A Proposal for the Evaluation of Adaptive Content Retrieval, Modification and Delivery

Killian Levacher , Seamus Lawless, Vincent Wade

School of Computer Science and Statistics, Trinity College Dublin, Ireland
{Killian.Levacher, Seamus.Lawless,Vincent.Wade}@scss.tcd.ie

Abstract. A key advantage of Adaptive Hypermedia Systems (AHS) is their ability to re-sequence and reintegrate content to satisfy a particular user's need, context or requirements. However, this requires large volumes of content, with appropriate granularities and suitable meta-data descriptions, representing a major impediment to the mainstream adoption of Adaptive Hypermedia. Open-corpus content is now widely available on the web, however, traditional information retrieval (IR) approaches are an inadequate means of incorporating these external content resources within AHS. This is due to the "one size fits all" content delivery paradigm offered by traditional IR. Slicing technology addresses these limitations by providing adaptive retrieval of open corpus resources, tailored to suit AHS specific content requirements. This is achieved through the on demand provision of tailored content called slices. This paper introduces slicing systems and details the objectives and challenges involved in the evaluation of such systems. A framework for the evaluation of slicing systems is presented along with a proposed experimental implementation.

Keywords: Adaptive Hypermedia Systems, Content Slicing, Evaluation

1 Introduction

Adaptive Hypermedia systems have traditionally attempted to deliver dynamically adapted and personalised presentations to users through the sequencing of pieces of information. While the effectiveness and benefits of such systems have been proven in numerous studies [4] [14], a major obstacle to their widespread adoption relates to their traditional closed corpus nature [19]; there is an inherent reliance upon bespoke, proprietary content [1]. The adaptivity that an AHS can deliver can be restricted by a lack of sufficient content in terms of volume, granularity, style and meta-data.

In parallel, a wealth of information has become accessible on the WWW and in digital repositories. This is often referred to as open corpus content and is typically accessed via traditional information retrieval (IR) systems. However, these systems only offer a "one size fits all", untailored delivery of results, with limited control over granularity, content format or associated meta-data. Open corpus material is very heterogeneous. It comes in various formats, languages, is generally very coarse-grained and contains unnecessary noise such as navigation bars, advertisements etc.

One of the key impediments to the incorporation of open corpus content is the absence of a standard representation of resources across AHS. Content is currently authored for particular systems in particular formats and using particular metadata schemas [2]. Metadata standards such as LOM [15] attempt to provide generic content packaging solutions, however they are domain specific and involve a considerable amount of manual effort [8][9] to produce. As a result, it is extremely difficult for AHS to incorporate externally authored content when generating adaptive offerings.

While some AHS do support the incorporation of open corpus content, they do so in a limited, and largely manual fashion[5][6][12]. It is also clear that even in the cases where external content can be used, the content returned by traditional IR systems is not suitable for incorporation by AHS [11].

Slicing technology addresses these limitations by providing adaptive retrieval of open corpus resources, tailored for consumption by individual AHS. Content is delivered in the format chosen by each AHS. Automating the production of adaptive content resources for AHS will lower content production time and costs while providing access to a large volume of content for a wide diversity of AHS. However, to the author's knowledge, it appears as if there is currently no agreed standard evaluation methodology or metrics for such technology, making Slicer evaluation a complex and time consuming task.

Contribution: This paper attempts to i) detail the main characteristics and evaluation challenges across various slicer implementations and ii) propose a standard evaluation methodology for slicing systems. While the experimental implementation and metrics presented are standard across the field of IR, the objective is iii) to present how such standard techniques can be used to support this new methodology. This paper is structured as follows: section 2 presents an overview of the constituents and operation of a slicer along with related work while section 3 presents the challenges involved with evaluating such systems and proposes an evaluation methodology. Section 4 outlines possible experiments, followed by a summary in the final section of this paper

2 The Anatomy of a Slicer and Related Work

Rather than providing a detailed explanation of a slicer's architecture [20] and with the aim of serving an understanding of the overall evaluation challenges involved, this section presents an overview of how a slicer implementation would typically function with respect to the stages involved in the process.

As illustrated in Figure 1 such a system is designed as a pipeline of successive modules, each analysing relevant content previously targeted and harvested by focused crawlers such as the OCCS [18]. As content passes through the pipeline, each component appends meta-data specific to the document being processed.

The first phase of any slicer consists of fragmenting documents into individual independent atomic segments (such as menus, article paragraphs, advertisements etc.). Many algorithms which attempt to infer different sections of documents based on their structure can be used for this task [16][7]. Once an initial set of fragments is produced, each is analysed by a set of standard NLP algorithms (such as entity extraction [21], passage retrieval [22] etc.) with the intention of producing topic maps of semantics contained within each fragment along with their location. The semantic meta-data produced might include, writing style, passage delimitations or even how such a content should be manipulated etc. At this stage of the pipeline, a slicer will contain both a set of fragments with associated structural and semantic meta-data. The resulting array of adjustments (extent of control over granularity, formats and annotation types) available to slice consumers (AHS systems requesting content) prior to slices being generated is referred to as a slicer's Content Adaptation Spectrum (CAS). Whenever slice requests are received, an intelligent slicing unit combines atomic fragments together along with relevant meta-data into tailored slices in an

attempt to map CAS features requested by each consumer. A slice therefore is defined as: "*Tailored content, automatically produced, consisting of fragment(s) from pre-existing document(s) (originally authored for a specific purpose) collated and combined with a set of meta-data, tailored to suit specific content requirements of one particular slice consumer with individual specific content reuse intentions*". Slices can contain other slices and fragments of various documents. They can be reused and collated within many slices.

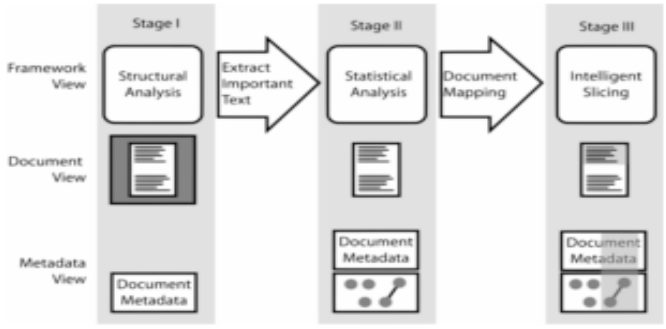

Figure 1 Slicer Pipeline

3 Slicer Evaluation Challenges and Methodology

3.1 Evaluation Challenges

The evaluation of slicing systems is difficult due to the complexity of such systems and the diversity of use cases which they must serve. The ability of a slicer to convert existing content into reusable slices must be evaluated independently to both the context of reuse of slices and the general usability of slice consumers. To the authors' knowledge, there is currently no agreed standard methodology for the evaluation of slicers. However, Adaptive Information Retrieval Systems, as defined by [17], appear to share many evaluation pitfalls with slicers. These include:

Difficulty in attributing cause. Is a measured effect caused by the produced slice or is it due to slice consumers functionality or design (e.g.: system usability)?

Difficulty in defining effectiveness of reuse. Slicing aims at supporting various reuse contexts. Defining what constitutes effective reuse, however, is a challenge.

Statistically Insignificant Results. The quality of a slicer is directly related to the percentage of slices produced being correctly used within different contexts. Evaluating slices in different usage scenarios can lead to wide performance variance, making statistically comparable results more difficult to obtain.

Insufficient resources. Due to the expected performance variances, evaluating slicers involves possessing many slice consumers and users.

Reuse context dependency. How dependent upon the context of reuse is the evaluation performed? Is there a risk of a slicer being over-fitted to a reuse scenario?

Pipeline dependency. Each module within a slicer pipeline affects the result of subsequent modules, creating a performance compounded effect. Hence, determining how components affect a slicers overall performance is challenging.

Any evaluation methodology adopted should therefore aim to take into account each of these challenges in addition to selecting adequate metrics which are comparable across slicer implementations and slice consumers.

3.2 Evaluation Methodology

Following the identification of major challenges involved with slicer evaluations, this section focuses in defining the major components which any slicer evaluation should consider. Slicers CASs can differ widely, however common evaluation concerns do exist. Hence we define: (i) the input of a slicer as a set of open corpus resources available over the web; and (ii) the output as a set of slices produced to fit various slice consumers preferences and content requirements. How these slices are presented and what task is being performed are outside the scope of a slicers evaluation.

Three major concerns are addressed across this evaluation methodology: (i) how does this slicer implementation affect the reuse of native content (original content) in comparison to other implementations?; (ii) what is a slicers performance with respect to its CAS?; and (iii) how does each slicer component perform individually and affect the results obtained in both (i) and (ii)?

From these 3 concerns can be derived 5 measurements categories:

1) Slice Reusability measures the overall performance of slices produced. The reusability of slices can be measured from three separate perspectives, namely: i) Quality; ii) Autonomy; and iii) Interoperability. The quality of a slice refers to properties such as the quality of annotations, the validity of the format, relevance to requests, including topic and style (reading difficulty, language etc.). The autonomy of a slice refers to the degree of success in decontextualising the original content fragments from their original documents. In other words, are parts of the original document needed in order to make sense of the content presented by a slice? Finally, the interoperability measurement refers to the ability for a slice to be consumed by diverse slice consumers with different content requirements (such as format).

2) Content Branching measures the extent to which a slicer can produce a diverse range of slices from a fixed collection of source content. It differs from interoperability measures whereby content branching measures content nature while slice interoperability measures slices form. Content branching is crucial as it measures the extent to which a slicer can re-purpose content.

3) Slicer Component Performance aims to measure the performance of each component within the slicer pipeline with respect to the output of the previous component. Standard metrics specific for each phase are used. Within the context of structural analysis for example, Adjusted Random Indexes can be used to measure fragmentation precision.

4) Component Impact upon Reusability consists in measuring the sensitivity of slice reusability measurements for varying performance levels of each pipeline component.

5) System Efficiency. Large scale slice production requires a variety of data processing and meta-data generation techniques. As slicers aim at reducing content production cost, processing cost and responsiveness must also be measured.

4 Slicer Experiment Outline and Metrics

This section proposes an implementation of the set of measurements specified above, using standard evaluation techniques, through three experiments which fulfil these evaluation criteria. The utility centric [13] experiments proposed in this section aim to combine user-based [3] and empirical [10] evaluations in order to provide a set of qualitative and quantitative measurements to support the analysis of a slicer implementation.

4.1 Experiment 1 - Assumption Validation

As using an automated content production system to support AHS is relatively novel, it is still unclear whether such a service is valid for all content consuming applications. Hence, this initial experiment aims to validate the assumption that automated content production in a selected environment can achieve relatively good results in comparison to their manually produced equivalent. Since manually created content by definition goes through a human curation process, our hypothesis is that the utility perceived by users of slice consumers will decrease when using automatically produced slices in comparison to their manual equivalent. However, sliced content doesn't necessarily need to perform as well as its manual equivalent in order to validate its use; as long as the utility experienced by the user is still sufficient enough to perform a particular task correctly. Validating this assumption proves a slicer could provide large scale content production at minimal cost. Content quality and autonomy are measured as part of this experiment. As illustrated by Figure 2, the experiment consists of a blind test whereby each user is asked to perform a same task twice with the two different sets of content.

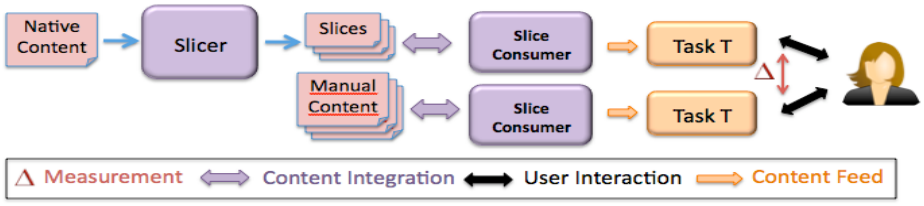

Figure 2 Assumption Validation Experiment

User effectiveness and efficiency can be measured quantitatively using domain-dependent metrics (time to perform a task, number of correct answers etc.) as estimations of content quality. Content autonomy can be measured by providing users with links displaying the fragments within their original context and recording the number of times users clicked upon this link in order to perform the selected task.

Furthermore, user satisfaction can be measured by enabling users to report inadequate slices manually or by using qualitative System Usability Scores (SUS) [3].

4.2 Experiment 2 – Slicer Efficacy Analysis

Once the assumption described above is validated, the efficacy of such a slicer with respect to its CAS along with its components must be tested. This experiment aims to measure each component's performance and its impact upon the reusability of the slices produced. This leads to heuristics illustrating the expected effect of each component upon the general slicers performance.

Since each component within the slicer pipeline is dependent upon the output of previous components, the approach proposed (illustrated in Figure 3) consists of successively switching/replacing individual components and measuring both their performance and the resulting impact upon the general reusability of slices produced. This approach isolates the component being tested from other confounding factors and enables empirical evaluations to be performed. Any resulting overall utility change on the slice consumer side can only be due to this switch within the pipeline.

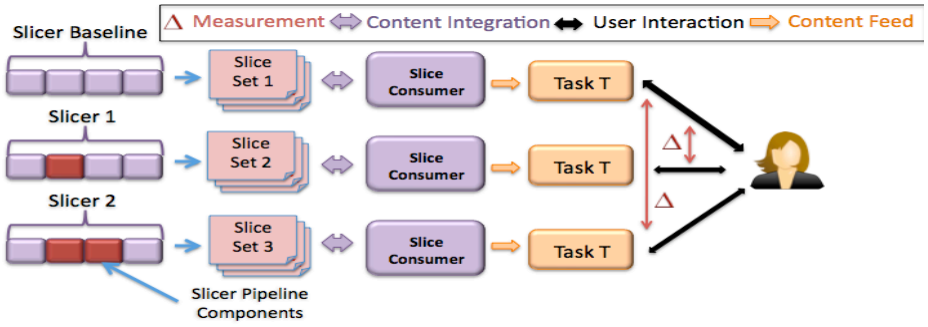

Figure 3 Slicer Efficacy Analysis Experiment

The component's performance can be measured using standard quantitative metrics used for each particular component while the impact upon reusability is measured by computing the difference in utility measured with experiment 1. Such an experiment can be repeated for each pipeline component combination desired and compared relatively to results obtained within experiment 1 as a baseline reference point.

4.3 Experiment 3 – Content Branching & Interoperability

The previous experiments aim to measure the efficacy of a slicer in producing content with minimal utility loss. However, these experiments are constrained within a specific usage context. Hence they will be specific to slice consumers used for the experiments, thus limiting an overall appreciation of a slicer's performance. An evaluation performed across various usage contexts is therefore necessary.

The aim of this experiment is to provide a usage context-free (or minimised) analysis of the slicer through an estimation of its content branching ability.

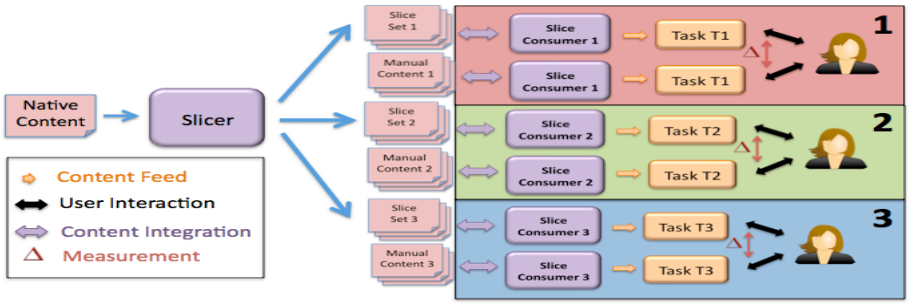

Figure 4 Content Branching Experiment

This slicer attribute represents the most difficult measurement to perform, nevertheless, Figure 4 describes how such an experiment could be performed. It is in

nature very similar to experiment 1 repeated across multiple slice consumers. A set of distinct users are divided into multiple groups each corresponding to a specific usage context (eLearning, tourist information etc.). Each usage context has a corresponding slice consumer serving a specific task. It is still unclear, how many slice consumers represent a statistically significant number.

A predefined set of content, with attributes (topic, language, writing style etc.) relevant across contexts is selected prior to the experiment. Slices produced by the slicer are then compared to manually generated content specific to each usage context. The difference in utility for each pair of content is measured as previously, and the average result across slice consumers is computed. The differences between slices produced from a common document (format/annotation differences etc.) should also be computed using quantitative metrics. High content branching measurements would be proportional to low average utility differences, high slice differences and high numbers of experimental contexts considered.

5 Summary

This paper presented an overview of the main challenges and concerns involved when evaluating a slicing system. The evaluation methodology presented in this paper proposed to assess/compare the ease of reuse of content produced, CAS performance and the impact of individual components upon the overall reuse and content branching performance of a slicer, irrespective of implementation used or content reuse context. Three utility-driven experiments using standard evaluation techniques and combining qualitative user-centric and quantitative metrics to support this methodology were also proposed.

An initial implementation of slicing technology has been developed as part of the Centre for Next Generation Localisation. Initial evaluations following the framework presented in this paper are currently underway with selected slice consumers within the fields of eLearning and customer support. These experiments will enable the framework to be validated or/and iteratively enhanced following the analysis of initial results.

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 Localisation.

References

[1] Aroyo, L., De Bra, P., Houben, G.J. “Embedding Information Retrieval in Adaptive Hypermedia: IR meets AHA!”. *In the Proceedings of the Workshop on Adaptive Hypermedia and Adaptive Web-Based Systems, at the Twelfth International World Wide Web Conference, WWW2003*, pp. 63-76, Budapest, Hungary. May 20th, 2003.

[2] Bra, P.M.E.D. Teaching Hypertext and Hypermedia through the Web. *Journal ofUniversal Computer Science, 2*, 12 (1996), 797-804.

[3] Brooke, J. "SUS: a "quick and dirty" usability scale". *Usability Evaluation in Industry*., (1996).

[4] Brusilovsky, P. and Pesin, L. Adaptive Navigation Support in Educational Hypermedia: An Evaluation of the ISIS-Tutor. *Journal of Computing and Information Technology*, (1998).

[5] Brusilovsky, P., Chavan, G., Farzan, R. “Social Adaptive Navigation Support for Open Corpus Electronic Textbooks”. *In Proceedings of 3rd International Conference on Adaptive Hypermedia & Adaptive Web-Based Systems, AH 2004*, P. DeBra, W. Nejdl (Eds.), LNCS, Vol. 3137. Berlin: Springer Verlag, pp. 24-33. 2004.

[6] Carmona, C., Bueno, D., Guzmán, E., Conejo, R. “SIGUE: Making Web Courses Adaptive”. *In Proceedings of 2nd International Conference on Adaptive Hypermedia & Adaptive Web Based Systems, AH2002*, Malaga, Spain, 29-31 May, 2002. LNCS, Vol. 2347. Berlin:Springer Verlag, pp. 376-379. 2002.

[7] Debnath, S., Mitra, P., Pal, N., and Giles, C.L. Automatic Identification of Informative Sections of Web Pages. *IEEE Transactions on Knowledge and Data Engineering 17*, 9 (2005), 1233-1246.

[8] Dieberger, A., Jose, S., and Guzdial, M. CoWeb - Experiences with Collaborative Web spaces. *In From Usenet to CoWebs: Interacting with Social Information Spaces*, (2002).

[9] Farrell, R., Liburd, S.D., and Thomas, J.C. Dynamic Assembly of Learning Objects. *Proceedings of the 13th World Wide Web Conference*, (2004), 162-169.

[10] Gena, C .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.

[11] Henze, N. and Wolfgang, N. 1999_Henze_Adaptivity in the KBS Hyperbook. *Workshop on Adaptive Systems and User Modelling on WWW*, (1999).

[12] Henze, N. and Nejdl, W. “Extendible Adaptive Hypermedia Courseware: Integrating Different Courses and Web Material”. *In the Proceedings of the International Conference on Adaptive Hypermedia & Adaptive Web-Based Systems, AH2000*, pp. 109-120, Berlin: Springer-Verlag, Trento, Italy. August 28th-30th, 2000.

[13] Herder, E. 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*, (2003), 25-30.

[14] Höök, K. Evaluating the utility and usability of an adaptive hypermedia system. *Proceedings of 1997 International Conference onIntelligent User Interfaces*, (1997), 179-186.

[15] IMS Global Learning, I.G.L. Standard for Learning Object Metadata LOM. 2006. http://www.imsglobal.org/metadata/mdv1p3/imsmd_bestv1p3.html

[16] Kohlschütter, C. and Nejdl, W. A Densitometric Approach to Web Page Segmentation. *CIKM, Proceeding of the 17th international conference on Information and knowledge management*, (2008), 1173-1182.

[17] Lawless, S., Connor, A.O., and Mulwa, C. A Proposal for the Evaluation of Adaptive Personalized Information Retrieval. *CIRSE 2010 Workshop on Contextual Information Access, Seeking and Retrieval Evaluation*, (2010), 1-4.

[18] Lawless, S., Hederman, L., and Wade, V. OCCS : Enabling the Dynamic Discovery , Harvesting and Delivery of Educational Content from Open Corpus Sources. *ICALT - IEEE International Conference on Advanced Learning Technologies*, IEEE Computer Society (2008), 676-678.

[19] Lawless, S. "Leveraging Content from Open Corpus Sources for Technology Enhanced Learning". *Ph.D. Thesis*, Submitted to the University of Dublin, Trinity College, 2009.

[20] Levacher, K., Hynes, E., Lawless, S., O’Connor, A., and Wade, V. A Framework for Content Preparation to Support Open-Corpus Adaptive Hypermedia. *Proceedings of 20th ACM Conference on Hypertext and Hypermedia*, (2009).

[21] Pennacchiotti, M. and Pantel, P. Entity extraction via ensemble semantics. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing Volume 1 - EMNLP ’09*, August (2009), 238

[22] Wade, C. and Allan, J. *Passage Retrieval and Evaluation*. University of Massachusetts, Amherst, 2005