

Dynamic Hypertext Generation for Reusing Open Corpus Content

Ben Steichen, Séamus Lawless, Alexander O'Connor, Vincent Wade

Centre for Next Generation Localisation

Knowledge and Data Engineering Group

School of Computer Science and Statistics

Trinity College, Dublin, Ireland

{Ben.Steichen, Seamus.Lawless, Alex.OConnor, Vincent.Wade}@cs.tcd.ie

ABSTRACT

Adaptive hypermedia systems traditionally focus on providing personalised learning services for formal or informal learners. The learning material is typically sourced from a proprietary set of closed corpus content. A fundamental problem with this type of architecture is the need for handcrafted learning objects, enriched with considerable amounts of metadata. The challenge of generating adaptive and personalised hypertext presentations from open source content promises a dramatic improvement of the choice of information shown to the learner. This paper proposes an architecture of such a dynamic hypertext generation system and its use in an authentic learning environment. The system is evaluated in terms of educational benefit, as well as the satisfaction of the users testing the system. Concluding from this evaluation, the paper will explore the future work necessary to further enhance the system performance and learning experience.

Categories and Subject Descriptors

H.5.4 [Information Interfaces and Presentation]: Hypertext/Hypermedia - *architectures, navigation, user issues*

General Terms

Algorithms, Measurement, Design, Experimentation

Keywords

Hypertext Generation, Adaptation, Personalisation, Open Corpus Content, Metadata Generation

1. INTRODUCTION

The challenge of providing appropriate learning content to formal or informal learners typically depends on the quality of content available and the accuracy of the content retrieval system.

It can be argued that a key problem is not necessarily the lack of publicly available quality content, as there are vast amounts of

content distributed freely on the web (as long as the topic of interest is not especially obscure¹). The key problems are that such open content is (a) probably not written for the specific purpose (learning need) of the individual learner, (b) is not sequenced in a way which makes the learning more directed to the learners need, (c) is not sequenced/linked in an appropriate manner to empower the learning process for that learner, (d) is difficult to identify/classify against the learner's need, (e) is rendered in different styles with a different "look and feel" [1][2].

In this paper we argue that the repurposing of open corpus content presents the real opportunity for Adaptive Hypermedia systems. In particular, AH can generate hypertext across dynamically identified open corpus content and generate learning opportunities which are customised to the learner needs and preferences. This can be achieved by the AH system generating hyperlinks and hypertext navigation either embedded in the content, or as indexes across open corpus content. However, in order to achieve successful adaptive open corpus systems they must provide (i) accurate open corpus harvesting/identification systems and retrieval approaches, (ii) generate dynamic personalised hyperlinks based on appropriate learning strategies and (iii) ensure a level of uniformity for presenting heterogeneous content.

The challenge of accurate open corpus content identification and selection (retrieval) has seen the emergence of many open source web-based information retrieval systems e.g. Lucene [3], Nutch [4] etc. However, a significant problem with just using such IR systems is that they cannot provide a detailed description of each page of content retrieved. They typically only identify the relative relevancy of retrieved web pages for the users query. In this paper we argue that a combination of customised web harvesting/retrieval system, automated indexing system and a crowd sourcing based annotation system can provide a necessary content quality and retrieval accuracy to support dynamic hypertext generation for learning.

Moreover, in order to enhance and add value to the learning experience, we argue that a successful adaptive open corpus

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

HT'09, June 29–July 1, 2009, Torino, Italy.

Copyright 2009 ACM 978-1-60558-486-7/09/06...\$5.00.

¹ It is difficult to provide a universally acceptable definition as to what subject areas are most appropriate for open corpus based learning as the WWW is ever expanding, hence our use of the term 'obscure' to characterise subjects which are not appropriate. Later in the paper we provide more guidance as the subject areas which are most appropriate for our approach.

system needs to be able to generate learning navigations, which suit the learner's goals and learning preferences. Such personalised navigations need to be informed by adaptive strategies, informed by the learners' (user) models and informed by pedagogic strategies. These adaptive strategies should allow a user to navigate personalised links generated for adaptively selected content (specifically tailored to his/her needs and preferences). The benefit of the personalisation is that the repurposing of open corpus content can be achieved via the recomposition of existing content using dynamic sequencing, spatial layout adaptations and dynamic narratives. Finally, we propose that a minimal portal, which can frame and provide navigation around the retrieved heterogeneous open content, can provide sufficient uniformity of the content presented to lessen the impact of the heterogeneity.

This paper recognises that there are key challenges regarding IP and copyright when attempting to reuse open-corpus content, even for educational use. Initiatives such as creative commons can potentially have a very positive impact on clarifying the copyright permissions for educational use. However, for this scientific technical paper, copyright and IP issues are out of scope.

The remainder of this paper is structured as follows. It first outlines the challenges in open-corpus based adaptive hypermedia systems. It then describes the architecture of a prototype system, focusing on open corpus harvesting, annotation and adaptive navigation (and presentation). The paper outlines an evaluation experiment using real learning environments performed in a university setting. A description of the learners experience is presented based on both learning effectiveness and learner satisfaction. The initial results are shown to be positive, with the prototype proving to be a good starting point for further research. The paper also contrasts the work with recent research in adaptive open corpus content. Finally the paper concludes with a summary of the open implementation issues identified by the experimental results.

2. CHALLENGES & RELATED WORK

Multiple challenges exist for the development of an open-corpus hypertext generation system. First of all, content has to be harvested and cached for further data analysis and to be accessed by the personalisation service. Uniformity of the corpus needs to be addressed to present the user with a consistent and continuous set of data. Secondly, markup (metadata) needs to be generated in order to allow a system to apply adaptivity and personalisation. Finally, sophisticated personalisation strategies are needed to provide accurate adaptivity for a specific user's needs and preferences. These strategies not only require the content to be richly marked up, but also need to possess accurate user and domain models to reason appropriately about content selection and sequencing.

2.1 Identification of learning corpora

As educators attempt to incorporate technology enhanced learning (TEL) in formal educational offerings, the lack of appropriate and accessible digital content resources can often present difficulties [1]. Quality TEL resources can prove expensive to develop and have traditionally been restricted to use in the environment in which they were authored [5]. As a result, educators who wish to

adopt TEL must spend time authoring educational content rather than on the pedagogical aspects of TEL design [6] [7]. The accessibility, portability, repurposing and reuse of educational resources therefore provide major challenges [8] [9] [10].

Digital content creation was traditionally a linear process, from authoring to publication. However, the aggregation and reuse of existing content from various disparate sources is becoming more common. This trend can be observed in the development of digital content repositories, learning objects, mash-ups and content aggregators.

National digital content repositories have been developed in an attempt to encourage cross institution sharing of learning resources. Merlot in the USA [11], Jorum in the UK [12] and the NDLR in Ireland [13] are just three examples of such repositories. Educational institutions have also used repositories to provide access to their learning resources. Examples of such repositories are OpenLearn [14] and OpenCourseWare [15]. However, these initiatives have encountered problems with user engagement, as a result of difficulties with content repurposing and reuse, digital rights management (DRM) and institutional culture [16] [17].

Learning objects (LO) are digital, self-contained 'chunks' of learning content [11]. LOs aim to enable educational content reuse outside the context in which it was authored. This provides the potential for dynamic, 'on the fly' sequencing of resources [19]. However, this potential has yet to be fully realised [20]. Specifications and standards which target content reuse have also been developed, such as the Shareable Content Object Reference Model (SCORM) and Learning Object Metadata (LOM). Authoring LOs has proven very time-consuming. More specifically LOs tend to be developed by professional companies for their own products and not shared openly. The amount of open corpus learning material available on the web is orders of magnitude greater and more accessible than the tiny percentage of LOs available on the web.

2.2 Open Corpus Harvesting

Open corpus educational content, which is varied in structure, language, presentation style etc., can be sourced via the WWW. However, such content has yet to be comprehensively exploited in the field of TEL [1]. In order to facilitate the utilisation of open corpus content in TEL, methods of surmounting the heterogeneity of web-based content must be developed. This includes integrated means of content discovery, classification, harvesting, indexing and delivery.

Web crawling techniques are the most efficient method of large-scale content discovery on the WWW. Focused web crawling extends these techniques by enabling the discovery of content, which meets pre-determined classifications. Focused crawling systems, such as Nalanda [21] and Combine [22], typically employ complex methods of crawl scope definition, which require a high level of technical knowledge to manipulate and configure. This reduces the accessibility of such tools to the non-technical user. This is a significant problem if such tools are to be used in mainstream education (by teachers, academics and tutors), as non-technical educators must be able to define the scope of each crawl.

Once content is identified and harvested it must also be indexed to make it more readily discoverable. There are numerous open-

source indexing solutions, such as Lemur [23] and Lucene [3]. Some indexing tools have been integrated with a web crawler to form an information retrieval tool chain, such as Nutch [4] and Swish-e [24]. However, these tool chains typically utilise general purpose, rather than focused, crawling techniques and can be limited by the indexing methods employed. Swish-e also functions most efficiently on small to medium-sized content collections of less than one million objects, which is not compatible with web-scale content discovery.

It can be argued that the process of content harvesting is infeasible at runtime due to several reasons. The process requires considerable amounts of resources in order to discover and index large volumes of data. A persistent document cache needs to be created in order to ensure reliable content delivery during the hypertext creation phase. If the system only stores pointers to the live source documents it would not be able to guarantee that (i) the source document is still available and (ii) the content still reflects the same concepts as the initially discovered source document.

2.3 Metadata Generation

There have been numerous approaches to the generation of metadata, all involving varying levels of manual effort. Some content authoring tools such as Microsoft Word and Macromedia Dreamweaver automatically generate metadata during the authoring process. Metadata harvesting approaches have been developed to extract this metadata for use in content discovery. The Open Archives Initiative's Protocol for Metadata Harvesting enables the harvesting of metadata from content stored in OAI compliant repositories.

Another approach has been to analyse existing content in an attempt to infer meaning and generate metadata automatically. Sementag was developed by IBM to perform automated semantic tagging of content using the TAP ontology [25]. Sementag uses a Taxonomy Based Disambiguation (TBD) algorithm to ensure the correct classification of content in its tagging. Klarity and DC.dot are further examples of text classification based metadata generation systems [26].

A third approach is to exploit the large community of new users available on the web, sometimes termed "crowd sourcing" [27]. In recent years, the active participation of web users in creating metadata/content descriptions to enhance existing content has become both popular and very successful. Examples of such community involvement have been evident in Facebook [28], Flickr [29], digg [30], etc. The key aspects to this approach are to identify (i) who will be interested to create such metadata, (ii) how to ensure the quality of such content descriptions and (iii) what type of metadata should be captured.

2.4 Adaptive Hyperlinking

As TEL applications attempt to support functionality such as personalisation, adaptivity [31] and dynamic learning object generation [32] [19], their inherent reliance upon bespoke, proprietary educational content is a considerable problem [33]. For Adaptive Hypermedia, there are many successful approaches to dynamically navigate over multimedia corpora, such as KnowledgeTree [34], AHA! [35] and APeLS [36]. However, such systems rely on closed corpus or specifically engineered content. Such handcrafted closed-corpus corpora contrast greatly with the

heterogeneity, mixed granularity and non uniformity encountered in attempts to reuse open corpus corpora.

Dynamic hyperlink generation between open corpus objects presents a challenge that has yet to be solved. Several types of adaptive techniques have been explored, ranging from language models [37], over manual [38] and community-based hypertext creation [39], to automatic keyword-based hyperlink creations [40].

The KBS-hyperbook [38] represents one of the first open adaptive systems, taking a manual indexing approach. Documents are indexed using an external domain model, with any new document (potentially from any data source) being added in the same manner. Although this allows the system to add new material, the KBS-hyperbook was developed with an existing handcrafted corpus, where new material was mainly supplementing this content space.

One of the most open adaptive systems is the Knowledge Sea II system [39], which uses a community-based hypertext creation. It performs the creation of a cell-based knowledge map, with similar websites being in adjacent cells. Using social-navigation, the map can be expanded without the need for document pre-processing. Although this system allows an easy integration of new resources, no educational strategy helps the user through the content in a guided manner.

With the emergence of the semantic web, new types of hypertext generation strategies make use of the rich expressiveness provided by semantic technologies [41]. Hyperlinks can be inferred from richer domain models using semantic reasoning capabilities. However, challenges still remain in integrating such technologies successfully to form complete dynamic hypertext generation systems, encompassing document harvesting, indexing, hypertext creation and result presentation. Additionally, the problem of reusing open source content is yet to be solved, as well as the problem of generalising adaptive behaviour and functionality.

3. ARCHITECTURE

This section details the design and architecture of the dynamic hypertext generation system. It is built on top of the Adaptive Personalized eLearning Service (APeLS) [36] and uses a Multi-Model approach for concept/content selection, sequencing and composition.

The hypertext generator facilitates students who wish to learn about specific concepts by allowing them to query the system with keywords selected from a predefined concept list. An intelligent response is then composed by the system to address the student's information need. The content used in generating this response is open corpus in nature and provided by the OCCS [42].

Firstly, the architecture of the OCCS is presented, followed by an approach to acquire fine-grained metadata annotations. We will then outline the different models used by the system to perform the desired adaptivity and personalisation. Finally, the overall hypertext generation process will be explained, followed by more detailed descriptions about the strategies and presentation mechanisms used.

3.1 Content Cache Generation

The OCCS enables the discovery, classification, harvesting and indexing of content from open corpus sources. It is provided as an autonomous service which can be used to generate subject specific caches of content for use in TEL. The OCCS uses a focused web crawler built upon Heritrix, the Internet Archive’s open source web crawler. Crawls are conducted which target content on the WWW by topic. This is an incremental process in which a URI is selected from those scheduled and the content at that URI is fetched and classified.

Classification involves filtering the content by language and conducting a comparison between the content and a subject domain model. The OCCS uses a language guesser called JTCL [43] and a text classifier called Rainbow. Rainbow must be trained in advance of each crawl to generate a classification model of the domain. A combination of keyword files and ODP categories are used to generate positive and negative training sets of content. The classifier uses these training sets to build a statistical model of the subject area. The OCCS then uses this model to ascertain the relevancy of crawled content to the scope of the crawl.

All of the content cached by the focused crawler is stored in ARC files. To achieve indexing, the OCCS incorporates NutchWAX [44], which is an open source indexing solution that adapts the fetcher step of Nutch to process ARC files. NutchWAX sequentially imports, parses and indexes the cached content. Upon completion an index is created for the entire collection of ARC files in the OCCS content cache. Wera is used in the OCCS to link the NutchWAX index with the content cache and allow the visualisation of the archived content.

3.2 Metadata Annotation

To enable the use of open corpus content by Adaptive Hypermedia applications, descriptive metadata needs to be authored or generated. Web-based content rarely has accurate or comprehensive associated metadata descriptions. This metadata is required to inform the AH system of the subject matter of each piece of content and some educational indicators, which subsequently influences content selection.

Our research focus is taking a hybrid approach of automatic indexing and community-based crowd sourcing. In our approach, we identified a small community working in the subject area. In order to ensure the quality, a reduced vocabulary was used for metadata choice. This had an effect of speeding up the actual time taken to annotate, as well as to reduce the complexity of the task for the individual. The metadata vocabulary (control vocabulary) and the indexing terms are aligned with a high-level ontology of the domain of the open corpus content. This allows more flexibility for both the retrieval/harvesting of open-corpus content, the annotation of the open-corpus material, and the adaptive navigation generation. As the hypertext generation system is educational in nature, it was essential to ascertain not only the concepts addressed by the content, but also the level of complexity and educational purpose of the content. This is currently beyond the capabilities of purely automated metadata generation approaches.

The size of the potential pool of metadata authors combined with the level of post-annotation auditing implemented dictates the

baseline level of expertise required among participants to achieve high-quality annotation. In extremely large communities, crowd sourcing can function without defined limits upon participant expertise. Frequency of occurrence, voting and user reputation are some of the measures that can be used to dictate version selection and filter out invalid annotations. In the case of this research, the potential pool of annotation participants was relatively small, and so a higher baseline level of knowledge of the subject domain was required.

The generation of metadata through content annotation is an iterative process. The participants were presented with a random, unannotated item from the content cache and asked to assign values for complexity and educational purpose. They were also asked to identify the concepts addressed in the content. Once completed another item was selected from the cache and presented for annotation. These metadata descriptions were then stored as XML which could be accessed by the dynamic hypertext generation system.

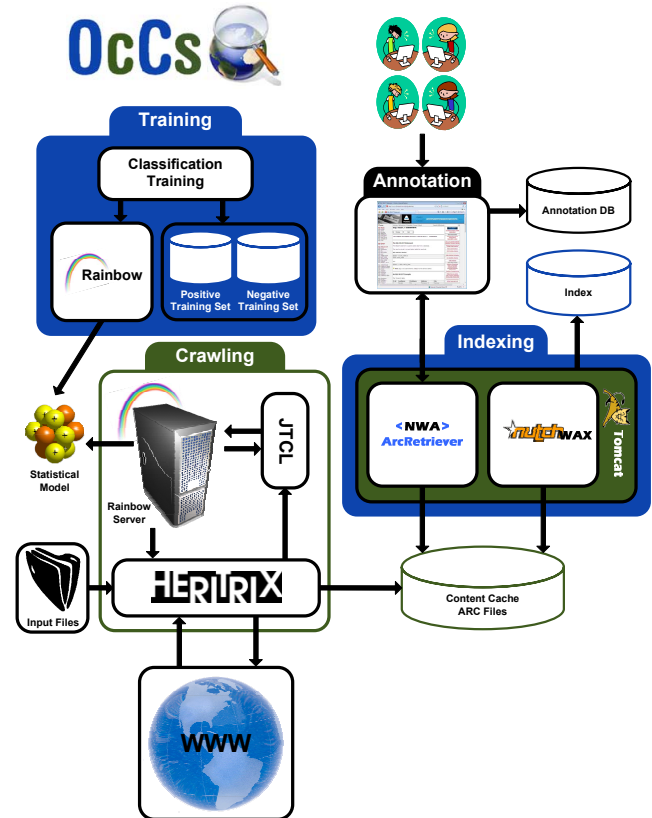


Figure 1. Content Harvesting and Metadata Generation system

3.3 Dynamic hypertext generation

At the end of the content harvesting and metadata generation process described above, several assumptions about the collected material can be made:

- A cache has been harvested on the specified domain from the open corpus web
- The cache is guaranteed to contain highly relevant content

- The cache contains reasonably accurate metadata descriptions of the content

The challenge of the dynamic hypertext generation process lies in the analysis of this data and the adaptive composition of an appropriate personalised response to a user query. The user interface needs to provide the necessary dynamically generated hyperlinks, generated across relevant content to address a user's learning needs. Moreover, a challenge lies in creating a uniform user interface, which (i) addresses user's goals and tasks and (ii) limits the presentational heterogeneity of open corpus content.

One could consider the overall architecture having 2 steps, consisting of adaptive query elicitation and adaptive response composition.

3.3.1 Query Elicitation

Firstly, the query interface allows users to compose a query from a set of domain-specific keywords (see Figure 2). Then, a domain-specific personal intent (goal) can be specified in order to provide the system with additional semantic information to compose an informed response. Additionally, a selected question type (what/how) indicates what type of response the user is hoping to receive. This helps the system adapt to the user by either choosing more of an explanation-based or a more tutorial/example-based response.

The system will generate a personalised response even if it only receives query keywords. However, including an intention and question type improves the results and presentation.

Figure 2. Query Elicitation

3.3.2 Response Composition

In order to provide an adaptive, personalised response, the dynamic hypertext generation system requires several models, which capture the actors and the domain involved.

The *user model* (UM) represents a user's preferences and prior knowledge in the subject area. In this experiment, we used a relatively simple user model, consisting of educational preferences (whether they prefer explanations, examples, etc.), media preferences (audio, text, etc.) and users' prior knowledge about domain specific concepts.

The *domain model ontology* (DM) captures the concept space, together with the relationships between the concepts. These include hierarchical relationships (e.g. the *is_a* relationship) and domain specific relationships. Additionally, concepts have

defined prerequisite concepts, i.e. a user should learn about a simple/general concept before attempting a more advanced concept. Using ontologies for the domain model allows the creation of a highly structured information space, together with the ability to reason about relationships between concepts.

Narratives encode the strategies that are used to select appropriate concepts and content for a user's query. Moreover, the encoded strategies define the appropriate sequencing and navigation for the selected content. This provides the learner with a guidance through the content, as the narrative adaptively builds a learning path according to an educational strategy. Furthermore, narratives perform personalisation using a learner's personal preferences encoded in the user model. The strategies encoded in the system are content independent, with the adaptation being performed only using concept-level information. Additionally, strategies are concept-domain independent, i.e. they are solely cognitive of abstract concept and relationship definitions. This property makes the strategies highly reusable across different concept spaces, as they are independent of the domain ontology used.

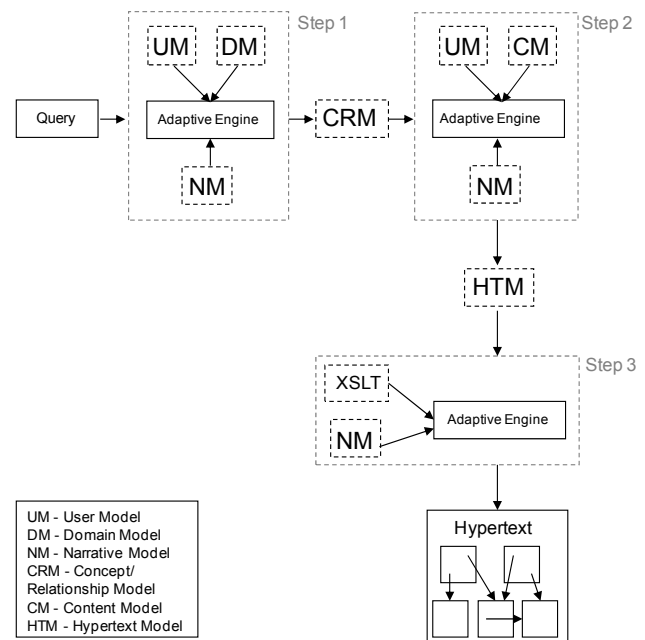


Figure 3. Hypertext Generation Process

For the purpose of explaining the adaptation system, we can consider the process consisting of 3 steps (see Figure 3).

- Concept level adaptation
- Hypertext Model Generation
- Result presentation

When a user specifies a particular keyword query, the system performs a concept level adaptation (step 1) using the domain model (DM), the user model (UM) and the narrative model (NM). The adaptive engine consolidates these models in order to produce a concept/relationship model (CRM), which represents the selection of concepts and relationships that should be shown to the user in the informed response. Step 2 uses this resulting model and performs content level adaptation, i.e. it selects and

sequences appropriate documents for the user. The resulting hypertext model (HTM) can then be used in a third step to generate a complete hypertext representing an informed response tailored to the user's input query and his/her preferences.

Concept level adaptation (Step 1)

As outlined above, the first step of the adaptation process is concerned with the selection of concepts from the domain model that match the user's request. It receives a user's input query, together with a chosen question type (what/how). The system queries the domain ontology for concepts matching these keywords, along with related concepts. The narrative then selects the appropriate concepts based on the user model. The selected concepts, together with their relationships are returned as a concept/relationship model (CRM), which represents a personalised selection of the concept space for a user's query and prior knowledge (see Figure 4).

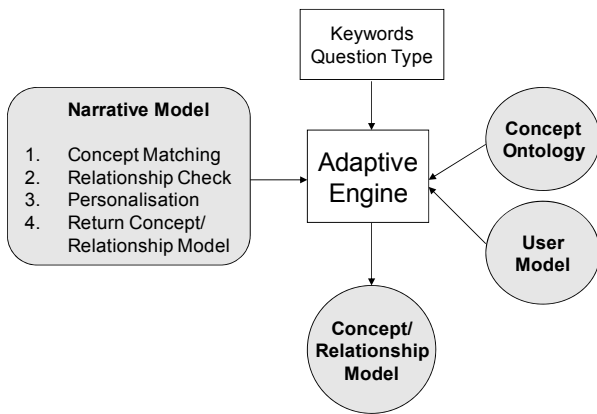


Figure 4. Step 1 creates a Concept/Relationship Model

Hypertext Model Generation (Step 2)

The second step uses the CRM returned by step 1 and sequences the different concepts according to specific strategy rules.

For the purpose of illustrating how this works, two simple strategies are outlined below:

- If a user's prior knowledge is low for a certain main concept, prerequisites are sequenced before the main concept.
- Related concepts are sequenced after the main concepts (including prerequisites if the main concept is well known).

Based on the question type given and the user's personal preferences, several educational purposes (e.g. explanation, example, etc.) are selected and sequenced for each concept.

For each concept-educational purpose combination, appropriate content is selected from the content space by matching the different preferences to the available document metadata.

After the second step, a full hypertext model is returned, containing the selected and sequenced concept space, together with the identifiers of the selected documents (see Figure 5).

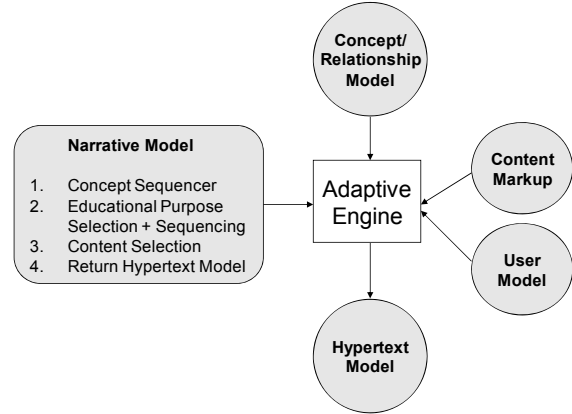


Figure 5. Step 2 creates a Hypertext Model

Result Presentation (Step 3)

The resulting hypertext model represents the different concepts, relationships and documents to be shown to the user. Using XSLT transformation techniques, the system transforms the hypertext model into a series of XHTML pages. A user is presented with a highly structured hypertext showing the different concepts to be learnt by the student, together with the associated documents providing the desired information.

The structure of the hypertext guides the student through the informed response by initially presenting a general overview of concepts to be explored. Related concepts of the main concepts are provided on hyperlinked pages. If a user wishes to explore a particular concept, there are several sequenced educational purpose options (e.g. introduction, explanation, example, etc...) available per concept. By choosing a concept and educational purpose (e.g. INTRODUCTION of PRIVILEGE), the user is presented with a choice of up to five documents providing the desired information (see Figure 6).

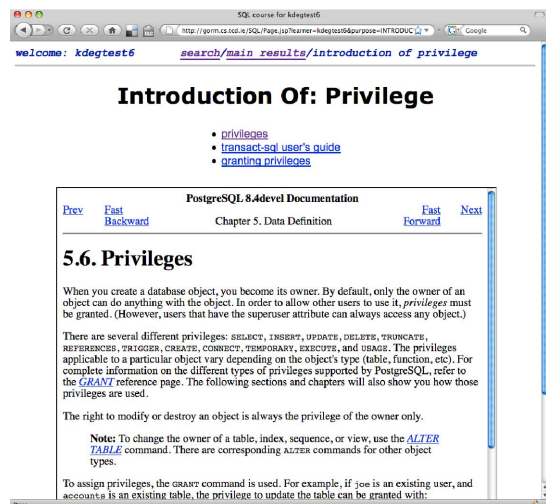


Figure 6. Content introducing the SQL Privilege concept

As the cached documents are heterogeneous in nature, an iFrame is used to separate the generated hypertext from the document text. Also, a decision was made to disable internal links in the documents. The concern here was that learners would link off

these sites and become "lost in hyperspace". Future versions will investigate the adaptive enablement of these links, based on a learner's preferences and experience.

Alternative Result Presentation

As outlined above, the proposed architecture creates a concept/relationship model (CRM) in a first step, which can then be used in a second step to create an adaptive hypertext model. However, there are other possibilities to use the CRM for alternative result presentations. For example, if the output is to be used for a standard information retrieval (IR) system, the relationships between concepts can be ignored and only the concepts in the CRM are used to expand an IR query. This second output possibility shows the versatility and reusability of the process for a system involving different content selection and presentation paradigms. The full CRM can be used by a second adaptive hypermedia step (as is the case for this system), whereas an expanded query could be used with any IR-type system. This is the subject of a separate study, which we hope to publish in the future.

4. EVALUATION

4.1 Experimental Definition & Setup

The dynamic hypertext generator was tested in an authentic learning environment in order to evaluate (i) the relevancy of the open-corpus cache, (ii) the educational benefit of the system and (iii) the usability from the students' perspective.

In order to carry out the experiment, a case study was chosen on the area of teaching SQL from an open-corpus content base. The experiment consisted of 3 stages, encompassing open-corpus cache harvesting, crowd-sourced metadata generation and finally the task-based usage of the hypertext generator by students in an authentic learning situation.

During the first stage, a focused document cache was harvested using the OCCS, yielding approximately 15000 documents in the SQL subject domain.

In a second stage, this cache was made available to researchers familiar with the domain using a web-based interface in order to view and annotate documents. Two categories of metadata were used to describe the content pages: the first category of metadata described the content of the document in terms of what were the primary and secondary concepts presented in the page, and which SQL commands these concepts described. The vocabulary used to describe/identify the concepts and commands were derived from the domain ontology, which described SQL. The second category of metadata described the document from the perspective of eLearning. This included an estimation of the prior knowledge required to understand a particular page and the type of document (e.g. tutorial, explanation, etc.).

A total of 20 annotators rated the cached documents according to these metadata categories. 9,249 pages from the cache were annotated, of which 1,525 pages were rated as valid documents to be used by the dynamic hypertext generator. The architecture of the web-based annotation tool permitted users to annotate a particular page in several sittings. The annotation tool recorded the start and end times of each annotation, but it was not possible to determine if, for example, the annotation tool was in the background while the user undertook another task. Because of this, a cut-off of 5 minutes was chosen in order to select valid

annotations for statistical analysis. This cut-off excluded less than two percent of the total annotations.

During the final stage, a class of 12 students was presented with a random selection of 3 tasks related to the subject of SQL (from a pool of 6 tasks). User tasks consisted of a question, the use of the tool by the user to learn about a specific topic, and the user answering the question. In these questions, the exact answer can only be formed by the student. The correct answer requires the student to synthesise new knowledge and interpretation by the student. An example task would be "What is a trigger? Explain how it can be used for automatically insuring integrity in a relational database. Give an example of a trigger command and explain how that example works." The process for each student consisted of four steps.

Firstly, students were presented with a questionnaire to indicate their personal preferences and their opinion of their prior knowledge in order to create user models. The second step involved a pre-test to capture the students' actual prior knowledge as opposed to their perceived competency from the questionnaire. This pre-test was specific to the particular set of 3 tasks that had been randomly assigned to each student, which would allow a reliable assessment of their knowledge gain. Thirdly, students used the dynamic hypertext generator in order to learn enough about SQL to complete their set of tasks. The search interface allowed students to indicate the type of task being performed (i.e. What/How), a query intention (e.g. Setting up a database), as well as one or more keywords to describe the question. The set of keywords was again derived from the ontology describing the SQL subject domain. The system then dynamically created a personalised hypertext according to the user query and preferences. During this phase, users' actions were tracked to identify particular trends in their search behaviour. The system collected the number of result pages viewed, as well as the time spent to complete a test question. Additionally, the query formulation was logged to identify the number of queries performed before answering a question, as well as the number of terms per query. Also, students' task answers were collected and corrected to measure the learning effectiveness of the system. Finally, following the completion of the task questions each student was presented with an evaluation questionnaire, involving a series of standard usability questions (SUS [45]), as well as free text questions to express particular likes and dislikes.

4.2 Results

An analysis of the annotation process revealed that the entire process took approximately 32 hours, with 90% of annotations being performed in less than 92 seconds (see Figure 7).

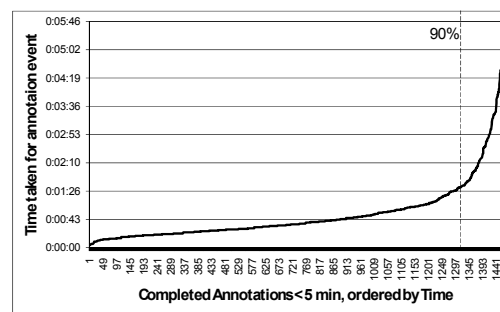


Figure 7. Duration of annotation events that produced complete descriptions in under five minutes

The analysis of the students' search behaviour revealed that the average number of queries was rather low at 2.13 queries per user per task (see Figure 8). On average, users chose 1.3 terms per query and looked at 4.3 documents presented by the hypertext generator.

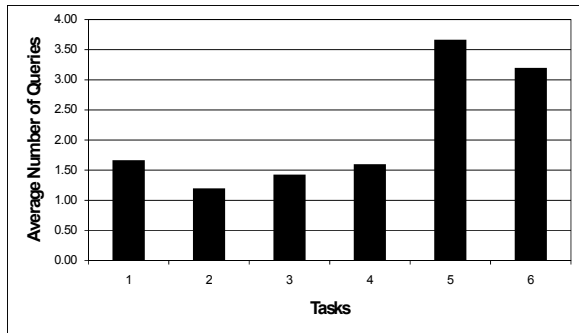


Figure 8. Average number of queries per task

From an educational perspective, the experiment has proven the dynamic hypertext generator very successful. To quantify the knowledge gain by the students using the system, we can compare the pretask scores (prior knowledge) with the task scores (post task knowledge). On a scale from 0 to 5, where 0 represents no knowledge of the task area, and 5 representing full knowledge needed to carry out the task, the average knowledge gain of the 12 students using the system was 4.25. This was calculated by scoring their pretask answers in the range 0 to 5 (0 representing complete absence of knowledge to complete the pretask and 5 representing successful completion of the pretask) and scoring the task completion itself in the range 0 to 5 (0 representing complete failure and 5 representing complete success). Although it can be argued that the range between 1 and 4 is not algorithmically calculatable (i.e. is subjective to the tutor marking the answers), a knowledge gain of 4.25 shows a significant educational impact. Figure 9 shows the total knowledge gain per student, with the different shades denoting the knowledge gain for each of the 3 tasks a student was assigned to.

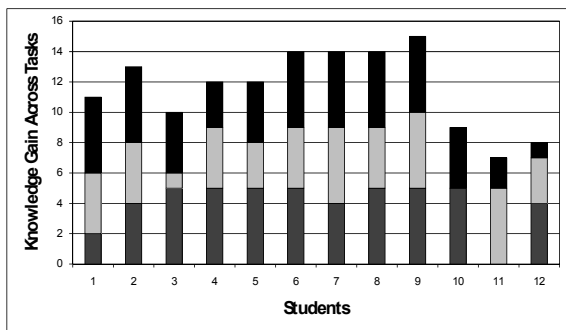


Figure 9. Knowledge gain of students

Additionally, although the knowledge gain was significant, the average number of queries was very low. These results would indicate that the quality and appropriateness of the presented documents was very accurate.

The usability evaluation questionnaire revealed that students were satisfied with the number of relevant search results returned by the system, with an average score of 3.92 (1=strongly disagree, 5 strongly agree). When asked about the system returning irrelevant

results for their query, the average score was only 2.5, suggesting that the system had a relatively low level of noise during content retrieval.

In terms of the dynamic hypertext structure, users gave a score of 3.1 when asked if the presentation of the search results was helpful. When given the opportunity to express particular likes and dislikes, 83% of the students chose to indicate the positive relevancy and presentation of the results (50% and 33% of users respectively). Major dislikes were mainly concerned with the system interface (41% of users), most notably the page design and the disabling of internal links in the presented documents.

5. CONCLUSIONS & FUTURE WORK

In this paper, we have presented an architecture for dynamic hypertext generation using open corpus content. The prototype system provides personalised responses to user queries, with the main components performing (i) harvesting of open corpus data, (ii) generation of metadata, and (iii) creation of adapted result presentations. The prototype combines recent advances in content retrieval, semantic technologies and adaptive hypermedia to provide a complete system for the desired reusability of open corpus content.

The system successfully harvested a large corpus of annotated documents using an open corpus content retrieval system and a metadata annotation tool. However, several aspects of this process were specific to the particular scope of the chosen subject domain.

First of all, for a subject area like SQL, there exists a wide range of content openly available on the World Wide Web. This is true for a great many subject areas. However, if we narrowed the subject of interest into very specific, almost niche areas, it could have an effect on the overall behaviour of the system. This needs to be further researched. Due to the fact that both domain experts (to create the domain ontology) and subject-experienced annotators were available, the area of SQL provided a good starting point to test all aspects of the system. However, changing the scope would require a generalised process for ontology creation and annotation generation. Additionally, no validation of annotations was performed, which assumes that the annotators gave reliable information. Changing the scope of the domain might require changing this model to either using expert users (as opposed to experienced users) or using large crowds of annotators with the facility to vote for the most popular annotation values per document.

The experimental results show that users generally responded positively to the system, with the number of documents viewed by students suggesting that it largely encouraged concept exploring. The fact that different educational purposes (e.g. introduction, example, explanation, etc.) were presented per concept introduced a more educational form of content delivery. Additionally, the usability questionnaire showed that students embraced the results presentation in the form of adapted hypertext presentations. However, the restriction to form a query from a fixed list of terms was regarded as cumbersome and reflected the users' desire to choose their own keywords.

From an educational perspective, the impact of the tool to support genuine learning for students in a focused domain has been very good. The students' success rate in gaining knowledge and being able to complete tasks, which they were previously unable to do indicates the benefit of such a tool. However, more

experimentation with different subject areas and different kind of tasks would provide a more comprehensive evaluation. Such evaluations are currently being planned. Also, a comparison of the focused system versus currently available query interfaces, such as Google are also planned.

Several areas of future work have been identified during the evaluation. First of all, changing the subject scope to a more refined domain would test the versatility of the system. Additionally, several variants of metadata generation could be employed to provide a comparison between manual, automatic and semi-automatic systems. The system could also provide greater support for the hybrid metadata generation approach involving automatic indexing and manual annotation.

An improved system will enhance the structure of the narratives, improve the adaptive presentation of the content and address the general usability concerns raised in the user evaluation. New design considerations have to be explored in order to provide a more aesthetically pleasing interface. Additionally, alternative ways of handling internal document links have to be researched.

Since this paper used a formal learning paradigm to test the system, a future iteration will also introduce the concept of informal learning. Finally, different types of test users (e.g. K12, adults, etc.) will be explored in order to develop more generalised strategies for personalised hypertext generation.

6. ACKNOWLEDGMENTS

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie <<http://www.cngl.ie>>) at Trinity College Dublin.

7. REFERENCES

- [1] Brusilovsky, P. & Henze, N. "Open Corpus Adaptive Educational Hypermedia". In *The Adaptive Web: Methods and Strategies of Web Personalisation*, Lecture Notes in Computer Science, vol. 4321, Berlin: Springer Verlag, pp. 671-696. 2007.
- [2] Lawless, S., Dagger, D., Wade, V. "Towards Requirements for the Dynamic Sourcing of Open Corpus Learning Content". In the Proceedings of the World Conference on E-Learning in Corporate, Government, Healthcare & Higher Education, E-Learn 2006, Honolulu, Hawaii, USA. October 13-17, 2006.
- [3] Apache Lucene is a full-featured text search engine library written entirely in Java. Available online at: <http://lucene.apache.org/java/docs/index.html>
- [4] Nutch is an open source web-search solution based upon Lucene. Available online at: <http://lucene.apache.org/nutch>
- [5] Boyle, T. "Design Principles for Authoring Dynamic, Reusable Learning Objects". In the *Australian Journal of Educational Technology*, vol. 19(1), pp. 46-58, 2003.
- [6] Cristea, A., Carro, R. "Authoring of Adaptive and Adaptable Hypermedia: An Introduction". In the *International Journal of Learning Technology*, Special Issue on Authoring of Adaptive and Adaptable Hypermedia, vol. 3(3), Inderscience. 2007.
- [7] Dagger, D., Wade, V., Conlan, O. "Personalisation for All: Making Adaptive Course Composition Easy". In *Special Edition of the Educational Technology and Society Journal*, IEEE IFETS, vol. 8(3), pp. 9-25. 2005.
- [8] Duval, E. "Standardized Metadata for Education: A Status Report". In the Proceedings of the AACE World Conference on Educational Multimedia, Hypermedia and Telecommunications, Ed-Media 2001, C. Montgomerie, & V. Jarmo (Eds), pp. 458-463, Tampere, Finland. June 25th-30th, 2001.
- [9] Koper, E. J. R. "Combining reusable learning resources and services to pedagogical purposeful units of learning". In *Reusing Online Resources: A Sustainable Approach to eLearning*, A. Littlejohn (Ed.), pp. 46-59, London: Kogan Page. 2003.
- [10] Littlejohn, A. "Community Dimensions of Learning Object Repositories". In the Proceedings of the 22nd Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education, H. Gross (Ed), pp. 3, Queensland University of Technology, Brisbane. 2005.
- [11] Multimedia Educational Resource for Learning and Online Teaching. Available online at <http://www.merlot.org>
- [12] Jorum, UK Higher Education Institutions Digital Repository. Available online at <http://www.jorum.co.uk>
- [13] National Digital Learning Repository. Available online at <http://www.learningcontent.edu.ie>
- [14] OpenLearn's LearningSpace provides free access to course materials from the Open University. Available online at <http://openlearn.open.ac.uk/>
- [15] OpenCourseWare is a free, web-based publication service for MIT's educational content. Available online at <http://ocw.mit.edu>
- [16] Zastrocky, M., Harris, M., Lowendahl, J-M. "Hype Cycle for Higher Education, 2006". Gartner Report G00139174. 30th June, 2006.
- [17] Ferguson, N., Charlesworth, A., Schmoller, S., Smith, N., Tice, R. "Sharing eLearning Content – A Synthesis and Commentary". London: JISC.
- [18] Wiley, D. A. "Learning Object Design and Sequencing Theory". PhD Thesis submitted to Brigham Young University. June, 2000.
- [19] Farrell, R., Liburd, S., Thomas, J. "Dynamic Assembly of Learning Objects". In the Proceedings of the Thirteenth ACM International Conference on the World Wide Web, WWW2004, pp. 162–169, Manhattan, NY, USA. May 17th-20th, 2004.
- [20] Weller, M., Pegler, C., Mason, R. "Putting the pieces together: What Working with Learning Objects means for the Educator". In the Proceedings of the Second eLearnInternational World Summit, Edinburgh International Conference Centre, Edinburgh, Scotland. February 18th-19th, 2003.
- [21] Chakrabarti, S., Punera, K., Subramanyam, M. "Accelerated Focused Crawling through Online Relevance Feedback". In proceedings of the Eleventh International World Wide Web

- Conference, WWW2002, Honolulu, Hawaii, USA. May 7-11, 2002.
- [22] Combine Web Crawler is an open source system for general and focused web crawling and indexing. Available online at <http://combine.it.lth.se/>
- [23] The Lemur Toolkit is an open-source suite of tools designed to facilitate research in language modeling and information retrieval. Available online at: <http://www.lemurproject.org>
- [24] Simple Web Indexing System for Humans – Enhanced (Swish-e), a flexible and free open source system for indexing collections of Web pages. Available online at <http://www.swish-e.org>
- [25] Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., Kanungo, T., Rajagopalan, S., Tomkins, A., Tomlin, J., Zien, J. “SemTag and Seeker: Bootstrapping the Semantic Web via Automated Semantic Annotation”. In the proceedings of the 12th International World Wide Web Conference, Budapest, Hungary, 178-186. 2003.
- [26] Greenberg, J. “Metadata Extraction and Harvesting: A Comparison of Two Automatic Metadata Generation Applications”. In the Journal of Internet Cataloging, vol. 6(4), pp. 59-82. 2004.
- [27] Berners-Lee, T., "The Two Magics of Web Science". Keynote at WWW 2007, Banff, Canada. 2007. Available online at <http://www.w3.org/2007/Talks/0509-www-keynote-tbl/>
- [28] Facebook is a free-access social networking website. Available online at: <http://www.facebook.com>
- [29] Flickr is an image and video hosting website, web services suite, and online community platform. Available online at: <http://www.flickr.com>
- [30] Digg is a provider of social bookmarks. Available online at: <http://digg.com>
- [31] Brusilovsky, P. & Peylo, C. “Adaptive and intelligent Web-based educational systems”. International Journal of Artificial Intelligence in Education, 13(2-4), 159-172, 2003.
- [32] Brady, A., Conlan, O., Wade, V. “Towards the Dynamic Personalized Selection and Creation of Learning Objects”. In the Proceedings of the World Conference on E-Learning in Corporate, Government, Healthcare and Higher Education, E-Learn 2005, G. Richards (Ed.), pp. 1903–1909, Vancouver, B.C., Canada. November, 2005.
- [33] 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 12th World Wide Web Conference, WWW2003, pp. 63-76, Budapest, Hungary. May 20th, 2003.
- [34] Brusilovsky, P. “KnowledgeTree: A Distributed Architecture for Adaptive E-Learning”. In the Proceedings of the Thirteenth International World Wide Web Conference, WWW2004, Alternate track papers and posters, ACM Press, pp. 104–113, Manhattan, NY, USA. May 17th-20th, 2004.
- [35] De Bra, P., Aerts, A., Berden, B., De Lange, B., Rousseau, B., Santic, T., Smits, D., Stash, N. “AHA! The Adaptive Hypermedia Architecture”. In the Proceedings of the Fourteenth ACM Conference on Hypertext and Hypermedia, pp. 81–84, Nottingham, England. August 26th-30th, 2003.
- [36] Conlan, O. & Wade, V. “Evaluation of APeLS - An Adaptive eLearning Service based on the Multi-model, Metadata-driven Approach”. In the Proceedings of the 3rd International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, AH2004, P. De Bra & W. Nejdl (Eds.), Berlin: Springer Verlag, pp. 291–295. 2004.
- [37] Zhou, D., Goulding, J., Truran, M., Brailsford, T. “LLAMA: automatic hypertext generation utilizing language models”. In the Proceedings of the 18th ACM Conference on Hypertext and Hypermedia, Hypertext 2007, Manchester, UK. 10th-12th September, pp. 77-80. 2007.
- [38] Henze, N., Nejdl, W. “Adaptivity in the KBS Hyperbook System”. In: Brusilovsky, P, Bra, P.D., Kobsa, A. (eds.) Proc. Of Second Workshop on Adaptive Systems and User Modelling on the World Wide Web. Vol. 99-07. Eindhoven University of Technology. pp. 67-74. 1999.
- [39] Brusilovsky, P., Chavan, G., Farzan, R. “Social adaptive navigation support for open corpus electronic textbooks”. In: De Bra, P., Nejdl, W. Proceedings of Third International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems. AH’2004. Lecture Notes in Computer Science, Bol. 3137. Springer-Verlag. 2004. pp.24-33
- [40] Dieberger, A. “Where did all the people go?” A collaborative Web space with social navigation information. 2000. Available online at <http://homepage.mac.com/juggle5/WORK/publications/Swik iWriteup.html>
- [41] Rutledge, L., Alberink, M., Brussee, R., Pokraev, S., "Finding the story: broader applicability of semantics and discourse for hypermedia generation." In the Proceedings of the 14th ACM conference on Hypertext and Hypermedia, Hypertext 2003, Nottingham, UK. 26th-30th August. pp. 67-76. 2003.
- [42] Lawless, S., Hederman, L., Wade, V. “OCCS: Enabling the Dynamic Discovery, Harvesting and Delivery of Educational Content from Open Corpus Sources”. In the Proceedings of the Eighth IEEE International Conference on Advanced Learning Technologies, I-CALT 2008, Santander, Spain. 1st-5th July, 2008.
- [43] The Java Text Categorizing Library is a Java implementation of libTextCat, a library for written language identification. Available online at: <http://textcat.sourceforge.net/>
- [44] Nutch and Web Archive eXtensions is a tool for indexing and searching web archive collections. Available online at: <http://archive-access.sourceforge.net/projects/nutch/>
- [45] Brooke, J. (1996) SUS: a "quick and dirty" usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester & A. L. McClelland (eds.) Usability Evaluation in Industry. London: Taylor and Francis.