An Online Framework for Supporting the Evaluation of Personalised Information Retrieval Systems

Catherine Mulwa[1], Luca Longo[2], Séamus Lawless[1], Mary Sharp[1], Vincent Wade[1]

Knowledge and Data Engineering Group[1], Distributed Systems Group[2]
School of Computer Science and Statistics,
Trinity College Dublin
{mulwac, llongo, seamus.lawless, mary.sharp , vincent.wade}@scss.tcd.ie

ABSTRACT. Scope - Personalised Information Retrieval (PIR) has been gaining attention because it investigates intelligent ways for enhancing content delivery. Web users can have personalised services and more accurate information. Problem - Several PIR systems have been proposed in the literature; however, they have not been properly tested or evaluated. Proposal – The authors propose a generally applicable web-based interface, which provides PIR developers and evaluators with: i) implicit recommendations on how to evaluate a specific PIR system; ii) a repository containing studies on user-centred and layered evaluation studies; iii) recommendations on how to best combine different evaluation methods, metrics and measurement criteria in order to most effectively evaluate their system; iv) a UCE methodology which details how to apply existing UCE techniques; v) a taxonomy of evaluations of adaptive systems; and vi) interface translation support (49 languages supported).

Keywords: *Personalised Information Retrieval, User-Centred Evaluation, Layered Evaluation.*

1. INTRODUCTION

The field of Personalised Information Retrieval (PIR) has been gaining momentum thanks to its ability to provide quantitative personalised content delivery. This research is at the intersection of the Information Retrieval (IR) and Adaptive Hypermedia (AH) research fields [1-2].

IR systems have the advantage of scalability when dealing with large document collections and performing large amounts of information processing. In this paper we will describe both PIR and AIR systems. PIR systems adapt the retrieval process to the individual whereas AIR systems aim at capturing and exploiting user context in the retrieval process. In general, "an adaptive system refers to a system which tailors its output, using implicit inferences based on interaction with the user" [3]. According to [4] an adaptive hypermedia system (AHS) "refers to any hypertext and hypermedia system which reflects some features of the user in a user model and applies this model to adapt various visible aspects of the system to the user". AH systems have the advantage of satisfying user needs. Evaluating PIR systems is a non-trivial task. In PIR, different stages of the retrieval process are adapted to the user. The vast majority of studies in the literature have focused on monolingual PIR and only little work has been done concerning cross-lingual PIR. Evaluation of IR systems has been an integral part of IR research from its early days with the Cranfield experiments [5]. 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. Personalisation in information retrieval aims at improving the user's experience by incorporating user subjectivity to the retrieval process.

In this paper, the authors propose a framework for supporting the evaluation of personalised information retrieval (PIR) systems. The main goal of this architecture is to provide comprehensive support to end-users to evaluate their systems. PIR software developers and evaluators can get recommendations on how to combine different evaluation methods, metrics and criteria while evaluating their systems along with the most suitable evaluation approach to use. Access to a repository of evaluation approaches is supported for geographically distributed users of any

nationality by facilitating dynamic translation of content.

The remainder of this paper is organized as follows: Section 2 presents a review of Personalised Information Retrieval Systems evaluation and evaluation techniques for AIR systems. Section 3 introduces the framework, with emphasis on the implementation of the recommender algorithm and overall evaluation process. Section 4 is aimed at describing the framework validation. Section 5 concludes the paper stressing future work and open issues.

2. A REVIEW OF EVALUATION OF PIR AND AIR SYSTEMS

The success of IR and AH fields [6-8] have made PIR research possible, PIR is motivated by the need to provide tailored information seeking to the individual, not one size fits all. This review focuses on the evaluation of PIR systems and on evaluation techniques for AIR systems. The hybrid systems that emerge from the combination of IR and AH are usually referred to as Adaptive Information Retrieval Systems (AIRs)[9]. In PIR, different stages of the retrieval process are adapted to the individual such as adapting the user's query and/or the results. Most PIR systems use both the user preference profile method and the filtering method, commonly adopted in recommendation systems [7]. The authors acknowledge that the aim of personalisation is to endow software systems with the capability to adapt any aspect of their functionality and appearance at run-time to the particularities of users, to better suit their needs. Personalisation can be performed on an individualised, collaborative, or aggregate scope [10-11]. Individualised personalisation occurs when the system's adaptive decisions are taken according to the interests of each individual user as inferred from their model. Collaborative personalisation occurs when information from several user models is used to determine or alter the weights of interests in other user models [12]. This approach is usually adapted to group users into a number of stereotype classes according to certain similarity criteria between their user models. This is useful for judging the relevance of a certain item or document to a user, based on information coming from other user models belonging to the same group. Stereotypes can be manually pre-defined or automatically learnt by using machine-learning techniques such as clustering. Personalisation can also be developed on an aggregate scope that means when the system does not make use of user models. In this case, aggregated personalisation is guided by aggregate usage data as exhibited in search logs (implicitly inferred general users' interests from aggregate history information) [13-14].

A recent review conducted by the authors [7], was aimed at summarising and comparing different personalisation approaches used in PIR systems (figure 1), and different evaluation techniques adopted in AIR systems (figure 2). A brief overview of this classification criterion is given:

(i) **Scope of Evaluation -** The first criterion is concerned with *what* is being evaluated in the PIR system. Different aspects of a system are subject to evaluation, such as the system's performance and its usability with respect to its users. A system's performance can be measured in terms of the effectiveness of its retrieval process [15] [11, 16] or in terms of how well it was able to depict the user's interests in the user model [17]. Instead, the usability of a system can be evaluated by usability questionnaires [18] or by measuring the user's performance in fulfilling certain tasks when using the system [19].

(ii) **Evaluation Metric & Instrument -** The second criterion is concerned with the different instruments and metrics used for evaluation which can be quantitative or qualitative. Examples of quantitative evaluation include measuring the precision or recall of the retrieved results using one of the well-known metrics in the IR community [13-14]. Similarly, measuring aspects related to a given search task to the user such as the time and number of actions needed to complete the task [19]. On the other hand, examples of qualitative evaluation include subjective questionnaires aimed at investigating the accuracy of the user model [17].

Application Area	Personalisation Scope	Personalisation Approach	Example Publication
Monolingual IR	Individualised	Result Adaptation (result re-ranking)	[20-22]
Monolingual IR & Information Filtering	Individualised	Result Adaptation (result re-ranking)	[8]
Monolingual IR	(1)Individualised & (2) Collaborative	Result Adaptation (result re-ranking)	[23]
Monolingual IR	Aggregate usage data	Result Adaptation (result re-ranking)	[24]
Monolingual IR	Aggregate usage data	Result Adaptation ((1) result scoring & (2) result re-ranking)	[25]
Information Filtering	Individualised	Result Adaptation (result scoring)	[26]
Monolingual IR	Individualised	Query Adaptation (query expansion using keywords from user model)	[27]
Structured Search on a Database	Individualised	Query Adaptation (query rewriting)	[28]
Cross-lingual IR	Aggregate usage date	Query Adaptation (query suggestions using similar queries from multiple languages)	[6]
Monolingual IR	Individualised	Query & Result Adaptation (query expansion using keywords from user model, and result re-ranking)	[29]

Figure 1: Comparison of Personalisation Approaches

In the literature, evaluations of PIR systems [30] have mainly used the system-centered approach. This focuses on the assessment of search algorithms by using statistical techniques and metrics such as precision and recall. Examples include projects such as Cranfield, SMART, STAIRS and TREC[30]. However, with the paradigm shift toward the cognitive and behavioural aspect of IR, there is a growing body of user-centered studies that focus on evaluating end-user satisfaction, performance and use of IR systems. Our framework fits this last class.

Scope of Evaluation	Evaluation Metric & Instrument	Example Publication
System Performance (retrieval process)	Quantitative (Precision at K, Recall at K, F-measure, Break-even point)	[24]
System Performance (retrieval process)	Quantitative (R-precision)	[23]
System Performance (retrieval process)	Quantitative (Normalised Discounted Cumulative Gain (NDCG))	[22]
System Performance (retrieval process)	Quantitative (Normalised Discounted Cumulative Gain (NDCG))	[15]
System Performance (retrieval process)	Quantitative (rank scoring based on explicit relevance judgments by users)	[21]
System Performance (retrieval process)	Quantitative (rank scoring based on implicit relevance judgments from clickthrough)	[20]
System Performance (retrieval process)	Quantitative(Precision at K(P@K), Normalised Discounted Cumulative Gain (NDCG), and Mean Average Precision (MAP))	[25]
System Performance (retrieval process)	Quantitative (11-point precision)	[6]
System Performance (user model & retrieval process)	Qualitative & Quantitative (questionnaires for users about how well the model depicted their interests & 11-point precision)	[31]
System Usability & Performance (usability & retrieval process)	Qualitative & Quantitative (usability questionnaire & 11-point precision, rank scoring based on explicit relevance judgments by users)	[8]
User Performance (task-based)	Quantitative (time and number of actions needed to complete search tasks)	[29]

Figure 2: Summary comparison of evaluation techniques

2.1 Challenges in the Evaluation of PIR Systems

A challenge encountered by researchers developing techniques for personalising search results, is the evaluation of their systems through relevance judgements. A relevance judgment

indicates the documents which are deemed relevant for a certain query by a certain individual. An excellent source of such information is personal query logs and click-through data. However, query logs are not always available to the wider research community due to privacy and monetary concerns. Moreover, the standard test collection in IR, namely the TREC datasets [13], cannot be used for evaluating personalised IR systems, since the topics (queries) and corresponding relevance judgments are not associated with particular users, but are consensus judgments.

Personalisation can indeed enhance the subjective performance of retrieval, as perceived by users, and it is therefore a desirable feature in many situations. However, it can easily be perceived as not appropriate or obtrusive if not handled and evaluated adequately[32]. The evaluation of PIR systems is challenged by the user effect, which is manifest in terms of users' inconsistency in relevance judgment, ranking and relevance criteria usage. In most cases, personalisation in PIR systems is performed by adapting the query and the results to the user's interests. A further concern in the field of personalisation technologies is reliability.

We also selected and analysed 56 publications on evaluations evaluation methodologies for adaptive systems conducted from 2000 to date, more specifically focusing on UCE[33]. Furthermore, reviews done in other areas of adaptive systems, such as Adaptive Educational Hypermedia systems (AEH) [34]. Adaptive Information Retrieval systems (AIR) [35] has lead the authors to propose a framework based on a user-centred evaluation approach (UCE), which is composed by three layers:

1. Requirements specification
2. Preliminary validation
3. Final evaluation phase.

The results from the analysed studies and advice from domain experts were used to design the framework.

3. THE FRAMEWORK

Several authors have emphasized and underlined the importance and the difficulties encountered by evaluators of personalised systems Some of the properties of personalised systems can lead to usability problems that may outweigh the benefits of adaptation (personalisation) to the individual user.. If these properties are not evaluated using the most appropriate evaluation methods and measurement criteria, also the outcomes turn to be not correct [34, 36-37]. The end users of our framework are: i) PIR software developers who want to evaluate these systems and ii) researchers of PIR systems.

3.1 Expected Benefits of the Framework

Several researchers acknowledge that one big issue, when attempting to evaluate adaptive systems, especially PIR systems, is the understanding of the adaptation. More specifically this refers to the benefits of the adaptation process and what would have been the outcome if a different kind of adaptation would have been adopted.

From the literature, it emerges that the evaluation of adaptive systems is a difficult task due to the complexity and the usability issues of such systems [38-41].

The expected benefits our framework can deliver are:

- A centralised repository which stores current UCE and layered studies of PIR systems, models and authoring adaptive technologies. Currently it seems to be very difficult for evaluators and researchers to find this information in a centralised place and reporting of these studies seems to be "sloppy" [42].
- Personalised recommendations; this reduces the time spent and the cost incurred while evaluating PIR systems, models and technologies.
- The ability to collaborate while globally distributed and learn faster.
- A methodology which illustrates how to use UCE techniques.
- A Taxonomy of evaluations of adaptive systems.
- Presented information is translated into 49 different languages to suit the user.

Furthermore the framework can be used to tackle existing difficulties encountered while evaluating PIR systems for instance time taken to identify which evaluation techniques, metrics and criteria to use. The information presented to the user are based on the following characteristics of the evaluated system: system name, URL link to the developer, evaluation approach used, evaluation purpose, brief description of the system, application area, evaluation methods (techniques) used, evaluation criteria, evaluation metrics, year the evaluation was conducted and what was improved by the Adaptation.

3.2 Architectural Design

The framework is designed as a web based 3-tier architecture, as can be seen in Figure 4, which consists of:

(i) *The presentation layer which* display information to the end user (figure 5).
(ii) *The business logic layer* which is pulled out from the presentation tier, it controls the frameworks functionality by performing detailed processing,
(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: the recommender section, a repository for current studies and search interface, a user-centred evaluation methodology and a taxonomy.

3.3 Recommender Algorithm

The recommender algorithm applies implicit recommendation techniques to personalise and recommend evaluation methods, metrics and criteria, as shown in figure 4.

{
- *Step 1:* The user selects the system variation type (adaptive hypermedia systems, personalized information retrieval systems and so forth).
- *Step 2:* In the case the user is a non-expert, the systems recommends an evaluation approach, otherwise the user can select an existing one.
- *Step 3:* Using the selected variation type of a system of step 1, the algorithm does the following:

1. Select all the systems belonging to the variation type selected in step 1.
2. Select all the evaluations that have been carried out on the systems of previous step.
3. Using the evaluation approach defined in step 2, the system retrieves all the methods, metrics and criteria are from database along with their evaluation results.
4. All the evaluation results for each method, metric and criteria are stored in a list.
5. The list is then ranked according to a combination of different factors (we do not provide further details here). The highest scores refer to the most appropriate methods, metrics or criteria.
6. If the methods, metrics and criteria in the list match the methods, metrics and criteria being used in the current evaluation then they are highlighted in the list.
7. Each result as a further flag indicating whether the evaluation was carried out specifically for the considered system or not.

}

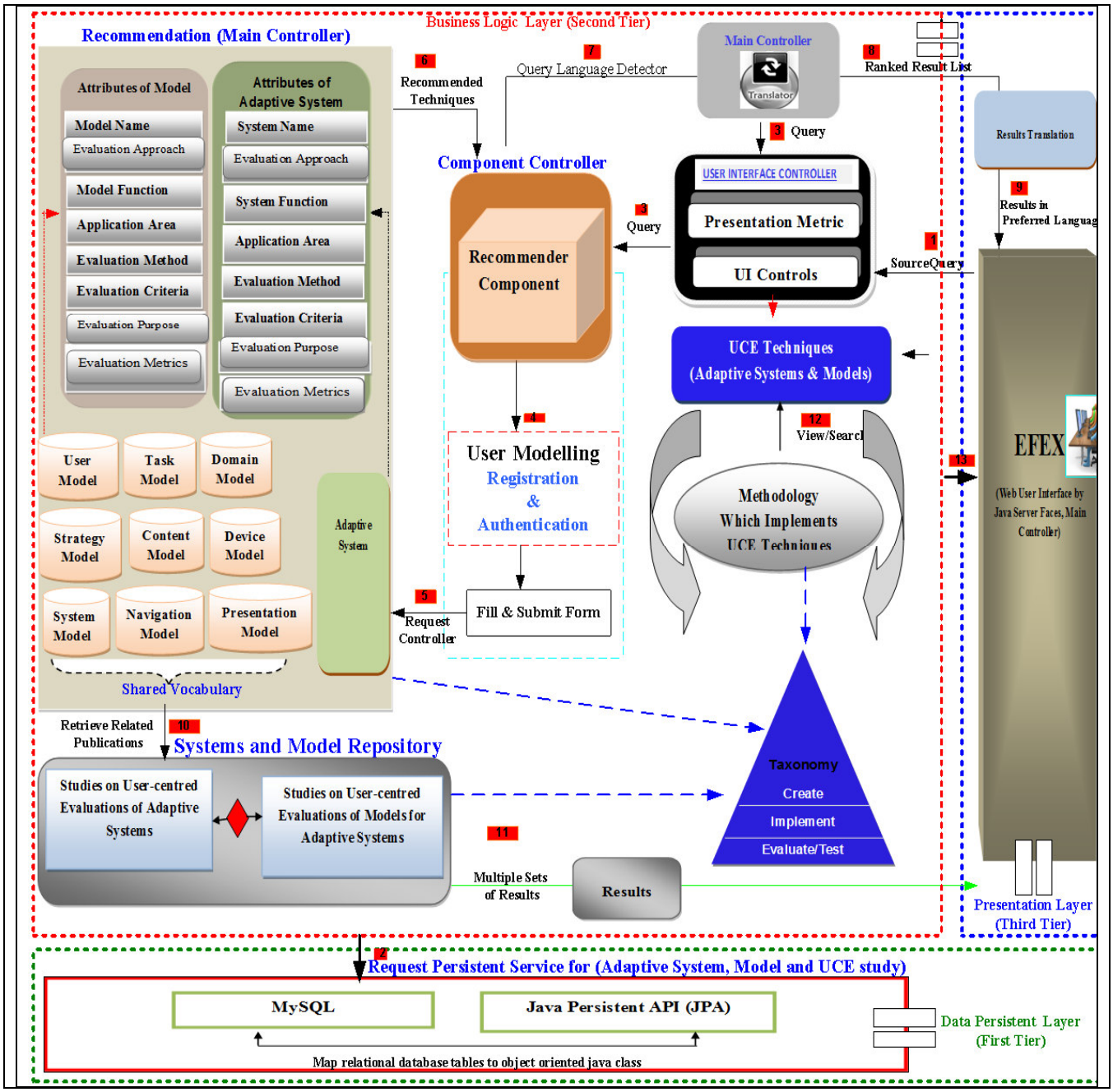

Figure 3: Architectural Design

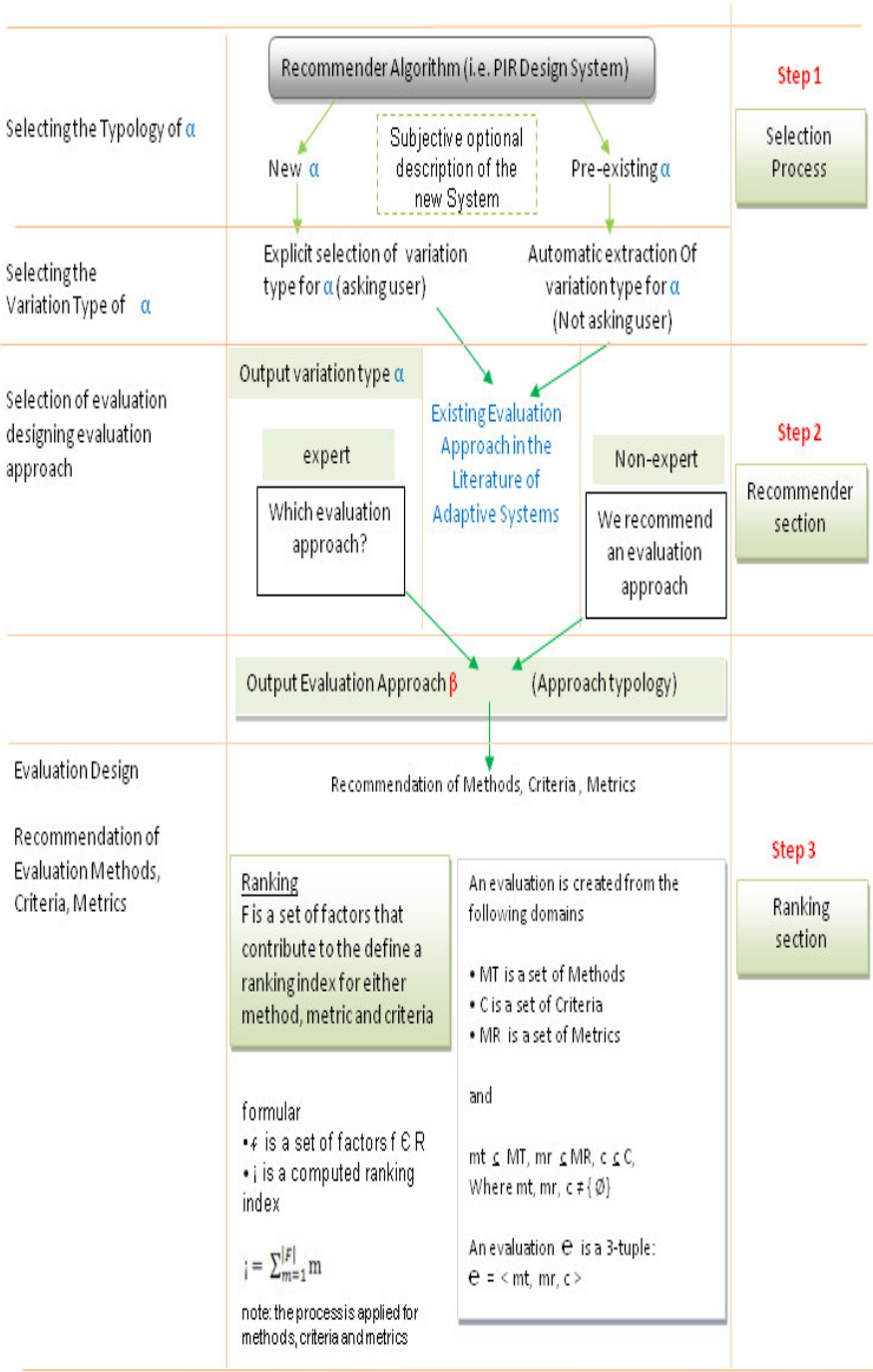

Figure 4: Process of Recommending evaluation methods, metrics and criteria for a PIR System

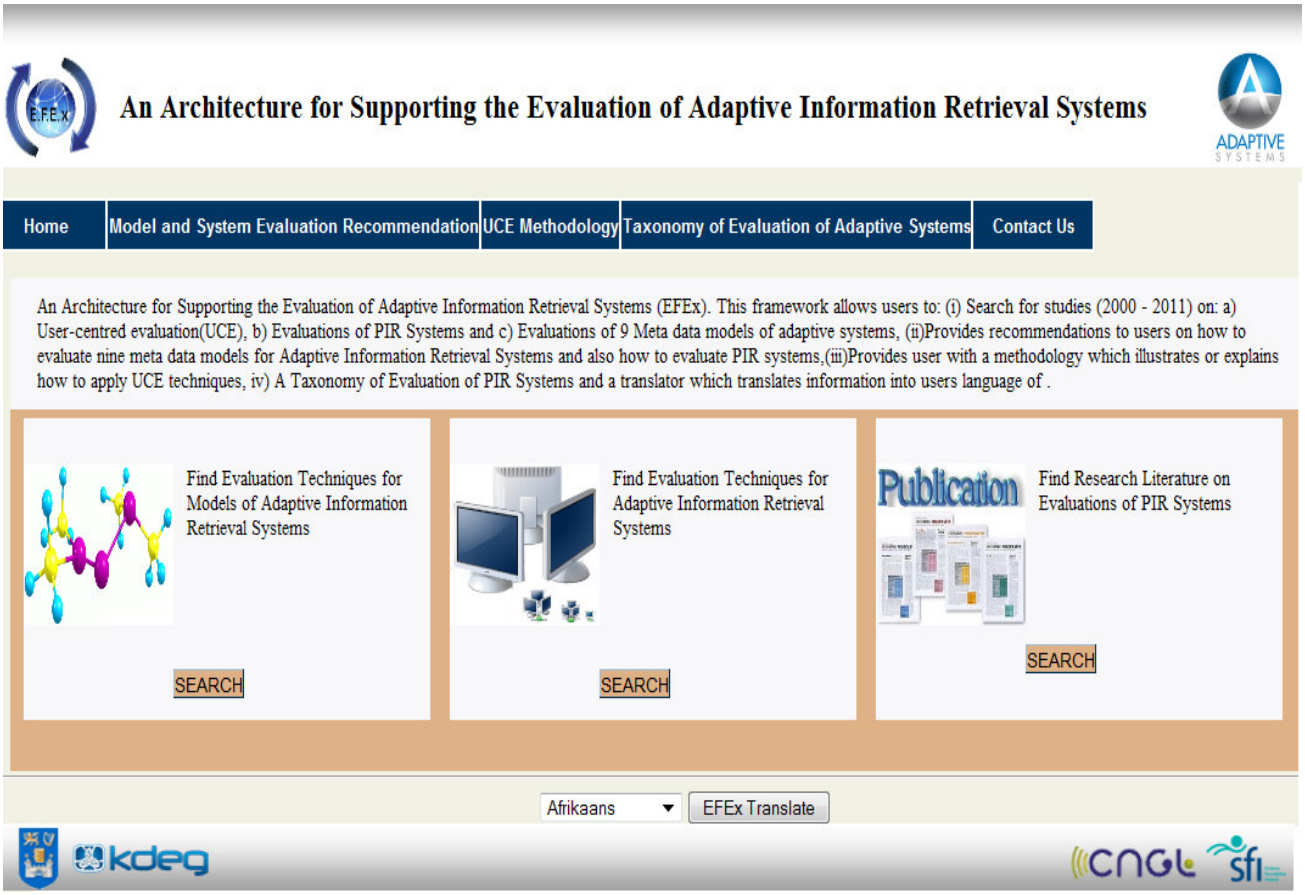

Figure 5: Screen Shot of the Index Page

4. VALIDATION OF THE FRAMEWORK

To validate the usefulness of our framework, for the preliminary evaluation, we interviewed 12 domain experts and conducted a task-based experiment. The use of interviews provided qualitative feedback on expert experience after using the framework.
The techniques adopted were 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 was;
(iv) efficiency, evaluating the capability of the framework to exhibit the required performance with regards to the amount of resources needed;
(v) maintainability, concerned with the framework's capability to be modified
(vi) portability, which will involve measuring the frameworks capability to be used in a distributed environment.

In order to assess the above characteristics, we are currently conducting an online survey. In the following, for instance, we propose a questionnaire dealing with characteristic 1 and 3 that means testing the functionality and usability of our frameworks.

- Have You Developed an Adaptive System in the Past (from 2000 to 2011)?
- (Possible answer: Yes, No)
- If You Have Developed an Adaptive System, What was improved by Adaptivity?
- What is the Variation Type of the Adaptive System You have Developed
- (Possible answer: PIR system, AIR system, AEHS system)
- Please Tick the Meta Data Models Your System Uses.
- (Possible answer: user model, domain model, task model, content model)

- If You Conducted a Whole-System Evaluation, What Evaluation Methods did you use?
- (Possible answer: task-based, interview)
- If you conducted a whole evaluation, what criteria did you use?
- (Possible Answer: Knowledge Gain, Usability, Perceived Usefulness)
- If You Conducted Evaluations of Specific Metadata Models of Adaptive System, What Evaluation
- Methods did you use? (For each model evaluated, please indicate which evaluation methods and criteria you used).

- During this Evaluation (Conducted in Question 5 and 6 above), What Metrics did You Use to Measure Performance against these criteria?
- (Possible answer: Accuracy of Recommendation, Accuracy of retrieval, Behavioural complexity).
- Which of the following features of EFEx Framework would you find (consider) useful? i) Recommendation on how to combine different methods, metrics and metrics to evaluate a PIR system, ii) repository of state-of-the-art review of UCE and layered evaluation of PIR systems, iii) A UCE methodology which illustrates or explains how to apply UCE techniques.
- (Possible answer: A repository)

We are in the process of designing further tests for the remaining characteristics along with a final general model for computing our framework usefulness degree.

5. CONCLUSION

After an overview of current approaches to PIR system evaluation and some of the main challenges for evaluating them, this paper proposes a framework to support designers in evaluating their PIR systems. The framework is based on user-centered approach. A preliminary strategy aimed at validating the framework has been presented. There are currently no standard evaluation frameworks for PIR systems. The framework presented in this paper will be a significant contribution to both the AH and IR scientific communities.

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.

6. ACKNOWLEDGEMENT

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.

6. REFERENCES

[1] S. Gauch, et al., "User Profiles for Personalized Information Access," in The Adaptive Web. vol. 4321, P. Brusilovsky, et al., Eds., 1 ed: Springer, 2007, pp. 54-89.

[2] A. Micarelli, et al., "Personalized Search on the World Wide Web," in The Adaptive Web. vol. 4321, P. Brusilovsky, et al., Eds., 1 ed: Springer, 2007, pp. 195-230.

[3] L. Van Velsen, et al., "User-centered evaluation of adaptive and adaptable systems: a literature review," The knowledge engineering review, vol. 23, pp. 261-281, 2008.

[4] P. Brusilovsky, "Methods and Techniques of Adaptive Hypermedia," User Modeling and User Adapted Interaction, vol. 6, pp. 87-129, 1996.

[5] C.-W. Cleverdon, et al., " Factors determining the performance of indexing systems," Design. ASLIB Cranfield Project. Technical Report, vol. 1 1966.

[6] W. Gao, et al., "Cross-Lingual Query Suggestion Using Query Logs of Different Languages.," presented at the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007), Amsterdam, The Netherlands: , 2007.

[7] C. Mulwa, et al., "Evaluation of Personalised Information Retrieval Systems through Implicit Recommendation," in Task Specific Information Retrieval (TSIR) Workshop being held in conjuction with the 19th International Conference on Conceptual Structures for Discovering Knowledge, Derby, UK, 2011, pp. 366-374.

[8] A. Micarelli and F. Sciarrone, "Anatomy and Empirical Evaluation of an Adaptive Web-Based Information Filtering System.," User Modeling and User-Adapted Interaction, vol. 14, pp. 159-200, 2004.

[9] S. Lawless, et al., "A Proposal for the Evaluation of Adaptive Personalised Information Retrieval," 2010.

[10] M. Speretta and S. Gauch, "Personalized Search based on User Search Histories," presented at the IEEE/WIC/ACM International Conference on Web Intelligence (WI 2005), Compiegne University of Technology, France, 2005.

[11] J. Teevan, et al., "Personalizing Search via Automated Analysis of Interests and Activities," presented at the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005), Salvador, Brazil, 2005.

[12] K. Sugiyama, et al., "Adaptive Web Search Based on User Profile Constructed without Any Effort from Users," presented at the 13th International Conference on World Wide Web (WWW 2004), New York, USA, 2004.

[13] B. Smyth and E. Balfe, "Anonymous Personalization in Collaborative Web Search," Information Retrieval, vol. 9, pp. 165-190, 2006.

[14] E. Agichtein, et al., "Improving Web Search Ranking by Incorporating User Behavior Information," presented at the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006), Seattle, Washington, USA, 2006.

[15] P.-A. Chirita, et al., "Personalized Query Expansion for the Web. ," presented at the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), Amsterdam, The Netherlands, 2007.

[16] P.-A. Chirita, et al., "Personalized Query Expansion for the Web," presented at the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007), Amsterdam, The Netherlands, 2007.

[17] A. Pretschner and S. Gauch, "Ontology Based Personalized Search," presented at the 11th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 1999), Chicago, Illinois, USA, 1999.

[18] A. Micarelli and F. Sciarrone, "Anatomy and Empirical Evaluation of an Adaptive Web-Based Information Filtering System," User Modeling and User-Adapted Interaction, vol. 14, pp. 159-200, 2004.

[19] J. Pitkow, et al., "Personalized Search," Communications of the ACM, vol. 45, pp. 50-55, 2002.

[20] M. Speretta and S. Gauch, "Personalized Search based on User Search Histories," presented at the IEEE/WIC/ACM International Conference on Web Intelligence (WI), Compiegne University of Technology, France, 2005.

[21] S. Stamou and A. Ntoulas, "Search Personalization Through Query and Page Topical Analysis " User Modeling and User-Adapted Interaction, vol. 19, pp. 5-33, 2009.

[22] J. Teevan, et al., "Personalizing Search via Automated Analysis of Interests and Activities.," presented at the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), Salvador, Brazil, 2005.

[23] K. Sugiyama, et al., "Adaptive Web Search Based on User Profile Constructed without Any Effort from Users. ," presented at the 13th International Conference on World Wide Web (WWW), New York, USA, 2004.

[24] B. Smyth and E. Balfe, " Anonymous Personalization in Collaborative Web Search. , , ," Information Retrieval, vol. 9, pp. 165-190, 2006.

[25] E. Agichtein and S. Dumais, "Improving Web Search Ranking by Incorporating User Behavior Information " presented at the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), Seattle, Washington, USA, 2006.

[26] A. Stefani and C. Strapparava, "Personalizing Access to Web Sites: The SiteIF Project," in 2nd Workshop on Adaptive Hypertext and Hypermedia Pittsburgh, Pennsylvania, USA, 1998.

[27] P.-A. Chirita, et al., "Personalized Query Expansion for the Web. ," presented at the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), Amsterdam, The Netherlands, 2007.

[28] G. Koutrika and Y. Ioannidis, "Rule-based Query Personalization in Digital Libraries," International Journal on Digital Libraries,, vol. 4, pp. 60-63, 2004.

[29] J. Pitkow, et al., "Personalized Search " Communications of the ACM, vol. 45, pp. 50-55, 2002.

[30] Y. D. Wang and G. Forgionne, "A decision-theoretic approach to the evaluation of information

retrieval systems," Information processing & management, vol. 42, pp. 863-874, 2006.

[31] M. Speretta and S. Gauch, " misearch " presented at the IEEE/WIC/ACM International Conference on Web Intelligence (WI) Compiegne University of Technology, France, 2005.

[32] P. Castells, et al., "Self-tuning personalized information retrieval in an ontology-based framework," in OTM Workshops, 2005, pp. 977-986.

[33] C. Mulwa, et al., "The Evaluation of Adaptive and User-Adaptive Systems: A Review," In the International Journal of Knowledge and Web Intelligence (IJKWI), 2011.

[34] C. Mulwa, et al., "Adaptive Educational Hypermedia Systems in Technology Enhanced Learning: A Literature Review," presented at the Association for Computing Machinery's Special Interest Group for Information Technology Education, H Hotel 111 W Main St. Midland,MI 48640, 2010.

[35] S. Lawless, et al., "A Proposal for the Evaluation of Adaptive Personalised Information Retrieval," in In the Proceedings of the CIRSE 2010 Workshop on Contextual Information Access, Seeking and Retrieval Evaluation held in conjunction with ECIR-2010 - European Conference on Information Retrieval, Milton Keynes, England, 2010.

[36] N. Tintarev and J. Masthoff, "Evaluating Recommender Explanations: Problems Experienced and Lessons Learned for the Evaluation of Adaptive Systems," presented at the User Modeling, Adaptation and Personalization, Trento, Italy, 2009.

[37] C. Gena and S. Weibelzahl, "Usability engineering for the adaptive web," *The Adaptive Web,* pp. 720-762, 2007.

[38] F. Missier Del and F. Ricci, "Understanding recommender systems: Experimental evaluation challenges," pp. 31-40, 2003.

[39] T. Lavie*, et al.*, "The evaluation of in-vehicle adaptive systems, User Modeling: Work on the EAS," pp. 9-18, 2005.

[40] S. Weibelzahl and G. Weber, "Advantages, opportunities and limits of empirical evaluations: Evaluating adaptive systems," *KI,* vol. 16, pp. 17-20, 2002.

[41] S. Markham*, et al.*, "Applying agent technology to evaluation tasks in e-learning environments," 2003, pp. 16–17.

[42] L. Van Velsen*, et al.*, "User-centered evaluation of adaptive and adaptable systems: a literature review," *The Knowledge Engineering Review,* vol. 23, pp. 261-281, 2008.