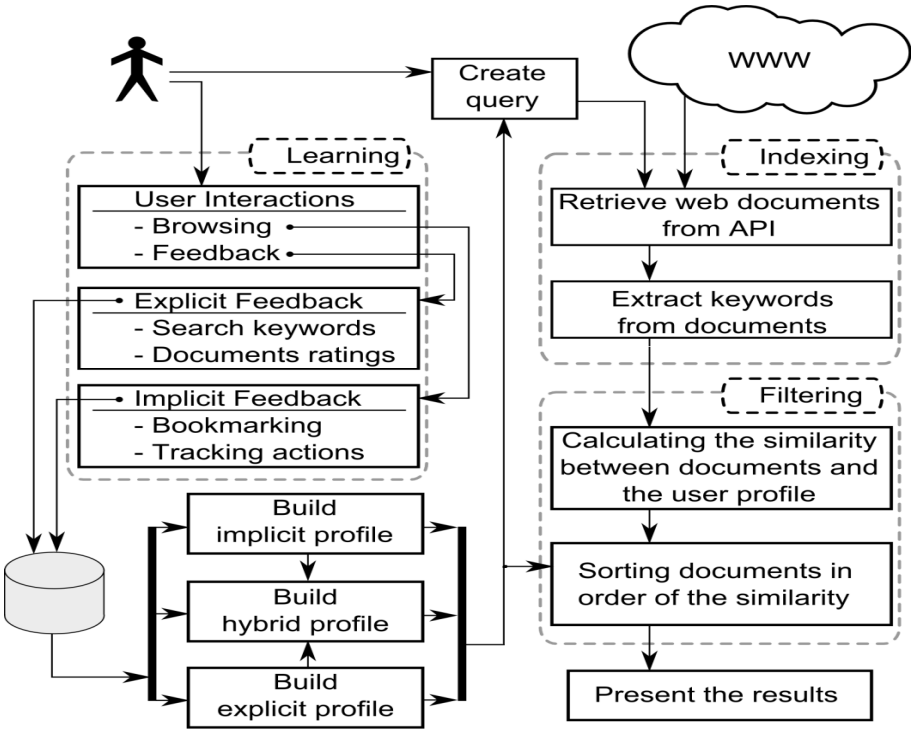


Fig. 1. Mediation framework architecture

As one of the objectives of the framework is to support the evaluation of different methods for user profile generation, the methods responsible for creating the user profile vector and for composing the profile vector with the entered query can be selected through the implementing of a single interface.

3.3 User Profile Generation

In order to evaluate the impact the framework, three different methods of profile generation have been investigated and implemented. In each case a user profile is represented in the VSM by a list of keywords with weights stored as a term vector.

An explicit user profile generation used for the evaluation is achieved by the submission of specific keywords by the users. The implicit profile is based on the observation of user behaviour and browsing history. The time spent on each page is assumed to be a good indicator of the user interest in a Web page. The system creates the profile vector by calculating the average time the user spent viewing each document, and adding together the vector representations for documents that were opened for longer than that time. The resulting vector is normalised in the final step.

In the hybrid profile the explicit and implicit profiles are generated independently and combined into a single term vector. For the purpose of this experiment the keywords from both vectors are simply added with equal weight.

The last editable part of the framework is the generation of a query. The associated method takes a list of keywords (as a query) entered by the user, and the generated user profile as parameters. Every keyword entered by the user is assigned a value equal to the highest value from the profile vector. Both vectors are then added and the result vector is truncated to 10 keywords. This vector is used by the framework to retrieve a list of documents from one of the base Web search API. Each document is in turn compared for similarity with the user profile.

4 Evaluation and Discussion

A comprehensive quantitative evaluation of the framework is presented. In the evaluation the performance of the three mediation systems is measured in terms of two metrics: precision and recall. The experiment was performed with 30 users with their own choice of keywords and areas of interests. To measure the system effectiveness the evaluation was conducted with Yahoo! and Google search APIs, and the mediation system with the three different types of profiling.

In the experiment, the system retrieves for each query 80 documents returned by each of the API; the framework was set to order these documents according to the similarity to the user profile and to return 20 documents with the highest similarity. For each base search APIs the first 20 returned documents were considered without any filtering.

4.1 Document Rating

To ensure that a consistent scale of scores is adhered to users were presented with an indication on how to assess a page depending on whether it was relevant or not. They were instructed to give 2 points to fully relevant documents, 1 point to documents containing relevant information as part of its contents, 0.5 point for documents that contained links to relevant information and 0 point if there was no relevant information [10]. The documents which could not be opened were ignored. The search results were presented in random order, and users were not aware of which search method generated the results. This process makes the scoring fairer, consistent and easier for the users.

4.2 Experimental Analysis

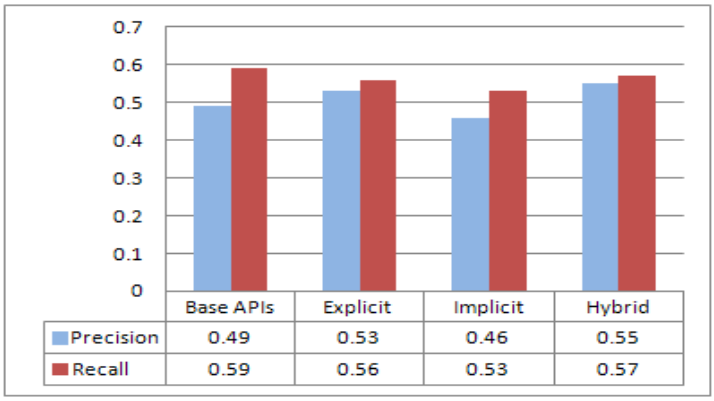
The experiment was conducted in two phases. In the first phase only a short time was given to build implicit profiles, while in the second phase this time was extended.

Experiment Phase 1. In the first phase the users were instructed to use the provided Web browser for 15 minutes so that the browsing behaviour could be recorded in the

database. After the browsing session each user proceeded to enter the keywords for the search. The documents returned by each of the implemented systems and by the base Web Search APIs were combined into one list, sorted randomly and pre-opened in a Web browser. This approach was designed to avoid the situation where the ratings given by the users would be affected by their opinions of the retrieval systems. The documents were then rated by users.

Experiment Phase 2. The second phase of the experiment was performed to check how the retrieval effectiveness changes when a system had additional time to learn from the user behaviour. Each user was allocated the same user name through all the experiments so that new information could be added to an already stored browsing history. The additional learning time was set to 15 minutes per user, so that the total time allowed for system learning was doubled. The users rated search results in the same way as in the first phase of the experiment.

First Phase Results. Figure 4 presents the average precision and recall calculated for each of the systems after the first phase of the experiment.

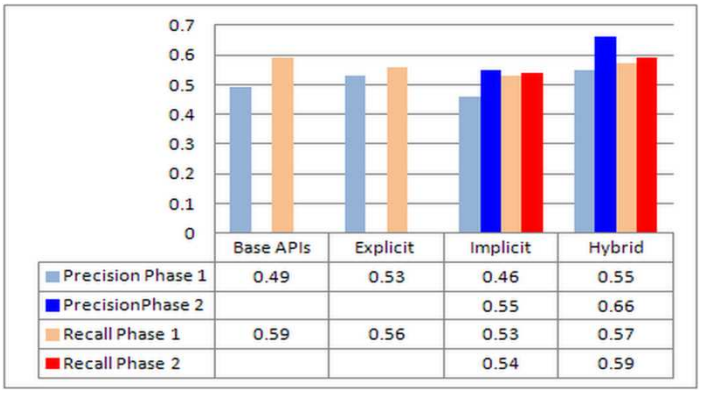


	Base APIs	Explicit	Implicit	Hybrid
Precision	0.49	0.53	0.46	0.55
Recall	0.59	0.56	0.53	0.57

Fig. 2. Average precision and recall after the first phase

Both explicit and hybrid systems have improved the search performance in terms of precision, while the implicit system performance is worse than the performance of the base API. The recall values are almost unaffected by mediation, with the exception of the implicit system. Even through the implicit system alone has a negative effect on the performance, the hybrid system which uses the implicit information as part of the user profiling performs better than the explicit system which does not use it. This leads to a conclusion that while the implicit profile alone suffers from 'the cold start' problem, it can still be beneficial (in terms of precision) to use it in the filtering process, even after very limited learning.

Second Phase Results. The second phase of the experiment was designed to assess how precision and recall are affected when more time is given to the system to learn. As the explicit system being evaluated or the base APIs would not benefit from the additional time, the precision and recall in this phase were only measured for the implicit and hybrid systems.



	Base APIs	Explicit	Implicit	Hybrid
Precision Phase 1	0.49	0.53	0.46	0.55
PrecisionPhase 2			0.55	0.66
Recall Phase 1	0.59	0.56	0.53	0.57
Recall Phase 2			0.54	0.59

Fig. 3. Average precision and recall after the second phase

As shown in Figure 6, in the second phase the precision improved for both implicit and hybrid systems. It can be concluded that the results generated by implicit systems are more precise once a browsing history is generated. The values for relative recall calculated after the second phase appeared to have improved.

The Student's t-test was conducted to determine whether these results are significant. The significance of the results was calculated by comparing the performance of the hybrid system with the other systems after the second phase of the experiment. The results show that the change in precision is significant with 99% confidence, while the change in recall is not significant. This confirms that hybrid personalisation generates more relevant documents, but that the pool of relevant documents is however similar to the other systems.

4.3 Discussion

This section puts the work in context and identifies issues for further work. The experimental results indicate that the system using the hybrid profiling has better and more accurate results than the base APIs without the profiling. The hybrid profiling combines the explicitly stated interests with the observation of user behaviour and it can offer an effective way of dealing with information overload.

Although the aims and objectives of this research were met, a number of limitations have been identified. Useful documents can be ignored by their linguistic constraint as the calculation of the similarity is performed by exact match only. Secondly, efficiency issues which are also important were not addressed. The overheads caused by the need to download multiple documents before the framework can presents the search results were not investigated.

The proposed framework appears to be a viable mediator between users and the Web, but there is still scope for enhancing its effectiveness. Further work will seek to generate more accurate profiles, e.g. by widening the criteria of implicit observation. Currently, the framework is based on VSM with exact match for keywords. Adding support for synonyms or ontological context could help to identify the terms that are related to those stored in a user profile.

5 Conclusion

This research is an integral part of the effort aimed at overcoming the limitations of the classic search engines. The investigation has led to the proposal of a mediation approach which was applied in the development of three evaluated systems.

Conducted experiments indicate that mediation frameworks can improve the quality of the Web search results, with the choice of the mode of mediation being an important factor in enhancing search precision. With the high improvement in terms of precision and insignificant change in the recall, the personalised systems can definitely provide better users` experience and reduce the time they need to spend on searching.

References

1. Salton, G., Singhal, A., Mitra, M. et al.: Automatic Text Structuring and Summarization. Information Processing and Management, vol. 33 pp. 193-207 (1997)
2. Rastegari, H., and Shamsuddin, S.M.: Web Search Personalization Based on Browsing History by Artificial Immune System. Journal of Advances in Soft Computing and Its Applications 3 (2), pp. 282-301 (2010)
3. Stegmann, R.: Improving Explicit Profile Acquisition by Means of Adaptive Natural Language Dialog. In: Anonymous Lecture Notes in Computer Science, vol. 3538, pp. 518-520 (2005)
4. Swapna, P., & Ravindran, R. B.: Personalized Web-page Rendering System. In: Das, G., Sarda, N., L. and Reddy, K.,P (eds.) COMAD, pp. 30-39. Computer Society of India, India (2008)
5. Claypool, M., Le, P., Waseda, P., and Brown, D.: Implicit Interest Indicators. Proceeding of the 6th international conference on intelligent user interface. held at Santa Fe, New Mexico, United States. pp. 33-40 (2001)
6. Aoidh, E. M., Bertolotto, M., Wilson, D. C.: Implicit Profiling for Contextual Reasoning about Users Spatial Preferences. 271 (2007)
7. Hussein, M., and Elsayed, T.: Studying Facial Expressions as an Implicit Feedback in Information Retrieval Systems (2008)
8. Hopfgartner, F., Hannah, D., Gildea, N., and Jose, J.M.: Capturing Multiple Interests in News Video Retrieval by Incorporating the Ostensive Model. 'Proceeding of the Second International Workshop on Personalized Access, Profile Management, and Context Awareness in Databases'. held at Auckland, New Zealand, pp. 48–55 (2008)
9. Paulson, P., & Tzanavari, A.: Combining Collaborative and Content-Based Filtering using Conceptual Graphs. Lectures Notes in Computer Science, pp. 168-185 (2003)
10. Kumar, S., B.T., & Prakash, J. N.: Precision and Relative Recall of Search Engines: A Comparative Study of Google and Yahoo. Singapore Journal of Library & Information Management, vol. 38 pp. 124-137 (2009)