A Framework to Provide Customized Reuse of Open Corpus Content for Adaptive Systems

Mostafa Bayomi
CNGL Centre for Global Intelligent Content, Knowledge and Data Engineering Group
School of Computer Science and Statistics, Trinity College Dublin, Ireland
bayomim@scss.tcd.ie

ABSTRACT
One of the main services that Adaptive Systems offer to their users is the provision of content that is tailored to individual user's needs. Some Adaptive Systems use a closed corpus content that has been prepared for them *a priori*, hence, they accept only a narrow field of content. Furthermore, the content is tightly coupled with other parts of the system, which also hinders its re-usability. To address these limitations, recent systems started to make use of open Web content to provide a wider variety of content. Previous approaches have attempted to harness the information available on the web by providing adaptive systems with customizable information objects. Since adaptive systems are evolving towards the Semantic Web and the use of ontologies, existing systems are limited by their ability to service these documents solely through keyword-based queries. In this research we propose a novel framework that extends existing content provision system, Slicepedia. Our framework uses the conceptual representation of content to segment it in a semantic manner. The framework removes unnecessary content from web pages, such as navigation bars, and then semantically reveals the structural representation of text to build a tree-like hierarchy. This tree can be traversed to obtain different levels of content granularity that facilitate content discoverability and adaptivity.

Categories and Subject Descriptors
H3.3 [**Information Search and Retrieval**]: Information Filtering; Retrieval Models; Selection Process;
H.5.4 [**Hypertext/Hypermedia**]: Architectures; User Issues;

Keywords
Open Corpus Content; Semantic Web; Content Semantic Slicing;

1. INTRODUCTION AND MOTIVATION
The amount of content on the World Wide Web is continuously growing. Several research fields have emerged that particularly focus on the challenges associated with this growing body of global content. These challenges include: how to identify, handle and retrieve content from different sources; how to search for information in multiple languages; and how to deliver this content in a personalized form that is most suitable for the user.

Various systems [6,9,16] have tried to address the challenge of producing adaptive compositions from open information sources in order to deliver content in a form that is most suitable to an individual user. Adaptive systems focus on providing such compositions based on a variety of user dimensions, such as user interests, prior knowledge, preferences or context. At the heart of these systems is the adaptive engine which deals with multiple loosely coupled models that are integrated as desired. Content, however, is still very tightly coupled to these engines and as a result strongly impedes the re-usability of this content.

Consider an E-Learning portal as an adaptive system that provides users with learning materials about specific subject. The portal has a user model that is responsible for personalizing content according to different dimensions, such as the user's interests, prior knowledge, preferences or preferred style of content (concise or detailed content). Since the dimensions are not the same for all users, the portal should: i) have various content resources, ii) be able to provide different levels of content granularity, iii) provide content at low production costs iv) and provide content that is amenable to be reused.

Early proposed adaptive systems have relied on closed document corpus with content specifically authored for their usage [7], hence they accept only a narrow field of content. As a result, Open corpus content is increasingly seen as providing a solution to these issues [3]. However, most systems incorporating open corpus content have mainly focused on linking such content with the internal content as a path for more content exploration [9]. Moreover, adaptive systems must strict to specific content structures to be able to make use of it which limits the amenability of content reuse.

The one-size-fits-all nature of Web content calls for automatic approaches that can tailor the content in a way that facilitates its reuse – whether in part or in full. This tailoring must be performed based upon various aspects, such as: granularity, content format and associated metadata [11]. Various systems have been proposed to address the challenge of producing adaptive content from open corpus content. These systems focused on separating the content from other models in the adaptive systems (such as domain and user models). Slicepedia [11], for example, was introduced as a service to process open corpus resources and extract content for reuse by right-fitting it to specific content requirements of individual adaptive system.

Since adaptive systems are moving towards the Semantic Web and the use of ontologies, Slicepedia and other systems are limited by the ability to service these documents solely through keyword-based queries. This means they only provide limited capabilities to capture the conceptualizations associated with adaptive system needs and content.

HT '15, September 1–4, 2015, Guzelyurt, Northern Cyprus.
 ISBN 978-1-4503-3395-5/15/09$15.00
DOI: http://dx.doi.org/10.1145/2700171.2804450

Hence, there is a need for a service that can provide adaptive content based on the conceptual needs of an adaptive system.

The research question of this study is: *To what extent can content be automatically adapted to intelligently meet the requirements of individual applications?*

Section 2 highlights the state of the art approaches and the key challenges. Section 3 presents the research proposal and the research objectives that are derived from the main research question. Section 4 presents the ongoing research work. Finally, section 5 gives the plan ahead of the research.

2. KEY CHALLENGES & RELATED WORK

As the vast quantity and diversity of content on the Web continues to grow, various systems [6,9,16] have tried to address the challenge of producing adaptive compositions from open information sources to deliver information in a form that is most suitable for an individual user. Adaptive systems attempt to perform such compositions using different methods: *manual* [9], *community-based* [2], *automatic linkage* [16] or *IR approaches* [14]. In manual methods, documents are manually incorporated within the adaptive system, however, these methods require a significant amount of time and effort due to the difficulty in identifying adequate content. Automatic linkage approaches attempt to improve this situation by providing guidance with respect to the relevant content that is available. However, these methods only provide an indication of the relative closeness of content and do not provide the details needed to support user guidance. The community-based approaches also tried to overcome the burden proposed by the manual approaches by analyzing the quantity of users stepping between various resources in order to derive this information. IR approaches provide a pluggable search service for the adaptive systems to support open corpus content identification and incorporation. The OCCS system [10] for example, uses focused crawling techniques to harvest large amounts of web resources and identify those most relevant to specific contexts of use, based on arbitrarily pre-selected topic boundaries. Various schemas such as LOM (Learning Object Metadata in e-learning) were developed to provide usage-agnostic solutions, however they require a lot of development effort thus prohibiting scalability [8].

All these approaches are limited in that they use harvested resources in their native form (web pages) as one-size-fits-all documents. As pointed out by Lawless [10], "*there is an inverse relationship between the potential reusability of content and its granularity*", i.e. the more granularity the content is, the more amenable it is to be reused.

As a result, various approaches have been proposed to utilize open corpus content to convert the wealth of information available on the web into customizable information objects. Slicepedia [11], for example, was introduced as a service to process open corpus resources and extract content for reuse by right-fitting it to the specific content requirements of individual adaptive system. Slicepedia converts original resources into information objects called *slices*. The concept of a *slice* is an abstract notion representing a stand-alone piece of information, originally part of an existing document, extracted and segmented to fulfil a specific information request. After slicing content, *slices* are stored in a repository to be provided and right-fitted to specific adaptive system requirements.

As adaptive systems are moving towards the Semantic Web and the use of ontologies [15], Slicepedia and other systems [13] are considered limited in that they are keyword-based content providing services, which means that they provide limited capabilities to capture the conceptualizations associated with adaptive system needs and content.

Hence, there is a need for a service that can provide adaptive content based on the conceptual needs of an adaptive system. By conceptual representation of content we mean the semantic metadata that describes it. Leveraging such semantic metadata is very important, but content published on the Web generally lacks the presence of such information [1].

Recent research looked at leveraging Semantic Web technologies in order to enrich content through annotation [4]. Such approaches use Named Entity Recognition to extract concepts (mostly people, locations and companies) from text. As a result, in our approach, instead of relying on the available semantic metadata for the open corpus content, we semantically annotate content to extract its conceptual structure and segment (slice) it based on the ontological relation between its constituents.

3. RESEARCH PROPOSAL

The research question posed in this study can be broken into three main objectives:

1- To investigate how existing content-provider systems, which are based on traditional keyword approaches, can be enhanced using semantic techniques.

2- To explore the effectiveness of semantic-based approaches in discovering and delivering content that best matches specific application requirements.

3- To explore the impact of the proposed approach in an industry case study, and measure the extent at which cost, time, and effort are saved when producing tailored content.

To achieve these objectives, we propose a service that: i) works as a pluggable content providing service; ii) separates the content model from other adaptive system models (such as domain and user models); iii) slices content based on its semantic (ontological) meaning to suit ontology based adaptive systems; iv) organizes content into tree-like structure. The benefits of the proposed service are a) improving content discoverability, b) facilitating its personalization, and c) maximizing its reusability.

4. Ongoing Work:

In order to investigate how existing systems can be enhanced using semantic techniques (objective 1) and to explore the effectiveness of the semantic approaches in discovering and delivering content (objective 2), we built a service that extends Slicepedia. The service is provided by an intelligent content provider framework which consists of the following modules (see Figure 1):

1) Harvester: Acquires open corpus resources, from the web, in their native form (HTML pages). Standard IR systems or focused crawling techniques [10] are used to gather relevant documents, which are then cached locally for further analysis. We will use the same harvester used by Slicepedia, the 80Legs[1] web crawler.

[1] http://www.80legs.com/

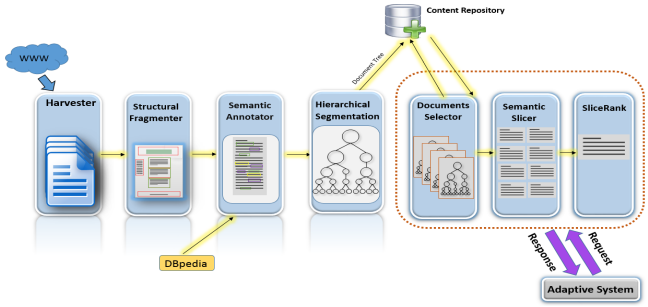

Figure 1. The Framework Structure

2) Structural Fragmenter: Once resources have been identified, the structural fragmenter starts to fragment the content into atomic pieces and the unnecessary fragments are removed. Plain fusion Densitometric Content Fragmentation (DCF) [9] is used by Slicepedia and also is selected as the structural fragmentation approach in our framework. The output of this module is a document that consists of plain text only.

3) Semantic Annotator: After removing the unneeded content from the web page, the remaining content is treated as one text document without barriers. The text is then semantically annotated using a named entity recognition algorithm, and text entities (names, places, organizations, etc.) are extracted. Each entity is then mapped to its class or classes in an ontology. For example, *Barack Obama*, as an entity, is mapped to DBpedia ontology classes: *["Person", "Agent", "Officeholder"]*. The text is then represented as a sentence-based vector-space where each sentence in the text is represented as a vector of entities, and each entity is represented by a set of classes that match the entity from the ontology. This vector space is then used as an input to the following phase. As our framework can work across different content domains, in this research we use *DBpedia*[2] ontology as the underlying knowledge- base as it is considered a cross-domain ontology. *DBpedia Spotlight*[3] is used as the named entity recognition system to extract entities from text.

4) Hierarchical Segmentation: After semantically annotate text and building sentence-based vector-space, text is segmented into hierarchically semantically coherent segments. Unlike traditional text segmentation approaches [5], our approach to text segmentation measures the similarity between text blocks based on the ontological similarity between them. A text block is considered the elementary unit of the segmentation algorithm, which could be one sentence or multiple sentences (paragraphs). To measure the ontological similarity between two text blocks, we measure the similarity between the classes of their entities using the **is-a** relation. In ontology structure, the **is-a** relations group the classes according to how they are conceptually related to each other. After measuring the similarity between text blocks, a Hierarchical Agglomerative Clustering (HAC) algorithm is applied. Conceptually, the process of agglomerating blocks into successively higher levels of clusters creates a cluster hierarchy (or dendrogram) for which the leaf nodes correspond to individual blocks, and the internal nodes correspond to the merged groups of clusters. When two groups are merged, a new node is created in this tree corresponding to this larger merged group. The output is a tree-like hierarchy of the text. This tree is then stored in a repository with its structure and with the entities that were extracted from the text. Figure 2 depicts a tree representation of a sample text of 10 sentences.

The benefit of this tree is that it represents different levels of granularity of the document, which means that the document can be sliced semantically at different levels of granularity. This is a powerful criterion in the hierarchical representation of text. In contrast to linear representation, in each level of the structure (tree), slicing with different levels of details could be obtained and can be usefully applied to different adaptive systems' needs.

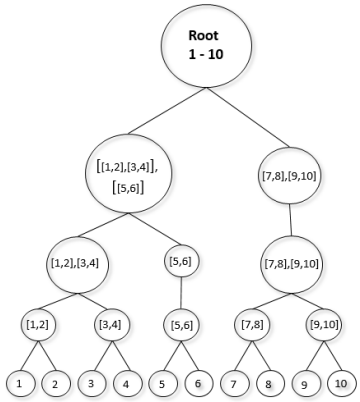

Figure 2. A tree representation for a text from 10 sentences

In the current state of the framework, we have built the text segmentation module that builds a structure of text based on the ontological similarity between text blocks. We conducted several experiments to evaluate this module on a well-known dataset [5] in the text segmentation field. The results show that text segmentation based on the ontological similarity is feasible with a low error rate. Table 1 reports the results of the best run of the experiments.

Table 1. Error rates for different subsets

Range of *n*	**3-11**	**3-5**	**6-8**	**9-11**
Error rate	**0.15**	**0.19**	**0.15**	**0.11**

5) Documents Selector: It is an ontology-based IR system that ontologically matches an adaptive system query with the documents in the repository. As the documents are hierarchically stored (as trees), this structural representation enhances the accuracy of the retrieval process. For example, suppose we have a query with two entities *E1* and *E2* and we have two documents *D1* and *D2*. *E1* and *E2* are found in the same branch of *D1*'s tree (i.e. they are near to each other in the hierarchy) while in *D2*'s tree the two entities are in different branches (i.e. they are far from each other in the hierarchy). Intuitively, *D1* is considered to be more relevant to the query as it contains the two entities in the same region of the document. The output is a set of documents that are ranked based on the relevancy between the entities in the query and the entities existence and location within the document tree structure.

6) Semantic Slicing: After retrieving and ranking the relevant documents based on how they match the request query, the slicing process starts. For each retrieved document, the semantic slicer starts to traverse its tree and based on the request, the slicer starts

[2] http://dbpedia.org/

[3] https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki

to slice the tree. For example, if the adaptive system requests a slice of six sentences that contains a specific entity in the request query. In a document tree like the one depicted in figure 2, suppose the entity is found in sentence number 4. In this circumstance, the algorithm needs to select five additional sentences that are related to sentence 4 in order to match the adaptive system request. From the document tree (Figure 2), it is clear that the most related sentences to sentence 4 are sentences from 1 to 6 as they are all clustered in one node based on the ontological relation between each other.

7) SliceRank: Bringing order into slices

The output from the previous step (the semantic slicer) is a set of slices that are extracted from different documents and match with the adaptive system request. At this point a question naturally arises. "Among slides created which one is the most appropriate slice to be retrieved to the adaptive system?" In a similar way in which a ranking mechanism can be used to rank documents, we apply what we call *SliceRank* to rank slices. Ranking slices is different than ranking documents in the sense that documents are ranked based on their relevance to the query, while slices are ranked based on the richness of information contained in them.

SliceRank is an unsupervised graph-based ranking model derived from TextRank [12] for text summarization. *SliceRank* ranks slices according to the ontological relation between their entities. Slices are represented as vertices in a graph; edges are created based on the ontological relation between entities within two slices. Edges are typically undirected and are weighted to reflect a degree of similarity between two slices. The idea behind *SliceRank* is that slices recommend each other based on the ontological similarity between their entities. A slice gets a high rank because it has entities that are ontologically similar to many entities found in the other slices. Which means that this slice has more information more than other slices even if it is extracted from a document that is not highly related with the request query.

In summary, the completed work so far involved identifying conceptually similar content fragments and structuring them into tree-like hierarchy which is the first step towards enhancing content discoverability and re-usability. In this step, the *Semantic annotator* and the *Hierarchical Segmenter* modules have been developed and evaluated. As for the ongoing work, it involves implementing the remaining modules of the framework.

5. Plan for Next Phase

Moving forward, future work will involve exploring the impact of the proposed approaches on industry and measuring the extent at which cost, time and effort are saved when employing our proposed service in producing tailored content (objective 3). The Documents Selector, Semantic Slicer, and SliceRank modules are yet to be evaluated. This will involve user studies and state of the art standard evaluation methodologies to measure the effectiveness of each module independently. Furthermore, a user study will also be carried out to evaluate the overall efficacy of the framework based on user evaluations within the context of adaptive systems and industry use cases.

Acknowledgements. This work is supported by Science Foundation Ireland (Grant 12/CE/I2267) as part of CNGL Centre for Global Intelligent Content (www.cngl.ie) at Trinity College Dublin. Special thanks to Prof. Séamus Lawless, Dr. Killian Levacher, and Dr. M. Rami Ghorab for their guidance, support, and fruitful advice.

6. REFERENCES

[1] Bizer, C., Eckert, K., Meusel, R., Mühleisen, H., Schuhmacher, M., and Völker, J.Deployment of RDFa, Microdata, and Microformats on the Web – A Quantitative Analysis. *The Semantic Web – ISWC, Springer Berlin Heidelberg*, 2013, pp. 17–32.

[2] Brusilovsky, P., Chavan, G., and Farzan, R.Social Adaptive Navigation Support for Open Corpus Electronic Textbooks. *Adaptive Hypermedia and Adaptive Web-Based Systems*, 2004, pp. 24–33.

[3] Brusilovsky, P. and Henze, N.Open Corpus Adaptive Educational Hypermedia. *The Adaptive Web SE - 22*. Springer Berlin Heidelberg, 2007, 671–696.

[4] Butuc, M.-G.Semantically enriching content using opencalais. *EDITIA 9*, 2009, 77–88.

[5] Choi, F.Y.Y.Advances in Domain Independent Linear Text Segmentation. *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference,* (2000), 26–33.

[6] Conlan, O., Staikopoulos, A., Hampson, C., Lawless, S. & O'Keeffe, I. *The Narrative Approach to Personalisation.* New Review of Hypermedia and Multimedia*,* 2013,132 - 157

[7] Dieberger, A. and Guzdial, M.CoWeb — Experiences with Collaborative Web Spaces. *From Usenet to CoWebs SE - 8*. Springer London, 2003, 155–166.

[8] Farrell, R.G., Liburd, S.D., and Thomas, J.C.Dynamic Assembly of Learning Objects. *Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers &Amp; Posters, ACM* (2004), 162–169.

[9] Henze, N. and Nejdl, W.Adaptation in open corpus hypermedia. *International Journal of Artificial Intelligence in Education* 12, 4 (2001), 325–350.

[10] Lawless, S.Leveraging Content from Open Corpus Sources for Technology Enhanced Learning. *Doctoral Thesis*, 2009.

[11] Levacher, K., Lawless, S., and Wade, V.Slicepedia: Providing Customized Reuse of Open-web Resources for Adaptive Hypermedia. *Proceedings of the 23rd ACM Conference on Hypertext and Social Media, ACM* (2012).

[12] Mihalcea, R. and Tarau, P.TextRank: Bringing Order into Texts. *Proceedings of EMNLP 2004, Association for Computational Linguistics* (2004), 404–411.

[13] Savić, G., Segedinac, M., and Konjović, Z.Automatic generation of E-courses based on explicit representation of instructional design. *Computer Science and Information Systems* 9, 2 (2012), 839–869.

[14] Speretta, M. and Gauch, S.Personalized search based on user search histories. *Web Intelligence* (2005), 622–628.

[15] Sudhana, K.M., Raj, V.C., and Suresh, R.M.An ontology-based framework for context-aware adaptive e-learning system. *Computer Communication and Informatics*, (2013).

[16] Zhou, D., Goulding, J., Truran, M., and Brailsford, T.LLAMA: Automatic Hypertext Generation Utilizing Language Models. *Proceedings of the Eighteenth Conference on Hypertext and Hypermedia, ACM* (2007).