AN EXTENSIBLE FRAMEWORK FOR AUTOMATIC KNOWLEDGE EXTRACTION FROM STUDENT BLOGS

Dr Andy M. Connor, Matthew Martin and Sam Joe
Auckland University of Technology, Auckland, New Zealand

ABSTRACT
This article introduces a framework for automatically extracting knowledge from student blogs and injecting it into a shared resource, namely a Wiki. This is motivated by the need to preserve knowledge generated by students beyond their time of study. The framework is described in the context of the Bachelor of Creative Technologies degree at the Auckland University of Technology in New Zealand where it is being deployed alongside an existing blogging and ePortfolio process. The framework uses an extensible, layered architecture that allows for incremental development of components in the system to enhance the functionality over time. The current implementation is in beta-testing and uses simple heuristics in the core components. This article presents a road map for extending the functionality to improve the quality of knowledge extraction by introducing techniques from the artificial intelligence field.

KEYWORDS
Blogging, Wikis, Knowledge Extraction, Enhanced Learning

1. INTRODUCTION
Blogs (a shortened form of “Web-logs”) are essentially an electronic journal that allows an individual to keep records of their thoughts and opinions by writing on a website. Such writings are normally ordered chronologically and often constrained to a particular topic of a timely nature. The writing of blogs (or “blogging”) dates back to the early 1990s with the term blog becoming popular after the late 1990s. One typical characteristic of blogs is that they contain selective content, often chronicling personal daily events or supporting contentious individual views. At first glance, blogs do not seem to fit into shared knowledge acquisition framework, however in recent years, the use of these electronic journals in education is increasing, giving rise to the opportunity to combine peer learning approaches [1] with natural language processing techniques [2] to create a shared knowledge base to enhance student learning.

This paper outlines an initial framework for the automatic extraction of knowledge from student blogs in the context of the Bachelor of Creative Technologies degree at Auckland University of Technology.

2. RELATED LITERATURE
The development of new media channels has created an emerging potential for the utilisation of innovative software and hardware in an educational setting. This has included smartphones, laptops, podcasts, tablets, wireless systems and Web-based learning environments. Most recently, however, there has been a growing interest on blogs, blogging and the use of Wiki based tools [3].
Ferdig & Tramell [4] provide an early perspective of the use of blogs in education as well as some of the pedagogical arguments for deploying blogs as an educational tool. They suggest that “Blogs are useful teaching and learning tools because they provide a space for students to reflect and publish their thoughts and understandings. And because blogs can be commented on, they provide opportunities for feedback and potential scaffolding of new ideas.” (pg. 2). They go on to suggest a number of practical suggestions for implementing blogs as well as four advantages of blogs supported by other literature, namely:

1) The use of blogs helps students become subject-matter experts.
2) The use of blogs increases student interest and ownership in learning.
3) The use of blogs gives students legitimate chances to participate.
4) The use of blogs provides opportunities for diverse perspectives, both within and outside of the classroom.

Recent research has argued that the use of blogs can be beneficial to students in a wide range of disciplines, including the health sciences [5], visual arts [6], computer science [7], engineering [8] and science [9] to name but a few. Many of these studies implement blog-based course components that support claimed benefits discussed by Ferdig & Tramell [4]. Clearly there is considerable interest in this area and much potential for individual students to enhance their individual learning experience through the use of blogs, irrespective of the discipline.

Many authors also suggest that collaborative learning can be augmented through the use of Wiki-based tools. For example, Coutinho & Bottentuit [10] suggest that learning can occur through social interaction generated by the exchange and sharing of information and opinions among a peer group in an online Wiki. Gruber [11] goes on to suggest that there are potential benefits from combining the best ideas from the Social Web and the Semantic Web, therefore integrating the ecosystem of participation associated with the former with the ecosystem of data associated with the latter, where value is created by the integration of structured data from many sources. Boulos, Maramba & Wheeler [12] also suggest that research could be undertaken to investigate the cross-operability and integration emerging Web 2.0 applications, particularly Wikis, blogs, and podcasts, and their respective and synergistic contributions toward the enhancement of student learning. They conclude their research by stating “Of particular importance is research into novel integrative applications, to serve as the ‘glue’ to bind the different forms of Web-based collaborationware synergistically in order to provide a coherent wholesome learning experience.”

Despite the identification of such potential and significant, most research to date seems to have focused on the use of blogs and Wikis in parallel, rather than undertaking any formal integration of the two. For example, Miyazoe & Anderson [13] examine the effectiveness of three different online writing activities in formal university education: forums, blogs, and wikis. They note in their work that few research studies have tested the simultaneous application of two or more online writing tools in a single course and as such only cite four studies that do so [14-17]. Whilst the scope of the research Miyazoe & Anderson [13] is limited to a single discipline, there is only a small body of literature that deals with the simultaneously implementation of blogs and Wikis.

Chen, Cannon, Gabrio, Leifer, Toye & Bailey [14] outline an approach they call “Folio Thinking”, an approach they describe as an instructional method grounded in the process of students creating learning portfolios. They go on to suggest that the pedagogical approach is designed to enhance self-awareness by enabling students to make their knowledge explicit and
visible for themselves as well as for others. In this study, the authors utilize the software Tikiwiki, which claims to integrate blogs and Wikis. However, whilst these features exist in the software they are still in part separate, which requires students to perform two separate processes – their individual blogging and reflection as well as the construction of a shared knowledge base via the Wiki. Tighter integration of blogs and Wikis has the potential to streamline the creation of a shared knowledge base and limit the possibility of lost knowledge that can occur when students delete content from their blogs or when they graduate from the programme.

The next section outlines the role of blogs and Wikis in the Bachelor of Creative Technologies (BCT) degree at Auckland University of Technology in the context of the philosophical underpinnings of the degree programme. The approach is similar to the Folio Thinking approach [14] but includes a flexible time-based progression that introduces new web-based tools at different phases of progress.

2. THE BACHELOR OF CREATIVE TECHNOLOGIES

The Bachelor of Creative Technologies degree is offered by a unique academic unit at Auckland University of Technology. The unit is Colab, the “collaboratory” for Design and Creative Technologies that exists purely in the research-teaching nexus. The unit combines the traditional academic structures associated with a School with the support and infrastructure of a Research Institute. The goal of Colab is to encourage researchers, students and stakeholders to imagine, construct, articulate and navigate rapidly changing social, economic, technological and career environments¹. The Bachelor of Creative Technologies (BCT) is a key enabler of this goal.

Historically, Colab emerged as a result of the creation of a new Faculty at Auckland University of Technology in 2005, where four existing Schools (Art & Design, Communications & Media Studies, Computing & Mathematical Sciences, and Engineering) were combined under a single umbrella. In 2007, the Faculty formed the Interdisciplinary Unit, a “virtual 5th school” to develop new experimental alliances, research collaborations and learning experiences across these overlapping disciplines. In 2012 the Interdisciplinary Unit was merged with a number of existing research centres to form Colab.

The BCT was originally designed with a flexible and experimental curriculum that encouraged the development of students as reflective practitioners. The curriculum was revised in 2013 to formalize the learnings from delivering an experimental curriculum, and introduce new innovations. In particular, the core of the degree was changed to introduce a formal research methods component and options were introduced to allow students to undertake a research practicum, essentially a research apprenticeship under the guidance of a research active member of staff. The curriculum structure is based around a number of Studio and taught courses. Students can then elect to take a number of subject courses which may also be selected from other academic schools. The structure of the degree is shown in Figure 1.

¹ For more information, please visit the Colab website: http://colab.aut.ac.nz/about/
Figure 1. Structure of the Bachelor of Creative Technologies (BCT)

The current list of subject courses in the second (Level 6) and third year (Level 7) are shown in Table 1, with elective papers being any course on offer throughout the University.

Table 1. Optional papers in the BCT

<table>
<thead>
<tr>
<th>Year/Level</th>
<th>Courses</th>
</tr>
</thead>
<tbody>
<tr>
<td>2nd (Level 6)</td>
<td>176002 Technology, Culture and Creative Practice; 176004 Entrepreneurship and Innovation; CTEC 601 Synthetic Realities; CTEC 602 Interactive Technology and Systems; CTEC 603 Digital Fabrication; CTEC 604 New Media Applications; CTEC 605 Creative Audio; CTEC 606 Experimental Imagery</td>
</tr>
<tr>
<td>3rd (Level 7)</td>
<td>177006 Virtual and Immersive Environments; 177007 Intelligent Agents and Environments; CTEC 700 Space, Image and Sound; CTEC 701 Mediated Performance; CTEC 702 Emerging Practices; CTEC 703 Interaction and Play; CTEC 704 Transmedia Narratives; CTEC 705 Creative Workflows; CTEC 706 Research Practicum I; CTEC 707 Research Practicum II</td>
</tr>
</tbody>
</table>

All of the courses offered as part of the BCT utilize web based collaborationware in terms of delivery and assessment. The following section outlines the overall framework for deploying such tools in the pedagogy of the degree.

3.1. The Role of the ePortfolio

An e-Portfolio can be produced by a student for many reasons. One example is the simple showcase of work in the digital sphere. Such e-Portfolios do not demonstrate the true value that can be achieved by using an e-portfolio as collaborative, personalized learning space. Such an e-Portfolio can assist building the skills essential for students in higher education through a process of generative learning. This learning approach involves incorporating existing
knowledge with new ideas based on experimentation and open-mindedness. From the educator’s perspective, the e-Portfolio can also facilitate assessment and feedback and therefore is a mechanism by which a developmental conversation is generated. Throughout the degree, students in the BCT develop an ePortfolio that is intended to help them develop the skills of collection, selection and reflection [18] that are required for students to take control of their own learning process.

Greenberg [19] argues that it can be helpful to think about ePortfolios in terms of when the work is organized relative to when the work is created. He goes on to present a classification of three types of ePortfolios:

1) The showcase ePortfolio: organization occurs after the work has been created.

2) The structured ePortfolio: a predefined organization exists for work that is yet to be created.

3) The learning ePortfolio: organization of the work evolves as the work is created.

Building on this classification, the usual process of collection, selection and reflection is extended to include an additional component of curation. The curation is a particular extraction of information from the showcase ePortfolio to suit a particular assessment event. The emphasis of the ePortfolio is that it contains an evolving body of work that is refined and repackaged to suit any particularly purpose.

Recent research suggests that the development of an ePortfolio does not significantly increase the cognitive load of Masters students [20], however there has been little research into the impact on first year undergraduate students in this regard. However both anecdotal evidence and other research suggest that the transition to University level education is a significant stressor for students [21]. To minimize the potential impact on cognitive load, the ePortfolio process is incrementally introduced to BCT students at different phases of their development.

3.2. Year 1: Blogs

Blogging is introduced in the first year of study to enhance education by encouraging reflective practice. Beale [22] argues that blogging provides advantages in terms of social and pedagogical perspectives. For example, the suggestion that blogging produces a sense of community amongst the students because they can read and comment on other students postings, and can learn from both experiences and from the insights of their peers is certainly borne out in practice. The role of the blog is primarily to develop a reflective habit, however it stealthily introduces the process of collection and selection. The former is driven by the need to provide relevant evidence of the student’s research whilst the latter is implicit by making their work public. Beale [22] also argues that the fact that students can see the sort of activities done by other students there is transparency in terms of the amount of work that is required as well as the quality of work being produced. Because others can also see their level of activity or inactivity, peer pressure should exert an influence and encourage them to maintain a degree of selectivity in terms of how the students present their work and their reflections.

3.3. Year 2: Portfolios

In their second year of the programme, students are introduced to the web development skills that allow them to develop their own ePortfolio website. The ePortfolio complements the blogging habit developed in the first year and encourages selectivity and discernment. In essence, the blog becomes a sandpit for experimentation from which successful elements are carried forward to the ePortfolio.
3.4. Wiki

The use of a Wiki is a relatively new feature of the programme, motivated by the desire to not lose knowledge as students progress through the programme and eventually graduate. When a student graduates, the link to their blog is lost and even if it was retained there is little incentive for current students to trawl through such blogs looking for particular knowledge embedded in a blog.

The use of Wikis in student-centric learning is increasing and there examples of successful [23] and unsuccessful [24] implementations in the literature, as well as studies that indicate some of the barriers to success [25]. The Wiki system has not been used for a sufficiently prolonged period of time to assert anything other than an anecdotal observation that the rate of content creation is relatively low. This can be attributed to the fact that the content creation is an additional task to be completed that is assessed. This observation is what motivates the work currently being implanted to automatically create Wiki content as a result of the blogging process.

4. AUTOMATED KNOWLEDGE EXTRACTION FRAMEWORK

The proposed framework for the automatic extraction of knowledge from student blogs is given in Figure 2.

![Automated Knowledge Extraction Framework](image)

The blogs and the Wiki are elements that already exist in the learning framework for the BCT degree, therefore only two components are required to complete the framework – the Blog Crawler and the Knowledge Engine.

4.1. Blog Crawler

The Blog Crawler is a relatively simple component and has already been successfully implemented in the prototype system, subject to some limitations in operation. The implementation of this component has been relatively straightforward as a result of two factors. The first of these factors is that all students are requested to use the same blogging platform (currently Tumblr) which therefore only requires one API to extract the content of any given blog in a structured format. The Tumblr API v1 easily allows the extraction of all content in an XML format from within a PHP script. Reading Tumblr data is easily by the Blog Crawler as links to all student blogs are collated on a simple HTML page. The Blog Crawler scans the HTML page searching for text that implies a link to a Tumblr blog, which is in the format http://studentblogtitle.tumblr.com, where “studentblogtitle” is the name of the Tumblr blog. The content of the blog could then be fetched by creating an HTTP query in the form
http://studentblogtitle.tumblr.com/api/read which returns a structured XML version of the blog content.

The second simplifying factor is achieved by imposing a process constraint on the blogging behavior of the students. If a student wishes a blog entry to be considered for inclusion in the Wiki, they are required to tag the blog post with only two tags. The first tag is “wiki” to indicate to the Blog Crawler to extract a given post and the second is a free text tag that specifies the subject of the post. This allows an alternative HTTP request to be formulated by adding GET elements to only return blog posts the student is intending for inclusion in the Wiki. The HTTP query is then in the form http://studentblogtitle.tumblr.com/api/read?tagged=wiki. The function of the Blog Crawler is therefore very straightforward, it parses the XML to extract the date of each post and the content of any posts dated since the last crawl are sent to the Knowledge Engine as for processing.

The Blog Crawler script can be initiated manually or automatically scheduled on the server using an appropriate mechanism, for example using cron on a Linux based server. It is currently limited to using Tumblr as a result of the specific API, however can in theory be extended to include other blogging platforms. A more immediate concern is the use of the Tumblr v1 API which has been superseded by V2, and the use of XML instead of the more lightweight JSON interface. Work is currently in progress to translate the Blog Crawler functionality to use Tumblr v2 API and JSON to improve performance.

4.2. Knowledge Engine

Within the overall framework for knowledge extraction a number of different models for the Knowledge Engine can be conceived. The current implementation uses straightforward text processing to inject content into the Wiki, however there is significant potential to enhance this through the use of approaches such as natural language processing [2] or word disambiguation [26] approaches that are common in information retrieval applications.

The current implementation of the Knowledge Engine is again relative simple and is based around the automatic injection of material that students feel is relevant in their blogging process. When the Knowledge Engine is fed a blog post from the Blog Crawler it queries the Wiki to determine whether a page already exists with the title of the subject tag in the blog post. This is again achieved by the formulation of an HTTP request of the form http://wikiserver/w/api.php?action=query&titles=tagToSearchFor&format=xml. The returned XML will have no values for pageid if no pages exist. If no such page exists, the Knowledge Engine creates a new page using the Wiki XML-RPC interface and populates it with the content of the blog post.

If a page with the same title already exists, the Knowledge Engine takes a pragmatic approach to ensuring that the content of the original page is not compromised. The Knowledge Engine creates a new page with a numeric suffix, again checking such a page does not exist and using an incremental number if it does, and then injects the new page in to the Wiki. The Knowledge Engine then injects content into each page to suggest that the full range of pages on the same topic be combined. Such notices are common on large scale Wikis such as Wikipedia.

The current implementation has two main limitations. The first of these is that each page is injected individually into Wiki by the Knowledge Engine as the Blog Crawler is running and sending data to it, which potentially slows down the process when many pages need to be created. The second limitation is that the Knowledge Engine has no capacity to intelligently merge text that relates to the same topic. The first of these limitations is being addressed by
integrating the MediaWiki Bulk Page Creator\(^2\) with the Knowledge Engine to allow offline structuring of the data and the creation of new pages after the Blog Crawler has completed its latest crawl. The resolving of best how to intelligently merge text in a coherent way is the focus of future research.

4. Further Research

The current implementation of the Knowledge Engine is functional, but clearly limited in terms of its ability to handle complex situations that arise with merging content that is duplicated or overlapping.

The overall framework shown in Fig. 2 is an abstraction of the actual architecture of the system that is based on a three layer model. This is a common model in computing systems where the presentation, application processing, and data management functions are physically separated in the software implementation. This allows each layer to be modified independently of the other layers, provided the interface between layers remains consistent.

Future work will first focus on the Knowledge Engine layer to improve the ability to merge documents in a meaningful way. This is a non-trivial task and will draw on previous work in the areas of natural language generation [27], multiple document processing [28] and latent semantic analysis [29].

In parallel to this a study will be designed to investigate the role the Wiki has in student engagement and learning, potentially using similar measures to Hazri & North [23]. This will allow the potential pedagogical value of the automated framework to be assessed against the value of just the Wiki itself.

Both the blogs and the Wiki offer potential for similarity clustering and text mining [30] to identify recurring themes. Similarly, there is potential to develop a more informed information search strategy that allows students to identify appropriate material that may be either the Wiki or the blogs using a contextual retrieval approach [31]. The potential value of a minable data repository will be fully explored.

4. Conclusions

This paper has outlined the ePortfolio process for students enrolled in the Bachelor of Creative Technologies degree at Auckland University. It has described the motivation for enhancing this process through the addition of an automated extraction framework that allows a shared knowledge base to be created with minimal overhead on behalf of the students. Progress to date in terms of the creation of a simple Knowledge Engine has been discussed and directions for further work have been identified.

References


\(^2\) http://meta.wikimedia.org/wiki/MediaWiki_Bulk_Page_Creator


Authors

Dr Andy M. Connor obtained a Ph.D. in Mechatronics in 1996 from Liverpool John Moores University in the United Kingdom. He is currently employed as a Senior Lecturer in Colab, the interdisciplinary research/teaching nexus at Auckland University of Technology (AUT) in Auckland, New Zealand. Prior to this he was a Senior Lecturer in the School of Computer & Information Sciences at AUT as well as being a Research Associate in the Engineering Design Centre at the University of Cambridge and a Research Officer in the Engineering Design Centre at the University of Bath. His research interests span Software Engineering, Artificial Intelligence and Design Automation. Dr Andy M. Connor is a Member of the Institution of Engineering & Technology.

Matt Martin obtained the Bachelor of Creative Technologies degree in 2013 from Auckland University of Technology in New Zealand. He is currently working towards his Master of Creative Technologies degree at Auckland University of Technology (AUT) in Auckland, New Zealand. His research interests relate to the role of technology in terms of generating experience.

Sam Joe obtained the Bachelor of Creative Technologies degree in 2013 from Auckland University of Technology in New Zealand. He is currently working towards his Master of Creative Technologies degree at Auckland University of Technology (AUT) in Auckland, New Zealand. His research interests involve the creation of collaborative workspaces using web technology in education.