A Framework for the Evaluation of Adaptive IR Systems through Implicit Recommendation

Catherine Mulwa[1], Seamus Lawless[1], M. Rami Ghorab[1], Eileen O'Donnell[1], Mary Sharp[1] and Vincent Wade[1]

[1]Knowledge and Data Engineering Research Group,
School of Statistics and Computer Science,
Trinity College Dublin,
{mulwac, ghorabm, odonnee, seamus.lawless, mary.sharp, vincent.wade}@scss.tcd.ie

Abstract. Personalised Information Retrieval (PIR) has gained considerable attention in recent literature. In PIR different stages of the retrieval process are adapted to the user, such as adapting the user's query or the results. Personalised recommender frameworks are endowed with intelligent mechanisms to search for products, goods and services that users are interested in. The objective of such tools is to evaluate and filter the huge amount of information available within a specific scope to assist users in their information access processes. This paper presents a web-based adaptive framework for evaluating personalised information retrieval systems. The framework uses implicit recommendation to guide users in deciding which evaluation techniques, metrics and criteria to use. A task-based experiment was conducted to test the functionality and performance of the framework. A Review of evaluation techniques for personalised IR systems was conducted and the results of the analysed survey are presented.

Keywords: Personalisation, Personalised Information Retrieval Systems, Implicit Recommendations, User-based Evaluation, Task-based Evaluation

1 Introduction

Evaluation has been an integral part of Information Retrieval (IR) research from its early days with the Cranfield experiments (Cleverdon et al. 1966) that used pre-defined queries that were run against a test collection in batch mode. One major problem with traditional IR systems is that they provide uniform access and retrieval results to all users, solely based on the query terms the user issued to the system. Evaluation frameworks for personalised information retrieval systems (PIRS) and adaptive information retrieval systems (AIRS) are necessary to "better interpret and give more exact hints and false inferences than a simple global vision, thus facilitating the improvement of applications and services, when required, as well as the generalisation and reuse of results"(Tobar 2003). In this paper we call systems which combine Adaptive Hypermedia (AH) and IR approaches, AIRS (Lawless et al. 2010).

The main aim of the framework described in this paper is to provide comprehensive support for users through implicit recommendations on which evaluation methods, metrics and criteria should be used to evaluate these systems and how to best combine these approaches. Access to this repository of evaluation approaches is supported for geographically distributed users of any nationality by facilitating dynamic translation of content. The authors acknowledge that personalisation in IR is aimed at improving the user's experience by incorporating user subjectivity in the retrieval process.

The rest of this paper is structured as follows: Section 2 presents a summary of a comparison of personalisation approaches and evaluation techniques in PIR systems. Section 3 introduces the proposed framework to evaluate AIRS systems. Also the methodology, architecture design, functionality and evaluation of the framework are also introduced. Finally section 4 concludes the paper and proposes future work.

2. A Review of Personalisation Approaches and Evaluation Techniques for AIRS

Personalised Information Retrieval (PIR) is a research area which has gained attention in recent literature and is motivated by the success of both areas, IR and AH (Gauch et al., 2007, Micarelli et al., 2007). IR systems have the advantage of scalability when dealing with large document collections and performing a vast amount of information processing. AH systems have the advantage of including the user in the process and thus the ability to satisfy individual user needs by modeling different aspects of the user. In PIR, different stages of the retrieval process are adapted to the user such as adapting the user's query and/or the results. This review focuses on personalisation approaches and existing evaluation techniques for PIR systems.

2.1 Overview of Personalisation Approaches

Personalisation can be performed on an individualised, collaborative, or aggregate scope. Individualised personalisation is when the system's adaptive decisions are taken according to the interests of each individual user as inferred from their user model (Speretta and Gauch, 2005, Teevan et al., 2005). Collaborative personalisation is when information from several user models is used to determine or alter the weights of interests in other user models (Sugiyama et al., 2004). This is usually used when a system groups the users into a number of stereotypes according to certain similarity criteria between their user models; at which point the system can judge the relevance of a certain item or document to a user based on information from other user models that belong to the same group. Stereotypes can be manually pre-defined or automatically learnt using machine learning techniques (e.g. clustering techniques). Personalisation can be implemented on an aggregate scope when the system does not make use of user models; in which case personalisation is guided by aggregate usage data as exhibited in search logs (i.e. implicitly inferred general users' interests from aggregate history information) (Smyth and Balfe, 2006, Agichtein et al., 2006).

The authors acknowledge that user-based evaluation of personalised IR systems is challenging because of the user effect in terms of the inconsistency in ranking and in relevance criteria usage. End-users are seen as the ultimate assessors of the quality of the information and of the systems as well as services that provide information(Barry and Schamber, 1998). User satisfaction is a composite term; amalgamating a cluster of "felt experience". Table 1 provides a comparison of the surveyed systems in the literature. The comparison focuses on the personalisation implementation stage of the surveyed systems, guided by the three classification criteria (i.e. individualised, collaborative and aggregate usage data).

Table 1: Comparison of Personalisation Approaches

Application Area	Personalisation Scope	Personalisation Approach	Published Study
Monolingual IR	Individualised	Result Adaptation (result re-ranking)	(Speretta and Gauch 2005), (Stamou and Ntoulas 2009), (Teevan et al.2005), (Pretschner and Gauch 1999)
Monolingual IR & Information Filtering	Individualised	Result Adaptation (result re-ranking)	(Micarelli and Sciarrone 2004)
Monolingual IR	(1)Individualised & (2) Collaborative	Result Adaptation (result re-ranking)	(Sugiyama et al. 2004)
Monolingual IR	Aggregate usage data	Result Adaptation (result re-ranking)	(Smyth and Balfe 2006)
Monolingual IR	Aggregate usage data	Result Adaptation ((1)result scoring & (2)result re-ranking)	(Agichtein et al. 2006)
Information Filtering	Individualised	Result Adaptation (result scoring)	(Stefani and Strapparava 1999)
Monolingual IR	Individualised	Query Adaptation (query expansion using keywords from user model)	(Chirita et al. 2007)
Structured Search on a Database	Individualised	Query Adaptation (query rewriting)	(Koutrika and Ioannidis 2004)
Cross-lingual IR	Aggregate usage date	Query Adaptation (query suggestions using similar queries from multiple languages)	(Gao et al. 2007)
Monolingual IR	Individualised	Query & Result Adaptation (query expansion using keywords from user model, and result re-ranking)	(Pitkow et al. 2002)

2.2 Evaluation Approaches for PIR Systems

The evaluation of PIR systems is challenged by user effect, which is manifested in terms of users' inconsistency in relevance judgment ranking and relevance criteria usage. Personalisation in PIR systems is generally performed by adapting the query and/or the results to the user's interests. Adaptation can either target specific individualized user needs, or target common needs of groups of users. Personalised systems involve information about users in the process and therefore adapt the retrieval process to the users' needs. In other words, a PIR system does not retrieve documents that are just relevant to the query but ones that are also relevant to the user's interests.

Table 1: Comparison of Evaluation Techniques

Scope of Evaluation	Evaluation Metric & Instrument	Experimental Setting	Example Publications
System Performance (retrieval process)	Quantitative (Precision at K, Recall at K, F-measure, Break-even point)	Controlled setting (47 users, 25 information needs per user, open web corpora via meta search engine)	(Smyth and Balfe 2006)
System Performance (retrieval process)	Quantitative (R-precision)	Controlled setting (20 users, 50 information needs per user, open web corpora via Google wrapper)	(Sugiyama et al. 2004)
System Performance (retrieval process)	Quantitative (Normalised Discounted Cumulative Gain (NDCG))	Controlled setting (15 users, 10 information needs per user, open web corpora via MSN Search)	(Teevan et al. 2005)
System Performance (retrieval process)	Quantitative (Normalised Discounted Cumulative Gain (NDCG))	Controlled setting (18 users, 4 information needs per user, open web corpora via Google wrapper)	(Chirita et al. 2007)
System Performance (retrieval process)	Quantitative (rank scoring based on explicit relevance judgments by users)	Controlled setting (11 users, 68 information needs per user on average, open web corpora via Google wrapper)	(Stamou and Ntoulas 2009)
System Performance (retrieval process)	Quantitative (rank scoring based on implicit relevance judgments from clickthrough)	Controlled setting (6 users, 2 information needs per user, open web corpora via Google wrapper)	(Speretta and Gauch 2005)
System Performance (retrieval process)	Quantitative(Precision at K(P@K), Normalised Discounted Cumulative Gain (NDCG), and Mean Average Precision (MAP))	Large-scale setting (12 million interactions by users, 3000 randomly selected queries out of 1.2 million unique queries, open web corpora using a major search engine)	(Agichtein et al. 2006)
System Performance (retrieval process)	Quantitative (11-point precision)	Large-scale setting (7 million unique English queries from MSN Search logs, 5000 randomly selected French queries out of 3 million queries from a French query log, 25 French-English query pairs, TREC-6 collection)	(Gao et al. 2007)
System Performance (user model & retrieval process)	Qualitative & Quantitative (questionnaires for users about how well the model depicted their interests & 11-point precision)	Controlled setting (16 users, 3 information needs per user, open web corpora via ProFusion)	(Pretschner and Gauch 1999)
System Usability & Performance (usability & retrieval process)	Qualitative & Quantitative (usability questionnaire & 11-point precision, rank scoring based on explicit relevance judgments by users)	Controlled setting (24 users, 15 information needs per user, open web corpora via AltaVista wrapper)	(Micarelli and Sciarrone 2004)
User Performance (task-based)	Quantitative (time and number of actions needed to complete search tasks)	Controlled setting (48 users, 12 information needs per user, open web corpora via Google wrapper)	(Pitkow et al. 2002)

3 The Proposed Personalised Framework

3.1 Methodology and Architectural Approach

The rational unified process (RUP) Methodology was used in the design and implementation of the framework described by this paper. The RUP methodology is significant with respect to: i) conducting iterative development, ii) requirements management, iii) designing a component-based architecture iv) visual modeling of the system, v) quality management and vi) change control management. The user-centred evaluation approach is used in order to verify the quality of an AIRS, detecting problems in the system functionality or interface, and supporting adaptivity decisions.

The framework is designed as a web-based 3-tier architecture, as can be seen in Figure 1, which consists of: *i) the presentation layer, ii) The business logic layer* which is pulled out from the presentation tier, it controls the frameworks functionality by performing detailed processing and *iii) the data persistence layer* which keeps data neutral and independent from application servers or business logic. The framework is divided into 4 major sections (i.e. the recommender, repository for current studies and search interface, and a user-centred evaluation methodology).

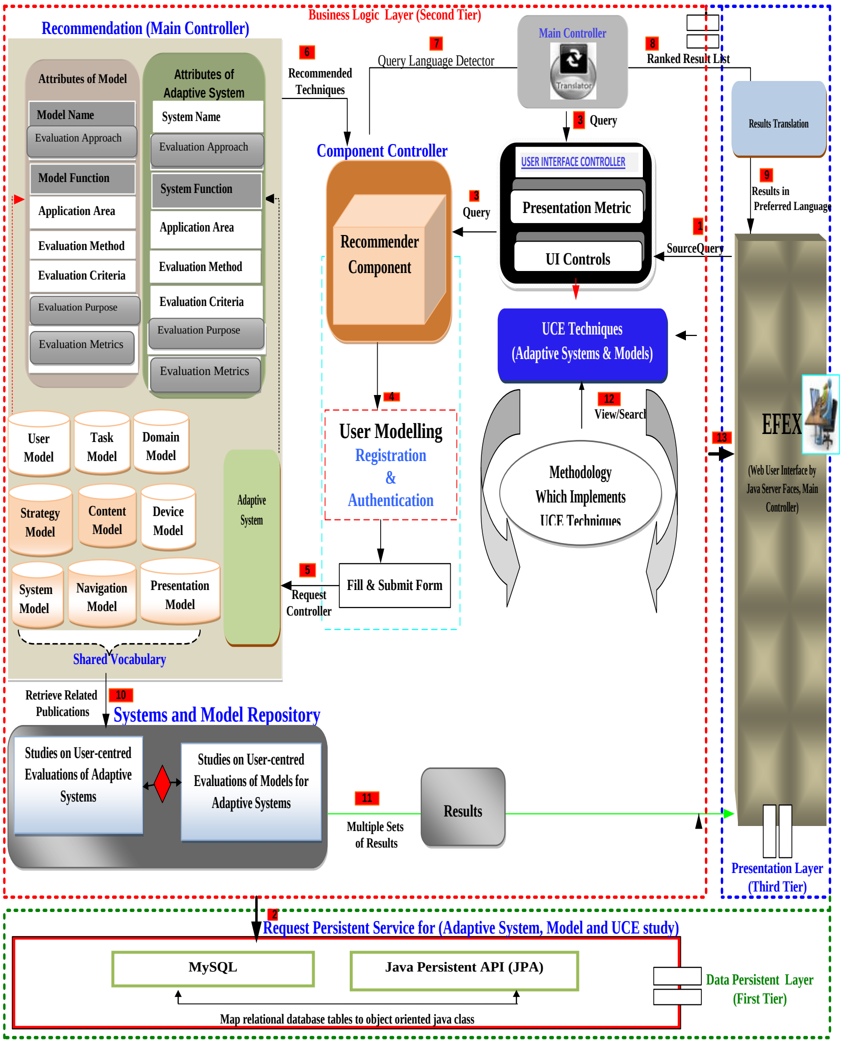


Figure 1: Architectural Design of Proposed Personalisation Framework

3.2 Implementation and Technologies Used

A combination of several technologies was used to implement the framework: **NetBeans 6.9** - platform, **Apache Lucene -** Search engine, **Apache OpenJPA** - To store and Retrieve data from the database, **Apache Tomcat** - server, **Myfaces-core** - Java Server Faces (JSF) used to display data on the Web, **MySql-win32** - MySql database server, **MySql-connector-java** - Connector for java to communicate to mySql, **Google Translate** – to translate the presented information into users choice of language, **Json** - To parse translations.

3.3 Proposed Implicit Recommendation Algorithm

The algorithm implemented in this framework applies implicit recommendation techniques to personalise and recommend evaluation methods, metrics and criteria. Suppose two types of users want to use the framework: i) **User A** wants to get recommendations on how to evaluate an AIRS system. The user does not know which methods, criteria or metrics to use; ii) **User B** wants to get recommendations on how to evaluate an AIR system he/she has developed. The user knows which methods, criteria or metrics to use, but is not sure whether they are the most appropriate ones. He/she wants recommendations on how to evaluate his system. Using the algorithm provided in Figure 2, the framework provides implicit recommendations to the users.

Figure 2: Functionality of the Recommender Algorithm

Start:

Step1: The user selects the system categories and approach in the initial steps.

Step 2. Using the categories selected the recommender does the following

a. Select all the systems belonging to these categories
b. Select all the evaluations that have been carried out on these systems
c. Using the approach of these evaluations all the methods, metrics and criteria are retrieved from database together with their evaluation results.
d. All the evaluation results for each method, metric and criteria are stored in a list.
e. Each result has a success score and a flag as to whether this evaluation was carried out specifically for this system or not. If it was it is given extra weight in the scoring process.
f. When all the results for each method, metric and criteria are collated they are added up and the list is sorted by score.
g. The results are presented as a percentage of the highest score in the list which will always have 100%
h. If the methods, metrics and criteria in the list match the methods, metrics and criteria being used in the current evaluation then they may be highlighted in the list.

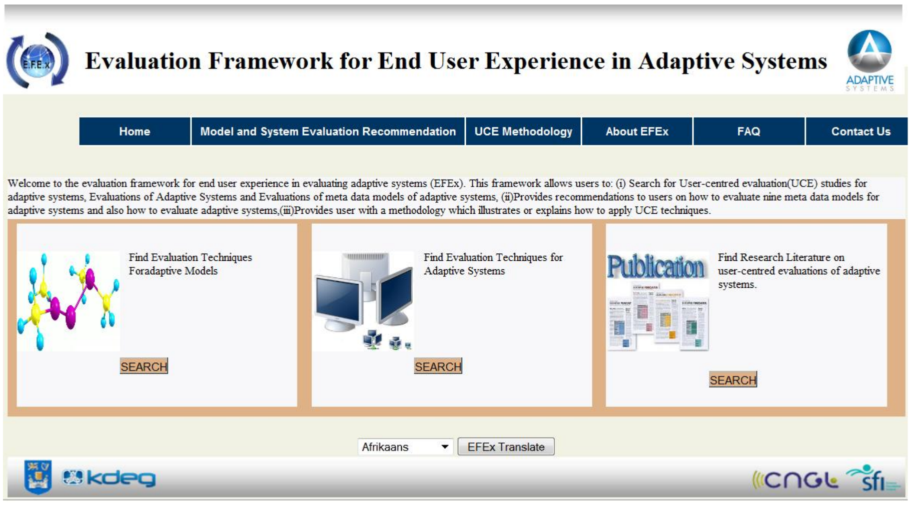
End

3.4 Benefits and Functions of the Framework

Users of the framework are provided with personalised information to suit the user’s requirements. In this case the framework considers the users interests and preferences in order to provide personalised services. Figure 3 presents the index page of the framework. Users are able to:

- Search for literature published from 2000 to date, such as user-centred evaluation (UCE) studies or evaluations of adaptive systems (i.e. adaptive hypermedia, adaptive educational hypermedia, adaptive e-learning, adaptive recommender, PIR and AIRS systems). The query results presented to the user are based on the following characteristics of the evaluated system: system name, developer, evaluation approach, evaluation purpose, system description, application area, evaluation methods, evaluation criteria, evaluation metrics, year of evaluation and finally what was improved by the adaptation.
- Get implicit recommendations on how to combine different evaluation methods, metrics and measurement criteria in order to evaluate a specific system.
- Translate the user interface into 49 different languages to suit the user.

Figure 3: A Web-based User Interface



Evaluation Framework for End User Experience in Adaptive Systems

Welcome to the evaluation framework for end user experience in evaluating adaptive systems (EFEx). This framework allows users to: (i) Search for User-centred evaluation(UCE) studies for adaptive systems, Evaluations of Adaptive Systems and Evaluations of meta data models of adaptive systems, (ii)Provides recommendations to users on how to evaluate nine meta data models for adaptive systems and also how to evaluate adaptive systems,(iii)Provides user with a methodology which illustrates or explains how to apply UCE techniques.

Find Evaluation Techniques Foradaptive Models

Find Evaluation Techniques for Adaptive Systems

Find Research Literature on user-centred evaluations of adaptive systems.

3.4 Task-based Experiments and User Evaluations

To evaluate the framework, three phases of evaluation were defined (requirements specification, preliminary evaluation and final evaluation phase). For each phase, the appropriate evaluation methods, metrics and criteria were identified. Currently, only the requirement specifications and preliminary evaluations have been conducted. This involved interviewing 12 domain experts and conducting a task-based experiment. The use of interviews provided qualitative feedback on user experience after using the framework. The experiment was designed based on a task-based problem scenario.

The task based experiment was significant in evaluating the overall performance and usefulness of the developed framework. In this case, 10 test users were presented with a list of tasks. The techniques adopted was based on internal quality estimation consisting of six characteristics: i) functionality, concerned with what the framework does to fulfil user needs; ii) reliability, evaluating the frameworks capability to maintain a specified level of performance; iii) usability, assessing how understandable and usable the framework is; iv) efficiency, evaluating the capability of the framework to exhibit the required performance with regards to the amount of resources needed; and v) maintainability, concerned with the framework's capability to be modified and finally portability, which will involve measuring the frameworks capability to be used in a distributed environment.

The results from the requirements specification and preliminary evaluation phase were used to improve the functionality of the developed framework. A major evaluation will be conducted for the final phase. This will involve a large number of users performing several tasks.

4 Conclusion and Future Work

This paper described a review and classification of personalised IR approaches and evaluation techniques for PIR systems in the literature. Future personalised IR systems could build on harnessing the benefits of both implicit and explicit approaches to gathering user information and feedback about the user's searches. There are currently no standard evaluation frameworks for AIRS systems. The framework presented in this paper will be a significant contribution to both the AH and IR scientific communities. Evaluators of AIRS systems should ensure that the correct evaluation methods, metrics and criteria are used while evaluating these systems. Two major evaluations of the framework will be conducted in future to test the: i) usability and performance of the overall framework and ii) end-user experience of using the framework.

Acknowledgements

This research is based upon works supported by Science Foundation Ireland (Grant Number: 07/CE/I1142) as part of the Centre for Next Generation Localization (www.cngl.ie). The authors are grateful for the suggestions of the reviewers for this paper.

References

AGICHTEIN, E., BRILL, E. & DUMAIS, S. 2006. Improving Web Search Ranking by Incorporating User Behavior Information. *29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006).* Seattle, Washington, USA: ACM.

BARRY, C. L. & SCHAMBER, L. 1998. Users' criteria for relevance evaluation: a cross-situational comparison. *Information processing & management,* 34**,** 219-236.

CHIRITA, P.-A., FIRAN, C., S. & NEJDL, W. 2007. Personalised Query Expansion for the Web. *30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007).* Amsterdam, The Netherlands: ACM.

CLEVERDON, C. W., Mills, J., and Keen, E. M. (1966) *An inquiry in testing of information retrieval systems (2 vols.)* (Cranfield, UK: Aslib Cranfield Research Project, College of Aeronautics).

GAO, W., NIU, C., NIE, J.-Y., ZHOU, D., HU, J., WONG, K.-F. & HON, H.-W. 2007. Cross-Lingual Query Suggestion Using Query Logs of Different Languages. *30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007).* Amsterdam, The Netherlands: ACM.

LAWLESS, S., MULWA, C. & O'CONNOR, A. Year. A Proposal for the Evaluation of Adaptive Personalised Information Retrieval. *In:* Proceedings of the 2nd International Workshop on Contextual Information Access, Seeking and Retrieval Evaluation, 28th March 2010 Milton Keynes, UK. CEUR-WS.org 4.

KOUTRIKA, G. & IOANNIDIS, Y. 2004. Rule-based Query Personalised in Digital Libraries. *International Journal on Digital Libraries,* 4**,** 60-63.

MICARELLI, A. & SCIARRONE, F. 2004. Anatomy and Empirical Evaluation of an Adaptive Web-Based Information Filtering System. *User Modeling and User-Adapted Interaction,* 14**,** 159-200.

PRETSCHNER, A. & GAUCH, S. 1999. Ontology Based Personalised Search. *11th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 1999).* Chicago, Illinois, USA: IEEE.

PITKOW, J., SCHUTZE, H., CASS, T., COOLEY, R., TURNBULL, D., EDMONDS, A., ADAR, E. & BREUEL, T. 2002. Personalised Search. *Communications of the ACM,* 45**,** 50-55.

SMYTH, B. & BALFE, E. 2006. Anonymous Personalised in Collaborative Web Search. *Information Retrieval,* 9**,** 165-190.

SPERETTA, M. & GAUCH, S. 2005a. misearch. *IEEE/WIC/ACM International Conference on Web Intelligence (WI 2005).* Compiegne University of Technology, France: IEEE Computer Society.

STAMOU, S. & NTOULAS, A. 2009. Search Personalised Through Query and Page Topical Analysis. *User Modeling and User-Adapted Interaction,* 19**,** 5-33.

STEFANI, A. & STRAPPARAVA, C. 1999. Exploiting NLP Techniques to Build User Model for Web Sites: the Use of WordNet in SiteIF Project. *2nd Workshop on Adaptive Systems and User Modeling on the World Wide Web.* Toronto, Canada

SUGIYAMA, K., HATANO, K. & YOSHIKAWA, M. 2004. Adaptive Web Search Based on User Profile Constructed without Any Effort from Users. *13th International Conference on World Wide Web (WWW 2004).* New York, USA: ACM.

TEEVAN, J., DUMAIS, S. T. & HORVITZ, E. 2005. Personalizing Search via Automated Analysis of Interests and Activities. *28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005).* Salvador, Brazil: ACM.

TOBAR, C. M. 2003. Yet another evaluation framework. *Second Workshop on Empirical Evaluation of Adaptive Systems is part of the 9th International Conference on User Modeling*