A Proposal for the Evaluation of Adaptive Personalized Information Retrieval

Séamus Lawless, Alexander O'Connor, Catherine Mulwa

Centre for Next Generation Localization
School of Computer Science and Statistics
Trinity College Dublin,
Dublin, Ireland

seamus.lawless@sccs.tcd.ie, alex.oconnor@sccs.tcd.ie, mulwac@sccs.tcd.ie

ABSTRACT

Personalisation in Information Retrieval is achieved using a range of contextual information such as information about the user, the task being conducted and the device being used. This information is used to devise the most suitable response for the individual's need. As such personalised Information Retrieval and response composition approaches become more widely used, traditional evaluation measures become less effective and applicable. This paper proposes that contextual, and specifically personalised, approaches to Information Retrieval could benefit from the experience of the Adaptive Hypermedia community. The paper details the approaches to evaluation commonly used by the IR and AH communities and proposes a means of combining and enhancing these disparate approaches in a unified framework for the evaluation of personalised IR systems.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.5 [**Information Interfaces and Presentation**]: Multimedia Information Systems; H.5 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia;

General Terms

Experimentation, Measurement, Performance

Keywords

Information Retrieval, Personalisation, Adaptive Hypermedia, Evaluation

1. INTRODUCTION

Contextual information is increasingly being used to facilitate personalisation in Information Retrieval (IR). The personalised identification, retrieval and presentation of resources can provide the user with a tailored information seeking experience. Such tailored experiences can produce a more informative response than a traditional ranked list approach.

As personalised information retrieval and response composition become more widely used, traditional approaches to IR evaluation become less effective and applicable. It is the contention of this paper that contextual, and specifically personalised, approaches to IR could benefit from the experience of the Adaptive Hypermedia (AH) community.

Recently, research has been undertaken exploring how to enhance and combine key aspects of AH research with IR research to provide advanced annotation, slicing, retrieval and composition of multilingual digital content drawn from corporate documents repositories as well as open corpus sources [1][2]. We call such systems, which combine AH and IR approaches to deliver personalised information seeking and access, Adaptive Information Retrieval Systems (AIRS).

AH systems have traditionally functioned using low volumes of content with associated rich metadata descriptions, in contrast to web-based IR systems which tend to function using statistical methods on very large scale collections of unannotated content. In order to facilitate effective personalised information retrieval at the scale required by web-based IR systems, a hybrid approach to both the implementation and evaluation of these systems is necessary. This paper proposes an approach to the evaluation of AIRS. Having introduced the disciplines of IR and AH, the paper details the approaches to evaluation commonly used by these communities. The paper then continues by proposing a means of combining and enhancing these disparate approaches to produce a framework for the evaluation of adaptive, personalized information retrieval systems.

2. EVALUATION APPROACHES

2.1 Information Retrieval

The Cranfield model [3] remains the dominant approach to the evaluation of IR systems. This evaluation model uses test collections to assess and compare the performance of IR systems. Typical test collections include a document corpus, a set list of search queries and manually assigned relevance assessments for each query. Such collections are utilised in IR communities such as the Text REtrieval Conference (TREC)[1], the Cross Language

Appears in the Proceedings of The 2nd International Workshop on Contextual Information Access, Seeking and Retrieval Evaluation (CIRSE 2010), March 28, 2010, Milton Keynes, UK.
`http://www.irit.fr/CIRSE/`

[1] The Text Retrieval Conference – http://trec.nist.gov

Evaluation Forum (CLEF)[2] and the National Institute of Informatics Test Collection for IR Systems (NTCIR)[3].

Precision and recall [4] are the metrics upon which the majority of traditional IR evaluation approaches are based, including the Cranfield model. Precision is the fraction of documents retrieved in response to a query that are relevant to the users information need when making that query. Recall is the fraction of the total set of relevant documents in a collection which are returned for a given query. Another common measure of performance is van Rijsbergen's f_1-measure [5], which calculates the weighted harmonic mean of precision and recall.

However, differences between the structure of the WWW and typical document collections and key differences between the users of web-based IR systems and traditional IR systems challenge many of the assumptions of these metrics. The primary concern is that both metrics are highly subjective, as relevancy can only be assigned based upon the users intent when submitting a specific query. In relation to the WWW, recall is an ineffective means of evaluation, as recall requires the entire collection of relevant documents to be known in advance of each query performed [6].

Whilst the precision measure for a set of results is valuable, it does not take into account the importance of ranking in modern, web-based, information retrieval. The average precision is a measure used to estimate the performance of an IR system by placing emphasis on relevant results appearing high up the ranked list. Average precision is calculated by finding the average of the individual precision calculations for each relevant result in the list, if the list were truncated directly after that result. The Mean Average Precision (MAP) for an IR system is then calculated by summing the average precision value for each query conducted and dividing by the total number of queries. This gives an indication of the overall performance of the system with relation to precision and ranking.

The Cranfield model of test collections aims to ensure controlled experimentation across systems, using metrics such as those introduced above, so that conclusions can be formed about the comparative performance of a variety of IR systems. However, this approach is limited as assumptions must be made about user intent and behaviour. Due to the mainstream accessibility of web-based IR systems, assumptions cannot be made regarding a user's level of knowledge with regard to either the system or the subject domain. The IR system must cope with users across the spectrum of knowledge and ability [7].

2.2 Adaptive Hypermedia

Adaptive Hypermedia (AH) is an alternative to the traditional "one-size-fits-all" approach to the development of Hypermedia Systems. Adaptive Hypermedia Systems (AHS) can potentially adapt on any attributes of a user, for example in Adaptive Educational Hypermedia Systems (AEHS) models can be constructed which describe the goals, preferences, knowledge or other attributes of each individual user. This model can then be used during any interactions with the user, in order to adapt aspects of the systems functionality or presentation strategy to the needs of that user [8]. AH approaches are applied in educational hypermedia, on-line information systems, on-line help systems, institutional hypermedia and systems for managing personalized views.

2.2.1 Current Evaluation of Adaptive Systems

In order to produce effective results, evaluation should occur throughout the entire design cycle and provide feedback for design modification [9].

User-centered approach: User-centered evaluation (UCE) can serve three goals: verifying the quality of an AHS, detecting problems in the system functionality or interface, and supporting adaptivity decisions [10]. These functions make UCE a valuable tool for developers of all kinds of systems, because they can justify their efforts, improve upon a system or help developers to decide which version of a system to release. The benefits of the user-centered approach are savings in terms of time and cost, ensuring the completeness of system functionality, minimizing required repair efforts, and improving user satisfaction [11].

Empirical approach: Empirical evaluations, also known as controlled experiments, refer to the appraisal of a theory by observation in experiments. These evaluations help to estimate the effectiveness, efficiency and usability of a system and may uncover certain types of errors in the system that would remain otherwise undiscovered. The key to good empirical evaluation is the proper design and execution of the experiments so that the particular factors to be tested can be easily separated from other confounding factors. This method of evaluation is derived from empirical science and cognitive and experimental psychology [12].

Layered approach: The layered evaluation approach [13][14] separates the 'interaction assessment' and the 'adaptation decision'. Both layers should be evaluated separately in order to effectively interpret the evaluation results. Evaluating AHS on a layer by layer basis has been recommended as a more comprehensive approach [14][15]. In contrast to approaches that focus on the overall user's performance and satisfaction [16], layered evaluation in particular assesses the success of adaptation by decomposing it into different layers and evaluating each layer individually. This has a number of advantages over other approaches, such as useful insight into the success or failure of each separate adaptation stage, facilitation of improvements, generalization of evaluation results and re-use of successful practices.

Utility-based approach: Current evaluation practices attempt to evaluate adaptation as a whole, with user satisfaction or performance as the overall metric for success, based on identified measurable criteria. In the utility-based approach the evaluation can be seen as a utility function X that maps a system, given some user context, to a quantitative representation of user satisfaction or performance. For example if one compares an adaptive system with its non-adaptive counterpart, the value of adaptation is the difference in utility between the two systems.

As described above, the main advantage of layered evaluation methods are that they break the utility function into several distinct functions. For example suppose there is a utility X_1 that maps the interaction assessment and the resulting user model to a real number that represents its correctness. Suppose there is also a

[2] The Cross Language Evaluation Forum - http://www.clef-campaign.org/

[3] The National Institute of Informatics Test Collection for IR Systems - http://research.nii.ac.jp/ntcir/

utility function U_2 that maps a system, given some user model, to a real number that represents user satisfaction or performance. In this case the whole utility function can be expressed as $X = X_1 X_2$. It is clear that the latter utility function better indicates the usability of an adaptive hypermedia system. Utility-based evaluation of adaptive systems [17] offers a perspective of how to reintegrate the different layers.

Heuristic approach: A heuristic is a general principle or rule of thumb that can be used to critique existing decisions or guide a design decision. An approach which integrates layered evaluation and heuristic evaluation has been proposed [18]. The use of heuristics ensures that the entire system can be evaluated in depth and specific problems can be discovered at an early design stage before releasing a running prototype of a system [19]. This approach can help evaluators by improving the detection and diagnosis of potential usability problems.

2.2.2 Challenges in Evaluating of Adaptive Systems

The evaluation of AHS is a difficult task due to the complexity of such systems, as shown by many studies [20][21]. It is of crucial importance that the adaptive features of the system can be easily distinguished from the general usability of the designed tool. Issues arise in the selection of applicable criteria for the evaluation of adaptivity. Many metrics can be used to measure performance, for example: knowledge gain (AEHS), amount of requested materials, duration of interaction, number of navigation steps, task success, usability (e.g., effectiveness, efficiency and user satisfaction). The evaluation of adaptive systems is not easy, and several researchers have pointed out potential pitfalls when evaluating adaptive systems. Examples of pitfalls [22] include:

- Difficulty in attributing cause: is the adaptation causing the measured effect or another aspect of system functionality or design (e.g. system usability).
- Statistically insignificant results: Adaptivity is typically used when individual users differ. However, differences in approach and preferences are likely to lead to a large variance in performance results, which makes it more difficult to produce statistically comparable results. In order to produce significant results, large volumes of queries and users are required. There are few general guidelines for the selection of these measurements.
- Difficulty in defining the effectiveness of adaptation: It can be difficult to define what constitutes a useful or helpful adaptation.
- Insufficient resources: To fully evaluate an adaptive system it is often necessary to have a large number of individuals interacting with the system. This is in part due to the expected variance between participants mentioned above.
- Too much emphasis on summative rather than formative evaluation: Evaluations often measure only how good or bad a system is rather than providing information on where the problems are and how a system can be improved.

The selection of the metrics to be used in the evaluation of AHS is crucial. There are currently no agreed evaluation methodology standards, thus making AH evaluation a difficult, complex and time consuming task**.**

3. EVALUATION METHODOLOGY

There are many methods of combining the techniques used in AH and IR. However, in order to define a common evaluation mechanism, we classify a content composition as the output from an AIRS. A content composition is an aggregated set of resources ordered according to the user's needs and preferences.

In order to sufficiently evaluate both the adaptive functionality and the retrieval performance of an AIRS, a hybrid approach is necessary. This involves user-centric assessment, layered evaluation of the personalisation which has been applied and quantitative performance metrics relating to the content delivered.

The layered approach to evaluation would allow the assessment of each individual aspect of personalisation applied within the AIRS to tailor the experience delivered to the user. By assessing each piece of functionality in isolation, it can be determined which provide the most value in relation to the experience delivered to the user. The layers which must be evaluated will differ for each system as the adaptation and personalization techniques used will vary depending on the range of models and adaptive presentation modalities available in the system.

There are a set of necessary elements of a hybrid AIRS evaluation model which can be defined irrespective of the system being evaluated. The key challenge is to be able to adequately combine the data-driven approach to assessing retrieval from IR evaluation with the more user-focused approach to evaluation from AH. The hybrid framework for AIRS evaluation proposed by this paper is as follows:

1. **Evaluating Query Formation.** Traditional IR methods are driven by keyword-driven queries. However, with the introduction of query term personalization and expansion, the effectiveness of these components must be evaluated. This can be achieved by assessing the correctness of the inference of a users' intention and information need.
2. **Evaluating Retrieval Effectiveness.** This is an assessment of the AIRS system in terms of traditional retrieval tasks. There are some additional constraints, however, as systems which use adaptive presentation methods are required to retrieve ordered sequences of content, which creates a more complex dependency than a list of independent candidate documents.
3. **User-Centric Evaluation.** A user-centric evaluation of the system is necessary to assess the assumptions made and adaptivity performed. Key metrics in user-centric approaches to evaluation include satisfaction, effectiveness and efficiency.
   a. **Evaluating Adaptivity Effectiveness and Efficiency**. The effectiveness of adaptivity performed by the system can be measured using domain appropriate performance indicators, for example, knowledge gain and knowledge retention in educational scenarios and information need fulfillment in IR. Different metrics can be assessed depending on the nature of the AIRS application and its use. This is necessarily a user-driven evaluation, because the behavior of the system depends on the properties of the user. Efficiency can be measured by examining user performance when conducting a set of defined tasks. Metrics measured can include time taken. Time taken can include two types of measurements; classical IR evaluations such as time to get the first relevant document and number of documents retrieved in exactly 10mn [23], number of queries needed etc.
   b. **User Satisfaction.** User satisfaction can be examined by eliciting direct user feedback. While there are many

means of eliciting such information, among the most widely used are qualitative questionnaires such as the System Usability Score (SUS) [24].

4. **System efficiency.** The creation of complex AIRS requires a wide variety of content and metadata. As such, the overall cost of provisioning a particular AIRS, and the performance of that system in terms of resource requirements and responsiveness must be assessed. This is particularly important as systems approach mass, web-scale use.

4. SUMMARY

This paper has presented an overview of the evaluation approaches commonly used in the IR and AH communities, with the objective of finding a means of combining these approaches to assess the next generation of personalization in IR. These approaches have traditionally varied in focus: the majority of IR evaluation is focused on the numerical performance and effectiveness of the retrieval of documents, while AH systems are evaluated with the users as an integral part.

The layered evaluation model proposed in this paper is an attempt to draw from the AH community's experience with the complex dependencies of different intelligent components in AHS. This is combined with traditional IR evaluation methods to gain a broad spectrum assessment of the entire system.

User-driven evaluation of IR is not new in itself, but the effects of personalization on creating reproducible, large scale experiments can, in the opinion of the authors, best be addressed by incorporating elements of AH evaluation techniques.

5. ACKNOWLEDGMENTS

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] G. J. F. Jones & V. Wade. Integrated content presentation for multilingual and multimedia information access. In F.C. Gey, N. Kando, C.Y. Lin and C. Peters (eds.), New Directions in Multilingual Information Access, vol. 40, pp. 31–39, 2006.

[2] B., Steichen, S., Lawless, A., O'Connor & V., Wade. Dynamic Hypertext Generation for Reusing Open Corpus Content. In the Proceedings of the 20th ACM Conference on Hypertext and Hypermedia, Hypertext 2009, Torino, Italy. 29th June – 1st July, 2009.

[3] C. W. Cleverdon, J. Mills, & M. Keen. Factors determining the performance of indexing systems. Vol. 1 - Design. ASLIB Cranfield Project. Technical Report, 1966.

[4] G. Salton. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, 1989.

[5] K., van Rijsbergen. Information Retrieval. London, England, Butterworths & Co. Ltd. 1979.

[6] S., Chakrabarti, M., van den Berg & B., Dom. Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery. In The International Journal of Computer and Telecommunications Networking, Vol. 31(11-16), Elsevier, pp. 1623-1640, May 1999.

[7] C., Hölscher & G., Strube. Web Search Behavior of Internet Experts and Newbies. In Proceedings of the 9th International Conference on the World Wide Web, WWW9, Amsterdam, The Netherlands. pp 337-346, 2000.

[8] P. Brusilovsky, E. Schwarz & G. Weber. A Tool for Developing Adaptive Electronic Textbooks on WWW, In the Proceedings of WebNet '96 World conference of the web society, pp. 64-69, 1996.

[9] C. Gena & S. Weibelzahl. Usability Engineering for the Adaptive Web, vol. 4321, LNCS, pp. 720-762, 2007.

[10] M. De Jong & P. Schellens. Reader-Focused Text Evaluation. An overview of goals and methods, vol. 11(4), pp. 402-432, 1997.

[11] J. Nielsen. Usability Engineering, Boston: MA: Academic Press, 1993.

[12] C. Gena. Methods and techniques for the evaluation of user-adaptive systems, The knowledge engineering review, vol 20:1, pp. 1-37, United Kingdom: Cambridge University Press, 2005.

[13] C. Karagiannidis & D. Sampson. Layered evaluation of adaptive applications and services, International Conference on adaptive hypermedia and adaptive applications and services, 2000.

[14] P. Brusilovsky, C. Karagiannidis & D. Sampson. The benefits of layered evaluation of adaptive applications and services, In the proceedings of the first workshop on empirical evaluation of adaptive systems, UM2001, Sonthofen, Germany, pp. 1-8, 2001.

[15] S. Weibelzahl & G. Weber. Advantages, opportunities and limits of empirical evaluations, Evaluating adaptive systems, vol. 3(2), 2002.

[16] D. Chin. Empirical evaluation of user models and user-adapted systems, pp. 181-194, 2001.

[17] E. Herder. Utility-based evaluation of adaptive systems, In the proceedings of the Second Workshop on Empirical Evaluation of Adaptive Systems, at the 9th International Conference on User Modeling, UM2003, Pittsburg, USA, pp. 25-30, 2003.

[18] G. Magoulas, S. Chen & K. Papanikolaou. Integrating layered and heuristic evaluation for adaptive learning environments, In the proceedings of the Second Workshop on Empirical Evaluation of Adaptive Systems, at the 9th International Conference on User Modeling, UM2003, Pittsburg, USA, pp. 5-14, 2003.

[19] L. Fu, G. Salvendy & L. Turley. Effectiveness of user testing and heuristic evaluation as a function of performance classification. Behaviour and information technology, vol. 21, pp. 137-143, 2002.

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

[21] T. Lavie, J. Meyer, K. Beugler & J. Coughlin, The evaluation of in-vehicle adaptive systems, User Modeling: Work on the EAS, pp. 9-18, 2005.

[22] N. Tintarev and J. Masthoff, "Evaluating Recommender Explanations: Problems Experienced and Lessons Learned for the," *UMAP 2009,* p. 54, 2009.

[23] E. Crestan, C. de Loupy, "Browsing Help for a Faster Retrieval; COLING 2004," 2004, pp. 576-582; Geneve, Suisse.

[24] J. Brooke. SUS: a "quick and dirty" usability scale. In Usability Evaluation in Industry, P.W. Jordan, B. Thomas, B.A. Weerdmeester & A.L. McClelland (eds.), London: Taylor and Francis, 1996