

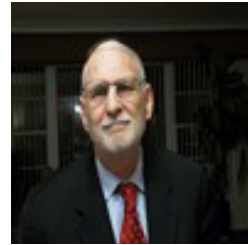
IRR
ODI

THE INTERNATIONAL
REVIEW OF RESEARCH IN
OPEN AND DISTANCE LEARNING

Technology-Enhanced Information Retrieval for
Online Learning

Vol 13
No 5

Editorial - Volume 13, Issue Number 5



Maiga Chang¹, Rita Kuo², Gene Loeb³ and Bolanle Olaniran⁴

¹Athabasca University, Canada

²Knowledge Square Inc., Taiwan

³Center for Technology and Mental Health of Elderly, USA

⁴Texas Tech University, USA

Technology for enhancing learning has developed rapidly in decades. Teachers can use computer technologies to enrich their learning materials, such as adding images and videos. Abundant teaching materials can be shared on the Web, and learners can easily access plenty of information through the Internet. When learners have questions, they can find answers or solutions by searching the textbook, learning materials, and the Internet. Learners can also ask tutors and peers questions as well as discuss problems with them.

Through sharing and communicating, learning information becomes large and complex. Teachers may experience difficulty managing the volume of teaching materials and assignments, while learners can also experience problems in retrieving and deciphering useful information efficiently and quickly. To solve this problem, information retrieval technology is important for teachers to manage and analyze learning materials and for learners to learn efficiently and effectively.

The purpose of this special issue is to explore how models, theories, and solutions of information retrieval and content analysis can be used in online learning and the benefits users (i.e., teachers and learners) are able to gain from such systems and agents. This special issue covers two aspects of information retrieval techniques – knowledge acquisition and information retrieval application – which can be applied in online learning.

For knowledge acquisition, Dr. Nguyen and Dr. Yang integrate statistics, data mining, and natural language processing techniques to construct a concept tree for the course Introduction to C Programming from text documents in Vietnamese. The method proposed in their article can help tutors reduce the time spent on building the ontology of fifty courses for a

recommender system which makes suggestions on course selections for students.

Natural language processing can be used to find models in qualitative data analysis also. As such, Dr. Tierney integrates graph theory to accelerate the speed of qualitative interview analysis and to find the categories of analyzed qualitative data. Also, Dr. Wen and colleagues use natural language processing to build an asynchronous question answering system for assisting teachers to respond to learners' messages automatically. The system is able to identify the colloquial nature of the messages and messages mixed with non-questions in order to eliminate question-irrelevant contents and to find the core questions in the messages. Such a system facilitates communication processes between teachers and learners as well as helps reduce learner's frustrations, attributable to the lack of real-time access to teachers in online learning environments.

Dr. Yu and colleagues, on the other hand, use image processing techniques to analyze instructional videos in order to automatically build an interactive index for learners who choose and watch lessons in videos. Learners can choose a snapshot region or image segment from the instructional video for further in-depth study. The researchers in their experiment found high accuracy of recall rate and precision rate compared to the index made by teachers themselves.

Regarding information retrieval applications, Dr. Tang and colleagues designed a personalized article recommender which can suggest additional readings for learners based on learner characteristics (i.e., interest, background knowledge, job nature, and learning expectation) and article features (e.g., paper topics). Their results show that such a context-aware article recommendation system could be effectively used for inexperienced learners, who are unfamiliar with a particular learning domain or with course content.

Dr. Wang and colleagues use the knowledge-based engine, InfoMap, to draw geometry figures dynamically to help high school geometry students. The accuracy of the figures drawn by the system is near 90%. Also, Dr. Baldiris and colleagues design two processes to solve the problem of searching for learning objects in distributed learning object repositories and placing the learning objects into adaptive learning design automatically to assist teachers' design work. The first process, Learning Object Repositories SEarcher (LORSE), a distributed learning object searching process, uses two different types of agents – the directory facilitator agent and the specific search agent – to support learners in the learning object search process; the second process, LOOK, a micro-context based positioning process, compares the micro-context of learning objects previously found in the repository and in the curricular structure to decide the best location for the learning object in the adaptive learning design.

Dr. Butakov and colleagues in an attempt to protect student's intellectual property submit student assignments to a plagiarism detection service offered by a third party. This is similar to the approach used by teachers when they see suspicious wording in a student's essay; they will search for the particular sentence(s) with conventional search engines such as Google and Bing. The proposed architecture only submits partial information to exter-

nal plagiarism detection services to protect students' copyright and to avoid intellectual property violations. Finally, Dr. Cheng and colleagues developed an e-book hub service on a free and open-source cloud computing platform – OpenStack. The e-book hub service is designed to help teachers produce, transform, and manage their e-books to fit the needs of students who are using different platforms including smart phones, tablets, and desktop computers.

Athabasca University 



Protecting Students' Intellectual Property in the Web Plagiarism Detection Process



Sergey Butakov

Concordia University College of Alberta, Canada

Vadim Dyagilev

Altai State Technical University, Russia

Alexander Tskhay

Altai Academy of Economics & Law, Russia

Abstract

Learning management systems (LMS) play a central role in communications in online and distance education. In the digital era, with all the information now accessible at students' fingertips, plagiarism detection services (PDS) have become a must-have part of LMS. Such integration provides a seamless experience for users, allowing PDS to check submitted digital artifacts without any noticeable effort by either professor or student. In most such systems, to compare a submitted work with possible sources on the Internet, the university transfers the student's submission to a third-party service. Such an approach is often criticized by students, who regard this process as a violation of copyright law. To address this issue, this paper outlines an improved approach for PDS development that should allow universities to avoid such criticism. The major proposed alteration of the mainstream architecture is to move document preprocessing and search result clarification from the third-party system back to the university system. The proposed architecture changes would allow schools to submit only limited information to the third party and avoid criticism about intellectual property violation.

Keywords: Intellectual property protection; plagiarism detection; system architecture; social issues; learning management systems

Introduction

The rapid development of the Internet along with increasing computer literacy has made it easy and tempting for digital natives to copy someone else's work and paste it into their own. Plagiarism is now a burning issue in the education, industry, and research community (Spafford, 2011). For example, one group of researchers estimated up to 90% of students in high schools are involved in different kinds of plagiarism (Jensen, Arnett, Feldman, & Cauffman, 2002). Other research has shown that students with extensive exposure to the Internet are more inclined to be engaged in copy-paste practices (Underwood & Szabo, 2003).

Distance and online education can be even more vulnerable to plagiarism because of its remote and asynchronous nature. Concern about such vulnerability is growing along with the increasing number of online programs (Heberling, 2002; Marais, Minnaar, & Argles, 2006). In online course settings, PDS should be used as a tool to combat plagiarism and educate students on proper research and writing practices. As an example, Jocoy and DiBiase (2006) showed how automated plagiarism detection tools could be of great assistance in helping educators to raise student awareness about plagiarism and improve course outcomes. Due to its increasing importance, there have been a number of research projects concentrated on plagiarism recently. In this study, we concentrate on the plagiarism detection process, particularly focusing on the architecture of software tools used for detecting possible sources of plagiarism.

A major problem that arises for anyone searching for the sources of a suspicious paper is the degree of effort necessary to perform a document search to compare the suspected plagiarism with all possible sources. The comparison base can be relatively small if the search has to be performed in the scope of a single learning community, such as one college. At the other extreme, if the same search has to be done against all publicly available web pages, then the searcher must consider billions of documents. To perform such a search at the individual school level, each institution would require its own web crawler. This option is prohibitively expensive for most of the universities worldwide. The other option—outsourcing the search—may lead to intellectual property (IP) violation charges from the students (Bennett, 2009).

In this paper we propose an improved way to build a PDS with an architecture that allows a school to use the applicable IP policy and yet also allows the PDS to maintain acceptable search capability. The proposed framework is based on the idea that only a limited amount of information from the original submission is required to locate potentially similar documents on the Web. The scope of the proposed approach is to build a PDS based on a conventional web search engine. The goal is not to disclose the entire suspicious submission to the third party—the outsourcing company that runs the PDS. This third party may or may not have its own web crawler. In the latter case, the PDS has to employ a conventional search engine to look for potential sources. This approach restricts the usage of such functions as the hashing of search queries to hide content because conventional search engines such as Google, Yahoo, or Bing do not work with hashed queries. The approach assumes that the

information will be transmitted to the PDS in an unencrypted form, and thus to preserve students' IP we have to limit the amount of information that can be transmitted to the external part of the PDS.

The rest of the paper is organized as follows: The second section, Related Works and Existing Solutions, describes the major options for how a general purpose PDS can be built and outlines why current architectures may be considered inappropriate from the IP protection point of view. The third section, Proposed Architecture, discusses a few of the legal cases against one of the major commercial PDS available and proposes a solution that would allow educators to avoid such cases. The fourth section, Experiments, provides more technical details about the proposed solution, outlining the modified client-server architecture for PDS. It also shares the experimental results that show how much of the original submission should be transmitted to PDS to locate similar documents on the Web.

Related Works and Existing Solutions

Plagiarism detection services are in demand in many areas, including but not limited to education, publishing, and research proposal evaluations. Some of these areas may be affected by IP protection legislation. Most likely in education, not all student works are subject to IP protection, but senior projects and master's and PhD theses could fall in to this category. The same applies to research papers, proposals, reports, and patent applications. In this paper we will be talking about IP protection during the process of plagiarism detection in student papers, but most of the concepts are applicable to other areas.

There are a number of research studies that deal with detecting duplicated material available on the Internet. The applied side of this research topic has evolved from earlier projects examining plagiarism detection in source code (Donaldson, Lancaster, & Sposato, 1981; Krsul & Spafford, 1997) to copy-paste detection in essays and program codes (Burrows, Tahaghoghi, & Zobel, 2007). A number of related studies on system architecture have been done to examine web indexing (Zaka, 2009), spam protection (Urvoy, Chauveau, Filoche, & Thomas, 2008), and writing style detection to identify individuals on anonymous Web sites (Abbasi & Chen, 2008).

Technical intellectual property protection is a well-developed area in digital rights management (DRM) systems. There are many ways that DRM systems can be implemented and managed. One of the areas that could be related to plagiarism detection is digital watermarking, especially implementations that establish accountability for copying the protected digital object (Arnold, Schmucker, & Wolthusen, 2003). DRM and PDS are similar in the sense that they allow users to identify the source. For example, the embedded watermark in an Oscar Academy award movie sent out for prescreening can reference the person who was given that particular copy (Cox, Miller, Bloom, & Fridrich, 2009), and a side-by-side comparison of two documents can reveal similarities in the texts. But the ability to implement watermarking or steganography on plain text files is very limited as changes will be visible.

In contrast to the well-established research field of DRM, at the moment there are not very

many studies on IP protection during the plagiarism detection process. If we look at the PDS currently available on the market, we can see that the PDS architecture for papers submitted locally is very straightforward. The school maintains a database of all student works and compares each new document with the existing ones upon submission. In terms of IP protection, the school informs students that their submissions will remain as digital files in the school database and will be used solely for PDS.

Conversely, the Internet search assumes that the paper must be compared against all possible sources on the open Web. As mentioned above, such a search can theoretically be performed on the school side or outsourced to a company that specializes in plagiarism detection. Performing such a search would require schools to maintain a web crawler similar to the ones used by major search engines. Thus this option is cost-prohibitive for most schools. Outsourcing, the second option, can be done in two major ways: outsourcing the whole process or outsourcing the most data-intensive parts of it. An overview of these approaches is displayed in Figure 1. There are two important points that should be highlighted here: (1) the complete student submission is transmitted to the PDS and (2) the PDS retains a copy of the document to use for comparison to other submissions in the future.

Figure 2 illustrates how information flows in the case of complete outsourcing: It starts with a student submission on the left side and ends with two similarity reports after passing through the university side. Such an approach is used by major players in the PDS market, such as iParadigms LLC, with its well-known Turnitin® service (www.turnitin.com), and BlackBoard's SafeAssign (www.safeassign.com).

The second way is to outsource the most difficult part of the detection process: the global search for candidate sources—documents that may be similar to the suspicious submission. In other words, this part of the process should narrow down the scope of the search from the tens of billions of documents available on the Web to just dozens of documents. As Figure 1b shows, the university portal in this case has to perform the detailed comparison, but it is not as intensive as a complete Internet search. Many smaller scale companies utilize public search engines to perform such searches. Crot, one example, utilizes a global search API from Microsoft's Bing search engine to perform this kind of selection. It uses a sliding window x words in length, thus sending to the search engine all the phrases from the document that have been formed by this sliding window and in doing so performs a very exhaustive search (Butakov & Shcherbinin, 2009).

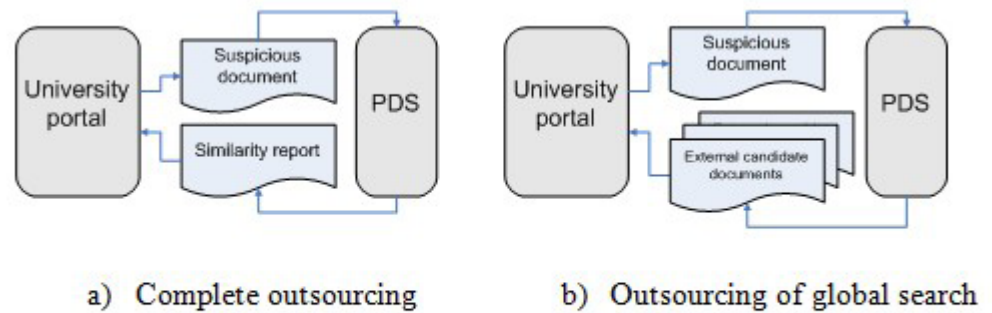


Figure 1. Two typical ways to outsource the plagiarism detection process.

Both of the plagiarism detection outsourcing approaches outlined in Figure 1 cause students to raise concerns that the PDS violates their IP rights. Up to now, four lawsuits have been filed against Turnitin by students. The company has won all of them, but in the latest lawsuit students were able to take their case to the point where the judge examined fair use policy tests outlined in US copyright legislation (Bennett, 2009). The students claimed that Turnitin PDS was making profits using their submissions, and therefore its owners should be liable for copyright violation. Although the service was not found guilty of this charge, such cases generate negative publicity for the schools involved. Also, similar cases could result in different decisions in other countries. For example, European Union law is known to be tougher concerning copyright protection.

Many universities are aware of these legal concerns. For example, the University of Maryland, University College suggests that faculty members consider the following questions when choosing PDS: “Will the [plagiarism detection] service archive all student submissions for further detection? Will all submissions be immediately destroyed after a report is generated?” (VAIL, 2012).

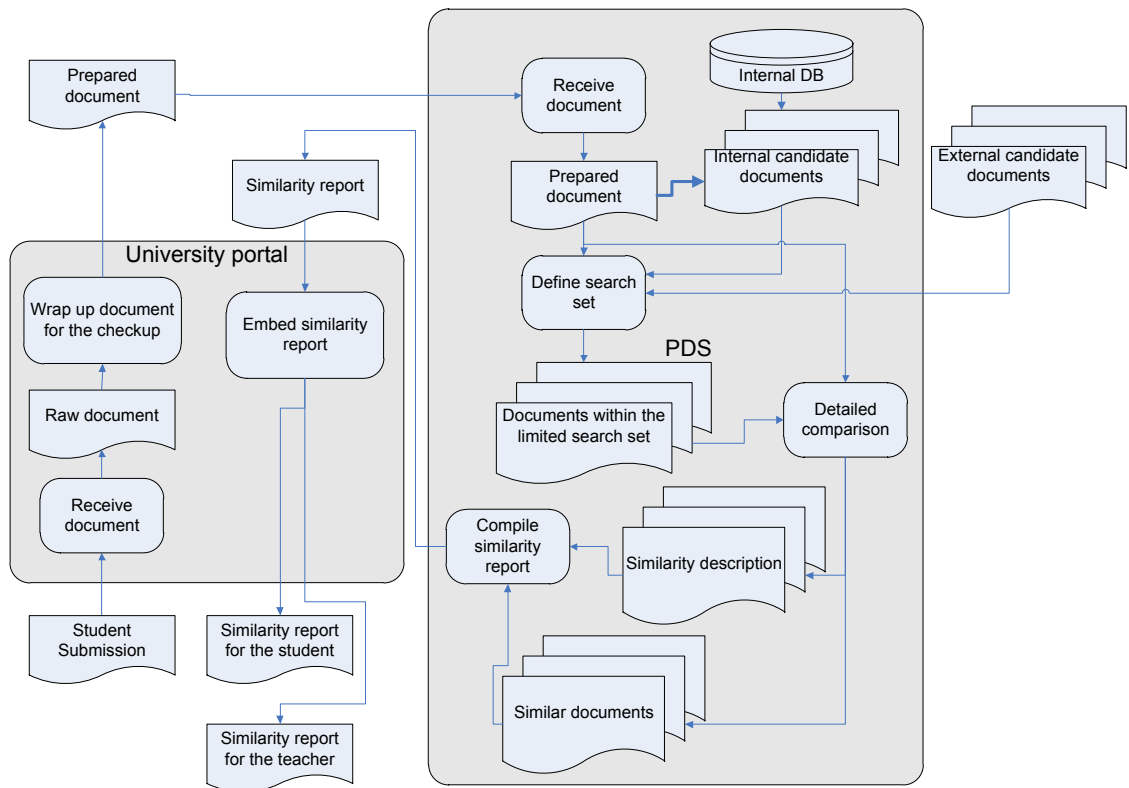


Figure 2. The complete outsourcing of the plagiarism detection process.

As illustrated by this example, from the IP protection point of view, using an external service to only narrow down the search scope is better than outsourcing the whole detection process because student submissions will not be stored and reused for profit. But there is still a concern because the suspicious submission goes to the third party. From a legal standpoint, some suggested using only a limited amount of information from the suspicious text to obtain potential sources from the Web (Butakov & Barber, 2012), but this left an open question about how such selection would affect the detection quality. Referring back to Figure 1, this approach means that the external PDS should not keep the submitted document and should not get it in its original form. The next section outlines the details of the architecture based on these principles and can serve as a blueprint for this sort of PDS development.

Proposed Architecture

Figure 3 shows the main concepts of the proposed architecture for PDS. The service itself is divided into an internal part running on university infrastructure and an external part running on a third-party system. The internal part plays a service role when it communicates with the university portal and a client role when it submits search requests to the external part of the service. Most of the data processing is done on university infrastructure. The only step in the process performed on third-party infrastructure is the preselection of can-

didate documents or locating the probable sources of the plagiarized paper on the Web. The major difference between the general approach presented in Figure 1b and Figure 3 is the form and amount of information that is transferred to the third-party infrastructure. The proposed architecture sends out essential queries, instead of sending the complete student's submission.

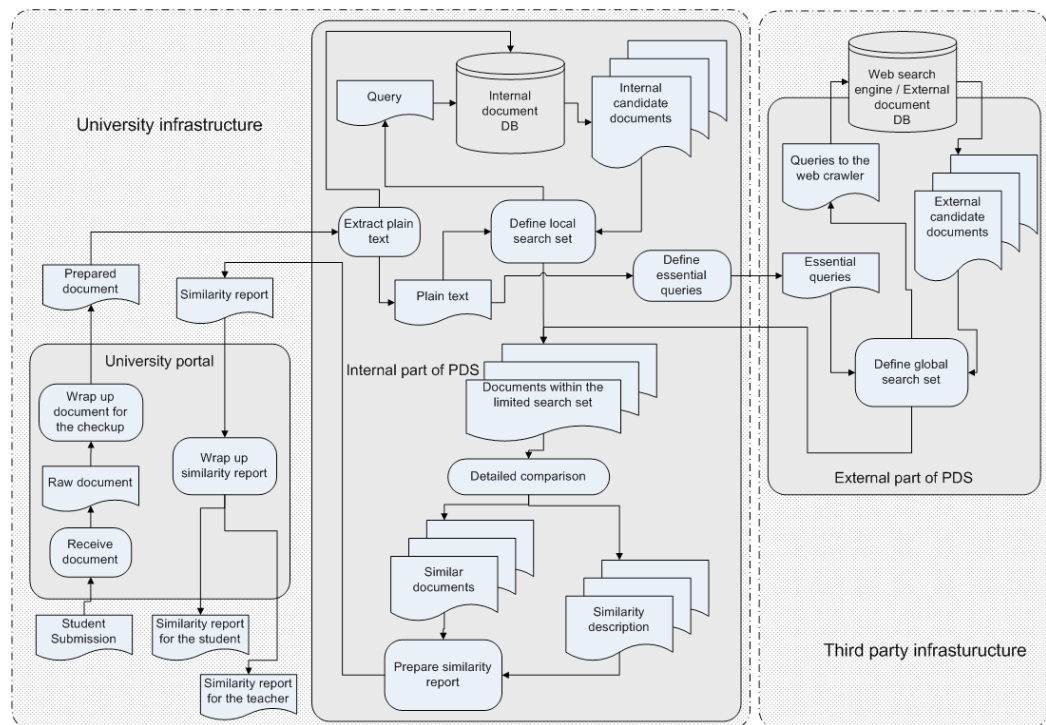


Figure 3. Outline of the proposed architecture for PDS.

To use the proposed architecture, we assume the whole suspicious submission is not required to select candidate sources. This assumption is based on the results of two reports that indicated an exhaustive search is not required to detect plagiarism. As Culwin and Child (2010) have illustrated, plugging exact phrases into a public search engine can be an effective way to locate the source of a suspicious paper, but the question of how to locate such an indicative phrase in a document remains. Butakov and Shcherbinin (2009) also indicated that if a significant part of the paper was plagiarized from the Internet, there was no need to send all possible queries to the search engine: Even as few as 10% of these can help to locate the source.

Like the typical architecture outlined earlier, the process starts with a student submission. The university portal prepares the document for the checkup, wrapping it with information about the course, assignment, type of required checkup, and so on. The internal part of the PDS checks the document against the local database and prepares queries for the external part. The distinctive feature of this proposed architecture is the way these queries are prepared. According to the legal requirements that protect IP, these queries should not contain enough information to recover the submission in its original form, and they should appear

significantly altered from the original document. Such a transformation would give the PDS protection from accusations of IP violation.

One of the more appealing ways to hide the content of the submitted document from the third party is to encrypt the queries and perform the search in an encrypted index (Song, Wagner, & Perrig, 2000; Goh, 2003). To maintain privacy during the search, these techniques require the search index to also be encrypted; therefore, it should be managed by the PDS. In many cases, private PDS outsource search index support. For example, Crot and SafeAssign employ Microsoft's Bing index to perform the selection of candidate documents. Such outsourcing makes it impossible for schools to use encryption mechanisms to hide a submission from the third party.

Experiments indicate that even limited numbers of properly selected search queries can help to locate plagiarism sources on the Web (Culwin & Child, 2010; Butakov & Shcherbinin, 2009). Essentially this means that the part of the PDS located on school infrastructure can prepare some queries from key parts of the text. These essential queries should be enough to locate the candidate sources. This technique is very similar to the approach used by many language professors when they see a grammatically perfect sentence written by a non-native student. They put the suspicious sentence in quotation marks and use conventional search engines to perform the search. There are a number of benefits to this approach. Besides offering a limited selection of results, these queries can be randomly shuffled as systems accumulate the results of search queries before selecting the ones with multiple appearances in the search results. Randomly shuffled, selected key phrases from a suspicious document can be considered a significant alternation from the initial submission. Such a query will not likely be subject to attack on IP infringement grounds because the complete student submission is not transmitted to the third party and cannot be completely restored from the information transferred for the search.

Compilation of the essential queries should consider the amount of information to be transferred to the third party. The sliding window algorithm that is implemented in the Crot PDS uses all possible queries that can be generated from the text. For example, for the Shakespearean quote "to be, or not to be: that is the question" with a window length $x = 4$, the algorithm will prepare seven queries: "to be or not," "be or not to," "or not to be," "not to be that," "to be that is," "be that is the," and "that is the question." Obviously if that text has y words, the total number of queries can be defined as $n = y - x + 1$. Since we know that y will be much larger than x , we can say the sliding window algorithm will form almost y queries. The initial student submission may be easily required from these queries because two neighboring queries, q_i and $q_i + 1$, have $x - 1$ common words. The total number of words that will be sent to the search engine will be about $y * x$. If we decide to select only y_i queries, and y_i satisfies the inequality (1), then we can guarantee it will be impossible to fully recover the initial student submission from the queries that will be sent to the external part of the PDS.

$$y_i < y/x \quad (1)$$

Of course, inequality (1) does not guarantee that parts of the document cannot be restored, but such a recovery may not be considered a violation of students' IP rights. The submitted

text has been significantly altered, and its recovery on the third-party side would be considered a deliberate attempt to gain access to the copyrighted material.

At the moment, related research has made no noticeable effort to protect student IP while performing the plagiarism detection process. The architecture we have proposed here is focused on this issue and leaves room for schools to be flexible in IP protection management. When it is necessary to establish a policy on how much of a student submission can go to the external PDS, the school may decide to accept the tradeoff between the granularity of search, that is, the probability of catching small-scale plagiarism (one sentence to a few paragraphs) versus sending copyrighted material to the third party. Inequality (1) can provide a boundary to help school officials make that decision.

The potential flaw of our proposed architecture is that theoretically it will not be able to catch a paper plagiarized from one submitted at a different school if this paper never appeared online and was not accessed by a conventional Internet search engine. The scale of this peer-to-peer copying can be assessed by finding out the actual percentage of such transfers between schools. Scanlon and Neumann (2002) indicated that the Internet is indeed the main source of plagiarized texts. Another study also indicated that about half of the surveyed students knew someone who had plagiarized from the Internet (Jones, Johnson-Yale, Millermaier, & Pérez, 2008). It also indicates that in many cases, the “deep” Web could be a source for plagiarized papers as the majority of students feel that academic papers on library databases are a reliable source of information. Even so, major search engines such as Bing and Google have access to subscription-based databases of research journals, and therefore outsourcing the search can help to locate these possible sources. Based on these two factors, we feel that the scale of peer-to-peer copying is not critical compared to the advantages provided by the proposed architecture for PDS. Moreover, local plagiarism within the school is taken care of by the database of local submissions outlined on Figure 3. Such a database is located on the university infrastructure and will not be subject to copyright claims from the students.

The following factors should be considered to assess the scalability of the proposed approach.

- The architecture increases the requirements of the available computational power and storage capacity of university infrastructure. Additional storage is required to keep digital fingerprints of the submitted documents. If typical fingerprinting algorithms (Schleimer, Wilkerson, & Aiken, 2003) or T9-like algorithms are used to compile fingerprints, then we can expect an additional 20 to 100% increase in required storage comparing to the space required for the plain texts. However, such an increase is insignificant, as plain text does not take up much space, and memory prices have followed a declining trend for years.
- Additional computational power is required to calculate a single document fingerprint, which is necessary for the fast comparison of documents. Since there are many algorithms available that are linear to the document size, such an increase can be also considered insignificant.

- Additional requirements will arise to perform one-to-many detailed comparisons in the internal PDS. These requirements will impose a high load on the database management system (DBMS), and therefore the cost of the DBMS licensing and maintenance on the university side will be the main factor that will affect the price of maintaining the proposed architecture. In most cases, universities already own and maintain some DBMS, and therefore licensing cost increases may not be significant. Hardware investments along with regular maintenance will have a major impact.
- The scalability of external search capacities could be an issue with the exponential growth of Internet content. But if the search is outsourced to one of the major search engines, the proposed architecture gets the whole power of this engine. Moreover, an OCLC report (2011) indicated that 90% of students begin their search for information using a search engine. Thus the probability that a particular source of plagiarism was indexed by a general purpose search engine is very high. If such a search engine is included in the architecture, the scalability and reliability of the PDS will be sufficient to detect copy-paste types of plagiarism.
- Increased requirements on the university infrastructure will ease heavy requirements for the DBMS on the third-party infrastructure. This factor can decrease the cost of third-party service and reduce the market entrance cost for start-ups, thus promoting competition among PDS.

The following section shows the results of our prototype of the four-component architecture for PDS: Learning Management System – Internal PDS – External PDS – Conventional Search Engine.

Experiment

The experiment was run on a set of papers that simulated different plagiarism levels from the Web. The main goal of the experiment was to check the applicability and practical implications of the proposed approach. The experiment was conducted to answer the following questions: (a) How does the amount of information used in the search affect the search quality? (b) What is the minimal amount of information required for guaranteed detection of different levels of plagiarism?

The experiment was conducted using a set of documents that were tailored to simulate different levels of plagiarism from Wikipedia pages. All 500 documents in the set were roughly the same length, 3,000 words. We used 50 different pages from Wikipedia as sources. The sources were grouped into the following categories: countries, objects from a specialized domain, social concepts and actions, and general objects. Wikipedia text was placed as one consecutive piece in a random place in the original document. Random placement of plagiarized passages was done to ensure the validity of simulated plagiarism since in real assignments, plagiarism can appear anywhere in the text. We used 10 different levels of plagiarism that took varying amounts (from 5 to 50%) of text from Wikipedia. In total, 500 documents were generated according to these standards. A 5% increase in the amount of

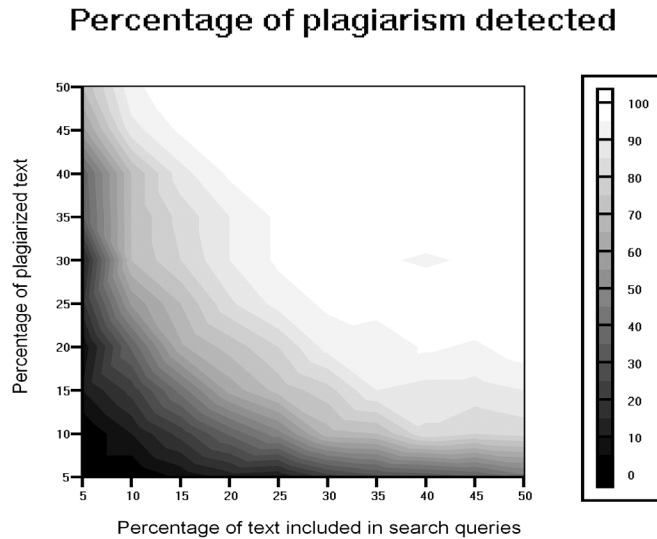
plagiarized text was used to maintain the balance between the granularity of the search results and the time required to run the experiment. The total amount of 500 documents and the coverage of four categories supported experiment result reliability.

The scan was scheduled to randomly select anywhere from 5 up to 50% of all possible phrases made up of six words from each document. For example, for a 5% selection of a 3,000-word document, the total number of queries sent to the search engine was 150, making the maximum possible value of y_i equal to 900 words. The selection of six words is based on two previous studies that have indicated a six-word running window is an effective search pattern for English language texts (Butakov & Shcherbinin, 2009; Culwin & Child, 2010). The Appendix contains tables showing the detailed information of plagiarism detection results for the different source categories mentioned above.

The minor inconsistency can be seen in the results when a higher percentage of search queries sent to the search engine resulted in actually lower detection level. This inconsistency could have been caused by network timeouts and the unavailability of certain resources on the Web at the time the experiment was conducted.

Figure 4 summarizes the results for all the documents. It displays the surface of the detection reliability on two axes: the percentage of text plagiarized and the percentage of queries sent to the search engine. Bright areas indicate better results. As can be seen from Figure 4, in the experiment's setting, reliable detection could be achieved for low y_i only if a significant portion of the suspicious text was plagiarized from the Web.

In other words, if only up to 15% of the queries are used from the suspicious document then this approach is applicable to the texts with plagiarism level of 45% and above. Or to rephrase, if more than 50% of the text is plagiarized then only 15% of the queries will be enough to reliably detect the plagiarism. Table 2 in the Appendix indicates that if the topic of the paper is very specific, then reliable detection can be achieved with 15% of the queries if about 25% of text is plagiarized. This can be explained by the fact that queries with more specific keywords draw better results from conventional search engines.



Note. Dark areas represent undetected plagiarism.

Figure 4. Summary of the plagiarism detection results.

The less reliable results for documents with lower levels of plagiarism could have been caused by the fact that these fragments were not covered by the random selection of the search queries. This means that better results for documents with less plagiarism can be achieved if, instead of randomly selecting queries, we can prescan a suspicious document and define segments that must be covered by the search queries. One possible way to do that is to implement an appropriate technique from the authorship detection and stylometry domain (Stamatatos, 2009). This issue will be addressed in subsequent research on the proposed architecture.

Conclusion

In this study, we concentrated on architecture for a PDS. The proposed solution contributes to a number of aspects of service architecture development. First of all, this novel architecture makes student copyright protection a main goal and guarantees that no third party directly or indirectly makes profit from student work. Limited and scrambled portions of student work that departs from the school's IT infrastructure cannot be used to fully recover the material.

A second distinctive feature is the decision to outsource the most time-consuming part of the plagiarism checkup to a third party, thereby reducing the workload on the university IT infrastructure. Such outsourcing removes the necessity for the PDS in each school to have its own private web crawler and allows different schools to rely on a common search engine for PDS. Such major search engines improve the probability of plagiarism detection because they have very high indexing capacities.

In future research projects, we are planning to work on improving the details of the proposed architecture. One possible direction to take might be to include stylometry in the

external part of the PDS to do preliminary checkups. This feature could lead to better scalability of the service, allowing the external part of the PDS to download suspicious sources of plagiarism with a higher probability rate and filter them before submitting the results to the internal part of the PDS.

References

- Abbasi, A., & Chen H. (2008). Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems*, 26(2). doi:10.1145/1344411.1344413. 7:2 – 7:29
- Arnold, M., Schmucker, M., & Wolthusen, S. D. (2003). *Techniques and applications of digital watermarking and content protection*. Norwood, MA: Artech House.
- Bennett, M. G. (2009). A nexus of law & technology: Analysis and postsecondary complications of A.V. et al. v. iParadigms. *LLC Journal of Student Conduct Administration*, 2(1), 40–45.
- Burrows, S., Tahaghoghi, S. M., & Zobel, J. (2007). Efficient plagiarism detection for large code repositories. *Software: Practice and Experience*, 37(2), 151–175. doi: 10.1002/spe.750
- Butakov, S., & Barber, C. (2012). Protecting student intellectual property in plagiarism detection process. *British Journal of Educational Technology*, 3(4), E101–E103. doi:10.1111/j.1467-8535.2012.01290.x
- Butakov, S., & Shcherbinin, V. (2009). On the number of search queries required for Internet plagiarism detection. In *Proceedings of the 2009 Ninth IEEE International Conference on Advanced Learning Technologies* (pp. 482–483). Washington, DC: ICALT, IEEE Computer Society.
- Cox, I., Miller, M., Bloom, J., & Friedrich, J. (2007). *Digital watermarking and steganography*. San Mateo, CA: Morgan Kaufmann.
- Culwin, F., & Child, M. (2010). Optimizing and automating the choice of search strings when investigating possible plagiarism. In *Proceedings of 4th International Plagiarism Conference, Newcastle*. Retrieved from <http://www.plagiarismadvice.org/>
- Donaldson, J. L., Lancaster, A.-M., & Sposato, P. H. (1981). A plagiarism detection system. In *Proceedings of the Twelfth SIGCSE Technical Symposium on Computer Science Education* (pp. 21–25). New York, NY: ACM. doi: 10.1145/800037.800955
- Goh, E. J. (2003). *Secure indexes* (Report 2003/216). Retrieved from Cryptology ePrint Archive website: <http://eprint.iacr.org/2003/216/>
- Heberling, M. (2002). Maintaining academic integrity in online education. *Online Journal of Distance Learning Administration*, 5(1). Retrieved from <http://www.westga.edu/~distance/ojdla/spring51/heberling51.html>
- Jensen, L. A., Arnett, J. J., Feldman, S. S., & Cauffman, E. (2002). It's wrong, but everybody does it: Academic dishonesty among high school students. *Contemporary Educational Psychology*, 27(2), 209–228.

- Jocoy, C., & DiBiase, D. (2006). Plagiarism by adult learners online: A case study in detection and remediation. *The International Review of Research in Open and Distance Learning*, 7(1). Retrieved from <http://www.irrodl.org/index.php/irrodl/article/view/242/466>
- Jones, S., Johnson-Yale, C., Millermaier, S., & Pérez, F.S. (2008). Academic work, the Internet and US college students. *The Internet and Higher Education*, 11(3-4), 165-177.
- Krsul, I., & Spafford, E. H. (1997). Authorship analysis: Identifying the author of a program. *Computers & Security*, 16(3), 233-257.
- Marais, E., Minnaar, U., & Argles D. (2006). Plagiarism in e-Learning systems: Identifying and solving the problem for practical assignments. In *Proceedings of the Sixth IEEE International Conference on Advanced Learning Technologies (ICALT '06)* (pp. 822-824). Washington, DC: IEEE Computer Society.
- OCLC (2011). *Perceptions of libraries and information resources: A report to the OCLC membership*. Dublin, OH: OCLC. Retrieved from <http://www.oclc.org/reports/2005perceptions.htm>
- Scanlon, P. M., & Neumann, D. R. (2002). Internet plagiarism among college students. *Journal of College Student Development*, 43(3), 374-385.
- Schleimer, S., Wilkerson, D., & Aiken, A. (2003). Winnowing: Local algorithms for document fingerprinting. In *Proceedings of the ACM SIGMOD International Conference on Management of Data* (pp. 76-85). New York, NY: ACM.
- Song, D. X., Wagner, D., & Perrig, A. (2000). Practical techniques for searches on encrypted data. In *Proceedings of the 2000 IEEE Symposium on Security and Privacy* (pp. 44-55). doi: 10.1109/SECPRI.2000.848445
- Spafford, E.H. (2011). Security, technology, publishing, and ethics (part II). *Computers & Security*, 30(1), 2-3.
- Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science*, 60, 538-556. doi: 10.1002/asi.21001
- Underwood, J., & Szabo, A. (2003). Academic offences and e-learning: Individual propensities in cheating. *British Journal of Educational Technology*, 34(4), 467-477.
- Urvoy, T., Chauveau, E., Filoche, P., & Thomas, L. (2008). Tracking web spam with HTML style similarities. *ACM Transactions on the Web*, 2(1), 3:2-3:28 doi: 10.1145/1326561.1326564
- VAIL (Virtual Academic Integrity Laboratory) (2012). Detection tools and methods: Stu-

dent copyright and detection services [Web page]. Retrieved from http://www.umuc.edu/cip/vail/faculty/detection_tools/services.html

Zaka, B. (2009). Empowering plagiarism detection with a web services enabled collaborative network. *Journal of Information Science and Engineering*, 25(5), 1391–1403.

Appendix: Plagiarism Detection Results

Table 1

Plagiarism Detection Results on all Simulated Documents

		Percentage of queries sent to the search engine									
		5	10	15	20	25	30	35	40	45	50
Percentage of plagiarized text	5	0	0	0	6	11	17	22	11	28	17
	10	0	6	11	28	33	56	50	72	67	72
	15	6	17	17	39	56	61	83	78	78	78
	20	6	22	50	61	72	89	89	94	94	94
	25	0	33	56	78	94	94	100	100	100	100
	30	6	61	61	78	94	100	94	89	94	94
	35	22	56	72	78	94	100	100	100	100	100
	40	39	67	78	94	94	100	94	100	100	100
	45	33	72	89	94	94	94	100	94	100	100
	50	50	83	94	100	100	100	100	100	100	100

Note. 500 generated assignments taken from 50 source documents.

Table 2

Plagiarism Detection Results for the Objects from a Specialized Domain Category

		Percentage of queries sent to the search engine									
		5	10	15	20	25	30	35	40	45	50
Percentage of plagiarized text	5	0	0	0	25	0	25	50	50	25	50
	10	0	0	25	75	75	100	100	100	100	100
	15	25	50	50	100	100	100	100	100	100	100
	20	25	25	75	100	100	100	100	100	100	100
	25	0	100	75	100	100	100	100	100	100	100
	30	0	50	100	100	100	100	100	100	100	100
	35	25	75	100	100	100	100	100	100	100	100
	40	0	100	75	100	100	100	100	100	100	100
	45	75	100	100	100	100	100	100	100	100	100
	50	100	100	100	100	100	100	100	100	100	100

Note. 40 generated assignments taken from 4 source documents.

Table 3

Plagiarism Detection Results in the Social Concepts and Actions Category

		Percentage of queries sent to the search engine									
		5	10	15	20	25	30	35	40	45	50
Percentage of plagiarized text	5	0	0	10	14	10	19	24	43	43	62
	10	0	10	29	29	48	67	81	90	81	86
	15	14	19	52	67	76	81	90	90	90	95
	20	0	48	67	67	81	90	86	95	90	100
	25	29	57	71	81	86	95	95	100	100	100
	30	19	71	86	95	95	90	95	95	95	95
	35	38	76	86	95	95	100	100	100	95	100
	40	29	57	95	95	95	100	100	100	100	100
	45	67	95	100	100	100	100	100	100	100	100
	50	76	100	100	100	100	100	100	100	100	100

Note. 210 generated assignments taken from 21 source documents.

Table 4

Plagiarism Detection Results in the General Objects Category

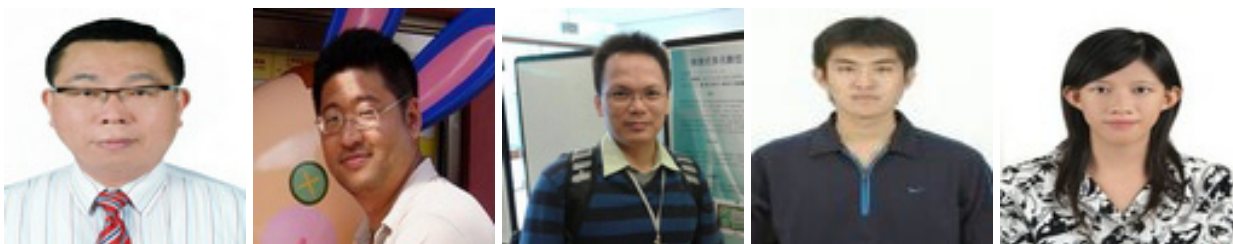
		Percentage of queries sent to the search engine									
		5	10	15	20	25	30	35	40	45	50
Percentage of plagiarized text	5	0	0	14	14	14	43	43	29	29	43
	10	0	29	29	57	71	86	71	86	100	86
	15	0	29	71	86	86	100	100	100	100	100
	20	14	43	86	100	86	100	100	100	100	100
	25	43	86	100	100	100	100	100	100	100	100
	30	14	100	100	100	100	100	100	100	100	100
	35	57	86	86	100	100	100	100	100	100	100
	40	71	100	100	100	100	100	100	100	100	100
	45	86	100	100	100	100	100	100	100	100	100
	50	86	100	100	100	100	100	100	100	100	100

Note. 70 generated assignments taken from 7 source documents.

Athabasca University 



A Rapid Auto-Indexing Technology for Designing Readable E-Learning Content



Pao-Ta Yu, Yuan-Hsun Liao, Ming-Hsiang Su, Po-Jen Cheng, and Chun-Hsuan Pai
National Chung Cheng University, Taiwan, Province of China

Abstract

A rapid scene indexing method is proposed to improve retrieval performance for students accessing instructional videos. This indexing method is applied to anchor suitable indices to the instructional video so that students can obtain several small lesson units to gain learning mastery. The method also regulates online course progress. These anchored points not only provide students with fast access to specific material but also can link to certain quizzes or problems to show the interactive e-learning content that course developers deposited in the learning management system, which enhances the learning process. This allows students to click on the anchored point to repeat their lesson, or work through the quizzes or problems until they reach formative assessment. Hence, their learning can be guided by the formative assessment results.

In order to quickly find the scene to index, some specific description of it was needed. Actually, most of the instructional videos were recorded by teachers and were part of their PowerPoint presentations. Based on the features of the PowerPoint slides in the instructional videos, such as the title or page number, the specified scene can be found. Since we used specific scene descriptions, it was easy to employ the rapid scene detection method using an image filter and Sobel mask. Finally, we applied an experimental design to check the precision of the scene detection and evaluate user satisfaction. The results showed that rapid scene indexing can definitely assist learners in their online learning; that is, it gives them better learning mastery and provides regulation for the online learning environment.

Keywords: Scene detection; scene indexing; instructional video; anchored or access point; mastery learning; regulated learning; e-learning; PowerPoint

Introduction

As a result of the rapid advance of asynchronous learning, there are not enough interactive learning objects available to meet the numerous demands of learners (Girasoli & Hannafin, 2008). The fastest and easiest way to provide an adequate amount of e-learning content is to record teachers' presentations in a classroom or studio and then directly put those recordings into a learning management system (LMS) (Mittal, Pagalthivarthi, & Altman, 2006). Using a high speed Internet connection and streaming technology, learners can repeatedly view the recorded instructional videos from anywhere, at any time (Blanchette, 2012). However, this kind of streaming data lacks flexibility and interactive capability. Therefore, a user-friendly interface is required to let students easily capture any segment of the recorded instructional videos (Liu & Kender, 2004).

Most users who spend a long period of time on a course Web site often lose their concentration and start surfing elsewhere. A regulating mechanism is required to pull students' attention back to their online learning. However, a mechanism to improve the interactivity of the recorded instructional videos is also needed. This motivated us to apply the technique of image indexing to the recorded instructional videos in order to divide them into segments with several entry points. Those entry points and the technique of image retrieval easily provide interactive features for the lessons (Zhang, Wang, Shi, & Zhang, 2010).

Most of the streaming data can run on a web browser in its corresponding embedded media player (Bouvin & Schade, 1999). This embedded player can interpret a given XML or HTML document and then follow the instructions to display the correct video content (Sun, Kim, & Kuo, 2005). In this experiment, we gave the media player a series of instructions about time segments, where each segment corresponded to a short clip of the recorded instructional video (Feng, Lu, & Ma, 2002). As we knew, mastery learning requires teachers to partition content into several small logically sequenced units. Then, teachers guide students on the learning path through one of these small units until students master that particular content (Block & Burns, 1976). Therefore, we also needed a pattern recognition method to determine the logical viewing order combined with the image indexing technique to find the access point of each unit of the recorded instructional video (Antani, Kasturi, & Jain, 2002). We will give a detailed description of this process in the Method section of this paper. After we found a set of access points, we assigned this information to an XML document using appropriate tags (Keyvnpour & Izadpanah, 2011). The media player can then follow the instructions of this XML document to provide some interactive options, including giving a student access to online supervision.

During the self-learning component of a distance education lecture, most learners cannot understand the complicated concepts or formulas in a short period by simply reading the e-learning content (Chiu, Chow, & Chang, 2007). This inability to grasp the lesson causes their learning progress to fall far behind the lecture schedule. Although the asynchronous e-learning environment permits learners to arbitrarily control the course process so they can repeat their lessons as needed, learners cannot successfully discipline themselves to follow the learning schedule just by reading monotonous content (Bergamin, Werlen, Siegenthal-

er, & Ziska, 2012). Therefore, we designed a mechanism of regular testing which requires learners to answer questions corresponding to pop-up information triggered when they click on an access point found by the indexing mechanism. If learners select the wrong answer, they have to play the video from the beginning access point to repeat the lesson until they reach the expected learning level. The behavior of repeat learning is also described in the XML document, and the video can be based on the document to play at a specific time.

This paper presents an automatic method for detecting the changes of the PowerPoint feature and embedding this instructional video in an online course as an interactive component. This is a simple and useful way to tailor the learning context and help learners to enhance meaningful learning. The rest of this paper is organized as follows: First, we present a literature review and the auto-indexing technology. Next, we discuss its application and the experiment's results. Finally, we present our concluding remarks.

Literature Review

Mastery Learning

Mastery learning is an effective way to make learners reach higher learning levels under the appropriate conditions (Bloom, 1982). Its proponents argue that all students can reach high levels of mastery of instructional material with the right support. Bloom (1968) stated that well organized teaching materials and effective management of a student's learning process are two factors that help individuals achieve successful mastery. According to Bloom (1976) and others (Block & Burns, 1976; Fuchs, Fuchs, & Tindal, 1986), mastery learning can be accomplished by following specific procedures. The first step is to divide the concepts and materials into relatively small and sequential learning units. Each unit is associated with concrete learning objectives, and the structure is organized by partitioning difficult content into several smaller units that are easier to grasp. After teaching each unit, the instructor conducts a formative assessment to determine whether the learners have reached the desired level or not; the assessment also provides students with feedback on their learning (Yang & Liu, 2006). The learners who have not mastered the unit begin a process of remedial activities or corrections to assist them in achieving this goal. This learning process is shown in Figure 1. Mastery learning is a suitable approach to use with students due to their weak discipline in self-directed learning settings.

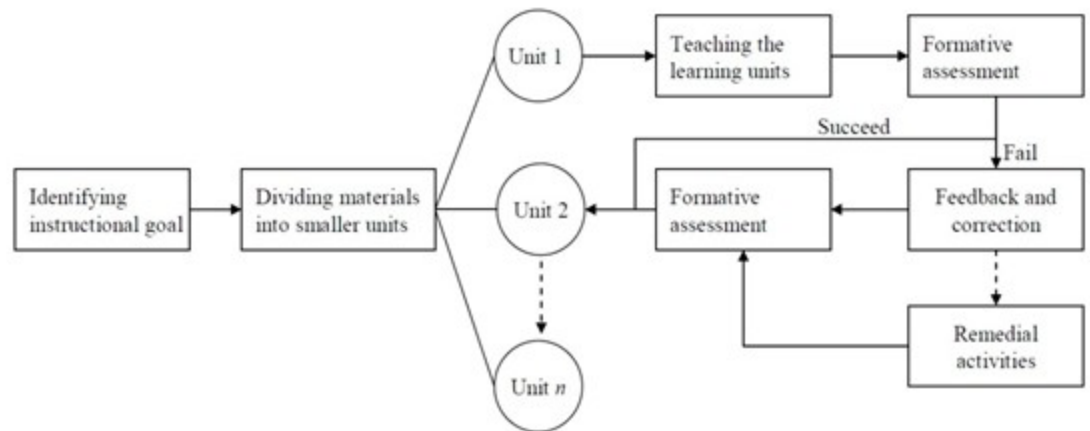


Figure 1. The strategy for mastery learning.

Detection Technology

Multimedia indexing and retrieval has become a challenge in light of the huge amount of data that must be organized. This is not a trivial task for large visual databases. Hence, dividing videos into brief segments might improve the likelihood of completing this task (Sakarya, Telatar, & Alatan, 2012). Enabling users to search all digital multimedia data in a library and access the relevant information requires efficient analysis, indexing, and re-organizing for suitable browsing (Naphade, Mehrotra, Ferman, Warnick, Huang, & Tekalp, 1998; Koprinska & Carrato, 2001; Gargi, Kasturi, & Strayer, 2000; Song & Ra, 2001; Hanjalic, 2002; Yuan, Wang, Xiao, Zheng, Li, Lin, & Zhang, 2007; Asan & Alatan, 2009).

With the development of the low-cost digital recorders, it is easy to obtain various kinds of recorded data, such as traveling or instructional videos. Since the instructional videos that we deal with usually include different lesson segments, the video files must be pre-processed first (Ahanger & Little, 1996). In this study, the proposed approach addresses the problem of instructional video segmentation by using scene detection. Each frame in the instructional video is a snapshot of a PowerPoint slide. The feature area selector is designed to choose features such as the title or page number of the slide. In this way, the automatic detection technology only examines the feature area of the image. The technology performs better compared to other methods; indeed, it is faster and simpler.

This detection technology is based on the idea of finding the dominant scene in the video according to the selected feature. The approach employs the Sobel mask operator, which is commonly used to find the edge of grayscale two-dimensional images. This has the ability to find the gradient of neighboring points. If we construct a line“image”using the z-coordinates of the scanned points as the gray scale, then a Sobel mask may detect any sharp changes in the gray values (Gonzalez & Woods, 1992).

Method

Auto-Indexing Technology

Our proposed method provides an automatic scene detection process to find the scenes of an instructional video and provide the corresponding access points. After all the data to be streamed is fixed with suitable anchor points, the rapid indexing technology can translate it to e-learning content in HTML format using those anchored points, as shown in Figure 2. For convenience, we call this process an RIT system.

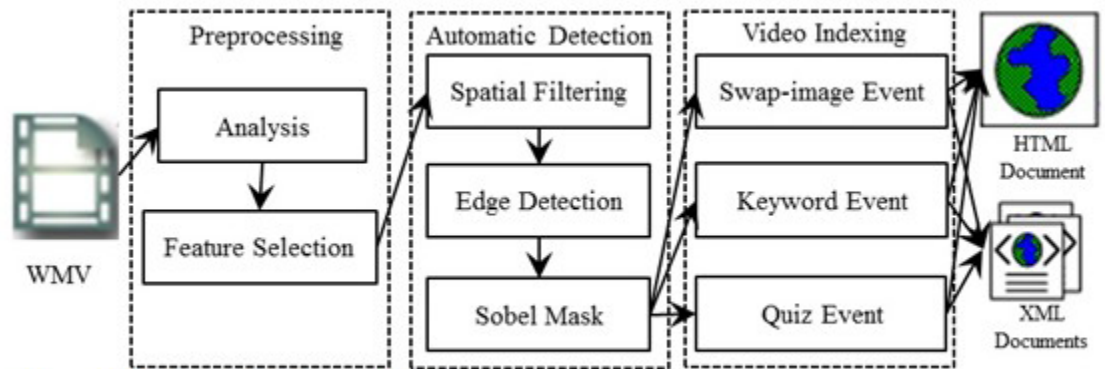


Figure 2. The auto-indexing method.

An illustrated example is shown in Figure 3. Based on snapshots obtained using automatic detection, the lecture is partitioned into four scenes and the access point of each scene can be easily found in the streaming data.

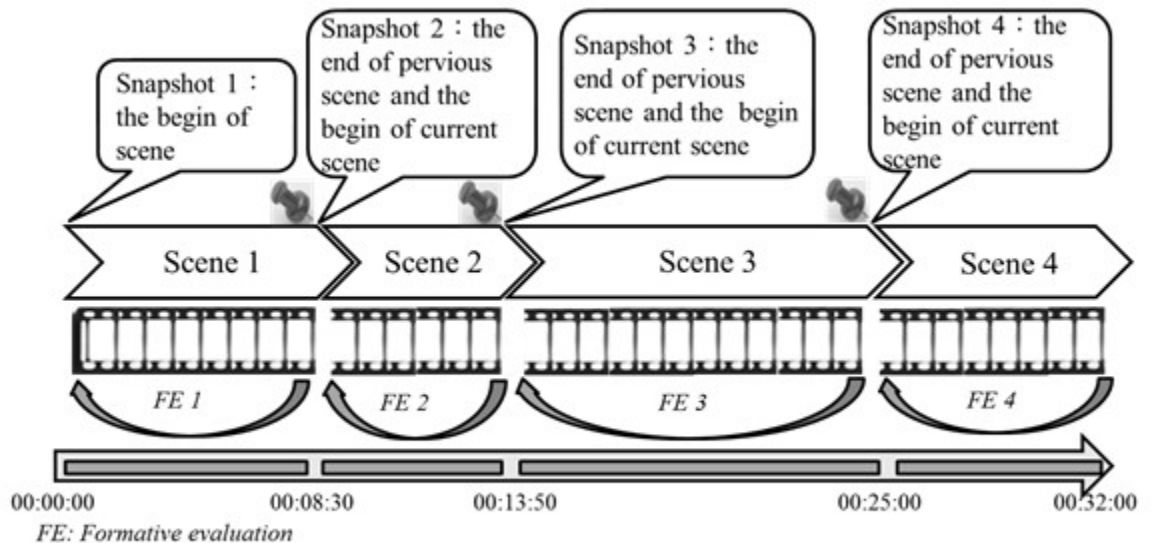


Figure 3. Diagram of slides and their divisions in the timeline.

As shown in Figure 4, we applied spatial filtering, edge detection, and a Sobel mask in our method to process the images from the instructional video, using a binary searching algorithm to find the anchored points. The basic rule is that if the comparison result of the first image and the last image in an interval is the same, the process will determine that the interval has no target image. Otherwise, the target image exists in the interval. This function is designed to provide a user with the feature selection area and can automatically detect the proper feature area.

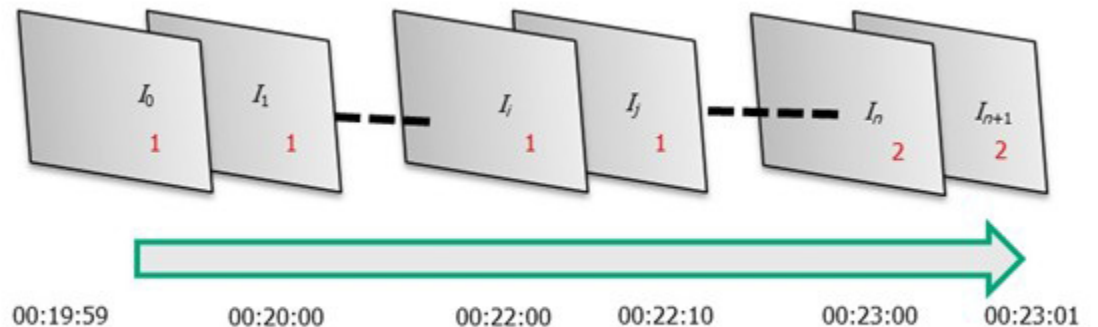


Figure 4. Automatic scene detection approach.

Video Indexing

Early video indexing methods were based on detecting shot boundaries and extracting key frames from which visual features like color, texture, shape, or edge were extracted to be used as indices (Smoliar & Zhang, 1994; Deng, Manjunath, Kenney, Moore, & Shin, 2001; Manjunath & Ma, 1996; Park, Jeon, & Won, 2000).

Whereas initial research in this area was directed at image databases, it did not take long for researchers to start using these methods to address the similar issues of retrieval and classification for video data. These tasks are guided by suitable indexing methods based on the content of the video itself or semantic descriptors that could be extracted from the data (Yi, Rajan, & Chia, 2005). Our video indexing technology created a script XML file within Windows Media Video (WMV) (Liao, Tsai, Su, Li, & Yu, 2011). The event information is defined by the time the event occurred and is embedded in the WMV files. Three types of event information are defined: the swap-image event, the keyword event, and the quiz event. The swap-image event consists of a shot image and access point for each slide. The shot image is the final snapshot of each slide, and the access point of each slide is its ending time. Included in the keyword event is both text content and the access point. The text content is made up of the keyword information, and the access point for the time of each keyword is used. The quiz event is made up of text content and an access point; the former consists of the questions and answers and the latter the time of the question period.

When the user moves forward through the lesson, the system will automatically initiate a swap-image event to change the current slide to the previous page, as shown in Figure 8. Then the keyword event will be triggered to display pop-up information, which always

provides keywords (or hints) and is edited by teachers. In the instructional video, the quiz event will be triggered to display questions in order to examine the learner's formative status.

In this study, the RIT system used XML tags to define the information elements. The information elements contain the type of event, the command of execution, the time the event occurred, and the playback function. The instructional video was played in accordance with the information elements, as shown in Figure 5.

```

- <Topic>
  <TopicTitle>基本理論</TopicTitle>
  <Keyword/>
  <TopicTime>00:00:50</TopicTime>
  <TopicContent>基本理論 2 高學會 什麼是認知理論及其操作流程? 什麼是多媒體學習理論? 那一些認知原則? 說說看, 你會學到什麼? </TopicContent>
  <TopicMemo/>
  <Thum> Thum000094000.jpg</Thum>
  <TopicThum> 2</TopicThum>
  <TopicPage> 2</TopicPage>
</Topic>
- <Topic>
  <TopicTitle/>
  <Keyword/>
  <TopicTime>00:01:34</TopicTime>
  <TopicContent> 3 認知理論(Cognitive Theory) 雙軌或雙頻道(Dual channels)是人類慣用的接收模式, 主要有 Auditory/Verbal channel
  Visual/Pictorial channel 有限的記憶能力 Working memory 主動的處理能力 Selecting Organizing integrating</TopicContent>
  <TopicMemo/>
  <Thum> Thum000376000.jpg</Thum>
  <TopicThum> 3</TopicThum>
  <TopicPage> 3</TopicPage>
</Topic>
- <Topic>
  <TopicTitle/>
  <Keyword/>
  <TopicTime>00:06:16</TopicTime>
  <TopicContent> 4 認知處理流程理論(Cognitive Processing Flow) </TopicContent>
  <TopicMemo/>
  <Thum> Thum000709000.jpg</Thum>
  <TopicThum> 4</TopicThum>
  <TopicPage> 4</TopicPage>
</Topic>

```

The image shows the XML content with annotations. A red box labeled "Information Element" points to the time value "00:01:34" in the second topic's XML. Another red box labeled "Tag" points to the closing tag "</TopicContent>" in the third topic's XML.

Figure 5. The content of an XML file.

Detection Process

Teachers select a video as a source and use the RIT system to retrieve the images of the slides from the video, as shown in Figure 6. The feature areas determined by the feature area selector serve as the parameters for retrieving images. After the images are retrieved with the RIT system, anchored points are created in the video. In addition, the system provides teachers with the ability to arbitrarily retrieve images from the video.



Figure 6. The interface of the detection process.

Embedding Interactive Information

The RIT system analyzes the access points at which it retrieved the slides and records the related information in a temporary XML document, which stores three types of embedded information: image, quiz, and keyword. Teachers can add to the content of a learning object or edit it by filling in the interface's fields; the RIT system not only saves this new content in XML file format but also records the embedded information. Finally, after teachers click the button to embed related information into a video, the system generates a postprocessed WMV file and a few XML documents. The system's outputs are integrated into a learning object by the predesigned HTML document.

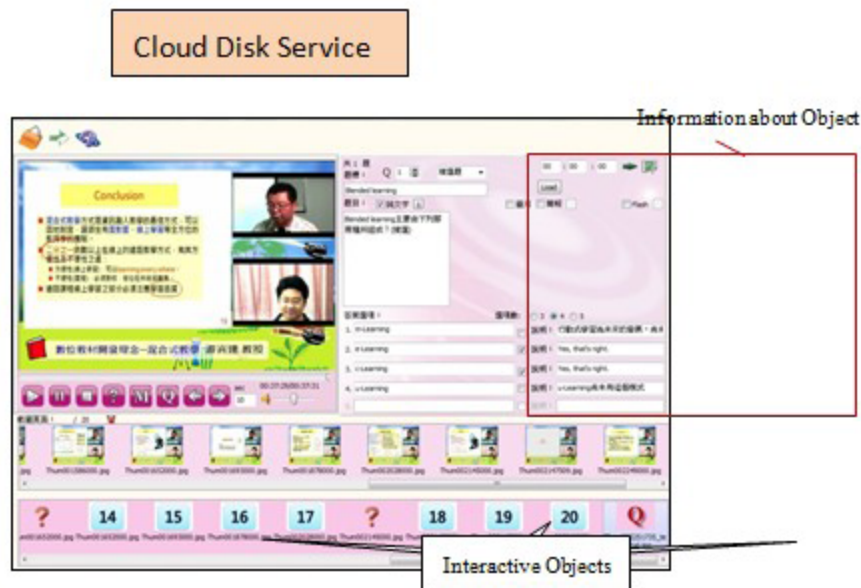


Figure 7. The interface of embedded interactivity information.

Learner Interface

Through the predesigned XML document, teachers can integrate the postprocessed video and several XML documents into a learning object. The learner's interface for e-learning courses is shown in Figure 8.



Figure 8. Learner interface of an e-learning course.

The instructional video component is the postprocessed video generated by the teacher, and the interactive index area shows the complete course content. The previous page component shows the previous slide or video image to assist students by reminding them of the content that came before. When the instructional video component sends the image message, the previous page component changes to the specified image. When the instructional video component sends the quiz message, the interactive object component will open a new window and the playback of the instructional video component will be suspended. If students answer the question correctly, the video will continue to play; otherwise, the media player will replay the related video clips. Figure 8 shows the quiz window, one of the interactive object components.

Learning Activity

When students' learning activities are ongoing, they can easily focus on the video content rather than skip it (as shown in Figure 9). When the quiz window appears, the video will be suspended. Students need to answer questions correctly in order to continue the learning activity. If they select a wrong answer, the video will be played from the start of the anchor point. The quiz window not only gives students feedback but also lets them review the content in stages. The keyword information was designed to allow teachers to place emphasis on the keywords of relevant hints or concepts. When the keyword window is open, students can continue to watch the video without interruption. The third interactive element shows the change of slide images. Students can refer to the previous slide to organize the discrete concepts.



Figure 9. The timeline of the learning activity.

As we have already mentioned, students spend their energy to concentrate on learning. Employing animation and interactivity principles (Mayer, 2005), we have provided the RIT system to instruct students using an interactive learning object. The interactive learning object is composed of an HTML document, a video, and several XML documents. Teachers can embed instructions into the video. When students use the learning objects, the video sends instructions according to the predesigned course and manner of progress. If students can continue to learn without frustration or distraction, they will be more likely to concentrate on their learning activities.

The Experiment

Detection Analysis

There are many scenes in the instructional videos, and each one represents a different unit of the course, arranged in logical order. In order to distinguish between different scenes, teachers can select a specified region of snapshot images from the instructional video to retrieve the desired features. The features retrieved from consecutive snapshot images in the same scene possess the same features. The detailed specifications are shown in Figure 10.



Figure 10. Different selection regions in an instructional video.

In order to get better detection results, we defined several selected regions for this study. Then, we tried the experiment in those regions to measure the accuracy of the retrieved features. We considered different sizes of the selected region: custom pixels, 80 x 60 pixels, 160 x 120, 320 x 240, 640 x 480, and full screen. The custom pixel size is appropriate for covering the feature region and is dynamic, in contrast to the fixed cases. We analyze the advantages of dynamic and fixed sizes under different considerations in the following section.

In the experiment, the performance was evaluated objectively by measures of precision, recall, f-measure, and misdetection rates (Shivakumara, Phan, & Tan, 2010). The scenes were classified into the following categories by our detection method.

- Truly detected key frame (TDK): The detected frames contain the corresponding key frame from instructional video. For a page from the course slide, a key frame is the frame in which the page content first appeared in the video.
- Falsely detected key frame (FDK): The detected frames contain the redundant frames besides the key frame.
- Key frame with missing data (MDK): The detected frames lose the key frames (MDK is a subset of TDK).

For each key frame in the instructional video, we also manually counted the number of Ac-

tual key frames (AKS), that is, the true key frames. The performance measures were defined as follows:

- $\text{Recall (R)} = \text{TDK}/\text{AKS}$
- $\text{Precision (P)} = \text{TDK}/(\text{TDK} + \text{FDK})$
- $\text{F-measure (F)} = 2 \times P \times R / (P + R)$
- $\text{Misdetecion Rate (MDK)} = \text{MDK}/\text{TDK}$

The experimental samples were obtained from four lectures: Digital Learning Design: Multimedia Learning Theory, Blended Learning, Mastery Learning Theory, and Design Experiences, respectively.

If the precision of the key frame detection was too low, it meant that a lot of unnecessary frames were recognized and search efficiency was poor. Moreover, if the recall rate was too low, it meant that many key frames were not detected.

Based on the difference of feature region, the precision, recall, and f-measure analysis were compared in Table 1. The precision in the custom feature region was higher than the other feature regions, with a recall rate of more than 94%.

The precision was related to the feature region. The precision and recall calculated by the f-measure proves that the custom region was better than all the other sizes. The experiment's results showed that the feature region was most effective, with better detection results, as shown in Table 1.

Table 1

Feature Region Results for Lectures

Feature size(pixels)	Lecture 1				Lecture 2				Lecture 3				Lecture 4			
	R	P	F	M	R	P	F	M	R	P	F	M	R	P	F	M
80 x 60	0.94	0.65	0.77	0.06	0.95	0.66	0.78	0.05	0.89	0.70	0.78	0.13	1.00	0.40	0.57	0.00
160 x 120	0.94	0.35	0.52	0.06	0.90	0.24	0.38	0.11	0.89	0.40	0.55	0.13	0.94	0.18	0.31	0.07
320 x 240	1.00	0.21	0.35	0.00	0.90	0.11	0.19	0.11	0.89	0.24	0.38	0.13	1.00	0.15	0.27	0.00
640 x 480	1.00	0.17	0.30	0.00	0.95	0.07	0.12	0.05	1.00	0.16	0.27	0.00	0.94	0.10	0.18	0.07
Full screen	1.00	0.06	0.11	0.00	1.00	0.06	0.11	0.00	1.00	0.05	0.10	0.00	1.00	0.07	0.13	0.00
Proposed	0.94	0.94	0.94	0.06	0.95	1.00	0.97	0.05	0.94	1.00	0.97	0.06	1.00	0.80	0.89	0.00

Note: R = Recall (Detection Rate); P = Precision (1 False Positive Rate); F = F-measure; M = Misdetecion Rate

Student Satisfaction with the System

Participants

We conducted the experiment to investigate how satisfied students were with the RIT system in the online learning course. In total, 46 college students in central Taiwan (20 men and 26 women, with a mean age of 33.1 years) who enrolled in the Digital Learning Design course participated in the experiment. The experiment ran for 12 weeks. Students watched one or two instructional videos online each week.

Instruments

Although information technologies can theoretically be useful to support the learning process, a careful evaluation of their benefits is needed (Yengin, Karahoca, & Karahoca, 2011; Wu, Tennyson, & Hsia, 2010). In this experiment, we evaluated the RIT system from the functionality and usability perspective. We created a questionnaire based on a modified and edited version of the questionnaire DeLone and McLean used in their study (1992, 2003) and the teachers' own experiences (Hwang, Tsai, Tsai, & Tseng, 2008). We collected data from the 18-item questionnaire that the students completed at the end of the experiment. The questionnaire includes questions about system quality, information quality, service quality, usage, user satisfaction, and the net benefits of the RIT system. Each item was assessed on a 5-point Likert scale, where 1 was *strongly disagree* and 5 was *strongly agree*. The Cronbach's alpha reliability coefficient for the questionnaire was 0.89, with a criterion-related validity of 0.78.

Procedure

One teacher and 46 students in a single class participated in a trial of the system. The class agreed to use the RIT system to support their lectures in the winter 2011 semester. The reading materials were uploaded to the learning management system (LMS). Students were able to watch the audiovisual materials produced by the RIT system. The procedure of the experiment, as shown in Figure 11, was as follows.

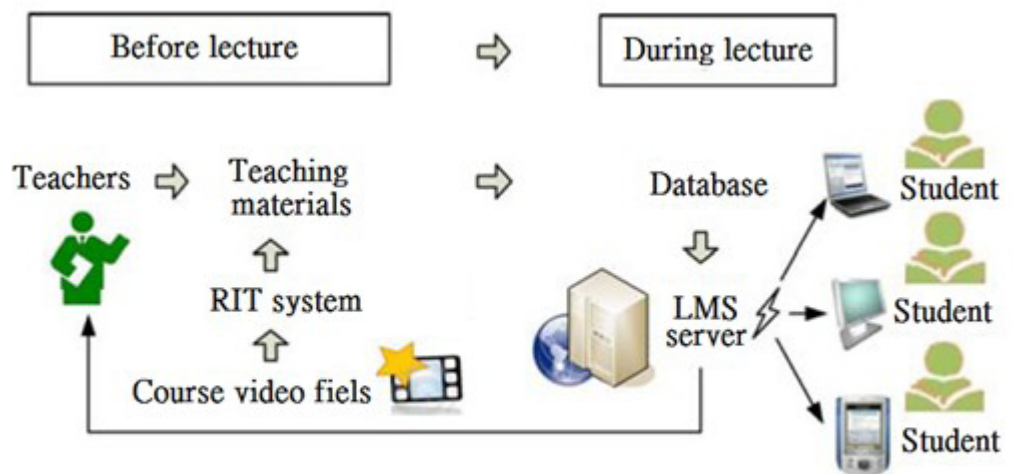


Figure 11. The RIT experiment procedure.

1. Before the lecture: Teachers prepare and produce the instructional videos, turning them into the teaching materials by using the RIT system. While preparing the course materials, teachers can subdivide lessons into multiple concepts according to the difficulty of the material. Then they can post questions for different concepts to test whether students understand them. Finally, teachers upload the materials to the LMS.
2. During the lecture: Students log in to the LMS system to watch the online materials. Once online, they can learn one concept step by step or review other lessons they have learned. In addition, students can discover which concepts they still don't understand by working through tests. Teachers can always amend course materials based on the situation of the students who are using them.
3. After the class: Students were asked to fill out the questionnaire.

User Satisfaction

Table 2 shows that the RIT is able to effectively construct a learning environment. Most participants (95.6%) found the application to be satisfactory; the mean satisfaction score was 3.81. Of the students, 4.3% were dissatisfied with this regulated learning, possibly because they didn't like the test conditions that determined their learning status. Most students, 73.9%, were satisfied that the system's interface was user friendly. Some of them suggested the quiz window should be locked onscreen to ensure that students will answer the questions.

Of the students, 71.7% were satisfied that the application could be used and accessed with ease. Furthermore, 69.6% were satisfied that their learning progress was integrated into the learning management system so that they could search for scores through one single system. The majority of the students, 76.1%, were satisfied that the application could be operated rapidly with the interface, and 84.8% of them were satisfied with the quality of the quiz questions, which made the content clearer. Of the students, 80.5% were satisfied with the services provided by the application, such as FAQ and security. Most, 69.6%, said they liked to use the application, and 78.3% stated they would often use it in the future.

Moreover, 82.6% stated that they found the application helpful and benefited from the learning process. Respondents who chose 4 or 5 (satisfied or very satisfied) said that they used the learning application to help them practice and remember important content, not only to watch the video. Thus we can conclude that the RIT system was satisfying for most students.

Table 2

Internal Consistency Results

	Construct	Non-existent	Poor	Fair	Good	Excellent	Alpha
System quality	Ease of use	0(0%)	2(4.3%)	11(23.9%)	29(63%)	4(8.7%)	0.871
	Integration	0(0%)	0(0%)	14(30.4%)	28(60.9%)	4(8.7%)	0.875
	Ease of access	0(0%)	1(2.1%)	10(21.7%)	29(63%)	6(13%)	0.875
	Interface quality	0(0%)	1(2.1%)	11(23.9%)	30(65.2%)	4(8.7%)	0.868
	Rapidity	0(0%)	1(2.1%)	10(21.7%)	31(67.4%)	4(8.7%)	0.867
Information quality	Information quality	0(0%)	0(0%)	7(15.2%)	31(67.4%)	8(17.4%)	0.868
Service quality	Service quality	0(0%)	0(0%)	9(19.6%)	32(69.6%)	5(10.9%)	0.868
Usage	Intensity	0(0%)	2(4.3%)	12(26.1%)	28(60.9%)	4(8.7%)	0.875
	Frequency	0(0%)	1(2.1%)	9(19.6%)	28(60.9%)	8(17.4%)	0.881
User satisfaction	User satisfaction	0(0%)	0(0%)	2(4.3%)	41(89.1%)	3(6.5%)	0.875
Net benefits	Helpfulness and benefit	0(0%)	0(0%)	8(17.4%)	35(76.1%)	3(6.5%)	0.872
Mean evaluation		3.831					

Conclusions

In this paper, we used an auto-indexing approach to create interactive learning objects. The RIT system provides automatic detection technology to assist teachers in producing learning materials. The automatic detection technology can retrieve slide images from a video and record access points. Then, the system transfers the record into information embedded in the video. Students can organize the concepts in meaningful ways, viewing the current slide with the previous one.

Students can take advantage of elements in this online course which allow them to repeat lessons until they reach a specific level. In addition, the anchored point of formative evaluation provides an appropriate way to test students and determine whether they have mas-

tered the content or not. The course data, including streaming data, text and pictures with anchored points can be integrated with the interactive learning objects in HTML format. Teachers can easily use the RIT system to give their instructional videos added value.

Finally, the results of our research show that students who use the RIT system have a more positive attitude toward the learning process. The RIT system benefits students, allowing them to review content asynchronously in the cyber classroom.

Acknowledgment

We would like to thank the National Council of Taiwan, R.O.C., for partially supporting this research under Contract No. NSC 100-2511-S-194-001-MY2 and No. NSC 99-2511-S-194-003-MY3.

References

- Ahanger, G., & Little, T. D. C. (1996). A survey of technologies for parsing and indexing digital video. *Visual Communication and Image Representation*, 7(1), 28–43.
- Antani, S., Kasturi, R., & Jain, R. (2002). A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video. *Pattern Recognition*, 35, 945–965.
- Asan, E., & Alatan, A. A. (2009, September). Video shot boundary detection by graph-theoretic dominant sets approach. *Proceedings of the 24th International Symposium on Computer and Information Sciences* (pp. 7–11). Washington, DC: IEEE Computer Society.
- Bergamin, P. B., Ziska, S., Werlen, E., & Siegenthaler, E. (2012). The relationship between flexible and self-regulated learning in open and distance universities. *The International Review of Research in Open and Distance Learning*, 13(2), 101–123.
- Blanchette, J. (2012). Participant interaction in asynchronous learning environments: Evaluating interaction analysis methods. *Linguistics and Education*, 23(1), 77–87.
- Block, J. H., & Burns, R. B. (1976). Mastery learning. In L. S. Shulman (Ed.), *Review of research in education* (Vol. 4, pp. 3–49). Itasca, IL: Peacock.
- Bloom, B. S. (1968). Learning for mastery. *Evaluation Comment*, 1(2), 1–5.
- Bloom, B. S. (1976). *Human characteristics and school learning*. New York, NY: McGraw-Hill.
- Bloom, B. S. (1982). *All our children learning*. New York, NY: McGraw-Hill.
- Bouvin, N. O., & Schade, R. (1999). Integrating temporal media and open hypermedia on the World Wide Web. *Computer Networks*, 31, 1453–1465.
- Chiu, M. M., Chow, B. W.-Y., & Chang, C. M. (2007). Universals and specifics in learning strategies: Explaining adolescent mathematics, science, and reading achievement across 34 countries. *Learning and Individual Differences*, 17(4), 344–365.
- Deng, Y., Manjunath, B. S., Kenney, C., Moore, M. S., & Shin, H. (2001). An efficient color representation for image retrieval. *IEEE Transactions on Image Processing*, 10(1), 140–147.
- Feng, S. F., Lu, H., & Ma, S. (2001). Mosaic representations of video sequences based on slice image analysis. *Pattern Recognition Letters*, 23, 513–521.
- Fuchs, L. S., Fuchs, D., & Tindal, G. (1986). Effects of mastery learning procedures on student achievement. *Journal of Educational Research*, 79(5), 286–291.

- Gargi, U., Kasturi, R., & Strayer, S. H. (2000). Performance characterization of video-shot-change detection methods. *IEEE Transactions on Circuits and Systems for Video Technology*, 10(1), 1–13.
- Girasoli, A. J., & Hannafin, R. D. (2008). Using asynchronous AV communication tools to increase academic self-efficacy. *Computers & Education*, 51(4), 1676–1682.
- Gonzalez, R. C., & Woods, R. E. (1992). *Digital image processing*. Reading, MA: Addison-Wesley.
- Hanjalic, A. (2002). Shot-boundary detection: Unraveled and resolved? *IEEE Transactions on Circuits and Systems for Video Technology*, 12(2), 90–105.
- Liao, Y.-H., Tsai, C.-Y., Su, M.-H., Li, H.-H., & Yu, P.-T. (2011, August). Digital learning video indexing using scene detection. In R. Kwan, J. Fong, L.-F. Kwok, & J. Lam (Eds.), *4th International Conference on Hybrid Learning 2011* (pp. 336–344). New York, NY: Springer.
- Liu, T., & Kender, J. R. (2004). Lecture videos for e-Learning: Current research and challenges. *Proceedings of IEEE Sixth International Symposium on Multimedia Software Engineering* (pp. 574–578). Washington, DC: IEEE Computer Society.
- Keyvnpour, M., & Izadpanah, N. (2011). A functional measure-based framework for evaluation of multi-dimensional point access methods. *Procedia Environmental Sciences*, 10, 784–789.
- Koprinska, I., & Carrato, S. (2001). Temporal video segmentation: A survey. *Signal Processing: Image Communication*, 16, 477–500.
- Mayer, R. E. (2005). *The Cambridge handbook of multimedia learning*. Cambridge, UK: Cambridge University press.
- Manjunath, B. S., & Ma, W.Y. (1996). Texture features for browsing and retrieval of image data. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 18(8), 837–842.
- Mittal, A., Pagalthivarthi, K. V., & Altman, E. (2006). Content classification and context based retrieval system for e-Learning. *Educational Technology & Society*, 9(1), 349–358.
- Naphade, M., Mehrotra, R., Ferman, A. M., Warnick, J., Huang, T. S., & Tekalp, A. M. (1998). A high-performance shot boundary detection algorithm using multiple cues. *Proceedings of the IEEE International Conference on Image Processing* (Vol. 2, 884–887). Washington, DC: IEEE Computer Society.
- Park, D. K., Jeon, Y. S., & Won, C. S. (2000). Efficient use of local edge histogram descrip-

- tor. *Multimedia 2000: Proceedings of the 2000 ACM workshops on Multimedia* (pp. 51–54). New York, NY: ACM.
- Sakarya, U., Telatar, Z., & Alatan, A. A. (2012). Dominant sets based movie scene detection. *Signal Processing*, 92, 107–119.
- Shivakumara, Phan, & Tan, (2010.)
- Skinner, B. F. (1999). The science of learning and the art of teaching. *Cumulative record, definitive edition* (pp. 179–191). Cambridge, MA: B. F. Skinner Foundation (Original work published 1954).
- Smoliar, S. W., & Zhang, H. (1994). Content-based video indexing and retrieval. *IEEE Transaction on Multimedia*, 1(2), 62–72.
- Song, B. C., & Ra, J. B. (2001). Automatic shot change detection algorithm using multi-stage clustering for MPEG-compressed videos. *Journal of Visual Communication and Image Representation*, 12, 364–385.
- Sun, X. S., Kim, C.-S., & Kuo, C.-C. (2005). MPEG video markup language and its applications to robust video transmission. *Journal of Visual Communication and Image Representation*, 16, 589–620.
- Wu, J.-H., Tennyson, R. D., & Hsia, T.-L. (2010). A study of satisfaction in a blended e-Learning system environment. *Computer & Education*, 55(1), 155–164.
- Yengin, I., Karahoca, A., & Karahoca, D. (2011). E-Learning success model for instructors' satisfactions in perspective of interaction and usability outcomes. *Procedia Computer Science*, 3, 1396–1403.
- Yuan, J., Wang, H., Xiao, L., Zheng, W., Li, J., Lin, F., & Zhang, B. (2007). A formal study of shot boundary detection. *IEEE Transactions On Circuits and Systems for Video Technology*, 17, 168–186.
- Zabih, R., Miller, J., & Mai, K. (1999). A feature-based algorithm for detecting and classifying production effects. *Multimedia Systems*, 7(2), 119–128.
- Zhang, D., Wang, F., Shi, Z., & Zhang, C. (2010). Interactive localized content based image retrieval with multiple-instance active learning. *Pattern Recognition*, 43(2), 478–484.

Athabasca University 



An E-Book Hub Service Based on a Cloud Platform



Jinn-Shing Cheng and Echo Huang

National Kaohsiung First University of Science and Technology, Taiwan, Republic of China

Chuan-Lang Lin

Hamastar Technology Co., Taiwan, Republic of China

Abstract

Due to the constant performance upgrades and regular price reductions of mobile devices in recent years, users are able to take advantage of the various devices to obtain digital content regardless of the limitations of time and place. The increasing use of e-books has stimulated new e-learning approaches. This research project developed an e-book hub service on a cloud computing platform in order to overcome the limitations of computing capability and storage capacity that are inherent in many mobile devices. The e-book hub service also allows users to automatically adjust the rendering of multimedia pages at different resolutions on terminal units such as smartphones, tablets, PCs, and so forth. We implemented an e-book hub service on OpenStack, which is a free and open-source cloud computing platform supported by multiple large firms. The OpenStack platform provides a large-scale distributed computing environment that allows users to build their own cloud systems in a public, private, or hybrid environment. Our e-book hub system offers content providers an easy-to-use cloud computing service with unlimited storage capacity, fluent playback, high usability and scalability, and high security characteristics to produce, convert, and manage their e-books. The integration of information and communication technologies has led the traditional publishing industry to new horizons with abundant digital content publications. Results from this study may help content providers create a new service model with increased profitability and enable mobile device users to easily get digital content, thereby achieving the goal of e-learning.

Keywords: E-learning; cloud computing; e-book; e-book hub; e-reader; digital content

Introduction

When an American company called SoftBook Press produced an e-reader called the SoftBook in 1998 (n.d.; Schilit, Golovchinsky, Tanaka, & Marshall, 1999), the device wasn't popular because there were just a few suppliers of e-books and e-readers at the time. After Amazon released the Kindle in 2007 and Apple the first iPad in 2010 (Amazon Kindle, n.d.), the e-book business began to flourish, and digital content became a new trend in the related publishing and education industries (Chou, Jay, Lin, & Hsieh, 2010; Shen & Koch, 2011; Yee, Chia, Tsai, Tiong, & Rajaraman, 2011; Rahman, Alam, & Saddik, 2011). According to statistics from the Association of American Publishers (Sporkin, 2012), the total volume of e-book sales in the American market has exceeded physical books since January 2011, when volume increased 116% over January 2010. Compared to the sales volume in February 2010, the range of increase reached 202%. Besides the e-reader, digital content providers, e-book transaction platforms, and network communication systems are needed to bolster the e-book market in order to create an e-book generation—one that is digital and paperless.

Recently, the number of users accessing e-book services has increased due to the latest improvements in the capability of cloud computing. Because of the storage limitations and computing capability inherent in many mobile devices (Haber, 1999; Sahu, Sharma, Dubey, & Tripathi, 2012), more and more companies store e-books in the unlimited and low-priced space offered by cloud servers. However, there is currently no appropriate solution to solve the load balancing issue when a cloud computing facility is used to maintain e-book systems.

The purpose of our research was to design and build an e-book hub service with multimedia editing capabilities based on open-source resources of cloud computing technology. The service we implemented successfully allowed content providers to easily edit and convert their e-books and enabled them to manage e-book hub services in an unlimited space offered by the scalable, robust, and safe cloud computing environment.

Cloud Platform Technology

The cloud platform technology we used in this research project is an open-source platform called OpenStack (2012a), freeware supported by industry insiders. OpenStack is a cloud operating system that adopted the authority clause of Apache Licence 2.0, and it may operate as a private, public, or hybrid cloud environment. OpenStack's extendable and flexible cloud environment can coordinate with two main public cloud providers, Amazon and Rackspace, and support other cloud environments in the future.

The three main kernels of OpenStack are Compute, Object Storage, and Image Service. Figure 1 shows the system's architecture, and we have illustrated the system.

Compute Infrastructure (Nova)

OpenStack Compute (2012b), code-named Nova, is open-source software designed to pro-

vision and manage large networks of virtual machines, creating a redundant and scalable cloud computing platform.

Storage Infrastructure (Swift)

OpenStack Object Storage (2012d), code-named Swift, is open-source software that creates redundant, scalable object storage using clusters of standardized servers to store petabytes of accessible data.

Imaging Service (Glance)

OpenStack Image Service (2012c), code-named Glance, provides discovery, registration, and delivery services for virtual disk images.

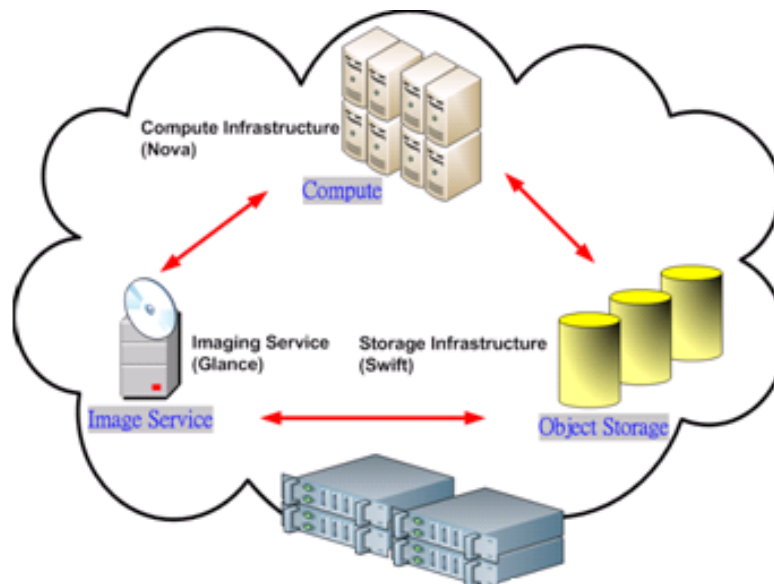


Figure 1. Architecture of OpenStack.

The Components of the E-Book Hub System

The architecture of the standalone client-server e-book hub system we developed in this research project is shown in Figure 2. We constructed the system with Microsoft IIS, .Net applications, and Microsoft SQL Server databases. Its components are divided into three parts: the access device, the physical network, and the APPs service.

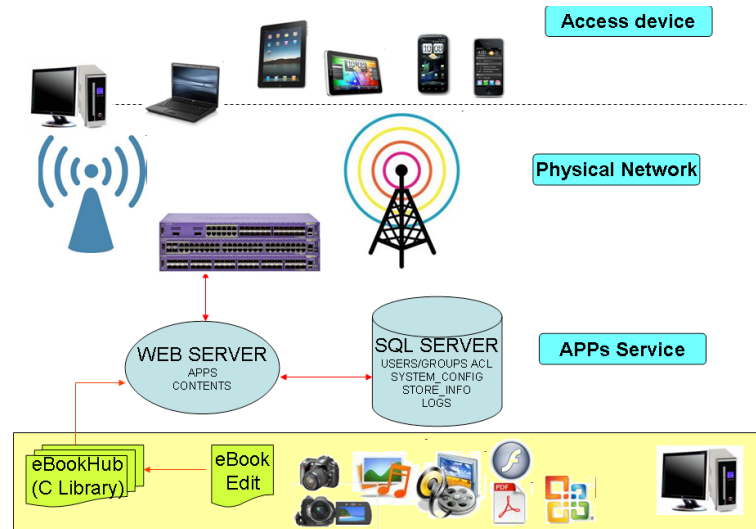


Figure 2. E-book system architecture.

The E-Book Editor

An e-book editor program called SimMAGIC eBook is user-friendly software for editing or converting word processor files into e-book formats. Users can use the editor to combine all the components—text, images, video, and audio—together to create their own e-books. The editor also allows users to transform supported files like PDF or PPT into e-books and then upload them to the e-book hub server. The development tools used to create the e-book editor included Visual Studio C#, Xcode, Eclipse, and Silverlight, among others. Its functions allow a user to define, edit, preview, simulate, and export e-book files. Snapshots of the function, editing, and operating interfaces in SimMAGIC eBook are shown in Figures 3, 4, and 5.



Figure 3. SimMAGIC ebook functions.

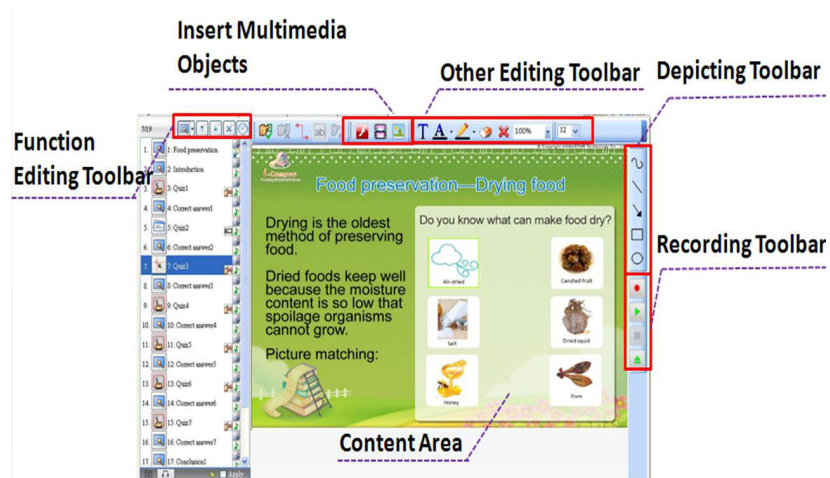


Figure 4. SimMAGIC e-book editing application layout.

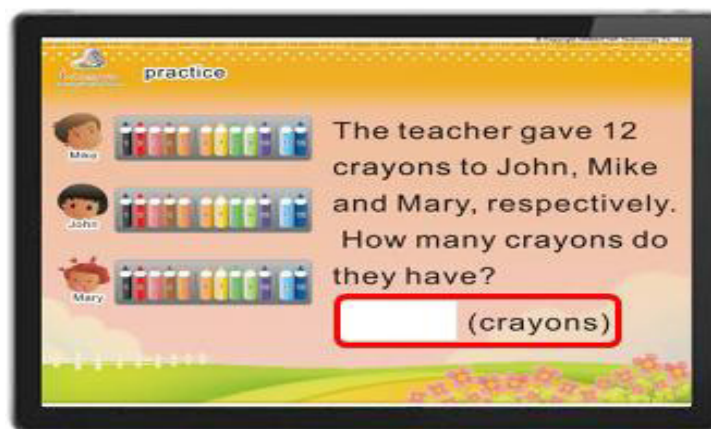


Figure 5. SimMAGIC e-book interactive operation.

Cloud Library

The cloud library called SimMAGIC C-Library is a combination e-book hub web server, content server, and database server that enables users to store, manage, and provide e-book content service. The development tools used to create the C-Library included Visual Studio 2010, .Net Framework 4.0, ASP.Net C#, JavaScript, HTML, and AJAX, among others. Its functions include the uploading and downloading of e-books, e-book hub classification and management, user registration and management, e-book catalogue searches, and statistical analysis. A snapshot of the web portal for the C-Library is illustrated in Figure 6.



Figure 6. SimMAGIC C-Library portal.

Mobile Device Applications

The application E-book Hub can be used on any type of mobile device with an iOS, Android, or Windows operating system to upload or download e-books. Figures 7 and 8 show snapshots of the application's interface.



Figure 7. E-book Hub on a device with an Android operating system.

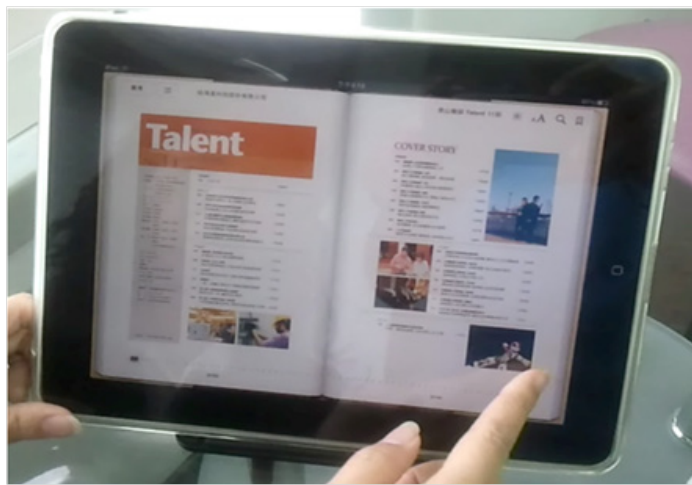


Figure 8. E-book Hub displaying an EPUB publication.

Cloud E-Book Hub System

The standalone e-book hub system we described in the last section has limitations in network traffic utilization, load balancing control, and resource management because it lacks an automatic adjustment control that would enable it to respond quickly to users' demands, and this problem makes it prohibitively expensive to set up the physical infrastructure and operating systems. This has limited the desire of companies to invest money in a standalone e-book hub system. Utilizing cloud computing technology with its high capability, fluency, extendibility, and security at a low cost may provide a solution to the problem.

To this end, we have designed and implemented a cloud technology-based e-book hub system in this study. To transform physical servers into a cloud platform, we used OpenStack, an open-source cloud platform in a private cloud environment, and it provided both infrastructure as a service (IaaS) and platform as a service (PaaS) utilities. We developed the cloud-based e-book hub system according to the framework depicted in Figure 9. We used OpenStack Compute Infrastructure to develop and manage the cloud e-book hub system, which included virtual servers for Web site hosting, content provision, multimedia edit-

ing, data backup, network traffic control, and load monitoring. Then we utilized OpenStack Imaging Service to develop, manage, and store images on the servers. Finally, we employed OpenStack Storage Infrastructure to provide an environment in which to store all the servers, e-book content, and images.

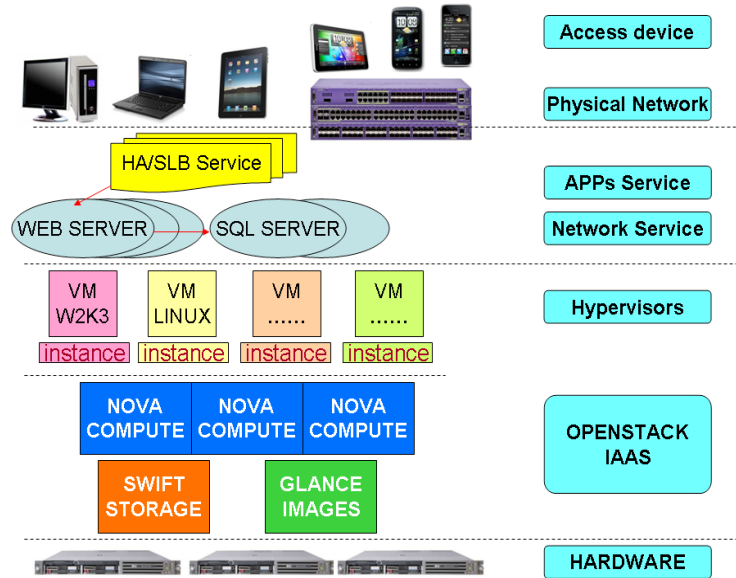


Figure 9. Our cloud e-book hub system framework.

The hardware for the whole system included physical servers, network infrastructure, and users' mobile devices. The physical servers used to build an OpenStack platform provided Cloud Control, Cloud Compute, Swift Storage, and Glance Image services.

Cloud Control – Cloud Compute Service

The Cloud Compute service controls the operation and resource management of a cloud system. It also constructs virtual machines in the environment, as depicted in Figure 10.

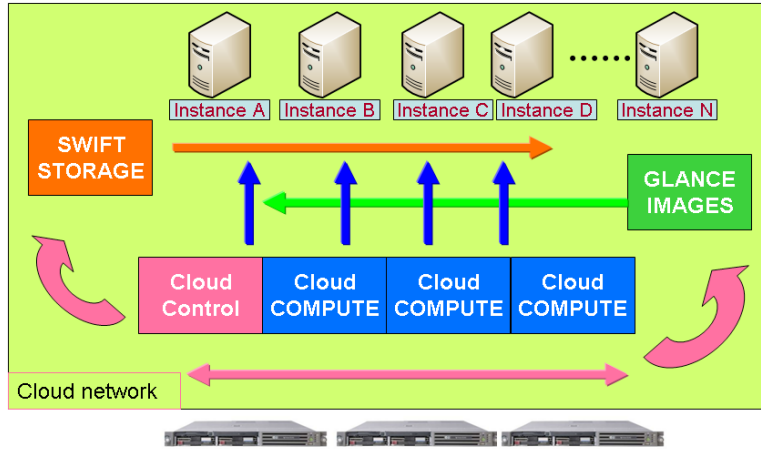


Figure 10. Cloud Control and Cloud Compute service.

Swift Storage Service

The Swift Storage service provides space for all the services in a cloud system. With the characteristics of unlimited storage, built-in replication, higher performance, and highly scalable read/write accessibility, the service enhances the reliability, completeness, and extensibility of the cloud system. The service is depicted in Figure 11.

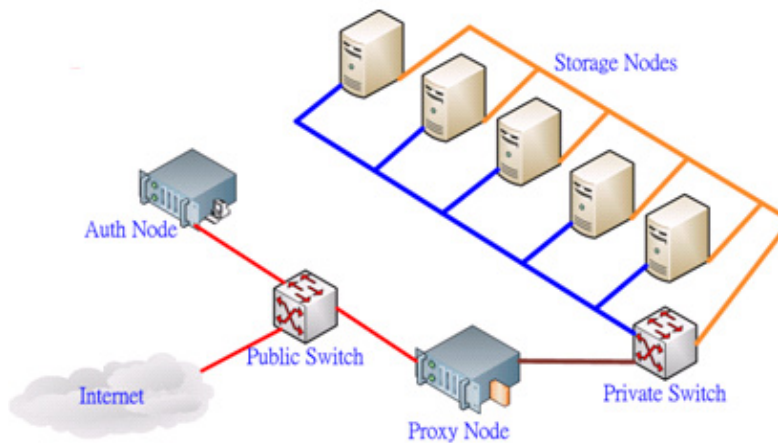


Figure 11. Swift Storage service.

Glance Image Service

The Glance Image service manages virtual machines in the cloud environment. For example, images for virtual web servers may be automatically launched or shut down to accommodate heavier or lighter web requests. The service architecture is shown in Figure 12.

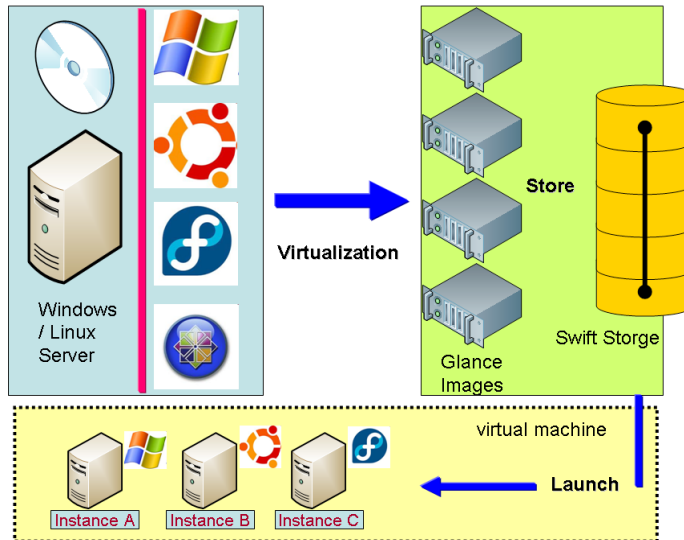


Figure 12. Glance Image service.

We used a high availability (HA)-server load balancing feature to provide high reliability, even flow distribution, and load balance for the virtual web servers. Our objective was to maintain the overall robustness of the system, so that when one of the cloud compute nodes was overloaded, the virtual web servers hosted on others continued working. Based on our desired strategy for traffic distribution, the system could use round-robin, least-connection, or weighted distribution algorithms to balance the traffic rate and decrease the chances of distributed denial of service (DDoS). In addition, SNMP and Windows management instrumentation (WMI) helped monitor the load and traffic rate on the server. With the automatic adjustment program, the Glance Image service increased or decreased instances of web servers according to predefined thresholds for loads and traffic. This automatic adjustment capability clearly demonstrated the advantage of cloud technology in resource management. All the virtual servers operated on instances constructed by cloud compute nodes, and their storage space was provided by the Swift Storage service. A framework for the entire cloud e-book system is depicted below.

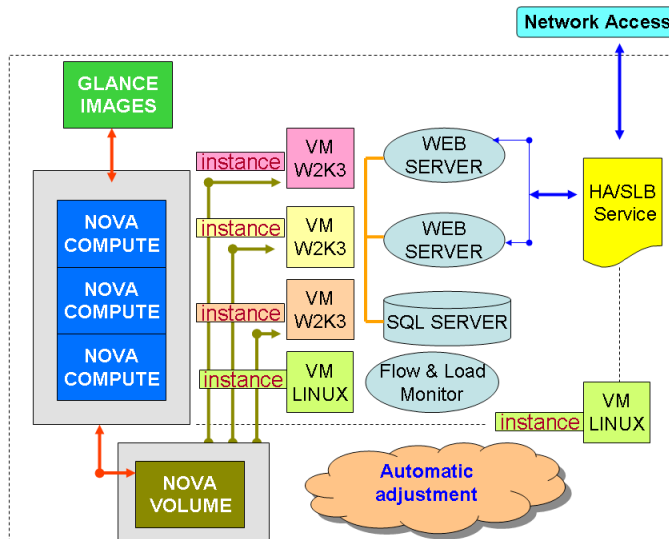


Figure 13. E-book cloud service.

System Performance Verification

We used the Apache Jmeter (2012), an open-source testing tool, to verify the performance of the systems we implemented in this study. The tool was used to measure the performance of a standalone e-book hub system and a cloud-based e-book hub system equipped with HA-server load balancing, as illustrated in Figure 13.

The elements that made up the test environment are shown in Table 1. To create a test environment close to the application’s real-life situation, we set up cloud services in the server farm of a school and used Apache Jmeter to test the system’s performance from places outside the campus. Figure 14 shows the network path we used for the tests. The tests showed that 20, 40, 80, and 160 users were concurrently sending read requests to the target system 500 times. The Apache Jmeter recorded the response time for each request, which included logging into either the standalone or cloud-based e-book hub system and reading an e-book from the system. The recorded response times served as the measure for the system’s performance.

Table 1

The Test Environment

Cloud node	
Vendor	HP
Module	ProLiant DL320 G5p
OS	Ubuntu 10.04.3 LTS /OPENSTACK
CPU	Intel Xeon X3210(Quad-Core) @ 2.13GHz *1

RAM	8 GB
HDD	HP WCASY6755765 ATA 500 GB
Network	Broadcom NetXtreme BCM5715 Gigabit Ethernet *2
Instance	
Application	APPs Server of E-book Hub Service
OS	Windows 2003 + IIS 6
Vcpu	1
Vram	2 GB

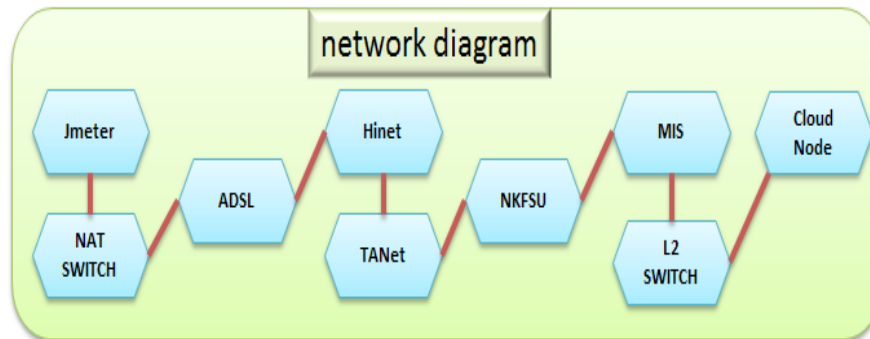


Figure 14. The network path for the test.

We have summarized the simulation results in Table 2. Here are the definitions for the fields in the table.

- Average – The average response time for a set of simulations.
- Median – The median (50%) response time for a set of simulations.
- 90% Line – The 90% quantile response time for a set of simulations. Only 10% of the simulations had a longer response time than this number.
- Min. – The minimum response time for samples with the same label.
- Max. – The maximum response time for samples with the same label.

Table 2

Simulation Results, with Response Times in Milliseconds

Single Virtual Machine (stand-alone system)						
Label	# Samples	Average	Median	90% Line	Min.	Max.
20 users	10,000	403	355	645	45	2,696
40 users	20,000	814	744	1,298	44	6,137
80 users	40,000	1,636	1,488	2,664	45	11,765
160 users	80,000	3,317	2,787	5,421	52	109,505
HA-Server load balancing (cloud-based system)						
Label	# Samples	Average	Median	90% Line	Min.	Max.
20 users	10,000	417	407	465	139	3,528
40 users	20,000	832	796	927	138	4,747
80 users	40,000	1,666	1,556	1,898	244	6,106
160 users	80,000	3,354	2,900	4,898	249	105,008

The results of our tests show that the cloud-based e-book hub system functioned better than the standalone e-book hub system when the system's load was increased gradually. Compared to the single virtual machine, the 90% quantile response time of the cloud-based system was shorter in each test environment. According to the 90% Line statistics, the heavier the load was on the system, the more stable it became.

Figures 15 and 16 present the results of 40,000 requests simulated through 80 users, with 500 read requests per user. As Figure 16 illustrates, the centralization of graphs for the cloud-based system with HA-server load balancing shows that it was more stable than the standalone e-book hub system. In this study, we not only designed a cloud-based e-book hub system, we also implemented it in the OpenStack cloud computing environment. By using cloud computing, the quantity of e-book-provisioning servers could be flexibly increased just in time to meet the sudden demand of a high load.

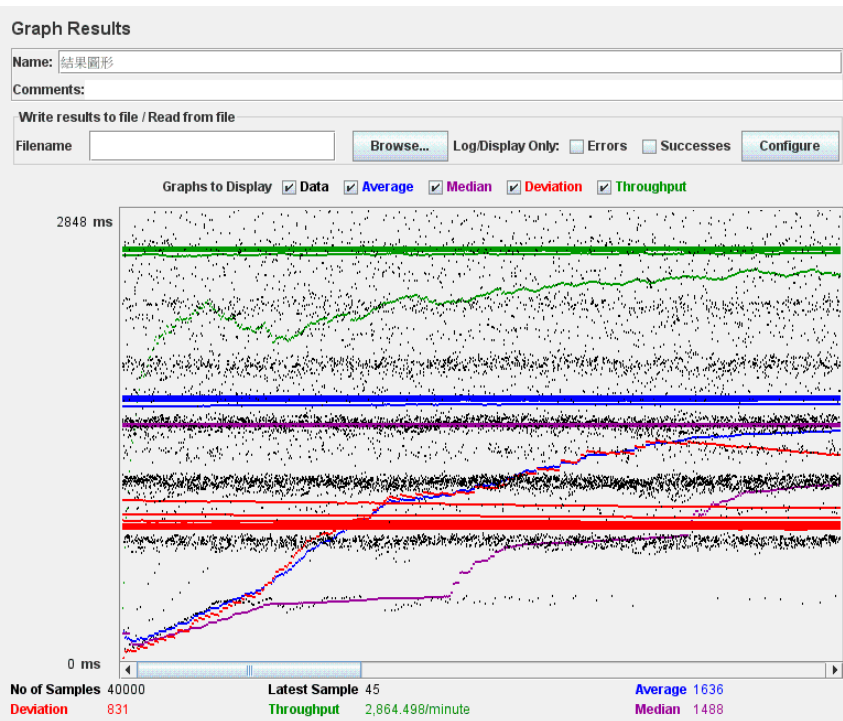


Figure 15. Single virtual machine.

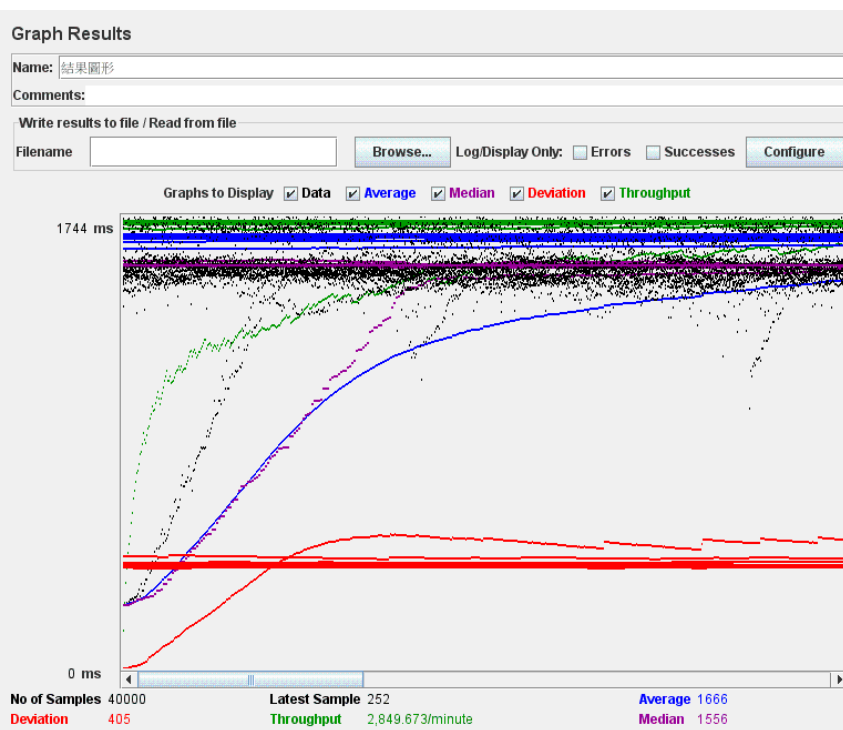


Figure 16. HA-server load balancing structure.

Conclusions

In this study, we designed and implemented an e-book hub service on the cloud computing platform to overcome the limitations of storage and computing capabilities inherent in many mobile devices. Our system allowed content providers to easily edit and convert their multimedia e-books and manage the e-book hub service in an unlimited capacity that offered scalable, robust, and safe cloud computing services. Results from this study may help content providers to create a new service model with increased profitability and allow mobile device users to easily access digital content, helping them to achieve the goal of e-learning.

We can take a few directions to extend the work we did in this study. So far, the e-book hub system can only operate on Windows servers. We can migrate the developed program to Linux servers to reduce the total cost. Besides, in order to achieve the goal of self-service cloud computing on demand, information from virtual server images should be automatically inserted and converted when it is constructed with the cloud-init component.

References

- Amazon Kindle (n.d.). In *Wikipedia*. Retrieved from http://en.wikipedia.org/wiki/Amazon_Kindle
- Apache Software Foundation (2012). Jmeter [desktop application]. Retrieved from <http://jmeter.apache.org/>
- Chou, S. C., Stu, J., Lin, Y. T., & Hsieh, W. T. (2010). Toward computer-based learning: A taxonomy of e-book readers. In *Proceedings of the 2nd International Conference on Computer Engineering and Technology (ICCET), Vol. 2* (pp. 431–434).
- Haber, L. (2011, August 29). Tablet storage limited, but options abound. *TabletPC Review*. Retrieved from <http://www.tabletpreview.com/default.asp?newsID=2524&news=tablet+storage+options+ipad+android+ios>
- Openstack: The Open Source Cloud Operating System (2012a). About Openstack [Web page] Retrieved from <http://www.openstack.org/projects/>
- Openstack (2012b). Compute [Web page]. Retrieved from <http://www.openstack.org/projects/compute/>
- Openstack (2012c). Image Service [Web page]. Retrieved from <http://www.openstack.org/projects/image-service/>
- Openstack (2012d). Object Storage [Web page]. Retrieved from <http://www.openstack.org/projects/storage/>
- Rahman, A. S. M. M., Alam, K. M., & Saddik, A. E. (2011). A prototype haptic eBook system to support immersive remote reading in a smart space. *2011 IEEE International Workshop on Haptic Audio Visual Environments and Games* (pp. 124–128).
- Sahu, D., Sharma, S., Dubey, V., & Tripathi, A. (2012). Cloud computing in mobile applications. *International Journal of Scientific and Research Publications*, 2(8), 1–9.
- Schilit B., Price, M., Golovchinsky, G., Tanaka, K., & Marshall, C. (1999). The reading appliance revolution. *Computer*, 32(1), 65–73.
- Shen, W., & Ute, K. (2011). eBooks in the cloud: Desirable features and current challenges for a cloud-based academic eBook infrastructure. *Digital Publishing and Mobile Technologies, 15th International Conference on Electronic Publishing* (pp. 80–86). Retrieved from http://elpub.scix.net/data/works/att/113_elpub2011.content.pdf
- SoftBook (n.d.). In *Wikipedia*. Retrieved from <http://en.wikipedia.org/wiki/SoftBook>
- Sporkin, A. (2012, May 18). US publishers see rapid sales growth worldwide in print and

e-formats (press release). Retrieved from the AAP, [Association of American Publishers](http://www.publishers.org/press/68/) Web site: <http://www.publishers.org/press/68/>

Yee, K. Y., Chia, Y., Tsai, F. S., Tiong, A. W., & Rajaraman, K. (2011). [Cloud-based semantic service-oriented content provisioning architecture for mobile learning](#). *Journal of Internet Services and Information Security*, 1(1), 59–69.

Athabasca University 



Contexts in a Paper Recommendation System with Collaborative Filtering



Pinata Winoto
Konkuk University, Korea

Tiffany Ya Tang
Kean University, USA

Gordon McCalla (not shown)
University of Saskatchewan, Canada

Abstract

Making personalized paper recommendations to users in an educational domain is not a trivial task of simply matching users' interests with a paper topic. Therefore, we proposed a context-aware multidimensional paper recommendation system that considers additional user and paper features. Earlier experiments on experienced graduate students demonstrated the significance of this approach using modified collaborative filtering techniques. However, two key issues remain: (1) How would the modified filtering perform when target users are inexperienced undergraduate students who have a different pedagogical background and contextual information-seeking goals, such as task- and course-related goals, from those of graduate students?; (2) Should we combine graduates and undergraduates in the same pool, or should we separate them? We conducted two studies aimed at addressing these issues and they showed that (1) the system can be effectively used for inexperienced learners; (2) recommendations are less effective for different learning groups (with different pedagogical features and learning goals) than they are for the same learning groups. Based on the results obtained from these studies, we suggest several context-aware filtering techniques for different learning scenarios.

Keywords: E-learning; pedagogy

Introduction

A recommender system (RS) can follow the steps of its user, observe the interests of a group of similar users, and pick items that best suit the user based on either items the user liked (content-based filtering) or implicit observations of the user's followers/friends who have similar tastes (collaborative filtering, or CF; McNee et al., 2002; Herlocker, Konstan, Terveen, & Riedl, 2004; Lekakos & Giaglis, 2006). In the majority of these approaches, the successful match of the recommended item is measured by its utility, usually given a numerical rating by the user based on how much he or she liked the item (Adomavicius, Mobasher, Ricci, & Tuzhilin, 2011), a single-dimensional RS. However, users' preference for an item may be influenced by one or many contexts (Tang & McCalla, 2009; Winoto & Tang, 2010; Adomavicius et al., 2011). For instance, say a user is looking for a movie that is suitable for a fun family activity, such as a "family-friendly" movie. Contexts considered in a RS would vary depending on the applications (e.g., movies, books, music, education, etc.) and tasks the system intends to support (Gunawardana & Shani, 2009).

In the field of e-learning, a RS can help a tutor or learner to pick relevant courses, programs, or learning materials (books, articles, exams, etc.), and the contexts include the user's learning goals, background knowledge, motivation, and so on. These contextual attributes can be injected into the recommendation mechanism during either the prerecommendation or postrecommendation filtering process (Winoto & Tang, 2010; Adomavicius et al., 2011). A context-aware RS is referred to as a multidimensional RS.

Table 1 presents an example of a user rating matrix in single and multidimensional RSs for books. Here, target user John's rating of *The Da Vinci Code* can be predicted (the cell with "?") based on the ratings of those sharing similar interests with him.

Table 1

An Example of a Single-dimensional (Top) and Multidimensional (Bottom) Book Ratings

	<i>Jane Eyre</i>	<i>Robinson Crusoe</i>	<i>Lord of the Flies</i>	<i>To Kill a Mockingbird</i>	<i>The Da Vinci Code</i>
Alice	8	7	6	8	5
Bob	6	7	6	6	6
Carol	6	5	7	7	6
John	5	5	8	?	?

	<i>Jane Eyre</i>			<i>Robinson Crusoe</i>			<i>Lord of the Flies</i>			<i>To Kill a Mockingbird</i>			<i>The Da Vinci Code</i>		
	story	style	impact	sto	sty	i	sto	sty	i	sto	sty	i	sto	sty	i
Alice	8	7	8	7	6	8	6	6	5	8	7	8	5	6	5
Bob	5	5	8	7	8	6	8	6	5	6	6	7	7	6	6

Carol	7	6	6	6	4	4	8	7	7	5	7	8	6	6	6
John	5	5	6	7	5	4	7	8	8	?	?	?	?	?	?

In the multidimensional RS, each book's overall rating is reflected by three sub-ratings: story, style (writing style), and impact. The overall rating, when represented by subtle and specific attributes, tends to be a more reliable predictor of whether users like the book or not. Excluding contextual considerations, Carol is the closest neighbor to John (Table 1, top); however, if we only consider the story of the books, Bob is his closest neighbor (Table 1, bottom). As such, in the first case the book *To Kill a Mockingbird* will be suggested to John (the predicted rating of the book is 7); while in the latter case, *The Da Vinci Code* will be recommended. From the situated cognition point of view, the context-aware perspective incorporated in multidimensional RSs can more appropriately capture human beings' information-seeking and cognitive behaviors (Rieh, 2002).

Tang and McCalla (2009) explored the factors that may result in a high rating for a paper in terms of its pedagogical benefits (whether the learner has gained knowledge from reading the paper will affect his or her rating). Results showed the importance of several pedagogical contexts in making a paper recommendation, particularly that learner interest is not the only dimension. Other contextual information-seeking goals, such as task- and course-related goals, are also important to learners' perceptions of the paper's value. Learners' willingness to peer-recommend a paper largely depends on how close the paper topic was to that of learners' own work. These observations can help tutors determine which items to select.

In this paper, we extend the research efforts of Tang and McCalla (2009) by performing two groups of experiments on both undergraduate and graduate students spanning a period of two years. The major reason we tested two distinct learner groups is twofold: First, it is known that RS performance is sensitive to users from different segments of the population, affected by factors such as demographic or socioeconomic status (Pazzani, 1999; Lekakos & Giaglis, 2006). Second, the two learner groups differed significantly in a number of key contexts which the recommendations should consider: pedagogical background, job-related experiences, learning goals, study practices, and so forth. Bernt and Bugbee (1993) pointed out that these contexts are good and persistent indicators of academic success in learning environments. In addition, people make judgments based on both the information and cognitive authority of the item (Rieh, 2002). The former is defined by the extent to which users think that the item is "useful, good, current and accurate," while the latter is the extent to which users think that they can trust the item (Rieh, 2002). Both of these judgment criteria motivated us to conduct two sets of studies on two pools of learners with different backgrounds. Specifically we report our findings from the comparative study of the two learner groups to address two key issues: (1) how would the recommenders perform when the target users are undergraduate students who have a different pedagogical background and contextual information-seeking goals, such as task- and course-related goals, from those of graduate students?; (2) Should we combine the pools of different learners into a single pool, combining graduate and undergraduate students into the same group, for example? Or

should we separate them into different pools for the collaborative filtering? Based on these extensive studies, we further make recommendations on several context-aware techniques for different learning scenarios.

The organization of this paper is as follows: In the Related Works section we discuss the earlier efforts of researchers exploring contexts in (educational) recommendation systems. The Recommendation Techniques and Experiment Setup section provides the details of the modified CF techniques for educational paper recommendation and shows the recommendation flow of our system. The Experiment I section documents our first study (focused on experienced learners) and highlights the performance of the CF under various learning scenarios, while the Experiment II section introduces our second study (using inexperienced learners) and compares the experiment results with those in the first study. Then we provide a general discussion of our study and conclude by describing what lessons we learned from our research.

Related Works

To the best of our knowledge, no researchers have studied the cross recommendation of learning material among two or more groups of learners, and very few have studied the contextual attributes of educational RS. In this section we will discuss related work from two perspectives: context-aware RS and educational RS.

Context-Aware Recommender Systems

Adomavicius et al. (2011) argued that dimensions of contextual information can include when, how, and with whom the users will consume the recommended items, which therefore directly affects users' satisfaction about the system's performance. Pazzani (1999) studied a demographic based-CF which identified the neighbors of a target user and made recommendations accordingly. Lekakos and Giaglis (2006) considered users' lifestyles (their living and spending patterns) when making recommendations. Winoto and Tang (2010) studied a mood-aware recommendation approach that considered a user's mood to find a like-minded group for recommendation. In our study, the contexts are a learner's background knowledge and learning goals. Since undergraduate and graduate students have different background knowledge and goals, we expect their satisfaction with the recommendations will vary.

(Context-Aware) Educational Recommender Systems

Despite researchers' recent efforts to incorporate contexts into the recommendation process, the majority of early efforts in educational RS have been based on learners' interests. For example, Recker, Walker, and Lawless (2003) studied the recommendation of educational resources through Altered Vista, a system that enables teachers and learners to submit comments on the resources provided by learners who are precategorized into different "pedagogical" groups. Brusilovsky, Farzan, and Ahn (2005) reported on their user study of Knowledge Sea III, which provided "annotation-based" social navigation support for making personalized recommendations. McNee et al. (2002) investigated the adoption

of CF techniques to recommend additional references for a specific research paper. A similar study conducted by Torres, McNee, Abel, Konstan, and Riedl (2004) utilized document titles and abstracts to make recommendations. Other recommendation studies made use of data mining to construct user profiles (Khribi, Jemni, & Nasraoui, 2009). These studies failed to consider whether the recommended paper is appropriate to support learning (goal-oriented RSs).

Recently, researchers have made efforts to identify and incorporate learners' pedagogical features (contexts) for recommendations. Nadolski et al. (2009) studied the effect of a learner's competence level, study time, and efforts on the performance of an educational RS. The contexts considered in Manouselis, Vuorikari, and Van Assche's study (2010) on recommending learning objects were similar to ours (learning goals, ease of use, etc.), although the target users were not students. Other similar efforts include Lemire, Boley, McGrath, and Ball (2005); Khribi et al. (2009); Gomez-Albarran and Jimenez-Diaz (2009); Manouselis et al. (2010); and Drachsler et al. (2007). Table 2 compares the contexts used in these studies with those used in ours.

Table 2

Various Studies in (Context-aware) Education RS

	Object to be recommended	Contexts	Evaluators (size)
Manouselis <i>et al.</i> , 2010	learning resources	Ease of use, facilitate learning, and topical relevance	Teachers
Khribi <i>et al.</i> , 2009	Learning objects	Learner goals and learner needs	Simulations
Nadolski <i>et al.</i> , 2009	Learning programs	Competence level, study time, and efforts	Simulation
Drachsler <i>et al.</i> , 2007	Learning resources	Learner features, learner preference, demographic data, prior knowledge	Prototype only
Lemire <i>et al.</i> , 2005	Learning objects	Title, date, and author	Prototype only
Our work	Reading materials	Facilitate learning, topical relevance, popularity, and ease of understanding	Graduate (25) Undergraduate (45)

Recommendation Techniques and Experiment Setup

Pedagogical Paper Recommendation: Techniques

Since providing recommendation in a pedagogical context differs from doing so in other settings, we have modified the traditional techniques according to the characteristics of the learning domain. Broadly speaking, these characteristics are (1) the limited number of users, (2) the large number of unrated or new items, (3) the likelihood of the learners having difficulty understanding the items, and (4) the numerous purposes of the recommendation.

When there are a limited number of users and large numbers of unrated/new items, our RS cannot rely solely on rating-based CF (a cold start problem). Therefore, we considered a user model-based CF that does not need many learner ratings. We also took into consideration paper popularity (the average overall ratings of a paper) in an attempt to start the recommendation process when there were not many ratings in the system (this technique is largely considered to be nonpersonalized). Factors considered in our multidimensional CF were mainly used to correlate one user with another. Specifically, we considered the following factors: a paper's overall ratings, popularity, value-added, frequency of peer recommendation (or *peer_rec*), and learners' pedagogical features, such as interest and background knowledge. Overall rating represents the total rating given to a paper by a user (using a Likert scale of 1 to 4). Value-added represents the knowledge the user learned from a paper, and peer recommendation is defined as the user's willingness to recommend a paper to other learners.

Regarding the numerous purposes for recommendation, a tutor may aim for overall learner satisfaction (the highest possible overall rating), to stimulate learner interest only (the highest possible interest rating), or to help the learner gain new information only (the highest possible value-added rating), and so on. It is thus both appealing and imperative to collect multiple ratings and study multidimensional CFs that can utilize them. Table 3 categorizes the various recommendation techniques used in our studies, which generally fall into three main categories: content-based, CF-based, and hybrid recommendation.¹

¹ Interested readers can refer to Tang (2008) for details on these algorithms.

Table 3

A Summary of the Various Recommendation Techniques

Category	Name	Remarks
Content-based	ContentF	Content-based filtering
CF-based	1D-CF	Uni-dimensional rating-based CF
	3D-CF	Multidimensional rating-based CF
	UM-CF (2D-CF)	User model-based CF
	PopUMCF	A combination of non-personalized and UM-CF
Hybrid	PopCon2D	A combination of non-personalized, user item content filtering and 2D-CF

Nonpersonalized Recommendation (Benchmark)

Note here that we regard the inclusion of paper popularity as a nonpersonalized method. That is, this type of recommendation technique generates items based on a group of users tastes. We treat all of the students in the same class as a group. For this equation, the average rating of each paper, k , among all same-grade learners (denoted as \bar{r}_k), will represent the paper's popularity.

1D-CF

Unidimensional, rating-based CF is the traditional CF that has been used in the literature (Herlocker et al., 2004; Adomavicius et al., 2011). First, we calculated the Pearson correlation between users a and b , using the formula

$$P(a,b) = \frac{\sum_{k \in K} (r_{a,k} - \bar{r}_a)(r_{b,k} - \bar{r}_b)}{\sqrt{\sum_{k \in K} (r_{a,k} - \bar{r}_a)^2 \sum_{k \in K} (r_{b,k} - \bar{r}_b)^2}} \quad (1)$$

where k is the rating by user i on item r , \bar{r}_i is the mean rating by user i for all items, and k is the set of items co-rated by both a and b . The estimated rating target user a gave for a paper, j , $r_{a,j}^e$ is then calculated with the target user's neighbors, denoted as B , using the following formula:

$$r_{a,j}^e = \frac{\sum_B P(a,b) \times r_{b,j}}{\sum_B P(a,b)} \quad (2)$$

In the learning domain, not many papers (less than 30 for each student in one semester) are

commonly assigned as part of the learning activities in a course. Thus, our research focus is on a limited number of co-rated papers, so $|K| \leq 5$, and the number of neighbors $|B|$ ranges from 2 to 15.

User-Model-Based Collaborative Filtering (UM-CF)

In rating-based CF, a target user needs to rate a few papers before we can find his or her neighbors, which is a major drawback especially at the beginning of each semester when ratings have not yet been provided by students. This issue has long been known as one type of a cold start problem, known as the new user problem (Schein, Popescul, Unger, & Pennock 2002). Fortunately, a user model-based CF (UM-CF) can be employed for users who previously have not rated any papers. The UM-CF extracts a user's interest and his or her background knowledge and injects these features into *Equation 1* in order to compare the similarity between each user. In other words, UM-CF compares users based on their interests and background knowledge, and makes recommendations accordingly.

Combinations of Nonpersonalized Recommendation and User Model-Based Collaborative Filtering (PopUM-CF)

PopUM-CF is a combination of UM-CF with the nonpersonalized recommendation method. It is used to overcome rating-based CF's reliance on co-rated papers.

Combinations of Content-Based Filtering, Nonpersonalized Recommendation, and User Model-Based Collaborative Filtering (PopCon2D)

Another hybrid method combines content-based filtering with nonpersonalized recommendation and user model-based CF, namely PopCon2D (it stands for popularity + content-based filtering + 2D user model-based CF). However, we normalized the closeness value by dividing each value with $\max_B(|\text{closeness}_b|)$ so that our closeness value is always between -1 and 1.

The Recommendation System at a Glance

The proposed paper recommendation is achieved through a careful assessment and comparison of both learner and paper characteristics. In other words, each individual learner model will first be analyzed in terms not only of learner interest but also pedagogical features. Paper models will also be analyzed based on the topic, degree of peer recommendation, and so on. The recommendation is carried out by matching learner interest with the paper topics, with the goal of ensuring that the technical level of the paper should not impede the learner from understanding it. Therefore, the suitability of a paper for a learner is calculated by whether this paper is appropriate to help the learner in general.

When the tutor first initiates the system, she or he will be requested to fill in briefly the learner model, including learning goals, interest, knowledge background, and so forth. The tutor will then see the user interface listing an initial set of recommended articles that matches the learner's profile. Figure 1 illustrates the overall architecture of the system.

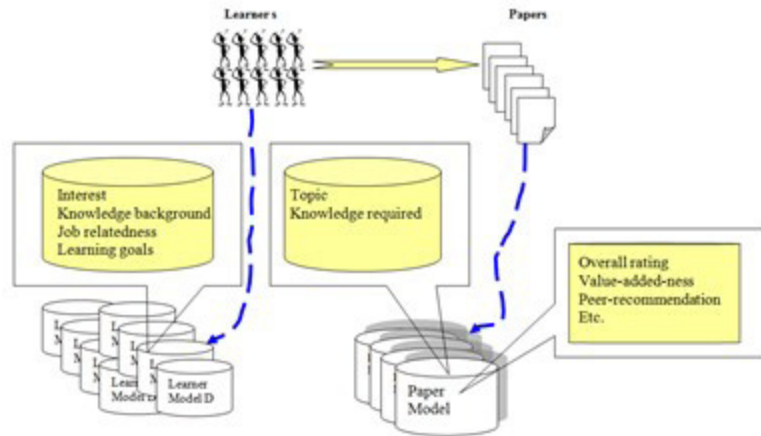


Figure 1. A closer look at the recommendation in our system.

The system consists of four main panes, showing the user model, paper model, paper rating, and recommendation (see Figure 2).

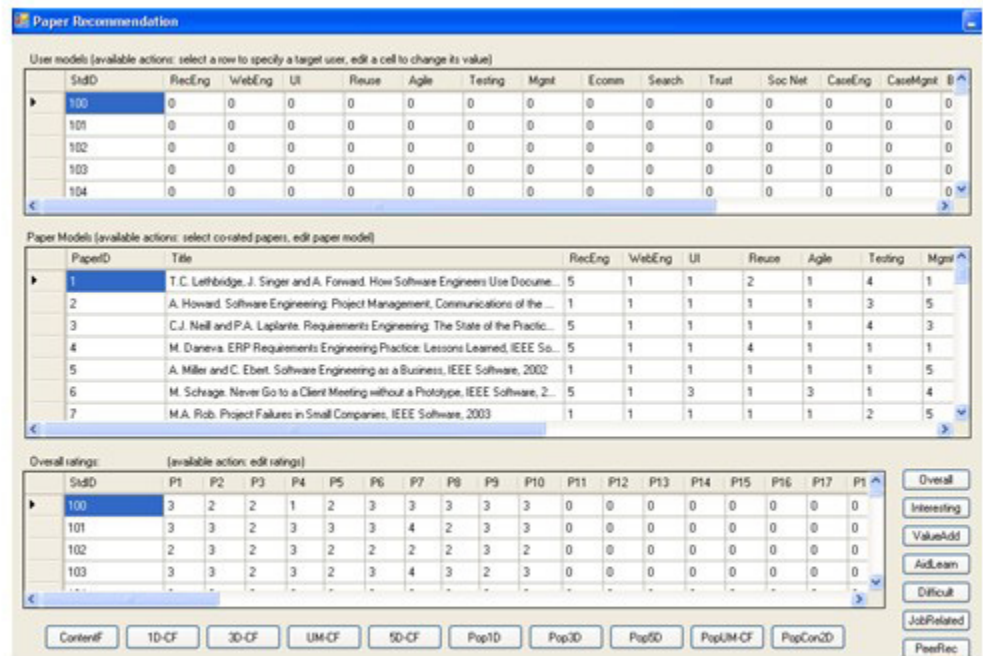


Figure 2. The initial system with ratings from previous users and paper models.

Figure 3 illustrates the results of an ensemble method, where we applied a weighted voting mechanism. The best three recommended papers from each applicable method (ContentF, UM-CF, PopUM-CF, and PopCon2D) are shown at the top of the figure, while at the bottom the calculation of users' weighted voting is visible. As a class proceeds, more papers will be rated by the learners, and therefore more methods can be used to make recommendations.

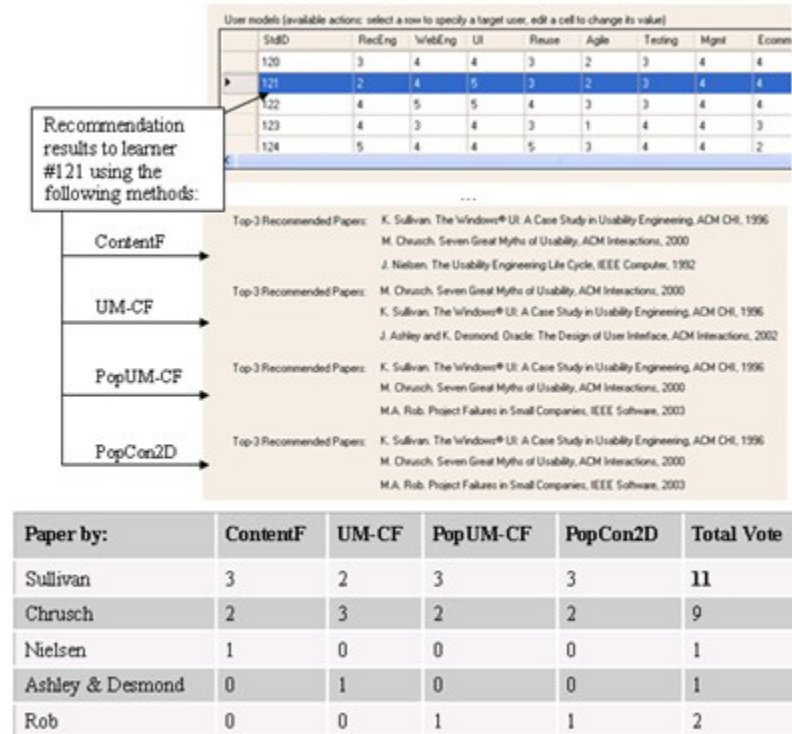


Figure 3. An illustration of an ensemble method.

As Figure 3 illustrates, the Sullivan paper is recommended to learner #121 after obtaining the highest votes from four applicable recommendation methods.

Experiment I

Data Collection

The first experimental study (hereafter Experiment I) was conducted in an introductory software engineering course for master's-level students. In total, 40 part-time students attended the course. Since in this pool students have at least one year of working experience in the IT industry or an IT-related field, we considered them experienced learners. During the class, 22 papers were selected and assigned to students as their reading assignments according to the curriculum of the course, without considering the implications the choices had for our research. The number of papers assigned each week varied according to their length. In total, 24 students agreed to participate in this experiment.

At the beginning, learner profiles were drawn from a questionnaire consisting of four basic categories: interest, background knowledge, job nature, and learning expectation. After reading each paper, students were asked to fill in a paper feedback form evaluating several features of each paper, including how difficult it was to understand, its degree of job-relatedness for the user, how interesting it was, its degree of usefulness, its ability to expand the user's knowledge (value-added), and its overall rating (on a 4-point Likert scale). Using

the collected ratings, we applied the recommendation techniques explained in the previous section to find the best one (top 1), the best three (top 3), and the best five (top 5) recommended papers for each target learner. Then we recorded the ratings given by the target learners to these recommended papers for our analysis.

Evaluation Metric

Evaluation protocols and methodologies should be designed to appropriately reflect the tasks that the RS supports and the users of the system (Herlocker et al., 2004; Gunawardana & Shani, 2009; Winoto & Tang, 2010). Our system is not intended for accurately predicting user ratings; instead, it helps learners understand the materials well in order to recommend the most appropriate papers. Therefore, our evaluation stresses learner acceptance rather than pure prediction accuracy (similar to those evaluations done by researchers in earlier works, including McNee et al., 2002, Torres et al., 2004, Brusilovsky et al., 2005, and Recker et al., 2003), which aligns with baseline measurement of a RS's quality (Herlocker et al., 2004).

Results

In some experiments involving CF methods, we chose 10 neighbors, selected after our analysis of a one-dimensional CF. Figure 4 shows some of the average overall ratings in 1D-CF with the number of neighbors, N , for the number of co-rated papers $|K|$ equal to 2, 4, and 8. It also shows the average overall ratings for all combinations of $|K|$ (from 2– to 15) and (top 1, top 3, top 5), labeled as “Total Average.”

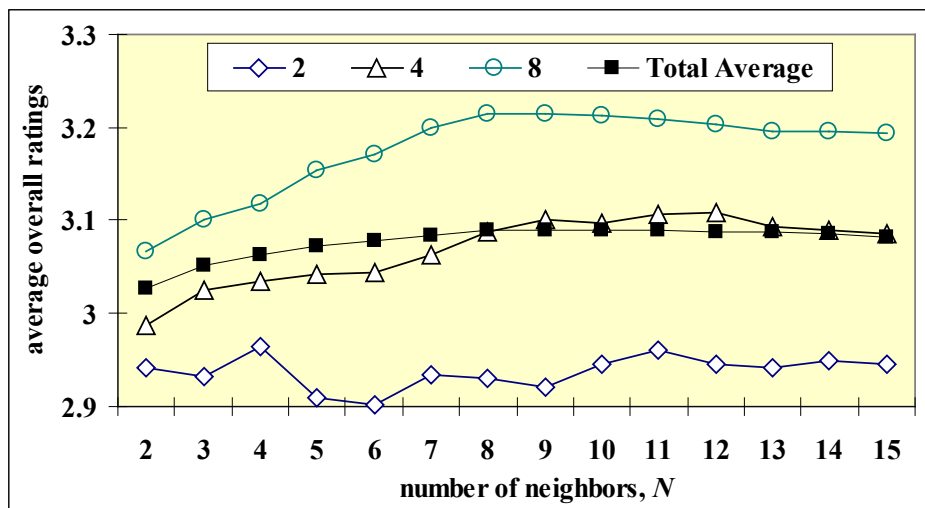


Figure 4. The average overall ratings with the number of neighbors used in 1D-CF.

When the number of co-rated papers is two, the performance is very unstable. The average ratings slightly decrease when the number of neighbors increases. This phenomenon indicates that pure CF-based methods rely on the quality of the neighbors and the denseness of the rating database. Therefore, the key issue is the similarity between the ratings. In our data, the total average reaches its maximum value when $N = 10$, which we used in our rating-based CF methods.

Discussion

The experiment's results are encouraging, especially because they confirm that making recommendations to learners is not the same as making recommendations to users in commercial environments such as Amazon.com. In learning environments, users are willing to accept items that are not interesting if they meet their learning goals in some way or another. For instance, our experimental results suggest that a user model-based CF works well with content-based filtering and nonpersonalized methods (such as paper popularity), such as PopCon2D. Although the computation required for PopCon2D is more sophisticated than other CF-based approaches, under certain circumstances it helps inform the recommender and therefore improve the recommendations.

Our findings illuminate learner satisfaction as a complicated element of learner characteristics, rather than a single issue of whether the paper topics matched learner interests. The results lead us to speculate that if there are a limited number of both papers and learners in the domain, considering other features, rather than relying on an overall rating and user interest, can help inform the recommendation. Table 4 summarizes our recommendations for adopting appropriate mechanisms based on various learning scenarios. Here, PopCon2D performs very well in three typical learning contexts for picking the single best paper, and the more complex PopUM-CF works well for making the best three recommendations. Due to its characteristics, PopCon2D can not only be used to start the recommendation but also to inform the recommendation (since it contains information such as paper popularity, paper content, user model of learner interest, and knowledge background, all of which can be used to generate recommendations without paper ratings). In dimensions such as this, with a limited number of both papers and learners (and other constraints, such as the course syllabus), we conclude that considering features other than just overall ratings and user interest can help inform the recommendation.

When the system does not have enough data on paper and user models, a content-based filtering method is appropriate because it matches the new user model and existing user and paper models. However, when there are not enough papers to perform the matching, some other features, such as popularity, r_j , need to be injected to inform the RS, as in PopCon2D and PopUM-CF. These methods define the features of pedagogical paper recommendation and reflect the reality that human judgments about scientific articles are influenced by a variety of factors, including a paper's topical content, its content appropriateness, and its value in helping users achieve their task (Custard & Sumner, 2005). They also highlight the importance of appropriately incorporating such factors into the recommendation process.

Table 4

A Summary of Suggested Recommendation Methods

Learning Scenario		Appropriate recommendation method(s)	
		Top one	Top three
When there are enough ratings and papers	The learner is new to the course	PopCon2D	PopUM-CF
	The learner is half way in the course	PopCon2D	PopUM-CF
When there are not enough ratings and papers	The learner is new to the course	PopCon2D	PopUM-CF

Experiment II

Experiment Setup and Evaluation Metrics

In this study (hereafter Experiment II) we collected data from 45 undergraduates who took a junior-level software engineering course. It is a mandatory course for full-time undergraduates in three majors: computer science, information technology, and a double-degree program of computer science and management. Most students claimed that they were inexperienced in practical software engineering when taking the course. Evaluation metrics were the same as those used in Experiment I in that we assessed recommendation performance in terms of three ratings: overall, aid_learning and value_added; however, we were more interested in ratings for both aid_learning and value_added because most undergraduates were inexperienced and this was their first software engineering course. Intuitively, undergraduates with less background knowledge may find more “new” information compared to the graduates. Hence, both aid_learning and value_added become more important recommendation goals in an undergraduate course.

Comparative Results and Discussion

The recommendation algorithms we used included PopUM-CF, PopCon2D, and 1D-CF, as suggested by Experiment I (Table 4).

Tables 5 to 7 compare the performance of PopUM-CF, PopCon2D, and 1D-CF for recommending the top 1 and top 3 papers to graduates and undergraduates. The values reported in the tables are the average ratings. The column “Grads” consists of the average ratings given by experienced learners (grads) to the recommended papers. The column “Undergrads” is divided into two columns: “From Grads” and “From UG”, which denote whether

the recommended papers were rated by experienced learners or inexperienced ones (undergraduates). We also included the best-case benchmark (recommending the most popular papers) after the solidus mark to compare whether the particular personalized recommendation is better (shown in bold font) or not. As we show in Table 7, all results from 1D-CF are significantly higher than those from the best-case benchmark, suggesting that when we have enough co-rated papers (at least eight), the recommendations for undergraduates work well. However, this is not always true for PopUM-CF (Table 5) or PopCon2D (Table 6), which means a pure, nonpersonalized recommendation may be useful in a cold start situation. It is not clear why both PopUM-CF and PopCon2D fail to provide a better outcome. One possible explanation is that most undergraduates, due to their lack of background knowledge, cannot specify their interests accurately; hence, adding personalized recommendations by matching less-accurate user models to papers or users could not improve the results.

Table 5

Average Ratings from PopUM-CF: Popularity Only

	Top 1			Top 3		
	Grads	Undergrads		Grads	Undergrads	
		From Grads	From UG		From Grads	From UG
Aid_learning	3.160	3.178 /3.111	3.111/3.178	3.000	3.030/3.089	3.067/3.089
Value_added	3.280	3.244/3.267	3.356 /3.267	3.240	3.170/3.185	3.296 /3.244
Overall	3.160	2.933/3.022	3.044/3.044	3.093	2.933/2.970	3.022 /2.978

Table 6

Average Ratings from PopCon2D: Popularity Only

	Top 1			Top 3		
	Grads	Undergrads		Grads	Undergrads	
		From Grads	From UG		From Grads	From UG
Aid_learning	3.160	2.978/3.111	2.933/3.178	3.040	2.963/3.089	3.089/3.089
Value_added	3.240	3.244/3.267	3.311 /3.267	3.173	3.178/3.185	3.259 /3.244
Overall	3.240	3.022/3.022	2.956/3.044	3.027	2.933/2.970	2.911/2.978

Table 7

Average Ratings from 1D-CF (Co-Rated Papers = 8): Popularity Only

	Top 1			Top 3		
	Grads	Undergrads		Grads	Undergrads	
		From Grads	From UG		From Grads	From UG
Aid_learning	3.170	3.291 /3.111 ($p = .015$)	3.344 /3.178 ($p = 0.027$)	3.075	3.153 /3.089 ($p = 0.086$)	3.184 /3.089 ($p = 0.027$)
Value_added	3.359	3.428 /3.267 ($p = .034$)	3.550 /3.267 ($p = 0.001$)	3.295	3.328 /3.185 ($p = 0.003$)	3.402 /3.244 ($p = 0.001$)
Overall	3.214	3.252 /3.022 ($p = .011$)	3.302 /3.044 ($p = 0.002$)	3.090	3.109 /2.970 ($p = 0.007$)	3.086 /2.978 ($p = 0.028$)

Another result indicated that the performances of PopUM-CF from graduates to undergraduates (column “From Grads”) are always worse than those from undergraduates to other undergraduates (column “From UG”) for both value_added and overall ratings (Table 5). The results suggest that collaborative filtering within the same group (from undergraduates to other undergraduates) works better than it does across groups (from graduates to undergraduates). This conclusion is supported by the results of 1D-CF in Table 7 where the recommendation across each group is mostly lower than those within each group (the only exceptional case is for the top 3 overall rating, where $3.086 < 3.109$).

With respect to 1D-CF, we observed that most results were significantly higher than the best-case benchmark as shown by the low p -value (< 0.05), which means 1D-CF can provide effective recommendations. In fact, the results here are better than those for graduates in terms of the overall ratings, which had less significant gains ($p = .33$ and $p = .29$ for top 1 and top 3 respectively, see also Table 5).

General Discussion

Due to the limited number of students, papers, and other learning restrictions, a tutor cannot simply require students to read many papers in order to stock the database. As a result, the majority of typical recommender systems cannot work well in the pedagogical domain. Our study attempts to bridge the gap by proposing a set of recommendation mechanisms that do work well in this domain. Through extensive experimental studies, we discovered three key findings to answer our three broad research questions:

1. It is worth the trouble of complicating the traditional single-dimensional recommendation by incorporating contextual information to inform the recommendations. This can be achieved by adopting approaches such as PopUM-CF for a number of learning contexts.

2. The simple 1D-CF performs equally well for both graduates and undergraduates. Hence, paper recommendation systems can be effectively used for inexperienced learners along with experienced ones.
3. Gathering recommendations from across different learning groups (with different pedagogical features and learning goals) is less effective than it is to gather them from within the same learning groups, especially with collaborative filtering.

Our two studies suggest that learners make judgments based on the information and cognitive authority of a paper (Rieh, 2002). Hence, appropriately designing a RS and evaluating its performance is key to improving system performance. Experiment results suggest that a user model-based CF works well with some nonpersonalized methods, including PopCon2D and PopUM-CF. Although the computations in PopUM-CF and PopCon2D are more sophisticated than other CF-based approaches, under certain circumstances they help improve system performance. Our experiments and evaluation also highlight the importance of appropriately designing a RS and evaluating its performance. As Herlocker et al. (2004) declared, “accurate recommendations alone do not guarantee users of recommender systems an effective and satisfying experience. Instead, systems are useful to the extent that they help users complete their tasks.”

This study attempts to bridge the gap between commercial recommendation systems and educational ones by proposing a set of recommendation mechanisms that work well in the learning domain. Through our experiments and prototypical analysis, we were able to draw a number of important conclusions regarding the design and evaluation of these techniques. In spite of this, more work needs to be done to further our understanding of this complex issue.

For instance, one of our biggest challenges was the difficulty of testing the effectiveness or appropriateness of a recommendation method due to a low number of available ratings. Testing the method with more students in two or three more semesters may not be helpful because the results would still not be enough to draw conclusions as strong as those found in other domains where there can be millions of ratings.

Hence, we are eager to see different institutions collaborate to use the system in a more distributed manner and on a larger scale (as it is very difficult to achieve accurate recommendations using only one class in one institution each time). Through this broader collaboration, our future work will include the design of a MovieLens-like benchmark database to use as a test bed on which more algorithms can be investigated, including ours.

In addition, as shown in the analysis, the papers are related to software engineering; hence, it is not appropriate to generalize the results to make recommendations to students in other classes. Papers may exhibit more technical difficulties due to their inherent features in some subjects (e.g., in artificial intelligence or data mining), and students may also be different when they begin to take a course, which in turn affects those pedagogical factors considered in the performance of the recommender system. For instance, in user model-based

CF, we intend to match user interest and background knowledge to a paper's topics, and the difficulty lies in making recommendations that ensure students have enough background knowledge to understand the paper. If, for example, the recommendation is made to students taking an artificial intelligence course, due to the nature of papers in this topic tutors should consider the technical difficulty of each paper carefully by giving more significant weight to paper difficulty (in the SE course, and due to the overall nature of the papers, the weight was suggested to be zero after performance comparisons) to reflect the importance of that variable.

In this study, we only investigated making recommendations for research articles. Nevertheless, we believe that the study can be extended to other educational resources, such as learning objects, chapters with different topics in a digital book, tutorial materials, and so on. In fact, almost all educational resources can be regarded as learning objects with different granularity, situated environments, and purposes. Hence, the various recommendation mechanisms can be extended to make personalized recommendations for learning objects to individual learners with different needs.

Conclusion

Obviously, finding a “good” paper is not trivial: It is not as simple as finding out whether the user will either accept the recommended items or not; rather, it is a multiple step process that typically entails users navigating the paper collection, understanding the recommended items, seeing what others like/dislike, and making decisions. Therefore, a future research goal we derived from this study is to design RSs for different kinds of social navigation in order to study their impact on user behavior and how over time user behavior gives feedback to influence the system's performance. Additionally, we realized that one of the biggest challenges is the difficulty of testing the effectiveness or appropriateness of a recommendation method due to a low number of available ratings. Testing the method with more students in two or three more semesters may not be helpful because the results would still not be enough to draw conclusions as strong as those from other domains where there can be millions of ratings. Hence, we are eager to see the collaborations between different institutions to use the system in a more distributed and large-scale fashion. Through this broader collaboration, our ambition is to create a MovieLens-like benchmark database for the education domain to use as a test bed on which more algorithms can be investigated.

Acknowledgments

We would like to thank three anonymous reviewers for their constructive comments and the editors for their time and valuable remarks. This work was supported by Konkuk University in 2012.

References

- Adomavicius, G., Mobasher, B., Ricci, F., & Tuzhilin, A. (2011). Context-aware recommender systems. *AI Magazine*, 32(3), 67–80.
- Bernt, F. L., & Bugbee, A. C. (1993). Study practices and attitudes related to academic success in a distance learning programme. *Distance Education*, 4(1), 97–112.
- Brusilovsky, P., Farzan, R., & Ahn, J. (2005). Comprehensive personalized information access in an educational digital library. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries* (pp. 9–18). New York, NY: ACM.
- Custard, M., & Sumner, T. (2005). Using machine learning to support quality judgment. *D-Lib Magazine*, 11(10). Retrieved from <http://www.dlib.org/dlib/october05/custard/10custard.html>
- Drachler, H., Hummel, H. G. K., & Koper, R. (2007). Recommendations for learners are different: Applying memory-based recommender system techniques to lifelong learning. In E. Duval, R. Klamma, & M. Wolper (Eds.), *Creating new learning experiences on a global scale: Second European Conference on Enhanced Technology Learning, EC-TEL* (pp.1-9). Crete, Greece.
- Gomez-Albarran, M., & Jimenez-Diaz, G. (2009). Recommendation and students' authoring in repositories of learning objects: A case-based reasoning approach. *International Journal of Emerging Technologies in Learning*, 4, 35–40.
- Gunawardana, A., & Shani, G. (2009). A survey of accuracy evaluation metrics of recommendation tasks. *Journal of Machine Learning Research*, 10, 2935–2962.
- Herlocker, J., Konstan, J., Terveen, L., & Riedl, J. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1), 5–53.
- Khribi, M. K., Jemni, M., & Nasraoui, O. (2009). Automatic recommendations for e-learning personalization based on web usage mining techniques and information retrieval. *Educational Technology & Society*, 12(4), 30–42.
- Lekakos, G., & Giaglis, G. (2006). Improving the prediction accuracy of recommendation algorithms: Approaches anchored on human factors. *Interacting with Computers*, 18(3), 410–431.
- Lemire, D., Boley, H., McGrath, S., & Ball, M. (2005). Collaborative filtering and inference rules for context-aware learning object recommendation. *International Journal of Interactive Technology and Smart Education*, 2, 179–188.
- Manouselis, N., Vuorikari, R., & Van Assche, F. (2010). Collaborative recommendation of e-learning resources: An experimental investigation. *Journal of Computer Assisted*

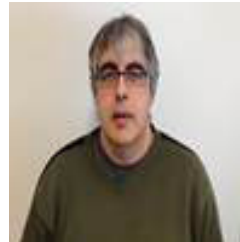
Learning, 26(4), 227–242.

- McNee, S., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S., Rashid, A., Konstan, J., & Riedl, J. (2002). On the recommending of citations for research papers. *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work, New Orleans, Louisiana* (pp. 116–125). New York, NY: ACM.
- Nadolski, R., Van den Berg, B., Berlanga, A., Drachslar, H., Hummel, H., Koper, R., & Sloep, P. (2009). Simulating lightweight personalised recommender systems in learning networks: A case for pedagogy-oriented and rating based hybrid recommendation strategies. *Journal of Artificial Societies and Social Simulation*, 12(14). Retrieved from <http://jasss.soc.surrey.ac.uk/12/1/4.html>
- Pazzani, M. (1999). A framework for collaborative, content-based, and demographic filtering. *Artificial Intelligence Review*, 1(5–6), 393–408.
- Recker, M., Walker, A., & Lawless, K. (2003). What do you recommend? Implementation and analyses of collaborative information filtering of web resources for education. *Instructional Science*, 31, 299–316.
- Rieh, S. Y. (2002). Judgment of information quality and cognitive authority in the web. *Journal of the American Society for Information Science and Technology*, 53(2), 145–161.
- Schein, A., Popescul, A., Unger, L. H., & Pennock, D. (2002). Methods and metrics for cold-start recommendations. *SIGIR '02, Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 253–260). New York, NY: ACM.
- Tang, T. Y. (2008). *The design and study of pedagogical paper recommendation* (Doctoral dissertation). University of Saskatchewan, Saskatoon.
- Tang, T. Y., & McCalla, G. (2009). A multi-dimensional paper recommender: Experiments and evaluation. *IEEE Internet Computing*, 13(4), 34–41.
- Torres, R., McNee, M., Abel, M., Konstan, J. A., & Riedl, J. (2004). Enhancing digital libraries with TechLens. *The Fourth ACM/IEEE Joint Conference on Digital Libraries (JCDL 2004)* (pp. 228–236).
- Winoto, P., & Tang, T. Y. (2010). The role of user mood in movie recommendations. *Expert Systems with Applications*, 37(8), 6086–6092.

Athabasca University 



Searching for and Positioning of Contextualized Learning Objects



Silvia Baldiris and Ramon Fabregat

University of Girona, Spain

Sabine Graf (not shown)

Athabasca University, Canada

Nestor Darío Duque Méndez

National University of Colombia, Colombia

Abstract

Learning object economies are marketplaces for the sharing and reuse of learning objects (LO). There are many motivations for stimulating the development of the LO economy. The main reason is the possibility of providing the right content, at the right time, to the right learner according to adequate quality standards in the context of a lifelong learning process; in fact, this is also the main objective of education. However, some barriers to the development of a LO economy, such as the granularity and editability of LO, must be overcome. Furthermore, some enablers, such as learning design generation and standards usage, must be promoted in order to enhance LO economy. For this article, we introduced the integration of distributed learning object repositories (DLOR) as sources of LO that could be placed in adaptive learning designs to assist teachers' design work. Two main issues presented as a result: how to access distributed LO and where to place the LO in the learning design. To address these issues, we introduced two processes: LORSE, a distributed LO searching process, and LOOK, a micro context-based positioning process, respectively. Using these processes, the teachers were able to reuse LO from different sources to semi-automatically generate an adaptive learning design without leaving their virtual environment. A layered evaluation yielded good results for the process of placing learning objects from controlled learning object repositories into a learning design, and permitting educators to define different open issues that must be covered when they use uncontrolled learning object repositories for this purpose. We verified the satisfaction users had with our solution.

Keywords: Learning design; learning objects economy; micro-context; similarity measures; word sense disambiguation

Introduction

Basic Concepts of Learning Objects Economy

Through the years, the concept of the learning object (LO) has been considered by many diverse and qualified people. The IEEE Learning Technology Standards Committee (LTSC, 2009), in its work on the Learning Object Metadata Standard (2002), defined a learning object as any element, digital or non-digital, that may be used for learning, education, or training. Such a definition categorizes almost everything as a learning object, but even so, not just anything is one. According to Polsani (2005), a LO needs to be accessible, reusable, and interoperable while also intended for a learning process.

Wiley (2000) reinforced the concept of reuse by introducing the definition for “object” from the object-oriented programming paradigm of computer science, where it is understood as a component that can be reused in multiple contexts. In this manner, a learning object is presented as a small instructional component that can be reused in different learning contexts when required. This definition is important to us because our study is based on the learning object economy (Duncan, 2004), where reuse is a key aspect.

Learning object economies are marketplaces for the sharing and reuse of LO. As in any economy, different actors play different roles. Ochoa (2008) identifies eight actors: market-makers, authors, resellers, publishers, teachers, end-users, assemblers, and regulators. Market-makers are researchers and trainers who provide support for LO interchanges with learning object repositories (LOR), open courseware sites, and learning object technologies. Authors, such as teachers or learning designers are LO creators. Resellers are those who have acquired the rights to exploit LO, for example, universities or private companies. Publishers put together and publish LO. Teachers use the LO for instructional purposes. End-users use LO for learning. Assemblers reuse small LO to construct more complex LO. Finally, regulators set the rules by which the sharing takes place.

Barriers to Assembling a Learning Object Economy

Offering a learning process that is available to all is a motivation for stimulating the development of the learning object economy. However, to ensure that this necessity becomes a reality, some barriers in the learning object economy must be overcome, as shown by Duncan (2004).

There are two main technical barriers to reusing LO: granularity and editability. Granularity refers to how complex a learning object should be. Wiley (2000) introduced two different viewpoints for deciding this: an efficiency and an instructional point of view. From the efficiency point of view, Wiley indicates that the decision regarding learning object granularity can be viewed as a trade-off: The possible benefits of reuse come at the expense of cataloguing. Conversely, from the instructional point of view, the major issues are the scope and sequence of the learning design.

Editability is important because any aspect of a learning object can be changed if it is avail-

able in a suitable form. If a LO is editable, its granularity can be modified. There are many distributed LO that are not editable; in fact, this is one of the most common excuses provided by teachers for not reusing LO.

Counting editable and open LO requires agreements among the LO economy actors. In particular, adequate author rights management would increase their confidence in distributing editable and open content. Implementing author tools to support LO editability, which would address the accessibility issues in the content, is one of the most important issues to meet for the successful establishment of this economy.

Barriers from the pedagogical view are basically related to the LO context. According to Dey and Abowd (2000), context is defined as any information that can be used to characterize the situation of an entity—in this case, the LO. Context in education is essential, but in practice, incorporating context in LO inhibits reuse. Addressing the context issues would allow instructors to use LO in different scenarios. Small granularity drives the context issues, and LO editability allows teachers to contextualize the LO according to the learners' needs.

Enablers of the Learning Object Economy

Along with the barriers, some enablers must be promoted in order to develop the learning object economy: learning design generation and standards promotion.

Learning design generation.

Learning design is a term coined by a pedagogical movement asking for more consistent approaches to describing and documenting teaching practices in order to facilitate communication and sharing, while also improving teaching practice. However, there is currently no standard definition for learning design (Koper & Yongwu, 2009). A well-accepted definition for the instructional design process is simple: the process that should be followed by teachers in order to plan and prepare instruction (Reigeluth, 1999). This process should address people's cognitive, emotional, social, and physical needs in an integral way. Given that LO are only content, to foster real learning experiences they need to be administered properly.

Adequate pedagogical theories and techniques need to be in place in order to insure that the LO have real impact (Koper & Yongwu, 2009).

Automatic learning design generation is an important topic in the research area of adaptive learning systems and technology-enhanced learning. Some researchers (Duque, Méndez, Ovalle Carranza, & Jiménez Builes, 2002; Morales, Castillo, & Fernández-Olivares, 2009; Ulrich & Melis, 2009; Karampiperis & Sampson, 2006; Hernández et al., 2009; Baldiris, Graf, & Fabregat, 2011) have proposed approaches to help teachers generate learning designs adjusted to user characteristics such as learning styles and competences, which is not an easy task, particularly for teachers. Actually, this problem implies that teachers need to know the different instructional theories; they also must be able to control the different user variables in the learning design construction, such as learning styles and competencies, among others. Furthermore, teachers need to know how to develop standardized

learning designs for the specific learning platform they use. Besides the personalization problem, another important issue for learning design generation is how to place learning objects from different learning object repositories into the generated designs.

Standards promotion.

If a global learning object economy is the goal, there must be common standards that every party agrees with to enable LO-sharing among heterogeneous systems (Ochoa, 2008). Important organizations and groups such as the IEEE Learning Technology Standards Committee (LTSC, 2009), the IMS Global Learning Consortium (n.d.), and the Dublin Core Metadata Initiative (n.d.) among others have proposed approaches for learning object standardization. Almost all elements, actors, and subprocesses of the educational process have been standardized. Baldiris, Santos, Fabregat, Jesus, and Boticario (2007) present an analysis of the different standards that have been accepted and validated internationally and the organizations involved in their creation.

Contributions and Outline of the Paper

In this paper, we aim to stimulate the enablers of the learning object economy to support the generation of standardized and adapted learning designs. Our investigation promotes LO reuse by encouraging instructors to access distributed learning object repositories (DLOR) as sources of LO with diverse granularity that could be elements in a generated learning design. Our proposal consists of two different parts: the distributed learning object metadata searching process (LORSE) and the micro-context-based positioning process (LOOK).

The distributed learning objects metadata searching process is a mechanism to promote reuse. It is supported by agent technologies, and its main purpose is to look for external LO that were not developed by the teachers which could be used as inputs in a learning design generation process. A micro-context-based positioning process analyzes a learning object's current micro-context (in the LOR) and future micro-contexts (in the learning design), using disambiguation techniques to establish the most promising micro-context for the LO in a learning design, and supports the placement of the object in its correct context.

The rest of this article is structured as follows. In Section 2, we introduce the distributed learning object metadata searching process. The third section describes the micro-context-based positioning process. In the fourth section we present the results of a layered evaluation. Finally, in the fifth section, we make some conclusions and comments on future research.

Section 2

LORSE: A Metadata Searcher of Open Learning Objects in Distributed Learning Repositories Based on Intelligent Agents

In order to facilitate the distributed learning object metadata search process, we developed LORSE, a distributed learning object metadata searcher, to promote reuse in the learning object economy. With LORSE, teachers, students, and external institutions can search in different learning object repositories using a unified interface. At the implementation level, LORSE (Baldiris, Bacca, Noguera Rojas, Guevara, & Fabregat, 2011) has been modelled as an independent set of JADE intelligent agents that collaborate to support users in the LO search process.

LORSE consists of two different types of agents: the directory facilitator agent and the specific search agent. The main purpose of this multiagent platform (Figure 1) is to deliver the most suitable LO according to the parameters provided by the user in a specific query.

The directory facilitator agent maintains a directory of tuples, where each item relates to one specific search service in a LOR with one specific search agent. Each specific search agent does the tasks of registering a new service in the directory facilitator agent and processing the requested services. When an external process needs to request a particular service on the platform, the external process must communicate with the directory facilitator agent to request the identifier of the agent in charge of a specific service. Specific search agents implement particular web clients by requesting search services in particular repositories. In Baldiris et al. (2011), we introduced an example of this application in three repositories (Merlot, Connexions, and UdG). In this article, we introduce an extension of LORSE that includes six additional services: DalSpace, Deep Blue, DLESE, ARIADNE, SMETE, and GATEWAY. The extended architecture of LORSE is shown in Figure 1.

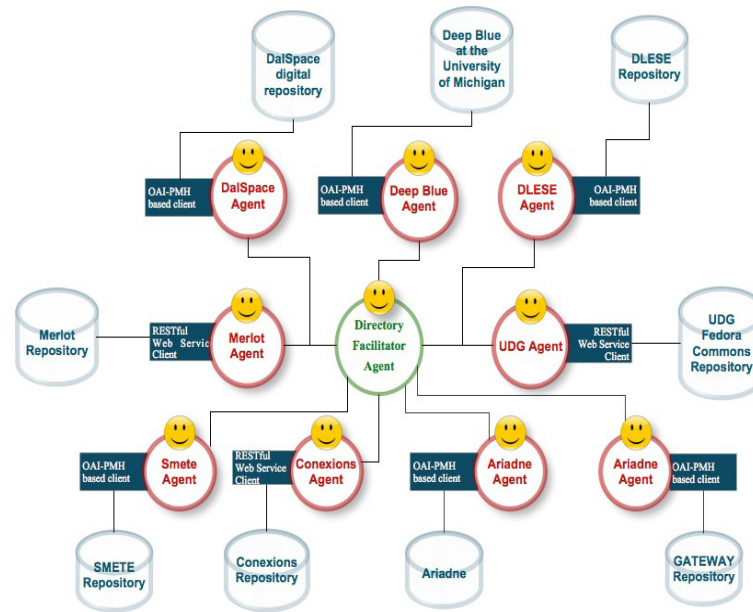


Figure 1. LORSE Multiagent platform.

When the Merlot agent (the specific search agent in charge of integration in the Merlot repository) is born, the Merlot search service is registered to the directory facilitator agent in order to allow other agents or processes to locate and send requests to it. The Merlot agent is activated when a search request is received. Merlot's agent implements a particular behavior, a client for the RESTful web service offered by the Merlot repository. When a request is sent to the agent, according to the terms and conditions of the query, the agent performs a connection with the service, sending the corresponding parameters, and then obtains a response as an XML document (metadata). The implementation of both the Conexions and UDG Agent is similar to the one for the Merlot agent; they have behaviors designed to interact with the RESTful web service offered by these applications.

To integrate the DalSpace digital repository, the Deep Blue Repository from the University of Michigan, the DLESE Repository, ARIADNE, SMETE, and GATEWAY into the multiagent platform, we created an intelligent agent for each. This agent presents indexer behavior, using the OAI-PMH harvester protocol to index the categories (catalogues) and records in the categories (resources) of each particular repository. Each metadata resource is stored in a database as a tree. In this manner, the information is available for a search process.

In order to test the extended version of LORSE independently, we integrated our development in an OpenACS/dotLRN learning environment. For the integration process, it was necessary to install the LORSE client package on this platform, which implements a web service client upon .LRN in order to send requests to the LORSE multiagent platform and process its responses. This package offers a user interface that provides functionalities allowing users to search several repositories in a transparent way. Therefore, when teachers use the learning environment, they are able to search for LO in those repositories to enhance the activities designed in the platform without leaving the learning environment.

Section 3

LOOK: Micro-Context-Based Positioning Process for Open Learning Objects

The main purpose of this section is to provide an introduction for the micro-context-based positioning process LOOK, which aims to place learning objects previously found by LORSE in learning designs.

To achieve this objective, two different sources of information are available: (1) the information from LOR, particularly the catalogue or indexed mechanism of the LO, and the LO metadata; and (2) the available information provided by the teacher for the competence definition, which defines the appropriate knowledge that a person should possess and show in a specific context. The competence definition consists of four categories of information: competence general information, which provides general data about the competence; competence elements, which are smaller learning purposes that provide more specific and concrete learning process outcomes; didactical guidelines; and the competence context of application.

Competence elements describe the essential knowledge that students should use in a specific context to demonstrate that they have acquired new information, and competence evidence is a mechanism that measures students' levels of achievement in each particular competence element. Schum (1994) explained how the evidence coming from different sources can be evaluated. In our case, analysis of the evidence is related to the relevance of the learning object that will address what the teacher is looking for, which he or she has defined in the competence definition of the course. In the following section, we introduce the main topics of relevance.

Learning Object Relevance

Borlund (2003) mentioned three central conclusions from the nature of relevance and its role in information behavior:

Relevance is a multidimensional cognitive concept whose meaning is largely dependent on users' perceptions of information and their own information need situations;

Relevance is a dynamic concept that depends on users' judgments of quality of the relationship between information and information need at a certain point in time;

Relevance is a complex but systematic and measurable concept if approached conceptually and operationally from the user's perspective.

Saracevic (1996) distinguished between five basic types of relevance: (1) system or algorithmic relevance, which describes the relation between the query (terms) and the collection of information expressed by the information object(s); (2) a topical-like type, associated with aboutness or criterion; (3) pertinence or cognitive relevance, related to the information need as perceived by the user; (4) situational relevance, depending on the task interpretation; and (5) motivational and affective, which is goal-oriented.

Ochoa (2008) used a modified version of Saracevic's categories (eliminating the motivational and affective dimensions) as the basis to define a set of complete metrics for LO relevance identification. These metrics are shown in Table 1.

Table 1

Features of Created Metrics

Type	Metric	Description	Inputs
Topical relevance	Basic topical relevance (BT)	Number of times the object has been previously selected from the result list when the same (or similar) query terms have been used	Queries the system keeps records of.
	Course-similarity topical relevance (CST)	Number of times that LO in the list have been used in the universe of courses.	Courses
	Internal topical relevance (IT)	The sum of the value of the courses where it has been used.	Courses

Type	Metric	Description	Inputs
Personal relevance	Basic personal relevance ranking (BP)	Analysis of the characteristics of the previously used LO, in particular the relative frequencies of the different metadata field values.	Metadata from the learning object used for a particular user.
	User-similarity personal relevance ranking (USP)	Number of times similar users have reused the objects in the result list.	Information about a learning object's use and its metadata.
Situational relevance ranking metrics	Basic situational relevance ranking (BS)	Cosine distance between the TF-IDF vector of contextual keywords and the TF-IDF vector of words in the text field of the metadata.	Description of the course, lesson, or activity and the learning object metadata.
	Context similarity situational relevance ranking (CSS)	Analysis of the objects that have been already used under similar conditions. Frequencies for different fields in the LO metadata.	Information about learning object use and its metadata.

Learning Object Relevance in the Micro-Context

Automatic word sense disambiguation (WSD) has been an interest and concern since the earliest days of computer language treatment in the 1950s. It is defined as the association of a given word in a text or discourse with a definition or meaning distinguishable from other meanings potentially attributable to that word (Ide, 1997).

All disambiguation work involves matching the context of the instance of the word to be disambiguated with either information from an external knowledge source (knowledge-driven WSD), or information about the contexts of previously disambiguated instances of the word derived from corpora (data-driven or corpus-based WSD).

The assignment of senses to words is accomplished by relying on two major sources of information:

- the context of the word to be disambiguated in the broad sense, including information in the text or discourse in which the word appears, together with extra-linguistic information about the text;
- external knowledge sources, including lexical, encyclopedic resources (among others), and hand-devised knowledge sources, which provide data useful to associate words with meanings.

Most disambiguation works use the local context of a word occurrence as the primary information source for WSD. Local or “micro” context is generally considered to be some small window of words surrounding a word occurrence in a text or discourse, from a few words of the context to the entire sentence in which the target word appears.

We consider the micro-context of a learning object to be a part of the curricular structure where the learning object should be placed (the learning design to be generated).

Consider the curriculum structure in Table 2 that belongs to a course teaching Unified Modelling Language (UML), which was generated based upon the competence definition provided by a teacher.

Table 2

Part of a Curricular Structure of UML Course

<ul style="list-style-type: none"> • Introduction to UML <ul style="list-style-type: none"> ◦ Concept ◦ Diagrams ◦ Relation of UML with the Unified Process of Development • The models <ul style="list-style-type: none"> ◦ Use case diagrams <ul style="list-style-type: none"> ◦ Actors ◦ Use cases ◦ Relations ◦ Class diagrams ◦ Sequence diagrams ◦ Activity diagrams

We need to place the LO, which can be obtained from a preliminary search based on the mechanism provided by the LOR, or according to the metrics described in Table 1, in the structure from Table 2.

We analyzed two different possible micro-contexts, the micro-context of the LO in the repository structure (catalogue), where the LO is placed, and the micro-context of the LO in the curricular structure, where the LO will be placed. Comparing these possible micro-contexts, a user can decide the best location for the learning object in the learning design.

Then, the first step is to define the micro-context of each learning object (LO) to be placed and also the possible micro-context in the curriculum structure.

The micro-context where a LO is placed in a LOR catalogue is provided by equation 1.

$$loMicroContext(D, C) = SuperCategories(D, C) \cup SubCategories(D, C) \quad (1)$$

In equation 1, LO is the learning object, and C is the catalogue in the LOR. loMicroContext defines the LO micro-context in a particular LOR catalogue.

Table 3 shows the loMicroContext of one LO, Introduction to OMG’s Unified Modelling Language.

Table 3

Introduction to OMG’s UML

Science and Technology <ul style="list-style-type: none"> • Computer Science <ul style="list-style-type: none"> ◦ Programming Languages <ul style="list-style-type: none"> -LO Introduction to OMG’s Unified Modelling Language
--

cuMicroContext defines the possible micro-context in the curricular structure (CS) provided by the teacher. These possible micro-contexts are given by equation 2.

$$csMicroContext = \sum_i^N cuMicroContext(leafes) \quad (2)$$

The number of leaves in the CS defines the possible micro-context of the curricular struc-

ture. Three of the nine possible micro-contexts from Table 2 in the CS are shown in Table 4.

Table 4

Possible Micro-Context in the UML Course

First possible micro-context in the learning design
Unified modelling language
Introduction to UML
Concept
Second possible micro-context in the learning design
Unified Modelling Language
The models
Use case diagrams
Actors
Third possible micro-context in the learning design
Unified Modelling Language
The models
Class diagrams

Now, the second step is to calculate the similarity between the different CS micro-contexts and the LO micro-context in order to place the LO in the structure. For this step, we proposed the use of different metrics to calculate the similarity between the TF-IDF (term frequency-inverse document frequency) inferred vectors in the analyzed micro-context (CS and LO). We used similarity measures that have been extensively validated in information retrieval: the Dice coefficient and cosine distance (Dice, 1945).

The Dice coefficient compares the similarity between two vectors (Q and D) from 0 to 1, where 1 indicates identical vectors and 0 orthogonal vectors. Equation 3 shows Dice coefficient.

$$s = \frac{2|Q \cap D|}{|Q| + |D|} \tag{3}$$

Cosine distance varies between -1 and 1, where -1 means exactly the opposite, and 1 means exactly the same, with 0 usually indicating independence, and in-between values indicating intermediate similarity or dissimilarity.

$$s = \frac{Q \cdot D}{|Q| |D|} = \cos(\theta) \quad (4)$$

Equation 4 presents cosine distance. θ represents the angle between Q and D . Based on the results of the algorithms for metrics implementation, the LO will be placed in the micro-context of the CS most similar to the micro-context of the LO in the repository structure (catalogue).

Section 4: Evaluation

Description of the Proposed Evaluation Process

After implementing our solutions for searching and locating LO, we conducted an evaluation of our developments. As we mentioned in the introduction, this article introduces our solution for looking up learning objects in distributed learning object repositories and positioning them in the most promising micro-contexts of learning designs that will be generated in the future.

Brusilovsky, Karagiannidis, and Sampson (2001) reported that the layered evaluation for adaptive hypermedia systems was a good approach to use to completely validate the elements for this kind of system. We used a layered evaluation process to measure the results in our research because the most important associated decision process (place a learning object in a learning design structure) supports an adaptive mechanism (adaptive learning design generation process based on students' and teachers' preferences). According to the adaptive system evaluation theory, different layers should be considered in order to test all the elements of the adaptive system (Brusilovsky et al., 2001; Karagiannidis & Sampson, 2000; Brusilovsky & Sampson, 2004). We define the following set of evaluation layers for our study:

- 1) The decision-making evaluation layer, where the question is, Are the decisions about where the learning objects should be placed valid and meaningful for teachers?
- 2) The user satisfaction evaluation layer, where the question is, Does the proposed solution match with the teachers' expectations?

Test Course: Object-Oriented Design with UML

Object-Oriented Design with UML is a course offered by the University of Girona in the formal education system. The course is supposed to establish student competence in UML: "The student will be able to design object oriented software using the unified modelling language (UML). The student will identify the most adequate diagrams to support the specification of each step in the object oriented development process."

To complete this competence, five different competence elements and the associated competence knowledge were defined.

- First competence element: Student defines Unified Modelling Language and identifies its main associated diagrams. Competence knowledge: Unified Modelling Language and its diagrams.
- Second competence element: Student understands the concept of use case diagrams and their associated concepts, such as actors, inclusion, extension, and generalization. Competence knowledge: Use case diagrams.
- Third competence element: Student understands the concept of class diagrams and designs class diagrams considering users' requirements. Competence knowledge: Class diagrams.
- Fourth competence element: Student understands the concept of interaction diagrams, particularly sequence and collaboration diagrams. He or she expresses the dynamic view of the software using these diagrams. Competence knowledge: Interaction diagrams, sequence and collaboration diagrams.
- Fifth competence element: Student understands the concept of activity diagrams to construct activity flows. Competence Knowledge: Activity diagrams.

For this course, 87 open learning objects were constructed. These learning objects were placed in an instance of the Fedora Commons Repository available at University of Girona. The set of learning objects that supported the learning process included diverse types of atomic resources with specific pedagogical intentions. These included exercises, simulations, diagrams, figures, graphs, indices, slides, tables, narrative texts, experiments, problem statements, lectures, questionnaires, exams, and self-assessments. Furthermore, each learning object had one associated LOM metadata where the most relevant information about the learning object was defined by a labelling process.

The Decision-Making Evaluation Layer

The main purpose of this evaluation layer is to validate our process for placing learning objects from different learning object repositories in the curricular structure of a learning design.

According to the typologies from McGreal (2008) and Sampson (2011) of the learning object repositories involved in our research and the character of previously obtained results, we divided the testing scenarios into two different environments, an uncontrolled and a controlled environment.

The uncontrolled environment consisted of repositories with diverse levels of labelling, where learning objects have different degrees of granularity. This environment permitted us to verify the possibilities and limitations of our approach in uncontrolled repositories where metadata labelling is not defined or supervised.

The controlled environment consisted of repositories available at University of Girona where the labelling was previously defined using relevant information and the granularity

of the learning objects was also defined. This kind of environment permitted us to verify more accurately the precision of the proposed algorithms in a controlled set of learning objects and their metadata. Both environments shared the same testing course and, for this reason, the competence definition and the analyzed micro-context associated with the competence were the same for both environments.

First Scenario: An Uncontrolled Environment

Description

We used this scenario to validate our proposal for locating learning objects from different learning object repositories in the curricular structure of a learning design. The uncontrolled environment considered different learning objects repositories linked through the same interface provided by LORSE. The involved repositories were ARIADNE, Merlot, SMETE, and GATEWAY. Some learning objects in these repositories were labelled with LOM, others with Dublin Core, but in general with a small amount of information defined by the market-makers.

Method

We looked for the catalogue provided for each defined repository. We performed different kinds of searches in the defined repositories using diverse search criteria. The criteria were defined with the information provided for the metadata in each repository and the searching mechanism provided each one. Then we selected the 10 most relevant LO for our study.

Using the previous information, we constructed the LO micro-context (loMicroContext) in the repository in two different ways. The first one was built as described in the LOOK section above. The second one also considered the LO metadata as a part of the LO micro-context. This was necessary since in many cases the LO micro-context based on the LO catalogue was not significant for our study; the LO micro-context did not support the proposed similarity analysis.

The next step was building the micro-context in the curricular structure (cuMicroContext). We defined six micro-contexts: five different micro-contexts according to the five competence requirements defined in the course competencies list and a general course micro-context. This general course micro-context consisted of the title, description, and all the knowledge associated with the competence requirements.

With all the micro-contexts involved (loMicroContext and the cuMicroContext), we proceeded to compare them, calculating the similarity measurements among the micro-contexts. We calculated the similarity of each learning object to each curricular structure micro-context. Then, we consolidated an average similarity, grouping the learning objects according to the repository where the LO were placed.

Results and Conclusions

Table 5 shows the most relevant results of this study. The first column defines the different criteria used for searching in the considered learning object repositories. The same criteria was used to define the LO micro-contexts. Additional columns represent the results of the average similarity consolidation for the general course micro-context.

Let us introduce an example: 0.2368 is the average similarity measure calculated among the 10 learning objects retrieved using the metadata—in this case, abbreviated keywords from Merlot. For each learning object, the similarity of its micro-context was calculated with respect to the general course micro-context.

We do not show the analysis of the other partial curricular structure micro-contexts considering the competence knowledge because the similarity measures were very small and extremely close. This did not permit us to determine the most promising micro-context for a learning object.

One of the most important conclusions we drew from this study was that using the definitions from the provided catalogue for uncontrolled repositories to define the learning object micro-context in a new learning design is very difficult. That can be seen in row six of Table 5. The reason is simple: The catalogue definition is too general for the LOOK positioning process to place the learning objects in a micro-context defined by the competence. The micro-context of the catalogue does not meet the micro-context extracted from the competence definition.

Table 5

Analysis of General Competence Micro-context and Learning Object Micro-Context

Search criteria	Merlot	ARIADNE	GATEWAY	SMETE
Metadata (abbreviation of a keyword)	0.2368	0.2140	0.0092	0.2164
Metadata (one keyword)	0.1915	0.2501	0.0518	0.1236
Metadata (keywords and abbreviation of a keyword)	0.1915	0.2737	0.0338	0.1236
Metadata + short categories	0.0986	NA	NA	NA
Metadata + categories	0.1337	NA	NA	NA
Categories	0.0081	NA	NA	NA

This situation led us to redefine the micro-context of the learning object, as is shown in Equation 5.

$$\begin{aligned}
 loMicroContext(D, C) = & \\
 & SuperCategories(D, C) \\
 & USubCategories(D, C) \\
 & UMetadata
 \end{aligned} \quad (5)$$

However, similarity measures for both micro-contexts do not show a strong relationship, although a manual analysis of the resource content shows strong relationships for the educational process.

Second Scenario: Controlled Environment

Description

In order to test our proposal in a controlled environment, we prepared a complete course of Object-Oriented Design with UML. The main objective of this study was to analyze our approach's capacity for adequately placing the learning objects into a specific course structure. The starting point was the "correct" classification developed by an expert teacher. This means that a teacher told us how he or she put the objects into the proposed curricular structure.

Method

According to the information provided in the competence definition, a structure for the course was defined, as shown in Table 2. The teacher manually placed the 87 available objects in the structure defined for the course. In this way, we have defined a point for comparison. Five micro-contexts associated to the UML course curricular structure (csMicro-context) were defined. The micro-contexts of each learning object (loMicro-Context) in the UML course were defined.

Average similarity measurements between each loMicro-context and each cuMicro-context were calculated. This means that for each LO in the course, we compared its micro-context to the five defined micro-contexts of the curricular structure. Grouping the LO according to the classification provided by the expert teacher, we calculated the average similarity for each csMicro-context. Then, we compared the similarity of each set of learning objects to each csMicro-context.

Results and Conclusions

Tables 6 and 7 present the LOOK system's precision, placing the LO in the best curricular structure micro-context. The obtained results came from calculating the average similarity for each set of learning objects previously placed by teachers in a particular csMicro-context. The results show a correspondence between the teacher's classification and the LOOK process classification, and indicate that in general, LOOK places the LO in the best csMicro-context according to the teacher's opinion.

In Tables 6 and 7, the rows show the identified csMicro-contexts (introduction, activity diagram, class diagram, use case diagram, and interaction diagram) and the columns represent the micro-context where the teacher classifies the set of learning objects previously. The values in the table indicate the average similarity between the micro-context for each set of LO previously classified by the teachers and each csMicro-context.

For example, in the first column, we calculated the average similarity of the set of LO previously classified by a teacher in the introduction micro-context and each csMicro-context. In this way, the similarities are (0.2222) for the introduction micro-context, (0.1379) for the activity diagram micro-context, and (0.1194) for the class diagram micro-context and so on. We observed that the average similarity for the set of LO placed by the teacher in the introduction micro-context coincides with the highest similarity calculated by LOOK to the introduction csMicro-context, 0.2222. In this way, the decision LOOK made to place these LO in the introduction micro-context corresponds with the teacher's decision to position these LO in the introduction micro-context.

Table 6

DICE Analysis Results

Curricular Structure Micro-contexts	Analysis for the set of LO previously classified by the teacher in each curricular structure micro-contexts				
	Introduction	Activity Diagram	Class Diagram	Use Case Diagram	Interaction Diagram
Introduction	0.2222	0.0350	0.0714	0.0833	0.12
Activity Diagram	0.1379	0.4262	0.1	0.0769	0.1481
Class Diagram	0.1194	0.0571	0.2898	0.1311	0.1587
Use Case Diagram	0.1481	0.0350	0.0357	0.4166	0.08
Interaction Diagram	0.1666	0.0634	0.1935	0.1111	0.3214

Table 7

COSINE Analysis Results

Curricular Structure Micro-contexts	Analysis for the set of LO previously classified by the teacher in each curricular structure micro-contexts				
	Introduction	Activity Diagram	Class Diagram	User Case Diagram	Interaction Diagram
Introduction	0.3292	0.1297	0.0564	0.1171	0.2240
Activity Diagram	0.1428	0.5439	0.0734	0.1307	0.2138
Class Diagram	0.1526	0.1657	0.3331	0.2221	0.2775
Use Case Diagram	0.1439	0.0992	0.0216	0.5379	0.1543
Interaction Diagram	0.1683	0.2011	0.1616	0.1617	0.5452

In particular, Table 6 presents the results applying DICE similarity measure. DICE analysis generates a precision of 100%, which means the process has localized 100% of the set of learning objects in the adequate curricular structure micro-contexts. Nevertheless, CO-SINE analysis generates 100% precision with respect to the classification provided by the teacher.

In general, the results of the study presented in Tables 6 and 7 show a strong correspondence between the classifications provided by the teacher and the possible classifications based on the similarity measures provided by the algorithms. The low values observed in Tables 6 and 7 are predictable because of the nature of the information available in the two different micro-contexts. Therefore, some labels in the competence definition as well as some in the LO metadata could contain irrelevant but comparable information because of the purpose of each kind of information. Only the relevant words for both micro-contexts are actually important, and the values shown in previous tables capture this relevance while selecting the best object for each particular micro-context.

User Satisfaction Evaluation Layer

Description

Our main objective in this evaluation layer was to develop a qualitative study (Hernández Sampieri & Baptista Lucio, 2004) that would permit us to achieve a better understanding of potential opportunities for improving our approach and show us more effective ways to support this task. The strategy we used was to develop case studies, which permitted us to concentrate on a particular situation—in our case, the use of distributed learning objects for creating learning designs.

The analysis was based on interviews with teachers, case studies where we applied a gap model instrument (Hernández Sampieri & Baptista Lucio, 2004) to evaluate their satisfaction level. The gap model allowed us to capture the difference between the teachers' expectations and the satisfaction that they really obtained from the offered service.

The gap model was applied in a particular instrument (a survey) to measure user satisfaction for four aspects of our proposal:

- satisfaction with the searching process (SEQ1), that is, the possibility of searching in different distributed repositories in a unique environment;
- the usability of the tool, developed on a dotLRN platform, to integrate LORSE (SEQ2);
- satisfaction with the results offered by the search process (SEQ3);
- satisfaction with the possible location of LO in a curricular structure available for testing (SEQ4).

The instrument was applied to 15 teachers (cases) at the University of Girona, Spain, as a part of descriptive research, where teachers had the opportunity to test our proposed application. These instructors teach different courses at the university from different areas of knowledge: pedagogy, economy, law, psychology, tourism, and administration science. Some of these courses are already supported by a virtual learning environment (Moodle).

Methodology

We arranged sessions with teachers from University of Girona. The main researcher introduced teachers to the learning object repository environment, showing them some of the most important ones. The main researcher introduced LORSE, its functionality and integration into the dotLRN learning management system as a *portlet*. The teachers had the opportunity to conduct some searches using the system. The LOOK process was described to the teachers, who observed the possible learning objects included in the test course. A session of discussions and brainstorming was proposed to every teacher in order to gain their opinions about our research. They were very motivated in this session.

Results and Conclusions

The results presented in Figure 2 show a very close relationship between the importance perceived by the users referred to the evaluated issues and their satisfaction with the solution. One of the most important parts of the descriptive analysis was the conclusions and opinions highlighted by the teachers: They all thought the reuse of learning objects was a possibility to facilitate the virtual learning process because efforts from teachers at different universities might be united.

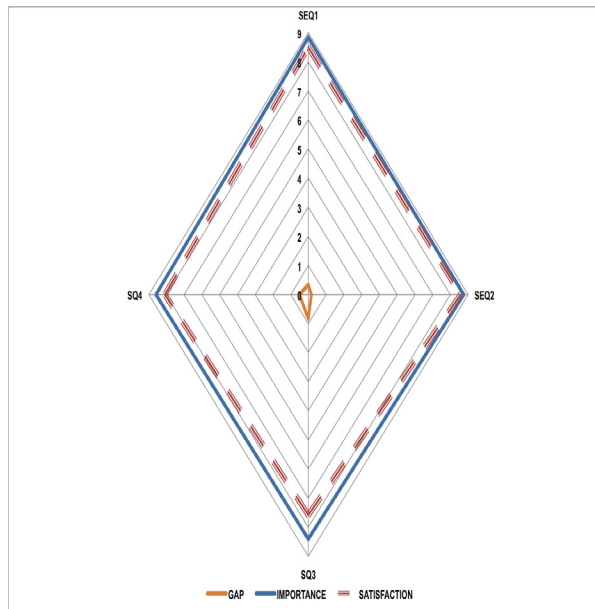


Figure 2. Teachers' Gap Model Results.

All the teachers emphasized the necessity of guaranteeing the quality of the selected learning objects to support learning design. For them, quality means that both the selected learning object should be contextualized for the teachers' and students' needs, and it must guarantee learning design quality.

According to the interviews from each teacher, we concluded that 60% of teachers consider it a good practice for universities to include in their strategic plans the creation of spaces to update teachers about the resources for learning and teaching available around the world and in their own institutions. Teachers think that much research and knowledge developed by important institutions is not well known in the academic context and, for this reason, their efforts may not be widely used by teachers. This is the case for the available and open learning object repositories.

Conclusions and Future Work

The main purpose of this article is to introduce our research into searching for learning objects in distributed learning object repositories and their positioning process in the most promising micro-contexts of future learning designs. Our solution includes the definition of two different processes: the distributed learning object metadata searching process (LORSE) and the micro-context-based positioning process (LOOK), which we introduced here.

We presented our results in two evaluation layers, the decision-making layer and the user satisfaction layer. The decision-making layer encouraged us to conclude that, on one hand, a search process for the LO over controlled LOR for feeding learning designs is a promising option. Learning objects selected and placed in the learning design meet the teachers'

opinions in a previous manual positioning process. In this process, the importance of the metadata labelling process and the competence definition has been demonstrated. On the other hand, the decision-making process for including learning objects from uncontrolled learning object repositories in semi-automatically generated learning designs is a difficult process. In fact, to achieve a viable solution with these repositories, the object metadata needs to be refined. Metadata available in the involved repositories currently has limited information.

To obtain a closer view of the teachers' satisfaction with our proposal, we used a user satisfaction evaluation layer. The results obtained with teachers from University of Girona permitted us to define some improvements from a user-centered design view. Although the results were promising and we obtained a high user satisfaction level, we also need to address some important elements.

Some teachers suggested improving the appearance of the learning design player because they believe it could be difficult to manage for the student. The teachers suggested simplifying both the LORSE and LOOK interfaces, in order to facilitate easy use of the programs and to improve the usability of our solution. Results obtained in the descriptive analysis stimulated the development of evaluation scenarios when the main issues were testing the usability and accessibility of the proposed solution.

Currently, our research interest is focused on some of the different issues identified in our research: A good way to improve our solution for uncontrolled learning objects repositories could be to develop a characterization for the learning object repositories using ontology. This will optimize the search process to obtain more contextualized LO. Characterizing learning object repositories using ontologies would allow us to add the necessary semantics that support the selection of the repositories for a specific design process. In particular, as a result of the evaluation we identified the necessity of the following knowledge: character and granularity of the LOR, technical details, and main knowledge areas (e.g., math and languages). Finally, we need to develop a usability and accessibility testing scenario in order to verify the facility of our solution to meet those user needs in more detail.

Acknowledgments

The authors acknowledge the support of NSERC and the European Commission for its support through the funding of the ALTERNATIVA Project (DCI-ALA/19.09.01/10/21526/245-575/ALFA III [2010] 88). They would also like to thank the Spanish government for its support through the funding of the Augmented Reality in Adaptive Learning Management Systems for All (ARrELS – TIN2011-23930) and the Catalan Government for its support through the European Social Fund (SGR-1202). Also thanks to the National Call of Strengthen Research Groups and Artistic Creation from the National University of Colombia 2010–1012, programme code 14163.

References

- Baldiris, S., Bacca, J. L., Noguera Rojas, A., Guevara, J. C., & Fabregat, R. (2011). LORSE: Intelligent meta-searcher of learning objects over distributed educational repositories based on intelligent agents. In *Frontiers in Education Conference (FIE), 2011* (pp. F1E1–F1E5).
- Baldiris, S., Graf, S., & Fabregat, R. (2011). Dynamic user modeling and adaptation based on learning styles for supporting semi-automatic generation of IMS learning design. *Proceedings of the Eleventh IEEE International Conference on Advanced Learning Technologies (ICALT '11)* (pp. 218–220).
- Baldiris, S., Santos, I.C. Fabregat, R., Jesus, G., & Boticario, G. (2007). Modelado de competencias en sistemas de gestión de aprendizajes. *3 Congreso Internacional sobre el Enfoque Basado en Competencias: Diseño Curricular por Competencias y Gestión de la Calidad del Aprendizaje*. Girona: University of Girona.
- Borlund, P. (2003). The concept of relevance in IR. *Journal of the American Society for Information Science and Technology*, *54*(10), 913–925.
- Brusilovsky, P., Karagiannidis, C., & Sampson, D. (2001). The benefits of layered evaluation of adaptive applications and services. In S. Weibelzahl, D. Chin, & G. Weber (Eds.), *Proceedings of Workshop on Empirical Evaluation of Adaptive Systems at the Eighth International Conference on User Modeling, UM2001* (pp. 1–8).
- Brusilovsky, P., & Sampson, D. (2004). Layered evaluation of adaptive learning systems. *International Journal of Continuing Engineering Education and Lifelong Learning*, *14*, 402–421.
- Dey, A. K., & Abowd, G. D. (2000, April 3). *Towards a better understanding of context and context-awareness*. Paper presented at the Workshop on the What, Who, Where, When, Why and How of Context-awareness, Conference on Human Factors in Computing Systems (CHI 2000), The Hague, Netherlands (pp. 1–6).
- Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*(3), 297–302.
- Dublin Core Metadata Initiative. About us. [Web page]. Retrieved from <https://dublincore.org/about-us>.
- Duncan, C. (2004). *Learning object economies: Barriers and drivers*. Paper presented at eLearnInternational, Edinburgh, Scotland, February 18–19.
- Duque Méndez, D. N., Ovalle Carranza, D. A., & Jiménez Builes, J. A. (2002). Artificial intelligence for automatic generation of customized courses. In E. Pearson & P. Bohman (Eds.), *World Conference on Educational Multimedia, Hypermedia and*

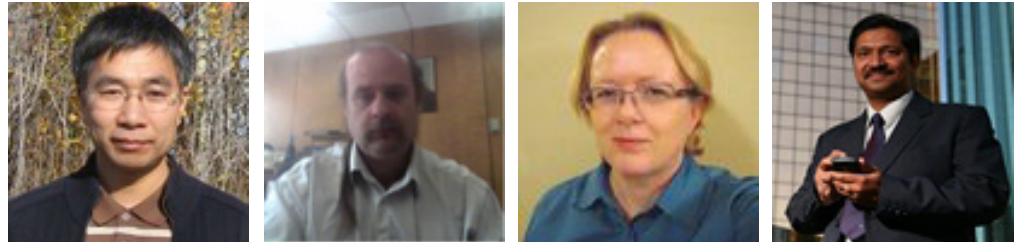
- Telecommunications* (pp. 2693–2698). Chesapeake, VA: AACE.
- Hernández, J., Baldiris, S., Santos, O., Huerva, D., Ramón, F., & Boticario, J. G. (2009). Conditional IMS LD generation using user modeling and planning techniques. *Proceedings of the Eighth IEEE International Conference on Advanced Learning Technologies (ICALT '09)* (pp. 228–232).
- Hernández Sampieri, R., & Baptista Lucio, P. (2004). *Metodología de la Investigación*. Mexico: McGraw Hill.
- Ide, N. (1997). Word sense disambiguation: The state of the art. *New York*, 1–41.
- IEEE, LTSC (Institute of Electrical and Electronics Engineers, Learning Technology Standards Committee). (2009, March 7). The IEEE LTSC temporary home page [Web log post]. Retrieved from <http://ieeeltsc.wordpress.com/>
- IEEE, LTSC. (2002). 1484.12.1-2002 IEEE Standard for Learning Object Metadata (version 1.2). Washington, DC: IEEE.
- IMS Global Learning Consortium. (n.d.) Background. [Web page]. Retrieved from <http://www.imsglobal.org/background>.
- Karagiannidis, C., & Sampson, D. (2000). Layered evaluation of adaptive applications and services. In *International Conference on Adaptive Hypermedia and Adaptive Web-based Systems, AH 2000* (pp. 343–346).
- Karampiperis, P., & Sampson, D. (2006). Adaptive learning objects sequencing for competence-based learning. *Proceedings of the Sixth IEEE International Conference on Advanced Learning Technologies (ICALT '06)* (pp. 136–138). Washington, DC: IEEE Computer Society.
- Koper, R., & Yongwu, M. (2009). Using the IMS LD standard to describe learning designs. In L. Lockyer, S. Bennett, S. Agostinho, & B. Harper (Eds.), *Handbook of research on learning design and learning objects: Issues, applications, and technologies* (pp. 41–86). New York, NY: Academic, 2009.
- McGreal, R. (2008). A typology of learning object repositories. In H. H. Adelsburger, Kins-huck, J. M. Pawlowski, & D. Sampson (Eds.), *Handbook on information technologies for education and training* (pp. 5–28). New York, NY: Springer.
- Morales, L., Castillo, L., & Fernández-Olivares, J. (2009, November). Planning for conditional learning routes. In A. H. Aguirre, R. M. Borja, C. A. Reyes Garcíá (Eds.), *MI-CAI 2009: Advances in Artificial Intelligence Lecture Notes in Computer Science, 5845* (pp. 384–396). New York, NY: Springer.
- Ochoa, X. (2008). *Learnometrics: Metrics for learning objects* (Doctoral dissertation,

- Katholieke Universiteit, Leuven). Retrieved from <http://ariadne.cti.espol.edu.ec/xavier/papers/ThesisFinal2.pdf>
- Polsani, P. R. (2005). Use and abuse of reusable learning objects. *Journal of Digital Information 3*(4), 1–10.
- Reigeluth, C. M. (1999). *Instructional design theories and models, a new paradigm of instructional theory*. Abingdon, UK: Laurence Erlbaum Associates.
- Sampson, D. (2011). From open educational resources to open learning design sharing educational practices in the knowledge cloud. In *Congreso Internacional de Ambientes de Aprendizaje Adaptativos y Accesibles | Hacia un sistema educativo comprometido con la diversidad*.
- Saracevic, T. (1996). Relevance reconsidered. In P. Ingwersen & N. O. Pore (Eds.), *CoLIS 2, Second International Conference on Conceptions of Library and Information Science: Integration in Perspective* (pp. 201–218). Copenhagen, Denmark: Royal School of Librarianship.
- Schum, D. A. (1994). *The evidential foundations of probabilistic reasoning*. Wiley Series in Systems Engineering and Management. Hoboken, NJ: John Wiley & Sons.
- Ullrich C., & Melis, E. (2009). Pedagogically founded courseware generation based on HTN-planning. *Expert Systems with Applications*, 36(5), 9319–9332.
- Wiley, D. A. (2000). Connecting learning objects to instructional design theory: A definition, a metaphor, and a taxonomy. *Learning Technology 2830*(435), 1–35.

Athabasca University 



Instructor-Aided Asynchronous Question Answering System for Online Education and Distance Learning



Dunwei Wen, John Cuzzola, Lorna Brown, and Dr. Kinshuk
Athabasca University, Canada

Abstract

Question answering systems have frequently been explored for educational use. However, their value was somewhat limited due to the quality of the answers returned to the student. Recent question answering (QA) research has started to incorporate deep natural language processing (NLP) in order to improve these answers. However, current NLP technology involves intensive computing and thus it is hard to meet the real-time demand of traditional search. This paper introduces a question answering (QA) system particularly suited for delayed-answered questions that are typical in certain asynchronous online and distance learning settings. We exploit the communication delay between student and instructor and propose a solution that integrates into an organization's existing learning management system. We present how our system fits into an online and distance learning situation and how it can better assist supporting students. The prototype system and its running results show the perspective and potential of this research.

Keywords: Automated question answering; natural language processing; information retrieval; LMS; distance education; online learning

Technology tools such as wikis, blogs, and podcasts can foster student interaction in online learning (Beldarrain, 2006). However the focus of instructional design should be on customization of content where the learner's needs become the center of attention (Reigeluth, 1999). Learners of all age groups are sensitive to the applicability of content and ease of use of an online learning portal. Denvir, Balmer, and Pleasence (2011) discovered that people aged 18-24, although they have grown up in the digital age, have reduced success in finding information online. Their results indicate that young people are less likely to utilize the Internet for obtaining advice, possibly due to frustration they experience when searching on the Internet. However, online learning has broadened the accessibility of education by the reduction of time-zone, employment commitment, and family obligation constraints, particularly with students who have left the traditional learning institutions. Nonetheless, the transition from traditional to online delivery isn't trivial for the student. Likewise, perceived usefulness and ease of use are critical factors in determining a teacher's acceptance and use of an educational portal (Pynoo, Balmer, van Braak, Duyck, Sijnave, & Duyck, 2012).

Question answering (QA) technology may be part of the solution. QA aims to automatically answer some of these questions in a way that the current search engines such as Google and Bing do not. A QA system attempts a deeper understanding of a query such that the retrieval of text/documents best answers the student's question. This is particularly difficult in the context of natural language where ambiguity in meaning is inherent in speech. Consider the following example query (Q) and the possible candidate answers (CA) available for selection.

Q: *Should I keep my dog away from eating flowers?*

CA1: More than 700 plants have been identified as producing physiologically active or toxic substances in sufficient amounts to cause harmful effects in animals.¹

CA2: Poisonous plants produce a variety of toxic substances and cause reactions ranging from mild nausea to death.²

CA3: Certain animal species may have a peculiar vulnerability to a potentially poisonous plant.³

CA4: A hot dog is one of my favorite things to eat.

Removing stop words and focusing on the core words of the query *dog*, *away*, *eating*, and *flowers*, one can see that CA4 would be the most likely answer when strictly a bag-of-words approach is used since none of the core words appear elsewhere. However, clearly CA4 is also the least desirable. Employing further mechanical techniques such as stemming and lemmatization would also strengthen the choice of CA4 as the best answer (the word "eating" would be stemmed to "eat").

1 www.humanesociety.org/animals/resources/tips/plants_poisonous_to_pets.html

2 www.calgaryhumane.ca/document.doc?id=78

3 www.treehouseanimals.org/Tree%20House%20Site-E-CPP.htm

This leads to the need of more thorough natural language processing (NLP). While many NLP techniques such as sentence segmentation, tokenization, name entity recognition, chunking, POS tagging, and syntactic parsing become necessary parts of QA systems, deeper NLP techniques including syntactic parsing and semantic role labeling (SRL) have increasingly attracted the attention of QA researchers. Semantic role labeling (SRL) is one method of achieving a deeper understanding of the query to better match the question with the candidate answers. Role labeling maps parts of a sentence to abstract themes (frames) and their supporting metadata (frame elements).

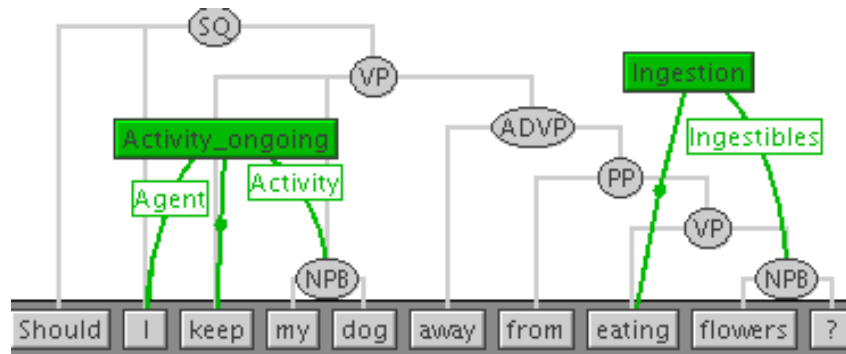


Figure 1. An example of SRL.

Figure 1 demonstrates the idea of SRL using the *FrameNet* (Baker, Fillmore, & Lowe, 1998) lexical semantic database. Two frames (concepts) have been identified in the query: [activity_ongoing] and [ingestion]. The frame elements provide supporting data for the frame. For example, flowers have been identified as the entity being ingested. However, contrary to intuition, these deeper NLP processing techniques are not always of much help as we will discuss below. Further research into the effective use of them in QA is needed. Moreover, as we can imagine, syntactic and semantic processing are much more computationally intensive (requiring more computing time and resources for performing the tasks) than other lower level NLP processing tasks and thus impose a big burden to online or synchronous QA that is supposed to provide answers to a question immediately.

Our current research is focused on these two problems and aims to reach our own solutions to produce more efficient and relevant answers particularly in an educational setting. To face the intensive computing challenge, we exploit the communication latency between student and instructor in online and distance education environments, and propose an asynchronous QA framework that makes the deep NLP analysis workable and acceptable in reality. This paper complements our recent work in exploiting semantic roles for asynchronous question answering (Wen, Cuzzola, Brown, & Kinshuk, 2012), with emphasis on the student and instructor interaction through the QA interface. We extend our contribution through new evaluation results using a large commercial corpus and compare the results with our alternative solution further validating our proposed technique.

An Asynchronous QA System

In this section, QA’s processing time issue when integrating deep NLP is examined. We begin by discussing the time complexity challenge then present an overview of the major parts of a QA system and finally offer the particulars of our proposed QA prototype system. We describe how our implementation of such a system can fit into an online learning situation despite this time obstacle.

The Trouble with Time

Modern QA systems incorporate NLP such as syntactic and semantic query analysis in an attempt to find the most relevant answers. Syntactic techniques examine the individual elements and structure of the sentence. These strategies include part-of-speech (POS) tagging where words are classed into their grammatical categories – noun, pronoun, verb, adverb, adjective, and so forth. Parse trees extend POS tagging with a structural representation of the sentence. In contrast, semantic techniques attempt to derive the context or meaning of a sentence. Such strategies include named entity (NE) recognition where parts of a sentence are generalized as entities such as country, person, date, temperature, weight, and so forth. Other more recent strategies involve labeling semantic roles and collecting them as reusable resources, such as *PropBank* (Palmer, Gildea & Kingsbury, 2005) that centers around the verb of the sentence and *FrameNet* which deconstructs the sentence into concepts (semantic frames) and supporting frame elements. The NLP research community has developed many very useful software tools to make the above techniques accessible and applicable to researchers for constructing new NLP projects. Such powerful tools include *ASSERT* (Pradhan, Ward, Hacioglu, & Martin, 2004), *SEMAFOR* (Das, Schneider, Chen, & Smith, 2010), and *Shalmaneser* (Erk & Pado, 2006), which are readily exploitable to those who can mitigate the computational time obstacle.

Table 1

Speed Comparison between ASSERT/SEMAFOR/Shalmaneser per 1,000 Sentences

ASSERT (PropBank)	SEMAFOR (FrameNet)	Shalmaneser (FrameNet)
19 Minutes	30 Minutes	77 Minutes

Table 1 shows the observed processing time required to semantically parse 1,000 sentences from the Reuters 21,578 corpus using PropBank and FrameNet. The speed comparisons were done using the software of ASSERT, SEMAFOR, and Shalmaneser on a 2.4 GHz Intel Core2 Duo with 8 GB RAM on a 64-bit Linux OS.

This overhead makes the use of semantic role labeling troublesome in synchronous QA where near-instantaneous answers are expected. Precomputation of the corpus in advance may avoid lengthy delays at the time of the query. The aforementioned Reuters corpus consisted of a total of 104,410 sentences which in the context of Table 1 may be feasible for

preprocessing and storage. However, a large corpus of millions of lines (and larger) makes beforehand labeling unpractical. Furthermore, this option is unavailable for dynamically changing content such as that accessible through search engines, Wikipedia, online forums, and similar World Wide Web artifacts.

Our research investigates the offline interaction between student and instructor that inherently involves a response delay between the time a question is asked and the time of an expected answer. It is within this delay where we propose a QA system that can benefit from semantic role labeling while side-stepping the time complexity problem.

The Components of a QA System

The components of a question answering type system are commonly comprised of four stages: *question analysis*, *query generation*, *search*, and *filtering/ranking*. A syntactic and semantic analysis of the question is performed to generate suitable queries for the search. The returned results also undergo syntactic and semantic evaluation to determine their relevancy with respect to the question and their ranking with respect to each other. Non-relevant answers are discarded (filtered). We utilized the open source framework *OpenEphyra* which provides the four stage pipeline in a modular and extensible implementation (Schlaefler, Giesemann, & Sautter, 2006). *OpenEphyra* marries *INDRI*'s very fast search engine for large lexicons with various syntactic capabilities that include part-of-speech tagging, sentence detection, and named entity recognition through the *OpenNLP* package (<http://opennlp.sourceforge.net>). Additionally, a lexical database of related concepts known as *WordNet* (Miller 1995) is also available for syntactic use. PropBank semantic analysis is performed through the ASSERT software. We are currently developing an interface module to either Shalmaneser or SEMAFOR that would give *OpenEphyra* the ability to perform subtler semantic analysis through FrameNet.

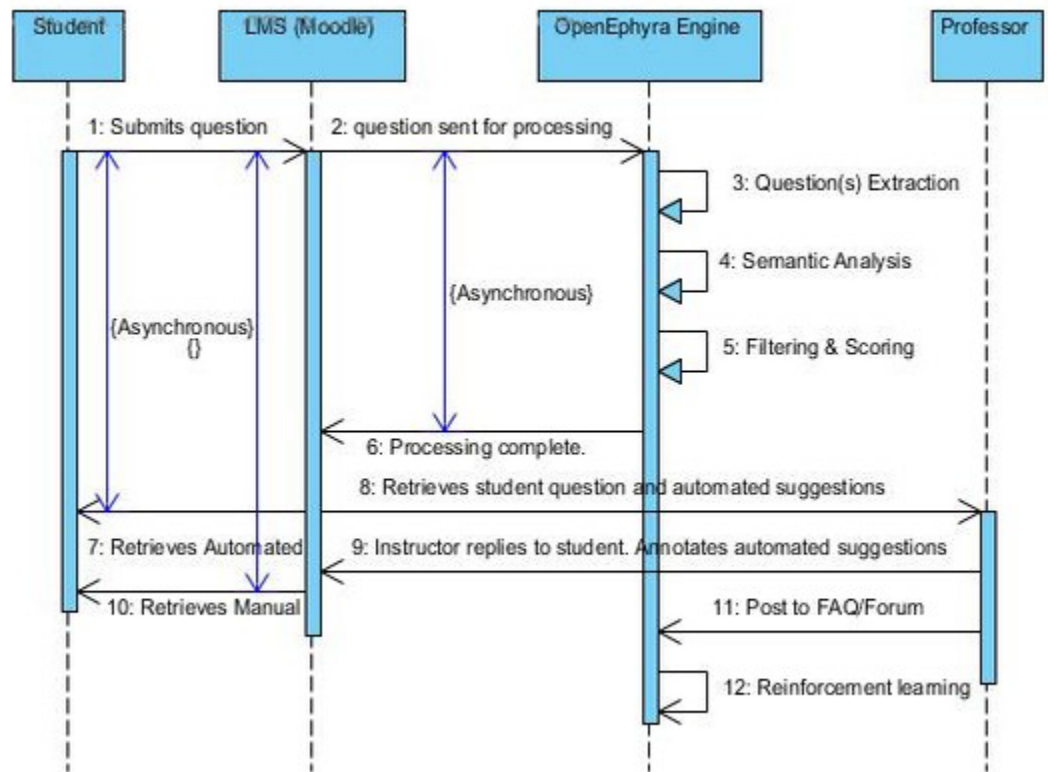


Figure 2. Asynchronous semantic QA model (Wen, D., Cuzzola, J., Brown, L. & Kinshuk, 2012).

Figure 2 outlines the proposed model. The primary actors involve the student, the learning management system (LMS), the OpenEphyra framework, and the course instructor (professor). A student first submits a question to the course instructor/tutor through an interface made available through the organization’s learning management system. This allows for seamless, integrated and familiar access for the student thus encouraging its use. We use Moodle (<http://moodle.org>) as the LMS, a popular choice for a significant number of educational organizations. Once a student’s question is posted, the OpenEphyra framework begins the process of question extraction, semantic analysis, and finally filtering and scoring of the top ten results (Steps 3-5). It is through this phase of the process that the proposed asynchronous solution is required. The time complexity of these three steps significantly exceeds the patience of any user expecting results as quickly as the traditional synchronous search engine (see Table 1). However, since this communication model between the time a student submits the question and the instructor’s response is anticipated to be delayed in an asynchronous learning environment, the problem of time complexity can be mitigated. Once the processing is complete, a ranked top 10 result, similar in output to that of a search engine, awaits the student and the course instructor in their LMS mailboxes for retrieval (Steps 7 and 8). The students may now investigate the automated results for potential answers to their questions while they wait for the instructors’ feedback. The instructor receives the student’s query as well as the automated suggested answers. The

instructor can give direct feedback to the student, annotate or modify the automated results list, and even give preference to specific result entries over other less-relevant automated suggestions. This instructor reply then awaits the student in his or her LMS mailbox as a follow up to the automated suggestions. The instructor may also decide that the revised questions/answers may be of use to other students in the class and can direct the LMS to post this information in the course FAQ or forum. Lastly, the model includes a reinforcement learning component for continuous improvement of the accuracy of the automated results (Step 12) by leveraging the annotations offered by the course instructor, a domain expert, to learn how the rankings of similar QA answers in the future should be adjusted.

Prototype Test Results and Discussion

This section introduces our proof-of-concept prototype of the proposed system and the preliminary results. This prototype was developed at Athabasca University for research and test purposes.

The QA Prototype

We have developed a prototype user interface as described in Figure 2 for Athabasca University. Figure 3 shows the Moodle plugin as displayed to the student. The plugin reveals the number of questions asked by the student that are still awaiting answers, the number of responses with unread automated answers (Step 7), and the number of replies from the course instructor yet to be read (Step 10).

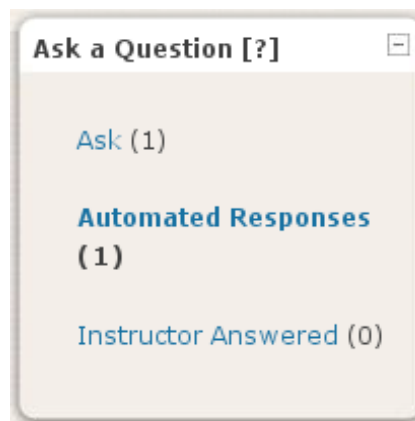


Figure 3. Moodle student plugin to QA system.

The student submits a question by clicking on “Ask” and entering a subject and message body to his/her question for the instructor. Figure 4 illustrates the question submission interface including a possible question asked by a student.

Athabasca University

Subject:

Essay Question

Enter your Question:

Hello Professor,

I hope you are well. I have a quick question with respect to our assignment. I am looking for reference material on global warming. Specifically, how does climate change affect the polar ice caps?

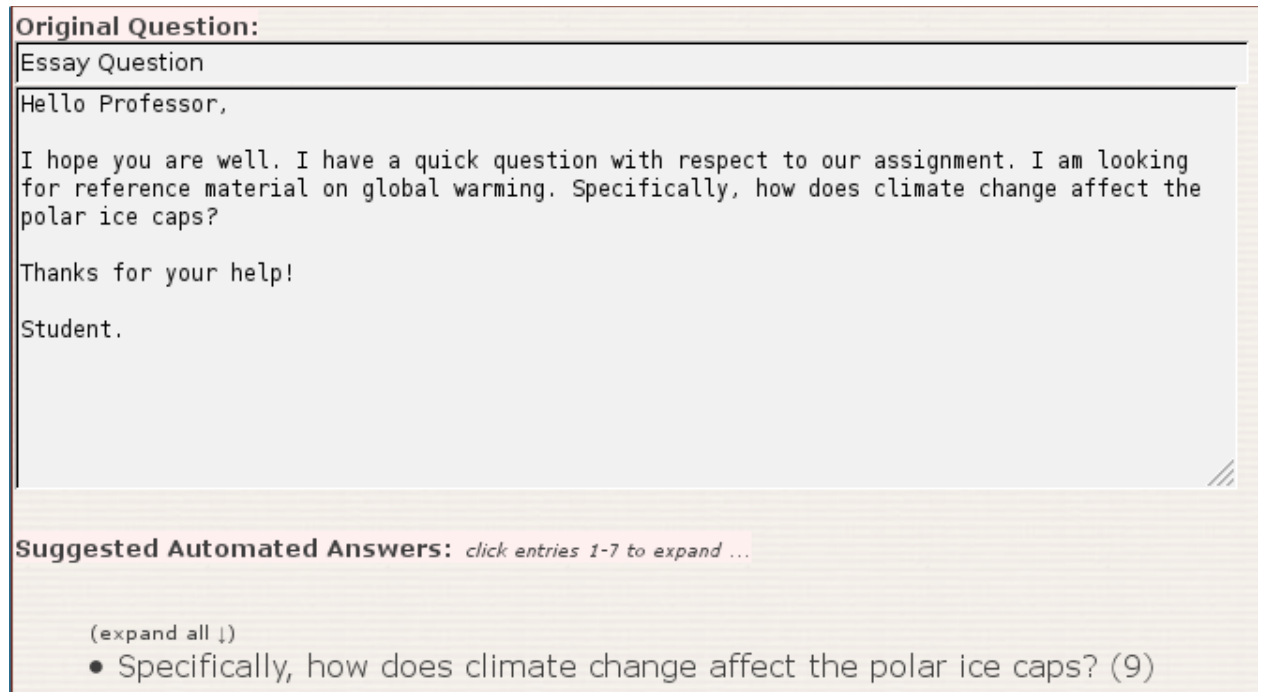
Thanks for your help!

Student.

SEND QUESTION TO COURSE INSTRUCTOR/TUTOR

Figure 4. Moodle interface for submitting a question to the QA system.

The simple interface resembles the composition form of a typical email program. Note that these types of questions pose an extra challenge to the QA system for a number of reasons. First, the colloquial nature of the message (i.e., “Hello Professor”) introduces informal language not often included in a query to a search engine. Second, multiple questions can be asked within a message that may be intermixed with nonquestions (i.e., “I hope you are well”). The system must identify the questions while ignoring the superfluous parts of the message. The task of precisely identifying questions is very much an open problem. Our implementation assumes that each sentence is a potential question. The message is decomposed into individual sentences using the sentence detector of OpenNLP; each sentence is then submitted for candidate answers. Once the automated suggested answers are formulated, the student is informed through the Moodle plugin (see Figure 3). The query portion of the message is identified and displayed to the student underneath his/her original message.



Original Question:

Essay Question

Hello Professor,

I hope you are well. I have a quick question with respect to our assignment. I am looking for reference material on global warming. Specifically, how does climate change affect the polar ice caps?

Thanks for your help!

Student.

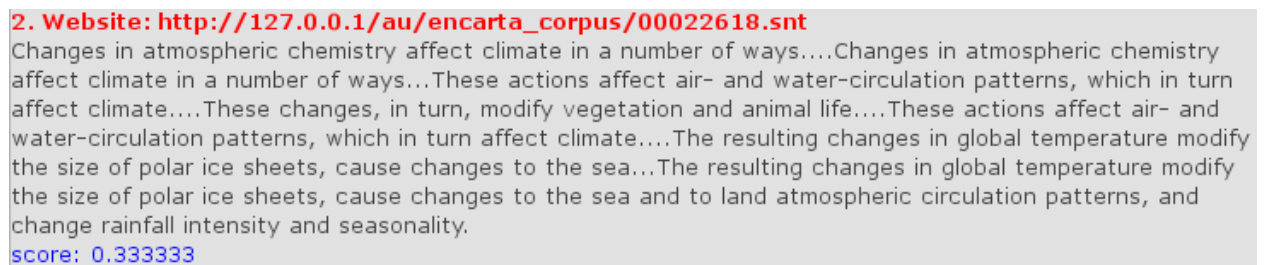
Suggested Automated Answers: *click entries 1-7 to expand ...*

(expand all ↓)

- Specifically, how does climate change affect the polar ice caps? (9)

Figure 5. Automated response returned by the QA system.

Figure 5 shows the primary user interface for retrieving the suggested answers from the automated system and from the course instructor. The number in the bracket (9) after the question indicates the count of retrieved documents that the system believes is relevant to the student's query. Clicking on the question reveals the list, which includes a summary paragraph, a relevancy score, and a link to the source document. Figure 6 shows the suggested second answer and its score from the list of nine answers.



2. Website: http://127.0.0.1/au/encarta_corpus/00022618.snt

Changes in atmospheric chemistry affect climate in a number of ways....Changes in atmospheric chemistry affect climate in a number of ways...These actions affect air- and water-circulation patterns, which in turn affect climate....These changes, in turn, modify vegetation and animal life....These actions affect air- and water-circulation patterns, which in turn affect climate....The resulting changes in global temperature modify the size of polar ice sheets, cause changes to the sea...The resulting changes in global temperature modify the size of polar ice sheets, cause changes to the sea and to land atmospheric circulation patterns, and change rainfall intensity and seasonality.

score: 0.333333

Figure 6. Suggested second answer from the list of nine.

A student can verify his/her question was fully considered by the QA system by clicking the "expand all" link which reveals the decomposition of the query, as shown in Figure 7.

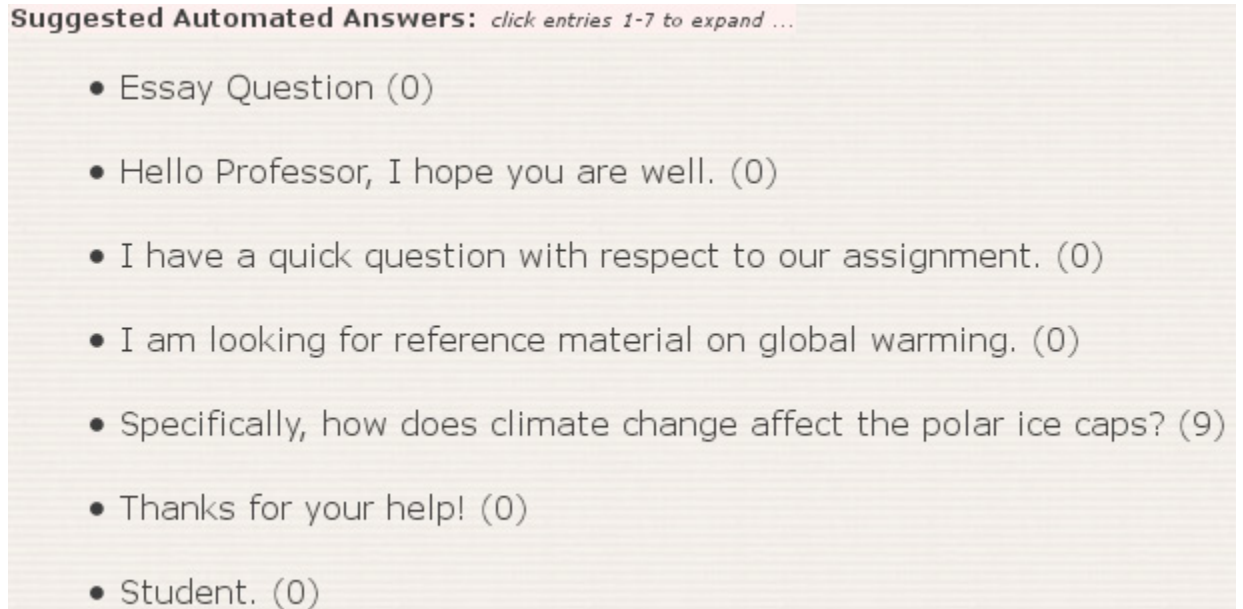


Figure 7. Submitted message split into sentences and answered by the QA system.

By default, sentences with zero answers are not displayed and this method is used as a simple filter to separate questions from anecdotal speech. In the future we plan to investigate state-of-the-art methods to differentiate questions from statements. In the interim, our simple no-zero answer filter policy has proven satisfactory. Also noteworthy is that the subject of the message is included as part of the query. This is intentional as in practice the subject line of a message often contains a synopsis of the message and usually excludes the colloquial extras that may cause confusion to the machine. Due to the concise nature of a subject line, we plan to incorporate the subject directly into each of the queries in an attempt to further improve the answers returned.

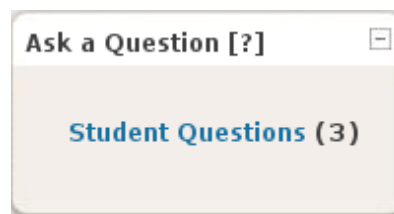


Figure 8. Moodle instructor plugin to the QA system.

Figure 8 shows the view of the Moodle QA from the perspective of the course instructor. A count of the number of unanswered questions pending is given in the brackets. The instructor can reply with his/her own message, annotate the automated answers, and/or post the reviewed solution directly to the course FAQ or forum.

QA Matching Theory and Methodology

With respect to the task of associating answers with questions there are typically two classes of queries: factoid based and information retrieval. Factoid queries give matter-of-fact type answers of a simple nature. Examples of such queries include “what are the first four

digits of PI?,” “who was the second president of the United States?,” and “what is the ratio of hydrogen to oxygen in water?”. In contrast, information retrieval questions attempt to locate supporting documents for a specific concept, topic, or idea. Such examples are: “how does the warming of the oceans affect weather patterns?” and “what is the link between smoking and heart attack?” For educational purposes, our focus is on information retrieval type queries. Unfortunately, OpenEphyra out-of-the-box is geared toward factoid answers only. Our first task was to modify the pipeline to allow for the latter. Our FrameNet inspired solution replaces the OpenEphyra pipeline. This pipeline includes our own answer filtering technique that involves incremental clustering of the top answers based on a dynamically adjusted boundary. Figure 9 shows the new pipeline as an alternative to the OpenEphyra pipeline.

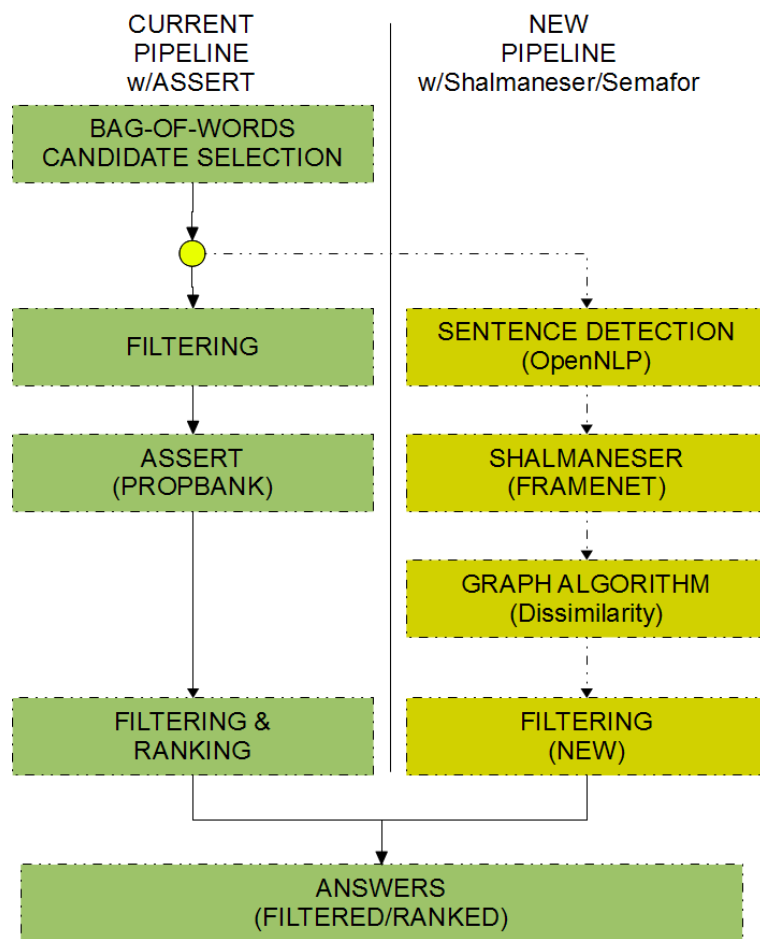


Figure 9. Alternative pipeline based on FrameNet.

The FrameNet lexical database consists of more than 1,000 semantic frames wherein each frame represents an event or concept. Associated with each frame are specific frame elements, sometimes referred to as roles, which represent the entities or participants of the frame. Additionally, the FrameNet database is relational in which frames may inherit from other frames.

Consider the act of eating which consists of what is being eaten (ingestibles) and who is doing the eating (ingestor). This act may or may not involve a utensil (instrument). In FrameNet, this event is represented by the frame [Ingestion] and includes frame elements ingestor, ingestibles, instrument, and others.

example: John_[Ingestion:ingestor] tried to eat his soup_[Ingestion:ingestibles] with a fork_[Ingestion.instrument].

Our graph algorithm involves locating matching frames, either directly or through inheritance relationships, in order to translate this intersection into a heuristic function that returns a dissimilarity measure between the query and candidate answers. To illustrate, consider the query “are potential and kinetic energy related” and two possible candidate answers: “there is ample data relating energy drinks to headaches” and “energy related to position is called potential energy.” Although both candidate answers contain the matching words *energy* and *relate*, the candidate answer “energy related to position is called potential energy” is clearly preferred. In order to resolve this ambiguity, our algorithm first performs a FrameNet semantic analysis of the query and candidate answers as shown in Figure 10. The scoring is obtained through a relative distance calculation between weighted paths from matching words of the query and candidate answer that share the closest intersecting frame. The further apart the intersection occurred, the less similar an answer is to the question.

These paths incorporate the inheritance model of FrameNet where each super-frame is a more generalized abstraction of a child frame concept. The edges are weighted with values αv^r where (v) is some arbitrarily chosen base raised to an incrementally larger exponent (r) at higher graph levels. Hence, weights at lower levels (more specific frames) are smaller than those at higher levels (generalized frames). A learning vector (α), derived through a supervised machine learning algorithm and a training set, provides weight adjustments for a better fitted model. Path values are computed through the product of the traversed edges and log normalized to account for large exponential values. Figure 11 illustrates such paths constructed from the semantic analysis of Figure 10.

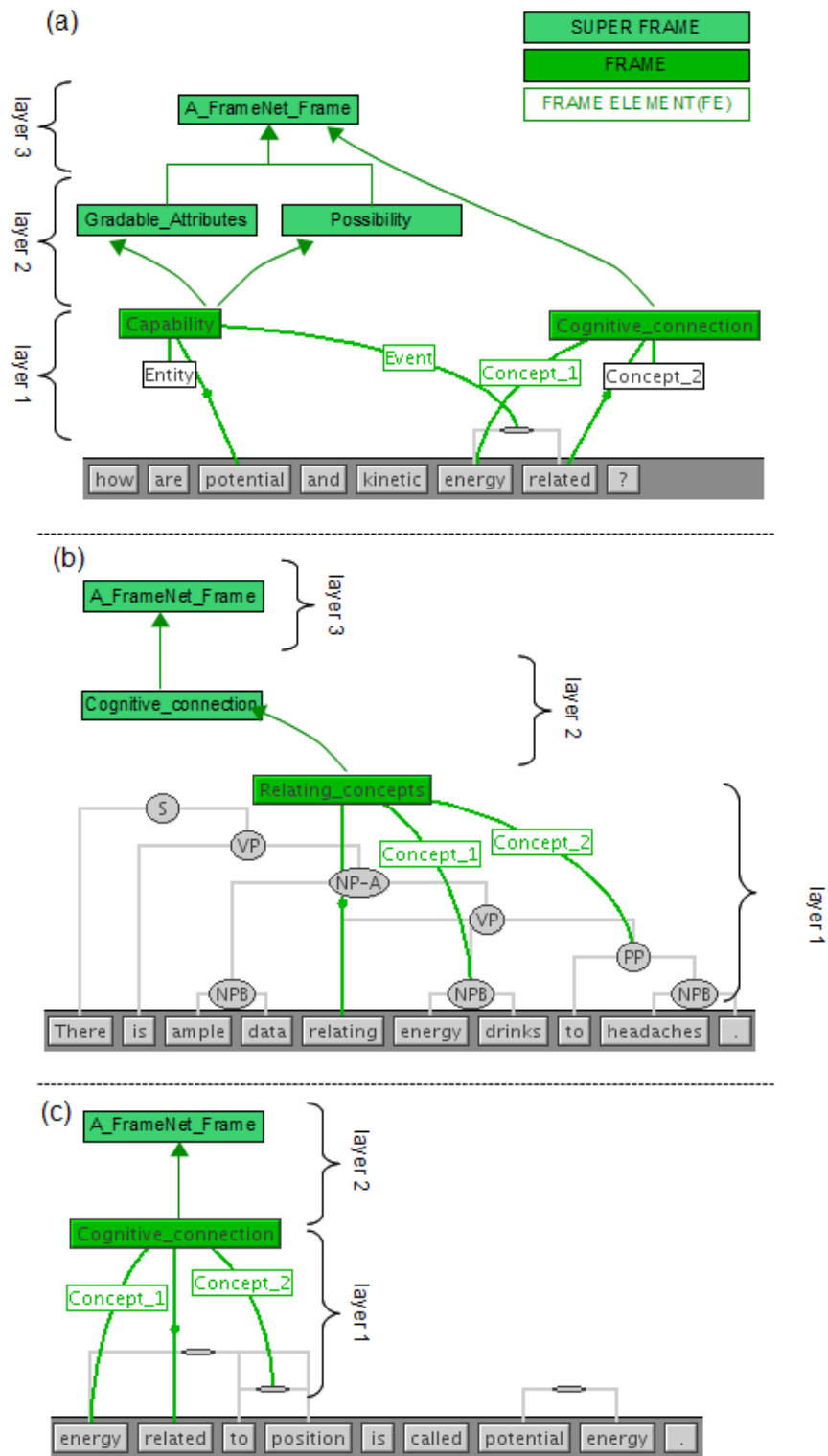


Figure 10. FrameNet semantic role labeling of query (a) and candidate answers (b), (c).

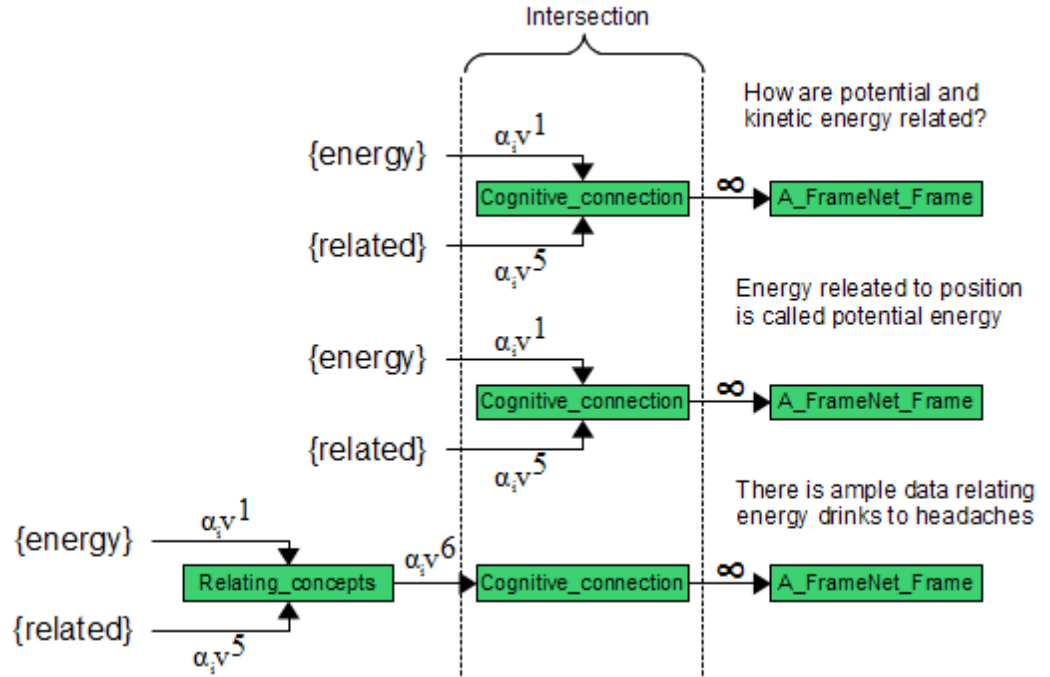


Figure 11. Paths with assigned edge weights from matching words to the topmost super-frame derived through Figure 10.

Notice that each edge weight along a path is exponentially larger than its predecessor with the exception of a special pseudo-frame called [A_FrameNet_Frame] which receives an edge weight of infinity. All frames are ultimately descendants of [A_FrameNet_Frame]. The infinity edge weight is assigned to discourage matching on this universal frame. The query and two candidate answers all share a common frame of [Cognitive_connection]. The heuristic value is computed from the product of the query's edges through the intersecting frame to the candidate's edges toward the matching words (energy/related). For example, the word energy in the preferred candidate answer has path value of $\alpha_i v^1 \alpha_i v^5$ compared to $\alpha_i v^1 \alpha_i v^6 \alpha_i v^5$ of the less desirable answer due to the extra frame of [Relating_concepts]. Since $\alpha_i v^1 \alpha_i v^5 < \alpha_i v^1 \alpha_i v^6 \alpha_i v^5$, the word energy in the answer “energy related to position is called potential energy” is favored over its use in “there is ample data relating energy drinks to headaches.” A similar calculation can be done for the matching word related. The goal is to find a minimal cost spanning tree between all constructed paths. In the next section of this paper, we test this algorithm on *Microsoft Question-Answering Research Corpus* (Microsoft Research, 2008) and test the PropBank solution with the *Introduction to Programming Using Java* online text (Eck, 2011). A more detailed technical introduction and experimental results of the algorithm will be reported in a separate paper.

QA Matching Results

We began preliminary testing on the effectiveness of OpenEphyra with PropBank style semantic filtering. We utilized the Microsoft Research QACorpus which was the basis of *Microsoft Encarta 98* digital encyclopedia published from 1993 to 2009. This corpus comprised 1,400 individual questions with each question containing multiple correct answers

for a total of 10,000+ QA pairs. The corpus is further divided into 32,715 individual files with each file containing a number of paragraphs. Overall, the corpus provided 163 megabytes of textual data for experimentation.

We evaluated our prototype by asking 33 information retrieval type questions and recording the hit rate (correct answers), miss rate (incorrect answers), and the positional ranking of the best answer. Table 2 gives the top and bottom 12 queries based on the number of relevant answers returned.

Table 2

Top/Bottom Twelve Questions

QUERY	HIT	MISS	BEST RANK
How is calculus used?	8	2	2
How do insects transmit disease?	8	2	1
How is pollution contributing to global warming?	7	3	1
How does diet improve your health?	7	3	8
what dogs originated in germany?	6	4	1
How is magnetism related to electricity?	4	3	3
↑ TOP 6 ↓			
↓ BOTTOM 6 ↓			
How is religion affected peoples attitudes toward birth control?	1	9	9
How has the space program contributed to everyday life?	0	9	N/A
Why do onions make your eyes water?	0	3	N/A
How is population explosion affecting the earth's resources?	0	10	N/A
How does the warming of the oceans affect weather?	0	4	N/A
How are potential and kinetic energy related?	0	0	No answers

Although the preliminary answers given by OpenEphyra with ASSERT showed promise, the overall precision needs improvement before deployment in an educational setting. Hence, our research is now focused on increasing the hit rate, lowering the miss rate, and improving the positional rankings of the best answers. These observations were the catalyst for our own filtering and scoring algorithms based on FrameNet semantic role labeling and graph theory matching.

To demonstrate we took the query “how are potential and kinetic energy related?” that produced no answers in the existing PropBank pipeline (see Table 2) and subjected it to our proposed FrameNet alternative. In the first step of the pipeline, 1,368 potential candidate answers to the query were selected through a simple bag-of-words approach. The candidates were syntactically parsed with frame elements assigned and inheritance chain calculated as described in the QA Matching Theory and Methodology section of this paper.

After completion of the similarity scoring, 29 answers survived our filtering stage of the pipeline. The top five answers and their dissimilarity scores are given in Table 3. For comparison, the ranking of the bottom five answers are also given; note that lower scores are better (less dissimilar) to the question asked.

Table 3

Top/Bottom Ten Answers to the Question “How are Potential and Kinetic Energy Related?”

RANK	ANSWER	SCORE
1	The law, which states that the sum of kinetic energy, potential energy, and thermal energy in a closed system remains constant, is now generally known as the first law of thermodynamics.	15.95594
2	Not until the action of molecules in a system was better understood by the development of the kinetic theory could this internal energy be related to the sum of the kinetic energies of all the molecules making up the system.	15.95594
3	Electric potential is related to electrical potential energy.	15.95594
4	Boltzmann Constant in physics, fundamental constant, designated k, that relates the average kinetic energy of particles in a gas to the temperature of the gas.	15.95594
5	Energy associated with motion is known as kinetic energy, and energy related to position is called potential energy.	29.77145
↑ TOP 5 ↑		
↓ BOTTOM 5 ↓		
1364	production, the excess energy is imparted to the electron pair as kinetic energy.	175.7519
1365	Endothermic reactions are always associated with the spreading, or the dissociation, of molecules.	175.7519
1366	If a gun is fired, the potential energy of the gunpowder is	186.738
1367	In the eye itself, winds are usually light and skies are partly cloudy.	186.738
1368	Such a body has an excess of electrons.	186.738

Our next phase of testing involved establishing how such a system, incorporated into the LMS of an educational organization, could improve learning particularly in the online/distance situations where face-to-face instructor interaction is infrequent or unavailable. In this test, the corpus consisted of the online introductory Java textbook comprised of 326 web pages freely available through a Creative Commons License. Table 4 gives an example of the types of questions a student may ask while taking this course. Such expected questions, given the nature of the course, may be “how do I compile my program?” or “how do I use annotations?”.

Similar to Table 2, the results show how many of the suggested answers are relevant to the question and the positional ranking of the best answer amongst the returned candidates. The queries were executed under the modified PropBank pipeline with mixed results with respect to hit, miss, and best rank. For example, when asked “how do I compile my program?,” 7 out of the 10 returned answers were considered satisfactory for the question (70% precision) but the best answer of the set was found (ranked) in the tenth position.

Nonetheless, a satisfactory (correct) answer was given somewhere in the suggested top-10. With the ability of the course instructor to annotate and/or modify the list, the student’s question should be adequately answered. We believe, as in our previous tests with the Microsoft Encarta corpus, that the alternate FrameNet pipeline will yield improvements. Further testing is ongoing.

Table 4

Commonly Asked Questions for an Introductory Course on Java Programming

Relevancy	Question -- [Rank] Best Answer
7 / 10	<p><i>Q1: How do I compile my program?</i></p> <p>[10] That section contains information about how to compile and run Java programs, and you might want to take a look at it before trying to write and use your own programs.</p>
8 / 9	<p><i>Q2: How do I use recursion?</i></p> <p>[8] Recursive algorithms work by reducing a complex problem into smaller problems which can solved either directly or by applying the same algorithm “recursively.” RGB</p>
6 / 10	<p><i>Q3: How can I use graphics?</i></p> <p>[2] If OSC is of type BufferedImage, then the method OSC.getGraphics() returns an object of type Graphics that can be used for drawing on the image....An off-screen canvas is a region of the computer’s memory that can be used as a drawing surface...</p>
1 / 5	<p><i>Q4: How do I use annotations?</i></p> <p>[1] Some annotations are used to make the programmer’s intent more explicit....For example, @Override is a standard annotation that can be used to annotate method definitions....Instead of writing that code by hand, a programmer could use an annotation to mark the variables that are part of the state that is to be saved....An annotation used in this way is an aid to writing correct programs, since the programmer can be warned about a potential error in advance, instead of having to hunt it down later as a bug</p>

Research Motivation

The motivation for this research is inspired by offline communication and forum posts that instructors frequently encounter with their students.

Hello Professor:

I need help with an assignment question:

“Analyze the UML model as described in Figure 4 of the textbook and give a solution.”

I could not find this figure in the textbook, nor the assigned reading. Where is it?

Figure 12. An example of a student’s forum post.

Figure 12 contains some key observations. First, the student's query is very specific. Although any popular search engine will provide links to resources on the topic of UML, the student's question is considerably more focused in asking for a seemingly missing figure related to a textbook reading assigned to this course. In order to accommodate this degree of specificity, deeper natural processing techniques such as semantic role labeling are necessary. Computational complexity, as shown in Table 1, would suggest this technology would normally be prohibitive. However, our proposed offline asynchronous implementation is tractable and consequently can provide an educational benefit within this context. A second observation is although the question is directed to the course instructor ("Hello Professor"), it has nonetheless been posted to the course forum. The reason is two-fold: to ask his peers in the event someone may know the answer or for the benefit of his peers in the circumstance that others may have the same question. Our prototype system specifically addresses this by giving the instructor the option to submit his/her comments along with the annotated auto-responses directly to the forum.

Our objective is to produce a QA system of sufficient accuracy of benefit to assist students and teachers. Our preliminary tests results have shown that the existing pipeline of the OpenEphyra framework is insufficient for this task. Consequently, our theoretical work focuses on a modified pipeline (Figure 9) that utilizes FrameNet SRL and weighted spanning trees (Figure 10/11). Systematically, we have introduced features in our implementation to further aid in the prototype's accuracy. The instructor has the option to not allow the system to respond to the student with its answers until the instructor has had an opportunity to review the system's suggestions and provide his/her corrections (see Figure 2 sequence no. 7 disabled). This allows the proposed system to operate either in unsupervised, semi-supervised, or fully supervised mode at the discretion of the instructor.

We have also incorporated a re-enforcement learning step (Figure 2 sequence no. 12) for perpetual accuracy improvements facilitated through the adjusted edge weight vector (α) of the model (Figure 11). Furthermore, the constrained domain of the course curriculum further aids in accuracy by limiting the search to only content available within the online offering.

Finally, it is worth noting that students are seeking assistance from search engines frequently, arguably more so in an online distance education situation than in the traditional face-to-face paradigm. Consequently, a QA system that is focused on the course material and moderated by the instructor offers advantages over the less restrictive, and sometimes inaccurate, sources of online information such as search engines and Wikipedia.

Related Works

A recent survey by Shuib et al. (2010) on 129 postgraduate computer science students found that students were having considerable difficulty in finding information appropriate to their learning style using the search tools available. It is also clear that not all available searchable content is created equal when measured against its educational value. In a case

study involving the use of a digital library in a middle school, Abbas et al. (2002) observed the usefulness of search systems varied based on the type of classes and differing student achievement levels. They conclude that an educationally useful search engine is more than seeking on-topic documents of interest but also is an organizational and collaboration tool that teaches iteration and refinement processes often leading to more than a single 'correct' answer.

Marshall et al. (2006) examined information retrieval in education using a digital library environment known as GetSmart. This system successfully augmented traditional search with concept mapping. Of 60 university students surveyed taking an undergraduate computing course, 86% reported that the marriage of concept mapping with search was "very valuable" or "somewhat helpful" in their queries.

Curlango-Rosas et al. (2011) developed an intelligent agent specifically for the retrieval of learning objects. A learning object (LO) was defined as "any entity, digital or non-digital, that may be used for learning, education or training [IEEE 2002]." Obviously, an LO is of special interest to educators over other forms of web content. Their "Learning Object Search Tool Enhancer" (LOBSTER) demonstrated that 96% of piloting teachers found suitable quality LOs compared to only an 80% success rate when using Google.

Martin and Leon (2012) proposed a digital library search for teachers that leveraged semantic and natural language processing. Their system made extensive use of case-based reasoning technology to construct a searchable ontology they named OntoFAMA. In the survey 50 engineering students were asked to rank the relevancy of suggested LOs by both OntoFAMA and Google; 85.4% of retrieved LOs from OntoFAMA were considered of acceptable or better quality compared to only 78.5% when using Google (a measure of precision). Only 14.4% of OntoFAMA's suggestions were considered poor quality compared to Google's 21.3% (a measure of recall).

A supervised learning approach to searching was investigated by Prates and Siqueira (2011). They used information extraction to create a training set that forms a baseline to the appropriateness of a retrieved document in a specific educational context. A teacher selects segments from available sources deemed as representative of the content he/she finds suitable. A student's query is expanded by using additional relevant terms as learned through the baseline before submission to a web search engine. Empirical tests showed that queries expanded in this manner gave better precision than their original nonexpanded counterparts.

In education a large body of information remains underutilized due to lack of an effective information retrieval system. Mittal, Gupta, Kumar, and Kashyap (2005) recognized this and they devised a QA system to search for information stored in PowerPoint slides, FAQs, and e-books. Feng et al. (2006) investigated a QA like discussion-bot, using a remote agent to provide answers to students' discussion board questions. Their results highlighted the importance of QA in online and distance education.

Feng, Shaw, Kim, and Hovy (2006) examined an intelligent robot that intercepted posts

to an educational forum and volunteered its own answers. This was similar to our own QA system, but lacked deeper semantic processing techniques. This gave quick answers to questions but the quality of the responses suffered. Cong, Wang, Lin, Song, and Sun (2008) used QA in their research to retrieve answers from online forums. Their technique took into consideration whether the answer was by a known domain expert and weighed his/her response accordingly.

Marom and Zukerman (2009) experimented on corpus-based techniques for the automation of help-desk responses, using request-response e-mail dialogues between customers and employees of a large corporation. Help-desk e-mail correspondence contains a high level of repetition and redundancy, and responses to customers contain varying degrees of overlap. This repetition and overlap is also commonplace in an educational environment where multiple students may have the same question. Hence, our prototype allows the instructor to post the annotated answers to the course forum.

Surdeanu, Ciaramita, and Zaragoza (2011) created a re-ranking model for non-factoid answer ranking using question-answer pairs from Yahoo! The results of the experiment show that semantic roles can improve ad hoc retrieval on a large set of non-factoid open-domain questions. Their paper provided compelling evidence that complex linguistic features such as semantic roles have a significant impact on large-scale information retrieval tasks and consequently can be a major benefit in online education.

Shen and Lapta (2007) introduced a general framework for answer extraction which exploited the semantic role annotations in FrameNet. The results were promising and show that a graph-based answer extraction model can effectively incorporate FrameNet style role semantic information. To the best of our knowledge, our proposed asynchronous QA system has not been implemented.

Conclusion

The Internet has provided educational opportunities in ways that were unavailable just 10 years ago. Online and distance learning has extended these opportunities past the traditional campus-based classes into students' homes, workplaces, and smart phones. Learning is now self-paced with the instructor acting as a facilitator and mentor. However, this anytime/anywhere access has presented a challenge in answering students' questions particularly in a medium where time zones are irrelevant. Our research contribution aims to develop an asynchronous QA system to fit student needs and support student and teacher participation in the learning process. Instructor feedback can validate the student's choice or provide instruction in choosing a suitable answer. Our system can assist with this validation. We also hope this research leads to improved communications between student and teacher and to lessen the frustrations a student may encounter in a distance learning environment where real-time access to the instructor may be absent. Our future work involves continued improvement to the question answering accuracy using semantic role labeling as well as enhancements to the Moodle user-interface to the extent that our working prototype

can be evaluated in an actual online course. Future papers will report on the progress of the QA accuracy and include evaluations from teachers and students who participated in the Moodle pilot.

Acknowledgement

The authors acknowledge the support of NSERC, iCORE, Xerox, and the research related gift funding by Mr. A. Markin. This work was also supported by a Research Incentive Grant (RIG) of Athabasca University.

References

- Abbas, J., Norris, C., & Soloway, E. (2002). Middle school children's use of the ARTEMIS digital library. In *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital libraries (JCDL '02)* (pp. 98-105). New York, USA: ACM.
- Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1* (pp. 86-90). Association for Computational Linguistics.
- Beldarrain, Y. (2006). Distance education trends: Integrating new technologies to foster student interaction and collaboration. *Distance Education*, 27(2), 139-153.
- Cong, G., Wang, L., Lin, C.-Y., Song, Y.-I., & Sun, Y. (2008). Finding question-answer pairs from online forums. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 467-474). ACM.
- Curlango-Rosas C., Ponce, G., & Lopez-Morteo, G. (2011). A specialized search assistant for learning objects. *ACM Trans. Web*, 5, 4, Article 21.
- Das, D., Schneider, N., Chen, D., & Smith, N. (2010). *SEMAFOR 1.0: A probabilistic frame-semantic parser*. Language Technologies Institute, School of Computer Science, Carnegie Mellon University.
- Denvir, C., Balmer, N. J., & Pleasence, P. (2011). Surfing the web - recreation or resource? Exploring how young people in the UK use the Internet as an advice portal for problems with a legal dimension. *Interacting with Computers*, 23(1), 96-104.
- Eck, D. J. (2011). *Introduction to programming using Java* (6th ed.) Retrieved from <http://math.hws.edu/javanotes/>
- Erk, K., & Pado, S. (2006). Shalmaneser - a flexible toolbox for semantic role assignment. In *Proceedings of LREC 2006* (pp. 527-532).
- Feng, D., Shaw, E., Kim, J., & Hovy, E. (2006). An intelligent discussion-bot for answering student queries in threaded discussions. In *Proceedings of the 11th International Conference on Intelligent User Interfaces* (pp. 171-177). ACM.
- Marom, Y., & Zukerman, I. (2009). An empirical study of corpus-based response automation methods for an e-mail-based help-desk domain. *Computational Linguistics*, 35(4), 597-635.
- Marshall, B., Chen, H., Shen R., & Fox, E. (2006). Moving digital libraries into the student learning space: The GetSmart experience. *J. Educ. Resour. Comput.* 6, 1, Article 2.

- Martin, A., & Leon, C. (2012, April). An intelligent e-learning scenario for knowledge retrieval. *Global Engineering Education Conference (EDUCON), 2012 IEEE* (pp.1-6).
- Microsoft Research (2008). *Microsoft research - question-answering corpus*. Retrieved April 24, 2012 from <http://research.microsoft.com/en-us/downloads/88c0021c-328a-4148-a158-a42d7331c6cf>
- Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, 38(11), 39-41.
- Mittal, A., Gupta, S., Kumar, P., & Kashyap, S. (2005). A fully automatic question-answering system for intelligent search in e-learning documents. *International Journal on E-Learning*, 4(1), 149-166.
- Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1), 71-106.
- Pradhan, S., Ward, W., Hacioglu, K., & Martin, J. H. (2004). Shallow semantic parsing using support vector machines. In *Human Language Technology Conference/ North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*.
- Prates, J., & Siqueira, S. (2011). Using educational resources to improve the efficiency of web searches for additional learning material. In *Proceedings of the 2011 IEEE 11th International Conference on Advanced Learning Technologies (ICALT '11)* (pp. 563-567). Washington, DC, USA: IEEE Computer Society.
- Punyakanok, V., Roth, D., & Yih, W. (2008). The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics*, 34(2), 257-287.
- Pynoo, B., Tondeur, J., van Braak, J., Duyck, W., Sijnave, B., & Duyck, P. (2012). Teachers' acceptance and use of an educational portal. *Computers & Education*, 58(4), 1308-1317.
- Reigeluth, C. M. (1999). What is instructional design theory? In C. M. Reigeluth (Ed.), *Instructional design theories and models: A new paradigm of instructional theory* (Vol. 2, pp. 5-29). Mahwah, NJ: Lawrence Erlbaum Associates.
- Schlaefler, N., Gieselmann, P., & Sautter, G. (2006). The Ephyra QA system at TREC 2006. *Proceedings of the 15th Text Retrieval Conference*.
- Shen, D., & Lapata, M. (2007). Using semantic roles to improve question answering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (pp. 12-21). Association for Computational Linguistics.

- Shuib, M., Liyana, N., Noorhidawati, A., Ismail, B., & Hafiz, M. (2010, Dec.). The use of information retrieval tools: A study of computer science postgraduate students. *Science and Social Research (CSSR), 2010 International Conference* (pp.379-384).
- Surdeanu, M., Ciaramita, M. & Zaragoza, H. (2011). Learning to rank answers to non-factoid questions from web collections. *Computational Linguistics, 37*, 351-383.
- Wen, D., Cuzzola, J., Brown, L., & Kinshuk (2012). Exploiting semantic roles for asynchronous question answering in an educational setting. In L. Kosseim & D. Inkpen (Eds.), *Advances in artificial intelligence - Canadian AI 2012, LNAI 7310* (pp. 374-379). Germany: Springer-Verlag.

Athabasca University 



Drawing Dynamic Geometry Figures Online with Natural Language for Junior High School Geometry



Wing-Kwong Wong,¹ Sheng-Kai Yin,² and Chang-Zhe Yang¹

¹National Yunlin University of Science & Technology, Taiwan, R.O.C.

²Mingdao University, Taiwan, R.O.C.

Abstract

This paper presents a tool for drawing dynamic geometric figures by understanding the texts of geometry problems. With the tool, teachers and students can construct dynamic geometric figures on a web page by inputting a geometry problem in natural language. First we need to build the knowledge base for understanding geometry problems. With the help of the knowledge base engine InfoMap, geometric concepts are extracted from an input text. The concepts are then used to output a multistep JavaSketchpad script, which constructs the dynamic geometry figure on a web page. Finally, the system outputs the script as an HTML document that can be visualized and read with an internet browser. Furthermore, a preliminary evaluation of the tool showed that it produced correct dynamic geometric figures for over 90% of problems from textbooks. With such high accuracy, the system produced by this study can support distance learning for geometry students as well as distance learning in producing geometry content for instructors.

Keywords: Natural language understanding; dynamic geometry; JavaSketchpad; geometry education; knowledge construction

Introduction

One approach to distance learning is to put learning content on web pages and ask learners, who may be in different geographical locations, to engage in some learning activities while browsing the pages. However, learning content takes a lot of time and human resources to produce, especially when the content involves geometric figures. Thus, producers of learning content would be very happy if they could use a tool to automatically convert an input text of domain concepts, such as a geometry problem, into a target output format such as some programming code to draw geometric figures on web pages. This article proposes a

methodology of how to convert input texts of domain concepts into some target output format. The methodology is illustrated by the construction of an actual system in converting input texts of geometry problems at a junior high school level into JavaSketchpad scripts to display dynamic geometry (DG) figures on a web page.

The National Council of Teachers of Mathematics (NCTM) has published two important documents on K-12 mathematics curriculum: *Curriculum and Evaluation Standards for School Mathematics* (1989) and *Principles and Standards for School Mathematics* (2000). The latter focused more on the skill of writing formal proofs of geometry (Knuth, 2002). Furthermore, mathematicians and educators agree that writing geometry proofs involves important skills that are difficult to learn (Koedinger, 1998; Whiteley, 1999).

Using geometry software increases the motivation of students in learning geometry. Some popular programs include Geometer's Sketchpad (GSP, <http://www.keypress.com/sketchpad/>), Cabri Geometry II (<http://www.cabri.com/v2/pages/en/index.php>), Geometry Expert (Chou et al., 1996), and Cinderella's Café (<http://www.cinderella.de/tiki-index.php>). These programs share one common focus on dynamic geometry figures. In a dynamic geometry figure, students can drag a geometry object such as a vertex of a triangle and change the figure dynamically while preserving the geometric invariants, which are the consequences of the given conditions. Thus these programs are commonly used to demonstrate geometry theorems.

Dynamic geometry can also be used by students to discover conjectures about a figure in some given conditions. Students can explore the variations of a dynamic geometry figure and try to discover conjectures about the figure on their own (Jones, 2000). In a study, students are given work sheets to fill in measurements of some properties of geometric objects, and they write down conjectures they discover in a dynamic figure (Furinghetti & Paola, 2003). There are two common problems in such learning activities. First, teachers and students have to learn to use the DG software and the learning process can be difficult. Second, there is no detailed record on how students manipulate the dynamic figures in order to support the findings of the study on these activities. Both problems make it difficult to evaluate the learning effectiveness of the students in using a dynamic geometry system.

To address the above two problems, we built an online dynamic geometry system that can "understand" geometry problems. With this system, the teachers and students can construct dynamic geometric figures on a web page for their learning activities, such as making geometry conjectures and proving a theorem. Furthermore, this tool can help researchers to design user interfaces that are capable of recording the data of students' interactions with the dynamic geometric figures. Based on the data, researchers can get more insights on how students make inferences from their interactions with the dynamic figures.

In addition to benefitting students, the online dynamic geometry system can be a valuable tool for instructors who struggle to produce the learning content of dynamic geometry. Specifically, two types of instructors can use the system in ways that suit their needs. The first type is instructors who do not want to work with JavaSketchpad (JSP). They can directly

use the web pages of learning content generated by the system or enhance those pages with other tools such as Adobe's Dreamweaver. The second type is instructors who want to learn JSP. They can input a problem text of their choice to the system and get a web page of JSP with the geometry content. Then they can study how the JSP code produces the dynamic geometry content they pick. This can be a good way to learn JSP by example. How effective this tool is for instructors and learners of geometry really depends on the accuracy of the system in producing correct geometry figures. Thus, an objective of this study is to evaluate the accuracy of the system empirically.

This system uses InfoMap (Hsu et al., 2001), which is a knowledge engineering tool for understanding natural language. InfoMap analyzes a geometry text by extracting geometric concepts and converting them into JavaSketchpad commands for drawing a dynamic figure. Then the system embeds the script in a web page, which can show the dynamic geometry figure on an internet browser. The article is organized as follows. The literature on attempts to solve the problem of learning JSP is first reviewed. Based on our previous study (Wong et al., 2007), a methodology and the core technologies of the system are presented. We also empirically tested the system with geometry problems from three main reference books for junior high schools in Taiwan. At the end of the article, some empirical results and conclusive remarks are reported, and future work is suggested.

Roles of DGE and Instructional Design in Geometry Learning

In research literature, there is yet no satisfactory explanation of how experimentation with dynamic geometry can help the acquisition of skills for formal theorem proving. Researchers think there is a big gap between the experience of dynamic geometry and the learning of formal proof production. Some studies also found that object dragging in a dynamic geometry environment (DGE) can reduce the gap between dynamic geometry experimentation and the generation of theorem proving ideas by learners. As a result, many researchers design some activities in a DGE and study what types of learning result from such activities and the nature of the learning process (e.g., Leung & Lopez-Real, 2003; Hoyles & Healy, 1999; Furinghetti & Paola, 2003; Christou et al., 2004).

Some studies indicate that when students explored conjectures in a DGE, they could explain the formal proof they wrote based on their experiences in the exploration (e.g., Holys & Healy, 1999). Furthermore, students would strengthen their beliefs in the geometry conjectures they made from their observation of the changes of dynamic figures in a DGE (de Villiers, 1996, 2003). In a DGE, students can drag geometric components and take measurements of geometric objects in a dynamic figure. Then they can notice the variance and invariance of conditions in a dynamic figure, deepening their understanding of geometry theorems (Laborde et al., 2006).

The ultimate goal of geometry learning is to design a DGE that helps students learn the skills of theorem proving in geometry. But current DGEs are not designed for this purpose since they do not provide any tools with instructional strategies for theorem proving. Although

a DGE is used in math classes in some schools, many math teachers and researchers think that it is difficult to design instructional materials for a DGE, and students might spend too much time in exploring without achieving the final learning objective. Therefore in the design of a DGE with specific teaching objectives, educators have to pay special attention to the design of interfaces for instructor to author guiding instructions and for learners to proceed to the final learning objective through guided exploration. Thus, we need to develop a system which can support the user interface design of a DGE for specific learning purposes.

Problems Using a DGE in Class

When a DGE is used in some learning activities, students need both basic geometry knowledge and the knowledge to work in a DGE. Sometimes, they need to add some geometric components in a dynamic figure. This can be an obstacle to some students, reducing the effectiveness of learning in a DGE (Talmon & Yerushalmy, 2004). Despite the potential benefits of using a DGE in geometry classes, a DGE is not available to some schools in Taiwan. We found that in four junior high and high schools, located in urban areas of Yunlin County, teachers have never used a DGE in their classes. We believe many other schools have similar experiences. Three reasons are suggested for not using a DGE in school. First, using a DGE might not contribute directly to the performance of students in public examinations for entering the next level of schooling. Second, funds are difficult to get for computer equipment. Third, too much time is required for teachers and students in learning how to construct dynamic figures in a DGE.

In order to address the third problem, some tools are developed to train users to learn to write JavaSketchpad (JSP, http://www.keypress.com/sketchpad/java_gsp/) script. On the web page <http://www.mathematik.uni-bielefeld.de/~lisken/jsp/>, Lisken provides a tool, `jsp.awk`, on which an author can write JavaSketchpad code. When coding is done, the tool embeds the code in an HTML file, which the author can view to check the dynamic figure with a browser. On the web page <http://home.wxs.nl/%7ehklein/jspgenerator/jsp-generator.htm>, Klein provides a similar tool called JSPGenerator, which is an authoring tool written in Javascript to generate JavaSketchpad files. After an author finishes writing Javascript code on JSPGenerator, he can choose to view the dynamic figure generated by the code in the same window just below the code. This previewing step is an improvement over `jsp.awk`. In Taiwan, Lin (2006) offers a similar tool called JavaGSP Editor, with which learners can write JavaSketchpad scripts and use these scripts in an online learning environment (Figure 1). The tool provides fancy graphical interface components such as buttons and menus for designing more interactive web pages. The above three tools simplify the task of producing dynamic figures on web pages. However, these tools still require users to learn JavaSketchpad's programming syntax and semantics, which can put off many instructors and students in high school.

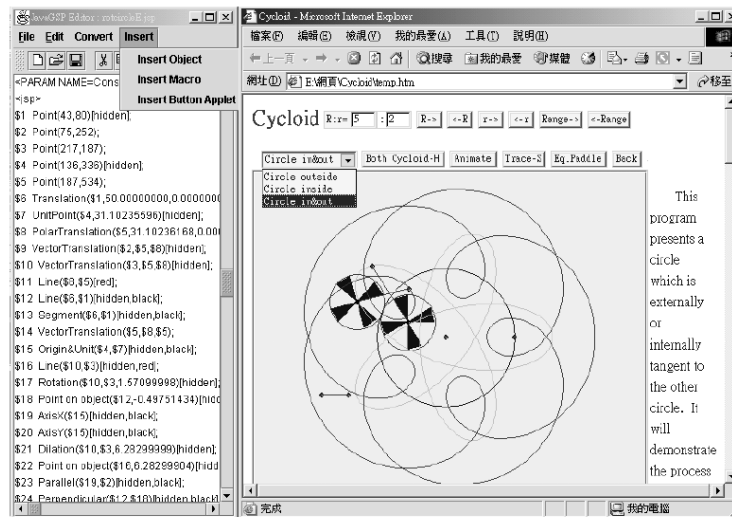


Figure 1. JavaSketchpad Editor (the left is Editor, viewer in the right (Lin, 2006)).

To make a DGE more accessible to teachers and students, we propose a system that can draw dynamic geometry figures by understanding texts of geometry problems. In this way, teachers and students are not required to spend so much time in constructing figures in a DGE. Moreover, if the system is available on a Web site, then there is no need to install any expensive commercial software in school, making a DGE more accessible to schools in poor school districts.

Natural Language Understanding for Computer-Assisted Learning

Natural language understanding is a challenging problem in the research of artificial intelligence. Some missions of research on natural language understanding are to analyze and comprehend human language and answer questions about given texts. With the progress of the research of natural language understanding, these technologies have been applied to various fields, including semantic web, Chinese speech processing, machine translation, concept modeling, knowledge engineering, and computer-assisted learning. In a study, users can use natural language instead of formal commands to work in the UNIX system (Lees & Cowie, 1996). In a study by Li & Chen (1988), an expert system was constructed for helping users learn concepts of computer science. In studies by Wong et al. (2007, 2008), a knowledge model of geometric concepts was used to understand word problems of geometry proofs.

Lees and Cowie (1996) proposed an enquiry system for training students to learn UNIX commands. The system provides a natural language interface for learners so that they can learn UNIX commands by themselves. After a learner inputs a sentence or a UNIX command, the system will parse the sentence with a chart parsing algorithm and generate a UNIX command as a response. The system checks with the learner in a dialog and executes the command if that is what the learner really wants to do with the input sentence. Li and Chen (1988) proposed a Chinese enquiry system about fundamental knowledge of computers. The system uses a linguist string parser to understand the question inputted by a learner, and then outputs an appropriate answer.

In Lu et al. (2005), a model is proposed to simulate the procedural knowledge of basic arithmetic operations. The model helps teachers design an appropriate curriculum and teaching strategy from the records of procedure that students used to solve arithmetic problems. The model is used in an intelligent tutoring system that can accumulate and reproduce the knowledge from teachers and students and help teachers build a good learning map for students. Furthermore, a student model, which is built from a collection of students' errors, can contribute to the design of more suitable teaching tactics.

In Wong et al. (2007), a LIM-G (learners' initiated model for geometry) system is used to understand geometry word problems and help elementary school students comprehend geometry word problems, which are about the area or circumference of various shapes. After a student inputs a geometry problem to LIM-G, the system understands the geometry problem using prebuilt geometry knowledge and constructs a figure for the problem. Geometry word problems from five textbooks published by five major publishers in Taiwan were used to evaluate the performance of LIM-G and about 85% of the problems were comprehended correctly.

In LIM-G, a cognitive knowledge base is constructed with an ontology-based knowledge engineering tool called InfoMap (Hsu et al., 2001), whose knowledge base includes generic template nodes for problem classes, problem-concept, lexical knowledge, lexicon, and so on. Using a template matching mechanism, InfoMap can extract the attributes and values of concepts in a given problem. After reviewing LIM-G and other similar studies, Mukherjee and Garain (2008a) point out that the methodology of natural language understanding technology and its application to understanding geometric problems in mathematics is mature for developing real applications. In our study, InfoMap is used to understand an input problem of geometry proof by extracting the geometric objects, resulting in a JSP script that draws the figure of the problem as a dynamic figure. The next section provides a methodology that other researchers can follow to develop similar systems for their learning areas.

Methodology

Applying the model of LIM-G (Wong et al., 2007), Mukherjee and Garain (2008b) implemented a tool for the automatic conversion of any input text about science and engineering into a concept map. In Mukherjee and Garain (2011), a knowledge base called GeometryNet was used in interpreting the geometric meaning of an input text to draw the corresponding diagram. These studies indicate the feasibility of a general approach in converting any input text about some subject domain, which is intended for some educational context, into a target output format such as JSP and concept map.

By generalizing and extending the model of LIM-G, we suggest a five-step methodology that researchers can adopt in automatic construction of figures of learning content in any subject domain (Figure 2). The first step is to construct a knowledge base (for example, to add geometric concepts and templates of geometric problems into InfoMap). The second is to adopt the basic mechanism of text understanding as shown in this article. Sometimes,

the researchers have to apply heuristics to improve the accuracy of text understanding. The third step is to analyze the output format (e.g., JavaSketchpad script and HTML). The fourth is to design learning activities in which learners use the content. The last step is to evaluate the learning effectiveness with empirical experiments. Based on the empirical results, the developer may need to add more concepts and heuristics for the first two steps. Then the cycle can be repeated until the researchers are satisfied. Following this methodology, we develop a system to convert texts of geometry problems into dynamic geometry figures embedded in web pages. The system architecture is described next.

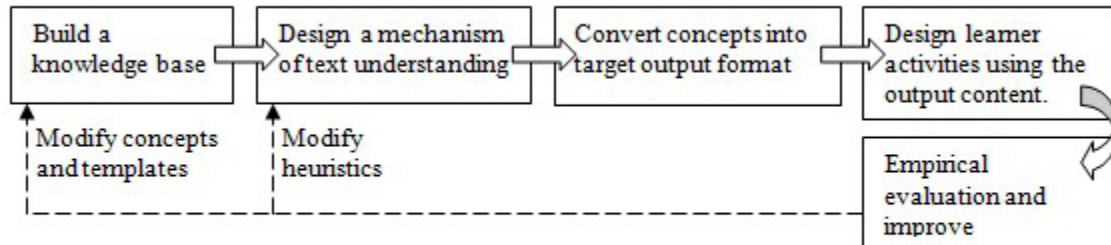


Figure 2. A five-step methodology for converting a text problem into a figure.

System Architecture

This section describes the overall architecture and user-interface of the system for generating dynamic geometry figures from input geometry problems. There are two main components in this system (Figure 3). The first component is the knowledge engine InfoMap. When a user enters a geometry problem in natural language, InfoMap analyzes the problem and extracts the attributes and values of the geometric concepts in the problem. This information is sent to the second component of the system, which is a script generator. This component generates a JSP script that draws a dynamic geometric figure of the problem embedded in an HTML document, which can be loaded by any web browser to display the dynamic geometry figure.

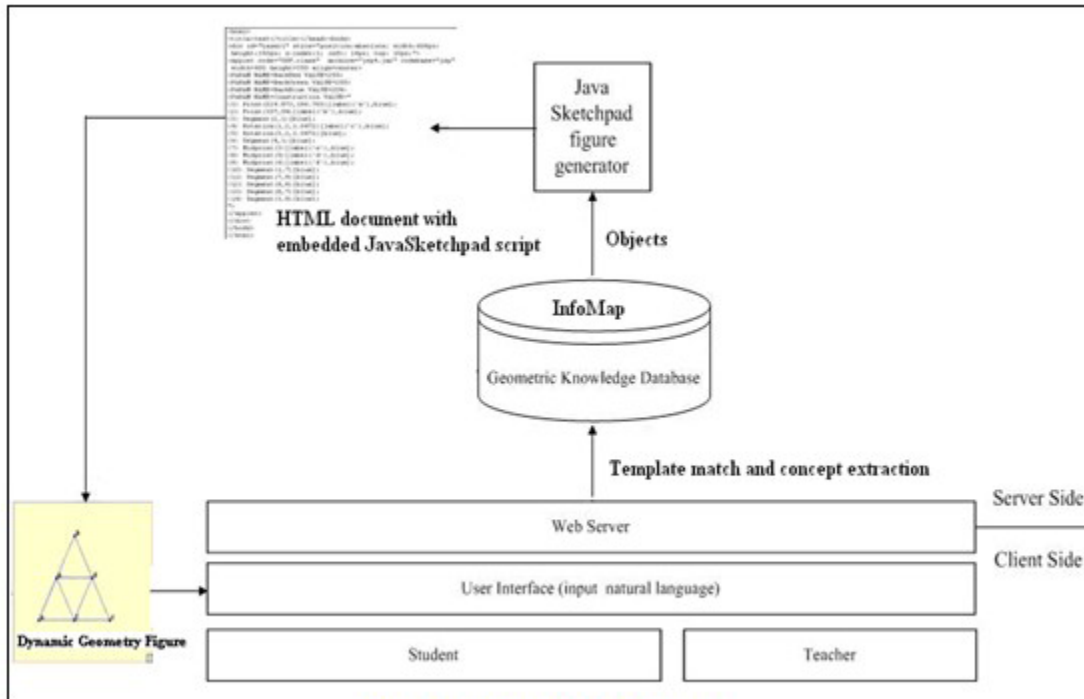


Figure 3. System architecture.

Figure 4 shows a snapshot of the user-interface in a web browser. A user can input the geometry problem in Chinese and simple mathematical symbols in the text area at the bottom. The canvas at the top displays the dynamic figure based on the system’s analysis of the input problem. The figure is drawn with a JavaSketchpad script embedded in an HTML document.

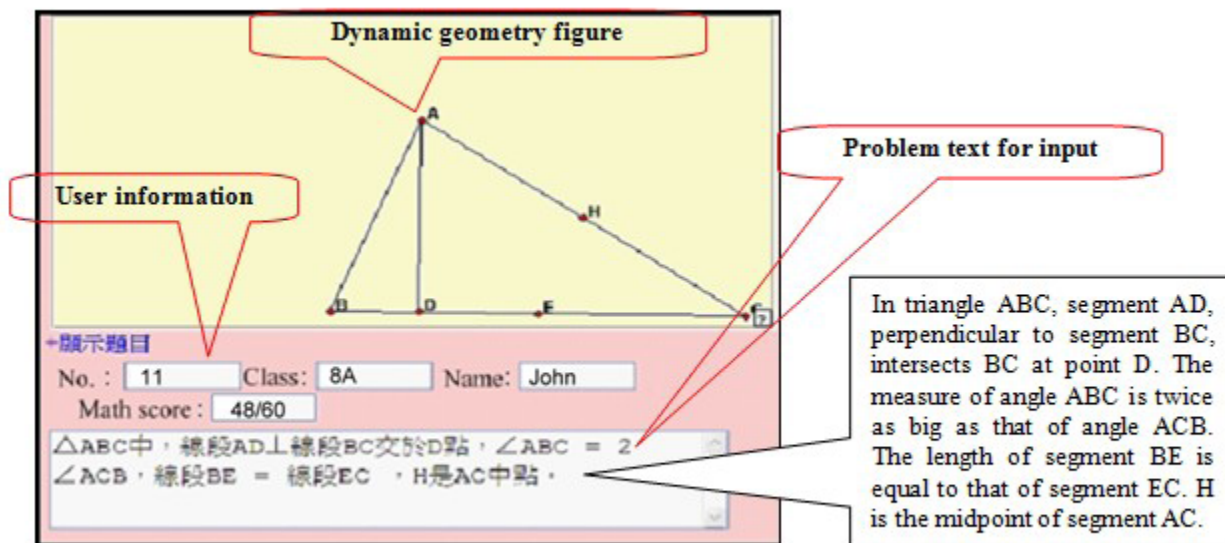


Figure 4. A snapshot of the user-interface.

InfoMap

InfoMap is a knowledge engineering tool provided by the Intelligent Agent System Lab, Institute of Information Science, Academia Sinica. InfoMap is an ontology-based system for knowledge representation and template matching (Hsu et al., 2001). InfoMap works as an agent by understanding texts in any domain and can answer questions about them if the needed domain knowledge is provided. In order to use the tool for this study, we must first build the basic knowledge for geometry problems.

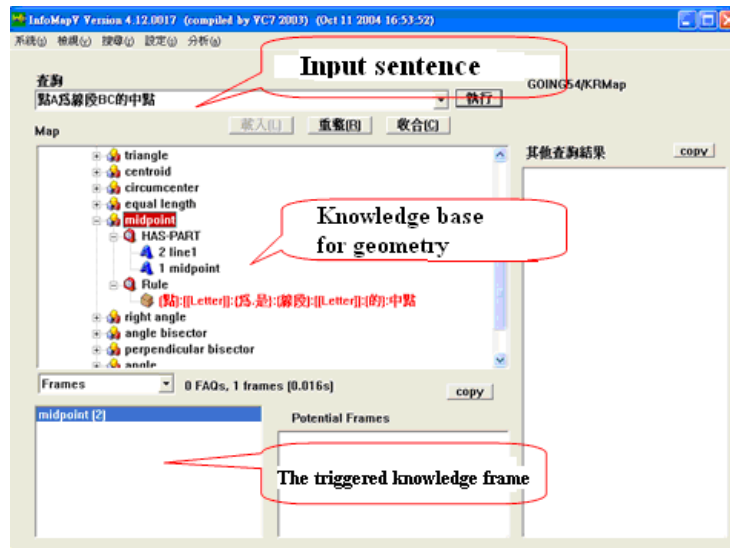


Figure 5. Inquiring a knowledge base of InfoMap.

In a knowledge base in InfoMap, nodes represent geometry concepts and each template of a node specifies the syntax of a sentence that involves the concept of the node. Templates are matched to input sentences in order to extract the concepts from the sentences. Figure 5 shows a tool for getting information from the knowledge base of InfoMap. Before the system can parse an input sentence, the system must load a knowledge base of geometry concepts in InfoMap. A user first inputs a sentence in the text area at the top. Then InfoMap triggers the templates that match the input problem, the parent nodes of the templates and the referenced nodes, extracting these nodes and their contents. Users can see the details of the knowledge base and the triggered nodes, which are colored in red (Figure 5).

Knowledge Base of Geometric Concepts

Before InfoMap can perform the understanding task, we need to build a knowledge base of geometry concepts first. Figure 6 shows part of the knowledge base, which includes many concepts (e.g., midpoint, pedal point, intersection, triangle, isosceles triangle, regular triangle, parallelogram, parallel line, point on line). In this study, we have built more than 50 nodes of geometric concepts. We can always add more concepts when needed.

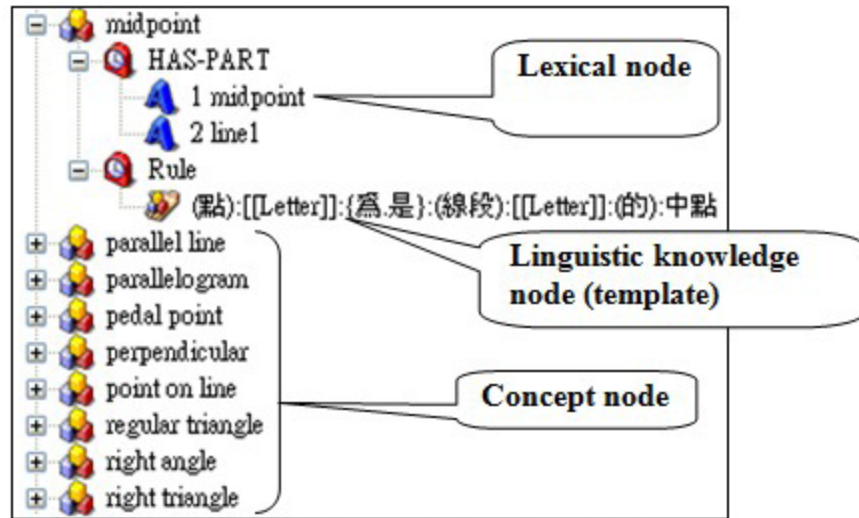


Figure 6. Knowledge base of geometric concepts.

In a knowledge base, a concept node is also a knowledge frame, which includes a rule node and two attribute nodes. The rule node generally includes multiple templates, which describe the syntax of possible sentences about the concept. The HAS-PART node specifies the component nodes that make up the concept. The component nodes can store the concepts and their names which are extracted from an input sentence. For example, the content of midpoint has two components, midpoint and line1 (Figure 6). The former component refers to the midpoint of the latter component line1.

Take midpoint as an example, when the user inputs a Chinese sentence meaning “Point A is the midpoint of segment BC” or “A is the midpoint of BC”. The sentence is matched against the InfoMap template of midpoint, and the node “midpoint” is triggered. InfoMap will extract the component concepts of midpoint A and segment BC and then label the component “midpoint” as “A” and component “line1” as “BC”. Table 1 shows the input sentence and its matched result. The templates in InfoMap are quite flexible in matching synonyms, extraneous words, and optional words. The construction of the templates is a time-consuming knowledge engineering task.

Table 1

Result of Concept Matching for Midpoint

Input sentence	Concept node	Lexical node
“Point A is the midpoint of segment BC”	midpoint	1 midpoint=A
or		2 line1=BC
“A is the midpoint of BC”		

Consider the template for matching sentences about midpoint for example. The template for an equivalent English sentence is “(Point) [[Letter]] is (the) midpoint (of) (segment) [[Letter]]”. So the sentences “Point A is the midpoint of segment BC” (sentence 1) and “A is the midpoint of BC” (sentence 2) both match the template of the “midpoint” concept. The matched results are shown in Tables 2 and 3. The label “NULL” means the word is missing.

Table 2

Result of Template Matching for Sentence 1

Geometry proof description	Point	A	is	the	midpoint	of	segment	BC
Template	(Point)	[Letter]	is	(the)	midpoint	of	segment	[[Letter]]

Table 3

Result of Template Matching for Sentence 2

Geometry proof description	NULL	A	is	the	midpoint	of	NULL	BC
Template	(Point)	[Letter]	is	(the)	midpoint	of	segment	[[Letter]]

JavaSketchpad

JavaSketchpad (JSP, http://www.keypress.com/sketchpad/java_gsp/) is a computer program with which authors publish dynamic geometry figures as a Java applet embedded in an HTML file so that users can interact with the figures with a web browser. The Geometer’s Sketchpad (GSP), on which JSP is based, is a DGE that can run on personal computers. Instructors can publish interactive, dynamic geometry content in learning activities so that students can participate over the Internet. GSP supports the web solution by publishing an HTML document embedding a Java applet containing the JavaSketchpad script into the document between the tags <body> and </body> with ordinal label {1} to {12} (Figure 7). The Java applet displays a dynamic geometry figure with the JSP code on a browser (Figure 8).

```

<!--This file created by The Geometer's Sketchpad 4.01-->
<HTML>
<TITLE>new_task5</TITLE></HEAD><BODY>
<H4 ALIGN=CENTER><APPLET CODE="GSP.class" ARCHIVE="jsp4.jar" CODEBASE="jsp" WIDTH=500
HEIGHT=300 ALIGN=CENTER><PARAM NAME=MeasureInDegrees VALUE=1><PARAM NAME=DirectedAngles
VALUE=0>
<PARAM NAME=BackRed VALUE=255>
<PARAM NAME=BackGreen VALUE=255>
<PARAM NAME=BackBlue VALUE=255>
<PARAM NAME=Construction VALUE="
(1) Point(100,40)[label('A'),red];
(2) Point(200,40)[label('B'),red];
(3) Segment(2,1)[blue];
(4) Translation(1,75,-150)[label('D'),red];
(5) Translation(2,75,-150)[label('C'),red];
(6) Segment(5,4)[blue];
(7) Segment(5,2)[blue];
(8) Segment(1,4)[blue];
(9) Midpoint(3)[label('E'),red];
(10) Midpoint(6)[label('F'),red];
(11) Segment(4,9)[blue];
(12) Segment(10,2)[blue];
">
</APPLET>
</H4>
</BODY>
</HTML>

```

The JavaSketchpad
script

Figure 7. JavaSketchpad script embedded in a HTML document.

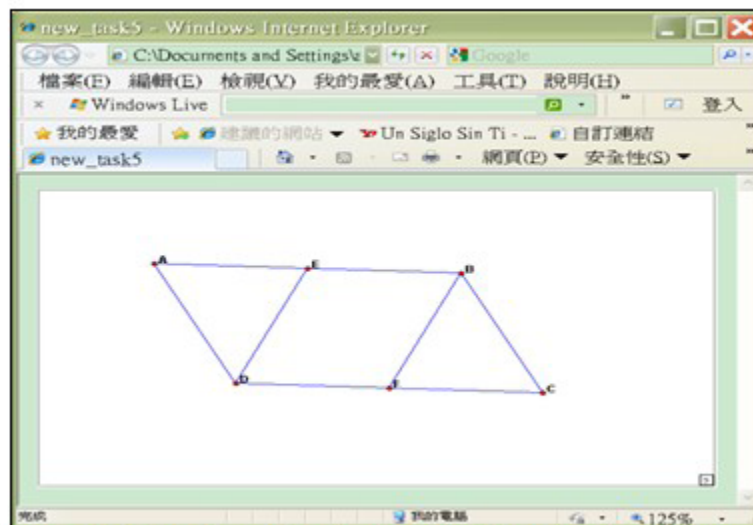


Figure 8. Figure drawn by the JavaSketchpad script in Figure 7.

Text Understanding

There are two different methods for parsing an input geometry problem text. The first method parses all sentences of a text at a time (Wong et al., 2007). This method is not flexible as it requires templates that cover all possible combinations of the sentences in the text. A better method is to parse one sentence at a time and then integrate the results for all sentences. We call this method sequential parses, in contrast to the single parse of an entire text in Wong et al. (2008). Consider the following problem: “Consider parallelogram ABCD. The point E is the midpoint of segment AB. F is the midpoint of segment CD. Prove the length of segment DE is equal to the length of segment FB.”

This text is segmented into four sentences. Each sentence is matched in InfoMap and the concepts of the sentences are extracted and mapped to JavaSketchpad statements, which are then generated. One problem occurs if the referenced object is not explicitly mentioned in the preceding sentences. For example, the second sentence mentions segment AB, which

is not mentioned in the first sentence explicitly. Rather, the segment AB is implied by the parallelogram ABCD mentioned in the first sentence. The next section explains how this problem is solved by the creation of geometric objects with JavaSketchpad statements.

Generation of a JavaSketchpad Script

The system needs to map the concepts of a sentence into one or more JavaSketchpad (JSP) statements to draw the concepts. The example which was described in the previous section, *parallelogram ABCD*, is mapped to the JSP statements in Figure 7. Since there is no parallelogram statement in JSP, the parallelogram is drawn by generating two points A and B, translating them with same displacement to get points C and D. Then the connection of each pair of points produces four segments AB, BC, CD, and DA. It is important to generate the segments AB and CD from the first sentence of the problem as both segments are referenced later by the second and third sentences respectively. Then, point E is created as the midpoint of segment AB in the second sentence, and point F is created as the midpoint of segment CD in the third sentence. In the last sentence of the problem, point E is referenced implicitly by the segment DE and point F is referenced implicitly by the segment FB. The final JSP script generated from the input text is shown in Figure 7 and the corresponding dynamic geometry figure is shown in Figure 8.

JavaSketchpad cannot draw some basic geometric objects directly, such as equilateral triangle, isosceles triangle, trapezoid, parallelogram, angle bisector, and arc. JavaSketchpad also cannot use some functions of GSP, such as step-by-step button, function graph, point of nonbasic geometric object, iteration, and the text area for input/output. To address this problem, we provide the scripts of some concepts with compass and straightedge constructions. These additional functions help to increase the number of problems that the system can handle. The following is an example of how to construct an angle bisector in JavaSketchpad:

Function *Make-AngleBisector(String angle(BAC)) return String*

Circle middle-Circle = Draw-JSPCircle(Get-Vertex(A));

Point intersect1(D) = JSP-Intersect(line1(AB), middle-Circle);

Point intersect2(E) = JSP-Intersect(line2(AC), middle-Circle);

Circle intersect1-Circle(X) = Draw-JSPCircle(intersect1(D), line3(AD));

Circle intersect2-Circle(Y) = Draw-JSPCircle(intersect2(E), line3(AD));

Point hiddenNode(H) = JSP-Intersect(intersect1-Circle(X), intersect2-Circle(Y));

Segment angle-bisector = JSP-Segment(Get-Vertex(A), hiddenNode(H));

return *angle-bisector*

Figure 9 shows the details of drawing an angle bisector. First, a circle is drawn with center A, intersecting AB at point D and AC at point E. Then a circle is drawn with center D and radius AD while another circle is drawn with center E and radius AE. These two circles intersect at point H as well as at A. Then AH bisects angle BAC.

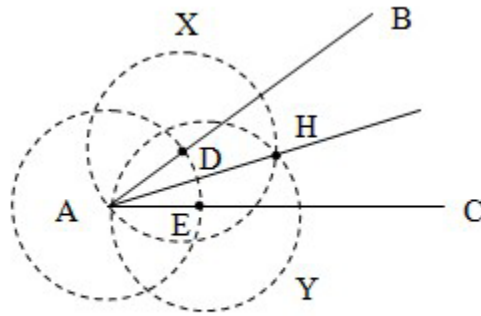


Figure 9. To construct an angle bisector by compass and straightedge.

Heuristic to Increase Correctness Rate: Reverse Understanding

A previous section shows JavaSketchpad has some limitations in constructing geometric objects. For example, the condition of two line segments equal in length or that of two angles equal in measure occurs in many geometry problems but these conditions cannot be constructed by JSP. Consider a typical problem: “Consider quadrilateral ABCD, line segments AB and CD are equal in length and so are line segments AD and BC. Prove the quadrilateral ABCD is a parallelogram.” In this example, the system will draw a quadrilateral ABCD. The system also extracts two segments AB and CD with equal length as well as two segments AD and BC. But there is no way of specifying the condition of two segments equal in length in JavaSketchpad.

In order to solve problems of this kind, we have analyzed geometry problems in junior high schools in Taiwan. About 10% of problems describe the necessary and sufficient conditions of a theorem and then ask students to prove the theorem. With a forward approach, the system will usually fail to draw the correct figure since the necessary and sufficient conditions are often complicated and the system may produce erroneous statements in JavaSketchpad. Fortunately, if the system skips the necessary and sufficient conditions and analyzes the goal condition directly, the system can draw the correct figure, resulting in a 10% increase of correctness.

This heuristic is called reverse understanding. Consider again an earlier example about a parallelogram. Figure 10 shows the process of reverse understanding. In processing the first condition of AB and CD with equal length, the system fails to find any statement to specify the condition. Then the system also fails to find any instruction to specify the second condition of AD and BC with equal length. Finally, the system finds that parallelogram ABCD is the goal to be proved. Since the system has a script to draw a parallelogram, this goal condition can be specified as a JavaSketchpad script. Moreover, this goal condition

entails the first and second condition. This reverse understanding is effective in drawing the figures of some problems, despite the failure of the default forward understanding method.

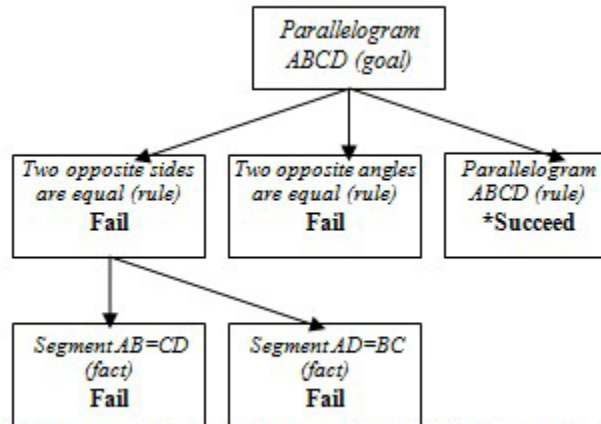


Figure 10. Reverse understanding: goal condition entails earlier conditions.

Evaluation Results and Discussion

In order to evaluate the correctness rate of this drawing system, we tested the system with problems from textbooks by three publishers, namely NanYi, ChienHong, and KangHsu-an. In order to keep the knowledge base within a reasonable size for this study, we chose problems involving only quadrilaterals or triangles. In geometry textbooks for junior high school, there are many geometry problems about quadrilaterals and triangles. In the selected textbooks, there were 61 geometry problems, including 34 on quadrilaterals and 27 on triangles. If a problem involving a circle is inputted to the system, the system will not understand the problem correctly because the circle concept is not included in the knowledge base. This is a restriction of concept coverage by the system.

After analyzing all the geometric problems in natural language, the system produced Java-Sketchpad scripts and constructed dynamic figures as web pages. Then the correctness of the produced dynamic figures was judged manually. Table 4 summarizes the evaluation results of the system. The correctness rates were 92% and 89% for 34 quadrilateral problems and 24 triangle problems respectively. The overall correctness rate was 90%.

Table 4

Empirical Results of Correctness

Type	Quadrilaterals	Triangles	Total
Correct	31	24	55
Incorrect	3	3	6
Total	34	27	61
Rate of correctness	92%	89%	90%

Before adopting text analysis with sequential parses, prebuilt scripts, and reverse understanding, the system achieved a 77% correctness rate of figure drawing with the method of Wong et al. (2008). In this study, the proposed methods increased the correctness rate from 77% to 90%. After analyzing the geometry problems whose figures cannot be generated from the input texts, we found two reasons to account for the 10% failure rate. First, JavaSketchpad can not specify two line segments with lengths in a given ratio. Second, for texts accompanied by figures, some texts do not explicitly state conditions that are obvious in the figures. As a result, there is no way to draw correct figures by processing the texts with missing information.

Consider the input text “Given equilateral triangles ADE and ABC. Prove line segments BD and CE are equal” accompanied by the figure in Figure 11. While the figure shows that points B, C, D are collinear, this condition is not stated in the input text. Naturally, the system cannot draw a correct figure by analyzing the text alone. In order to increase the correctness rate, we can prebuild more concept scripts to make up for the limited vocabulary of JavaSketchpad. To solve the missing information problem, we must rewrite the input text by adding the missing conditions from the accompanied figure.

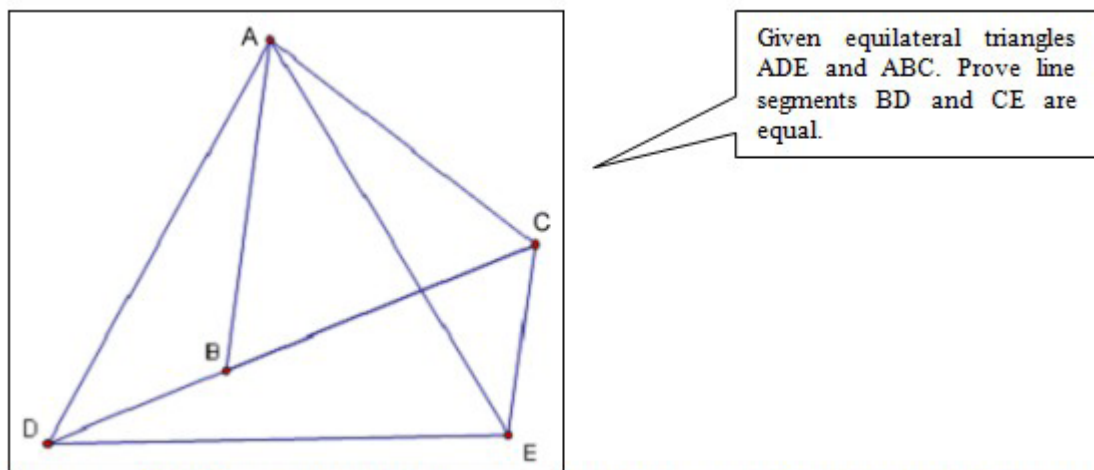


Figure 11. Conditions obvious in an accompanied figure are missing in the problem text.

Conclusion

Building upon the results of previous studies (Wong et al., 2007; Mukherjee & Garain, 2008a, 2008b, 2011), we propose a methodology of converting input texts about domain concepts into a target output format such as JavaSketchpad and concept map. The methodology is illustrated by the construction of a system that automatically produces dynamic geometry figures from input geometry problems. A dynamic geometry environment such as Geometer's Sketchpad is recognized as a tool with great potential educational value. In a DGE, students can observe invariant conditions, among other changing conditions, under given premises. Unfortunately, it can be difficult for instructors and students to use tools in a DGE to construct dynamic figures. We propose to address this problem by drawing dynamic figures automatically from input problem texts. A system was built for this purpose, using a knowledge base of basic geometric concepts and a knowledge inference engine, InfoMap, to translate problem texts into JavaSketchpad scripts. A JavaSketchpad script embedded in an HTML document can be viewed by a browser on the Internet. Empirical experiments indicated that about 90% of problems from textbooks for junior high school could be analyzed correctly to produce dynamic figures. With such high accuracy, real learning activities can be designed to use the learning content generated by the system.

Mukherjee and Garain (2008a) indicate that the method of natural language understanding in Wong et al. (2007, 2008) is a mature technology that can analyze simple text problems of areas and perimeters for elementary schools. By improving the original method, this study increased the correctness rate from 77% to 90%. The most significant improvement in the new system is the approach of text analysis with integration of parsing results from sequential sentences, while the original method parsed an entire text as one very long sentence. This was possible because the text problems of areas and perimeters in elementary school were simpler than geometry proof problems, and the complicated rule templates of the knowledge base for these simpler problems could still be constructed manually. In addition, the heuristics of reverse understanding and prebuilt scripts of geometric concepts derived with compass and straightedge construction also contribute to the increase in correctness rate.

Future Work

The dynamic geometry figures generated by our system were used in two follow-up empirical studies. In one study, students were asked to make conjectures they could find in some dynamic geometry figures produced by our system (Figure 12). In another study, students were asked to prove theorems with the resources of the corresponding dynamic geometry figures (Figure 13). The empirical results of these studies would indicate the utility value of our system in generating practical and usable geometry content for educational purposes.

At the beginning of this article, we also claim that the system can support instructors who want to learn JSP in constructing dynamic geometry content. In another follow-up study, we could design a distance learning environment for learners of JSP. Activities of writing

JSP codes would be designed for these learners, who could then solve their problems by inputting their hand-picked geometry problems as natural language texts and receive corresponding JSP codes from the system. Empirical experiments are needed to evaluate the effectiveness of this approach based on a learner's initiative.

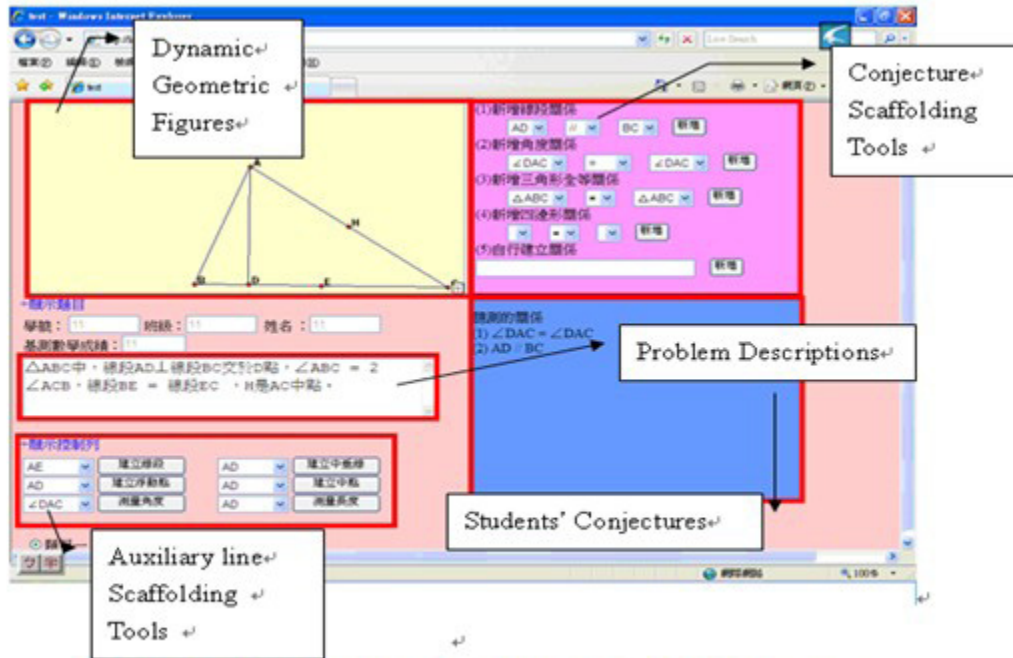


Figure 12. A user interface for making geometric conjectures.

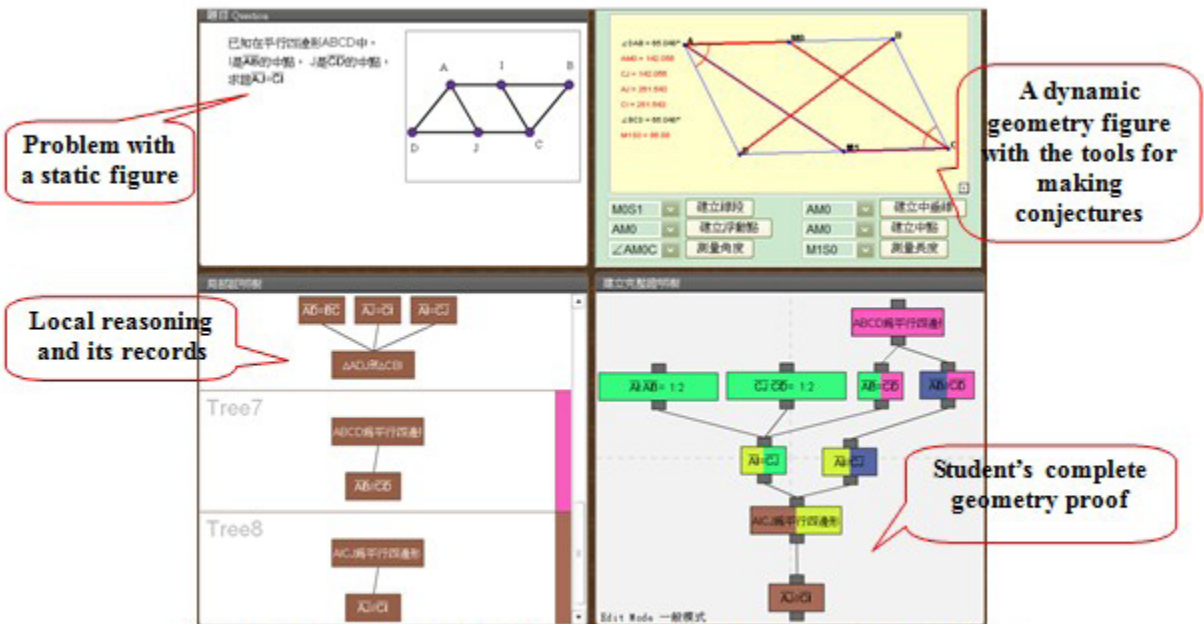


Figure 13. A user interface for learning geometry theorem proving.

Acknowledgements

This study is supported by the National Science Council, Taiwan (NSC 98-2511-S-224-005-MY3 and NSC 98-2511-S-224-004-MY2). InfoMap is provided by the Institute of Information Science, Academia Sinica, Taiwan.

References

- Chou S. C., Gao X. S., & Zhang J. Z. (1996). An introduction to Geometry Expert. *Lecture Notes in Computer Science (Proceedings of International Conference on Automated Deduction New Brunswick, CADE-13), 1104*, 235-239.
- Christou, C., Mousoulides, N., Pittalis, M., & Pitta-Pantazi, D. (2004). Proofs through exploration in dynamic geometry environments. *Proceedings of the 28th Conference of the International Group for the Psychology of Mathematics Education*, 2, pp. 215–222.
- De Villiers, M. (1996). *Some adventures in Euclidean geometry*. University of Durban-Westville (now University of KwaZulu-Natal), South Africa.
- De Villiers, M. (2003). *Rethinking proof with Geometer's Sketchpad 4*. Emeryville: Key Curriculum Press, USA.
- Furinghetti, F. & Paola, D. (2003). To produce conjectures and to prove them within a dynamic geometry environment: A case study. In N.A. Pateman, B.J. Dougherty, & J. T. Zilliox (Eds.) *Proceedings of the joint meeting PME 27 and PMENA, 2* (pp. 397-404).
- Hoyles, C., & Healy, L. (1999). Linking informal argumentation with formal proof through computer-integrated teaching experiences. In Zaslavsky (Ed.) *Proceedings of the 23rd Conference of the International Group for the Psychology of Mathematics Education* (pp. 105-112). Haifa, Israel.
- Hsu, W. L., Wu, S. H., & Chen, Y. S. (2001). Event identification based on the information map – INFOMAP. *Proceedings of the 2001 IEEE Systems, Man, and Cybernetics Conference* (pp. 1661-1672). Tucson, Arizona, USA.
- Jones, K. (2000). Providing a foundation for deductive reasoning: Students' interpretations when using dynamic geometry software and their evolving mathematical explanations. *Educational Studies in Mathematics*, 44(1-2), 55-85.
- Knuth, Eric J. (2002). Teachers' conceptions of proof in the context of secondary school mathematics. *Journal of Mathematics Teacher Education*, 5(1), 61-88.
- Koedinger, K. R. (1998). Conjecturing and argumentation in high-school geometry students. In R. Lehrer & D. Chazan (Eds.), *Designing learning environments for developing understanding of geometry and space* (pp. 319-347). Mahwah, NJ: Lawrence Erlbaum Associates.
- Laborde, C., Kynigos, C., Hollebrands, K., & Strasser, R. (2006). Teaching and learning geometry with technology. In A. Gutierrez & P. Boero (Eds.), *Handbook of research*

- on the psychology of mathematics education: Past, present and future (pp. 275-304). Rotterdam, The Netherlands: Sense publishers.
- Lees, B., & Cowie, J. (1996). Applying natural language technology to the learning of operating systems functions. *Proceedings of the 1st Conference on Integrating Technology into Computer Science Education (ITiCSE)* (pp. 11-13). NY, US.
- Leung, A., & Lopez-Real, F. (2003). Properties of tangential and cyclic polygons: An application of circulant matrices. *International Journal of Mathematical Education in Science and Technology*, 34(6), 859-870.
- Li, P. Y., & Chen, J. D. (1988). A computer training tool using chinese natural language. *Proceedings of the 1st International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE)*, 2 (pp. 899-904).
- Lin, P. P. (2006). JavaSketchpad editor used in the design of mathematical teaching and learning aids. *Science Education Monthly (In Chinese)*, 290, 48-57.
- Lu, C. H., Wu, S. H., Tu, L. Y., & Hsu, W. L. (2005). Ontological support in modeling learners' problem solving process. *Educational Technology & Society*, 8(4), 64-74.
- Mukherjee, A., & Garain, U. (2008a). A review of methods for automatic understanding of natural language mathematical problems. *Artificial Intelligence Review*, 29(2), 93-122.
- Mukherjee, A., & Garain, U. (2008b). Automatic diagram drawing based on natural language text understanding. *Lecture Notes in Computer Science*, 5223/2008, 398-400.
- Mukherjee, A., & Garain, U. (2011). Intelligent tutoring of school level geometry using automatic text to diagram conversion utility. *Proceedings of the 2nd International Conference on Emerging Applications of Information Technology (EAIT)* (pp. 45-48).
- National Council of Teachers of Mathematics (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: NCTM.
- National Council of Teachers of Mathematics (2000). *Principles and standards for school mathematics*. Reston, VA: NCTM.
- Talmon, V. & Yerushalmy, M. (2004). Understanding dynamic behavior: Parent-Child relations in dynamic geometry environments. *Educational Studies in Mathematics*, 57, 91-119.
- Whiteley, W. (1999). The decline and rise of geometry in 20th century North America. In J.

G. McLoughlin (Ed.) *Canadian mathematics education study group* (pp. 7-30). St. John's, NF: Memorial University of Newfoundland.

Wong, W. K., Hsu, S. C., Wu, S. H., & Hsu, W. L. (2007). LIM-G: Learner-initiating instruction model based on cognitive knowledge for geometry word problem comprehension. *Computers and Education Journal*, 48(4), 582-601.

Wong, W. K., Li, Y. H., Hsu, S. C., Yin, S. K., & Yang, H. H. (2008). An authoring tool for preparing online theorems and proofs with a dynamic geometry environment. *Proceedings of International Conference of Computers in Education* (pp.159-164). Taiwan, Taipei.

Athabasca University 



A Semi-Automatic Approach to Construct Vietnamese Ontology from Online Text



Bao-An Nguyen

Tra Vinh University, Viet Nam

Don-Lin Yang

Feng Chia University, Taiwan

Abstract

An ontology is an effective formal representation of knowledge used commonly in artificial intelligence, semantic web, software engineering, and information retrieval. In open and distance learning, ontologies are used as knowledge bases for e-learning supplements, educational recommenders, and question answering systems that support students with much needed resources. In such systems, ontology construction is one of the most important phases. Since there are abundant documents on the Internet, useful learning materials can be acquired openly with the use of an ontology. However, due to the lack of system support for ontology construction, it is difficult to construct self-instructional materials for Vietnamese people. In general, the cost of manual acquisition of ontologies from domain documents and expert knowledge is too high. Therefore, we present a support system for Vietnamese ontology construction using pattern-based mechanisms to discover Vietnamese concepts and conceptual relations from Vietnamese text documents. In this system, we use the combination of statistics-based, data mining, and Vietnamese natural language processing methods to develop concept and conceptual relation extraction algorithms to discover knowledge from Vietnamese text documents. From the experiments, we show that our approach provides a feasible solution to build Vietnamese ontologies used for supporting systems in education.

Keywords: Ontology; concept discovery; conceptual relation; text mining; lexical pattern; natural language processing

Introduction

An ontology is a formal, explicit specification of a shared conceptualization (Noy & McGuinness, 2001). Ontologies that belong to a specific domain are constructed from knowledge about domain concepts, their properties and instances, and the conceptual relations between them. In recent years, many semantic-based intelligent systems such as searching systems, recommender systems, and question answering systems have used ontologies as their knowledge bases. In education and e-learning, many researchers have built learning support systems that take advantage of ontologies. Li and Rui (2005) proposed a novel way to organize learning content into small “atomic” units called learning objects and systemized them together with their ontology into knowledge bases used for a recommendation mechanism. Ana et al. (2009) built a recommender system in which a domain ontological model is presented as support to Venezuelan students’ decision making for study opportunities. Saman et al. (2012) developed a knowledge-based and personalized e-learning recommendation system based on ontology to improve the quality of an e-learning system. These ontologies were constructed manually by using expert knowledge obtained from many resources and documents.

Due to their availability and abundance, text documents are one of the most popular types of knowledge sources for experts to construct their domain ontologies. Many research studies have been done on text mining and ontology construction using concept/entity extraction and conceptual relation discovery. Text mining is a subsection of data mining which could discover useful and hidden patterns or information from text. It has been used widely in many fields such as information retrieval, linguistics, knowledge engineering, and bioinformatics. Among text mining tasks, concept/entity extraction (concept mining) is applied extensively in many applications such as document summarization, question answer systems, taxonomy construction, and ontology construction. Most concept mining methods are based on linguistic rules, statistics, or a combination of both (Zhou & Wang, 2010). Other research studies also use frequent pattern mining and association rule mining for discovering concepts and conceptual relations from text (Maedche & Staab, 2000, 2001; Zhou & Wang, 2010; Chen, Zhang, Li, & Li, 2005).

Ontology construction requires efforts to uncover and organize relevant domain knowledge in a suitable structure according to the purpose of the ontology’s usage. This can be done manually or by using automatic or semi-automatic methods, in which learning methods and knowledge engineering are applied to extract concepts and conceptual relations from domain documents.

In manual construction approaches, domain experts play an essential role. Many tasks are done by these experts: covering domain terms (concepts), defining classes and class hierarchies, creating class slots (properties), filling slot values, and generating instances (Noy & McGuinness, 2001). Since every task is executed and verified by humans, the constructed ontologies tend to have a high level of accurate, reasonable, and adequate context. However, it requires a large amount of human effort and time, especially for large-scale domains such as the semantic web.

By contrast, fully automatic ontology construction methods try to learn and extract knowledge from domain documents without human supervision. For instance, Christian and Alfonso introduced an automatic ontology construction using bibliographic information (Blaschke & Valencia, 2002). Maedche and Staab presented an ontology learning framework from the semantic web through ontology import, extraction, pruning, refinement, and evaluation mechanisms (Maedche & Staab, 2001). Lee et al. (2007) presented an episode-based ontology construction mechanism from text documents and used a fuzzy inference mechanism for Chinese text ontology learning. Unfortunately, these methods are usually difficult to implement and limited in specific domains since many domain-specific decisions must be made to adequately specify the domain of interest (Jaimes & Smith, 2003). In addition, learning the knowledge base from unconstructed data is cognitive work that needs many supporting studies, and the concept hierarchy acquisition is one of the largest challenges.

Summarizing the above approaches, a semi-automatic ontology construction method is the most common approach in which information extraction techniques are used under the supervision of humans. Such methods include the learning modules to extract concepts and conceptual relations from domain documents. They require expert knowledge to verify the obtained information and decide which information should be included in the ontology. In English, many frameworks and plugins have been built to help users construct ontologies semi-automatically. For example, TextToOnto proposed by Maedche and Staab (2000) used generalized association rules to find out the co-occurrences between items and relations between them. OntoLT is a Protégé plugin that extracts concepts and relations from annotated documents for ontology construction.

Typically, taxonomy is needed in ontology acquisition tasks to construct the concept hierarchical structure of the ontology. In English, taxonomy-based approaches often use WordNet as a super taxonomy to determine the conceptual relations between concepts. In Chinese, HowNet has been used with the same role. When taxonomies are not available (or for other reasons), a nontaxonomy approach is considered using learning algorithms (e.g., Lee, Kao, Kuo, & Wang, 2007; Maedche & Staab, 2001; Blaschke & Valencia, 2002).

To extract candidate terms, the well-known statistical measurement TF-IDF can be used (Lee, Kao, Kuo, & Wang, 2007; Zheng, Dou, Wu & Li, 2007). Association rules or frequent patterns are mined to discover the co-occurrences and semantic relations between terms (Maedche & Staab, 2000, 2001; Zheng, Dou, Wu, & Li, 2007). Linguistic rules were also used in research (e.g., Zhou & Wang, 2010; Chen, Zhang, Li, & Li, 2005) in which predefined lexical patterns were used to extract candidate concepts by a bootstrapping mechanism. In Vietnamese, Nguyen and Phan (2009) proposed a hybrid approach which combines lexical rule-based and ontology-based methods to extract key terms and phrases from Vietnamese text.

In this research, we propose a semi-automatic approach to extract concepts and conceptual relations from Vietnamese text documents by using a combination of text mining techniques and statistics-based methods. Concepts will be discovered not only based on the TF-

IDF measure but also by applying lexical patterns and association rules mining. The reason to use a combination of various techniques is shown in Table 1. We also aim to compare the performance of various concept discovery algorithms and the combination of them to determine the best extracting approach for Vietnamese text documents.

Table 1

Comparison of Used Techniques

Techniques	Based on	Weaknesses
Statistics-based	Importance of terms – TFIDF	Easily affected by noises
	Co-occurrences of terms in documents	Does not consider semantic aspect of documents
	Association rules	
Lexical rule-based	Predefined linguistic rules	Hard to build a complete rule set covering all language cases
Combination of statistics-based and lexical rule-based	Taking into account both statistics and linguistic characteristics of terms.	

Proposed Method

In this section, we present our proposed system, called Vietnamese Text To Ontology, or ViText2Onto, along with learning techniques to discover concepts and conceptual relations from Vietnamese text documents. Our contribution can be stated as follows: Given a set of Vietnamese text documents in a specific domain, our system can support the user to construct an ontology using a semi-automatic approach. The resulting ontology contains concepts and instances organized in an appropriate hierarchy.

System Architecture

To construct an ontology, domain knowledge must be discovered and organized in a conceptual hierarchical structure. ViText2Onto employs a semi-automatic approach where discovery methods are used in combination with human supervision. From this perspective, an interactive mechanism is established between the system and users where the construction process is iterative and cyclic. After each iteration, the conceptual hierarchy is extended and verified by users such that the users can incrementally discover more concepts and relations based on the assessed concepts. The system architecture is shown in Figure 1, which includes the following components.

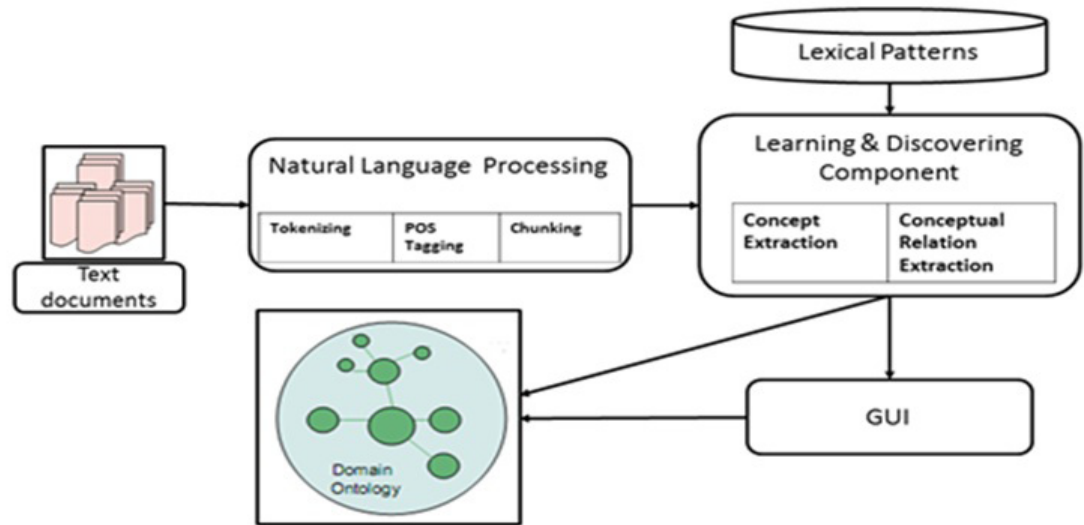


Figure 1. System architecture.

(1) A Vietnamese natural language processing module is used to make the Vietnamese text documents ready for extraction algorithms. It is a set of Vietnamese processing tools to perform tokenizing, part of speech tagging (POSTagging), and chunking. The output of this module is annotated documents being stored in text files. A small convertor is created to read these files and converts them into a compatible format that can be used by GATE.

(2) A learning and discovering component is used for extracting concepts and conceptual relations from annotated documents. We use various learning and discovering algorithms, including pattern-based, statistics-based, and association-based approaches. To implement pattern-based learning, we use JAPE (Java Annotation Pattern Engine) which is an element of the GATE framework. JAPE provides finite state transduction over annotations that let us extract predefined patterns based on rules written in a specific grammar.

(3) Lexical patterns contain lexical rules written using JAPE grammar which are used for pattern-based learning. They are constructed based on Vietnamese syntactic rules. Applying these rules on the corpus using JAPE, we can extract concepts and conceptual relations from the matched patterns.

Vietnamese Language Processing

We use Vietnamese language processing tools provided by the project of Building Basic Resources and Tools for Vietnamese Language and Speech Processing (VLSP) for preprocessing Vietnamese textual corpora. The processing components have the following features.

Vietnamese word tokenization.

Due to the characteristics of Vietnamese, a word might contain only one individual word (one morpheme) or a compound of two or more individual words (many morphemes). This tool identifies words and tokenizes sentences into separate tokens. Resulting tokenized

documents are used for further analysis tasks.

Vietnamese part of speech tagging (POS tagging).

As discovered concepts are mostly nouns, proper nouns, and noun phrases, POS tags play an essential role in both syntactic- and semantic-based learning for ontology acquisition. The POS tagger uses tokenized documents as input and assigns a POS label for all tokens.

Vietnamese chunking.

Chunking is used to divide each sentence into frames containing one or more words where each frame has a specific grammatical role in the sentence. Segmenting sentences into chunks helps determine grammatical roles of elements in sentences; hence, it is useful for learning and extracting. In our extraction rule sets, noun phrases and verb phrases are used as majority units of the patterns. Chunking frames are also used in association rule mining where phrases are used as input.

Stop words removing.

There are many words having high frequencies of occurrence in Vietnamese text while they contribute very little to the subject of sentences. To avoid noises caused by these words, we apply stop word removing when computing the TF-IDF of terms.

Learning Algorithms

In this research, the purpose of the ontology learning task is to discover concepts and conceptual relations. We use a combination of lexical pattern-based, frequent sequence-based, and statistics-based methods to overcome some drawbacks in each of the individual methods. Figure 2 shows the model of learning and discovery components.

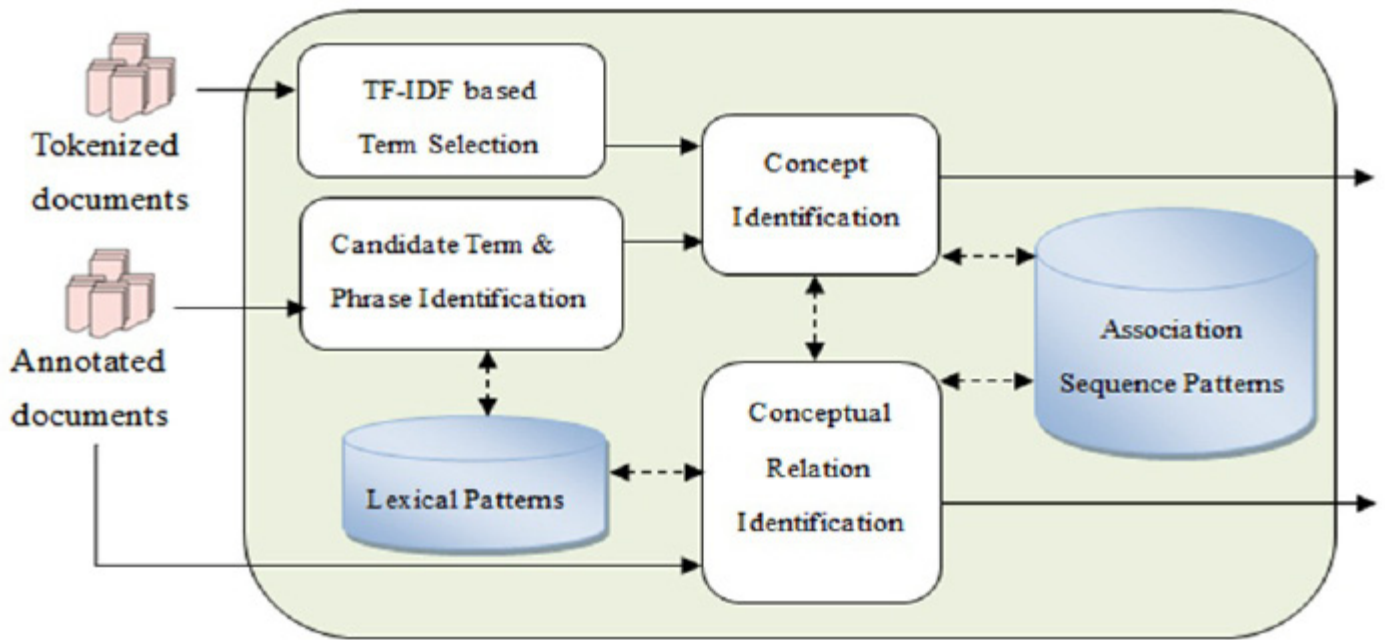


Figure 2. Learning model.

Overall Construction Process

The overall ontology construction process in the proposed system is illustrated in Figure 3. Initially, a user prepares an input corpus. Then the text files are put into the Vietnamese natural language processing module for tokenizing, POS tagging, and chunking. Processed documents are converted into the specific document format of GATE using our own converter. These Vietnamese text documents are ready for the learning process.

Firstly, candidate concepts are extracted and presented to our user interface. Note that users can specify the TF-IDF threshold and minimum support to the extraction algorithm. Concepts will be sorted in descending order by the TF-IDF score to help users select important concepts. At this step, users may only have a small number of concepts to select as seed concepts. Then a prefix-based discovery algorithm will be run to generate concept trees. Again, users will select relevant concepts as input to the relation extraction phrase. In this phrase, pattern-based and association rule-based relation extraction methods are executed. Matched patterns and association rules satisfying the minimum support and confidence are sent to the user interface in the form of a relation between a pair of concepts.

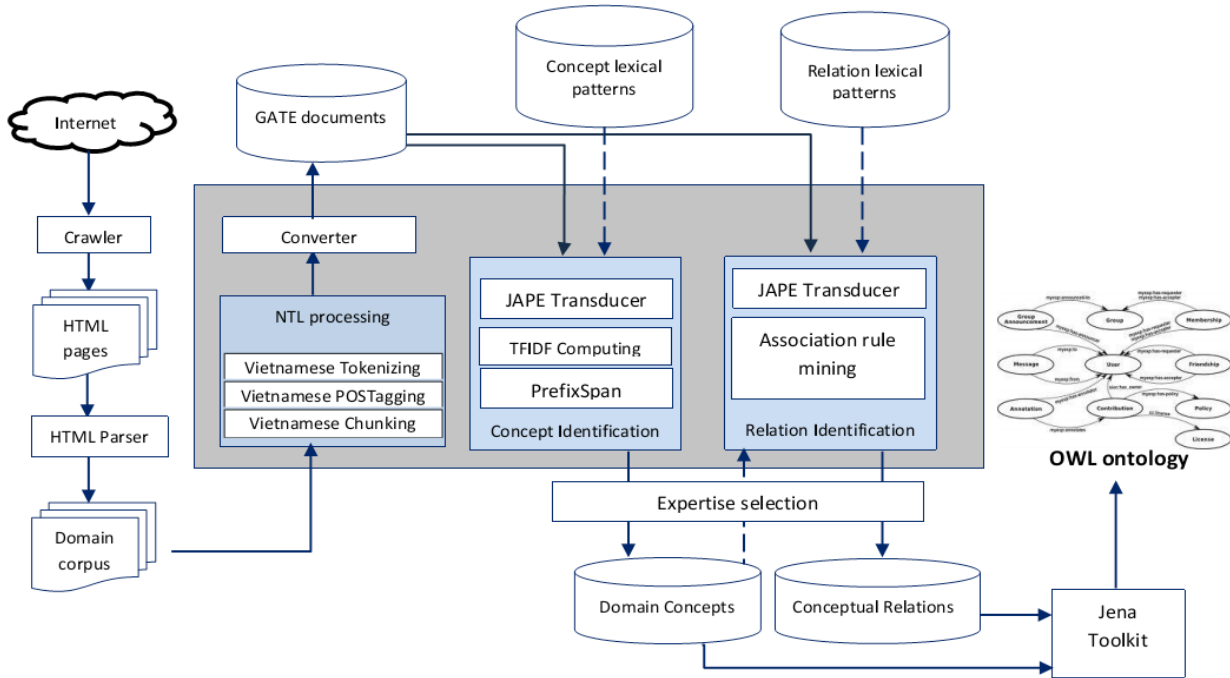


Figure 3. Overall construction process.

Users will make the final decision to select relevant concepts and conceptual relations. The selected result is exported to the ontology in OWL (web ontology language) by using the Jena toolkit.

In this process, expert knowledge is embedded into the ontology in two steps where concepts and relations are presented to users for selection. The resulting ontology contains all selected concepts and relations that can be edited easily using some ontology editing tools such as Protégé to meet user expectations.

Concept Discovery

TF-IDF-based candidate term selection.

The well-known term weight TF-IDF is used to measure the importance of individual terms contributing to documents. Important terms, such as terms having higher TF-IDF scores, will be selected based on a user-defined threshold. TF-IDF of term T_i in a document d_j is computed by the following equation:

$$tfidf(T_i, d_j) = tf(T_i, d_j) \times \log\left(\frac{|D|}{|d: T_i \in d|}\right)$$

where $tf(T_i, d_i) = \frac{n_{ij}}{\sum_k n_{kj}} tf(T_i, d_i) = \frac{n_{ij}}{\sum_k n_{kj}}$ is the term frequency of a term T_i in a document d , n_{ij} is the number of occurrences for T_i in d_j , the dominator is the size of d_j , and $\log\left(\frac{|D|}{|d: T_i \in d|}\right) \log\left(\frac{|D|}{|d: T_i \in d|}\right)$ is the inverse document frequency of T_i .

We construct a set S_{TFIDF} of candidate terms whose TF-IDF values exceed the threshold δ :
 $S_{TFIDF} = \{t_i | TFIDF(t_i) \geq \delta\}$

where t_i refers to one term in the documents.

Lexical pattern-based candidate term and phrase selection.

TF-IDF can be used to select important individual terms. However, a Vietnamese concept often consists of multiple terms. To discover multiple-term concepts, a lexical pattern-based approach is used. We built a set of lexical rules based on JAPE grammar to discover proper nouns and noun phrases. Input sentences are processed by the finite state transducer provided by JAPE, in which matched patterns will be discovered and annotated.

According to Vietnamese grammatical characteristics, the following patterns can be used for noun phrase and proper noun identification:

Noun+ Noun

Noun* (Noun | ProperNoun) (Adj | Noun)*

Noun+ Verb (Adj | Noun)+

Noun* ProperNoun+ Number

where “|” means or, “+” means one or more occurrences, and “*” means zero or more occurrences. The last pattern listed above is used for identifying proper nouns which end with a number, such as Windows Mobile 6.0 and iPhone 4.

Here is an example of a noun phrase: [Công ty]_N[trách nhiệm]_N[hữu hạn]_{Adj}[VinaCom]_{NP} (means [Limited]_{Adj} [Company]_N [VinaCom]_{NP}).

Applying these patterns to the input documents using GATE, we obtain a list of candidate noun phrases and proper nouns S_p .

Sequential pattern mining.

Lexical pattern-based learning is practical and appropriate for deep knowledge discovery, but the competence of its results depends strongly on the completeness of the set of lexical rules. To overcome this weakness, we adopt the advantages of frequent sequence mining in natural language. Based on the assumption that a concept might be a phrase or a part of a phrase in which element words usually appear in fixed orders, we consider a concept as a sequence of ordered words. Concepts might be obtained by mining frequent sequential patterns from the documents where each noun phrase is considered as a transaction.

In our research, we use segmented sentences with chunking labels as input for sequential pattern mining. As the input sentences are segmented into frames which have specific grammatical roles, a concept often belongs to only one frame. We consider each frame as a sequence and each word as an item. By mining frequent sequences we can obtain word

sequences that frequently occur together in a frame, and, hence, they can become candidate concepts. For example:

- **Input:** Điện thoại Iphone 4 màu trắng chưa được sản xuất
- **Meaning:** White Iphone 4 has not been produced yet
- **Chunks:** [Điện thoại Iphone 4 màu trắng]_{NP} [chưa được sản xuất]_{VP}

Assuming that one of the frequent sequences is “Iphone 4,” we can see the candidate concept completely belongs to the chunk [Điện thoại Iphone 4 màu trắng]_{NP}.

After the mining stage, we only use maximal frequent sequences as candidate phrases. The set of frequent sequences is denoted by S_F .

Concept identification.

By executing the above candidate concept discovery processes, we develop a list of candidate terms and phrases for concept identification. We need a filter mechanism to select relevant concepts for the ontology. The filter algorithm aims to merge three sets of candidates into a unique set of concepts, in which the candidates with lower TF-IDF scores are removed. The steps of concept identification algorithm are shown in Figure 4.

```

Input:
 $S_{TFIDF}$ : Set of candidate terms being selected based on TFIDF
 $S_P$ : Set of candidate phrases and proper nouns resulted by using lexical rules
 $S_F$ : Set of maximal frequent sequences resulted by using PrefixSpan
Output: Set of concepts C

// filter by using frequent patterns
for (every sequence  $p_j \in S_P$ )
    if  $\exists f_k \in S_F, p_j = f_k$  //  $p_j$  is a frequent pattern
         $C \leftarrow p_j$  // add  $p_j$  into set C
// filter by using TF-IDF threshold
for (every sequence  $s_i \in C$ )
    // for every term of sequence  $s_i$  having a TFIDF score less than the threshold
    if  $\nexists$  term  $t_j \in s_i$  and term  $t_k \in S_{TFIDF}, t_j = t_k$ 
        remove  $s_i$  from C
    endif
endfor
return C
    
```

Figure 4. Concept identification algorithm.

Conceptual Relation Discovery

In this phase, we use the combination of pattern-based and association-based learning to discover the relation between concepts. Using lexical rules in conceptual relation discovery

has high reliability since lexical rules were predefined by humans based on linguistic rules; however, predefined grammatical rules may not cover all cases of language usage. We use the advantage of association rule mining to overcome this weakness, in which relations between concepts are mined without considering the semantic aspect of sentences.

Lexical pattern-based conceptual relation discovery.

In this process, we take into account the semantic relations between elements in sentences, as used by Nguyen and Phan (2009). According to Vietnamese grammatical characteristics, some rules of relations between nouns or noun phrases are as follows:

- Rule 1: {Noun phrase A} “là một” {Noun phrase B} --> A is a B
- Rule 2: {Noun phrase A} {Proper noun B} --> B is an instance of A
- Rule 3: {Noun phrase A} “có” {Noun phrase B} --> A has a B
- Rule 4: {Noun phrase A} “của” {Noun phrase B} --> B has an A
- Rule 5: {Noun phrase A} “thuộc” {Noun phrase B} --> A is a subclass of B
- Rule 6: {Noun phrase A} “bao gồm” {Noun phrase B}, {Noun phrase C} --> B and C belong to A
- Rule 7: {Noun phrase A} (“và” | “hoặc”) {Noun phrase B} --> A (and | or) B

Based on these rules, we build a set of extraction rules using JAPE grammar. When the matching process is invoked, matched concept pairs and the relations between them will be discovered. In this research, we focus on finding subsumption relations and instances of concepts. The set of lexical rules contains many language usage cases that imply isA and hasA relations. Building a complete set of rules is not feasible; however, the rule sets can be enriched in further study.

Heuristic for Concept and Conceptual Relation Discovery

Context implication.

If A is a concept and B appears with A in the context {A} (“và” | “hoặc”) {B}, we can infer that 1) B is also a concept and 2) A and B have the same level of abstraction. For example:

- Điện thoại HTC Hero và Motorola Milestone đều được cài đặt hệ điều hành Android 2.1.
- Both HTC Hero and Motorola Milestone are installed with the operating system Android 2.1.

In this context, if we already know HTC Hero is a concept being recognized as a kind of mobile phone, we can easily infer that Motorola Milestone is also a kind of mobile phone.

Incremental learning.

Obviously, each domain has some cornerstone concepts, said seed concepts, which occur in many documents within the corpus. The learning phase should start from these seed concepts to discover concepts at a lower level of abstraction and repeat the process in an incremental manner.

For example, in the domain of mobile phone, we can start by some commonly used concepts such as mobile phone, keyboard, screen, and operating system. Based on Vietnamese characteristics, a child-concept often begins with its parent-concepts. We define the prefix-based concept and conceptual relation discovery algorithm below:

Given a concept C_{seed} as the seed concept selected by a user, if a concept C_i begins with C_{seed} , C_i might also be a relevant concept that should be selected by the user and C_i is a child-concept of C_{seed} (in the ontology, C_i becomes a subclass of the class C_{seed}).

By executing this algorithm on seed concepts, we can incrementally obtain a tree of concepts. This tree can be used as a part of the concept hierarchy for the ontology containing “isA” relations between child-concepts and its parents. An example of using this approach is shown in Table 2.

Table 2

An Example of Child-Concepts Generalization using Seed Concepts and Prefix-Based Concept Discovery Algorithm

Seed concepts	1_child concepts	2_child concepts	Meaning
màn hình			screen
	<u>màn hình cảm ứng</u>		touch <u>screen</u>
		<u>màn hình cảm ứng điện trở</u>	resistive <u>touch screen</u>
		<u>màn hình cảm ứng điện dung</u>	capacitive <u>touch screen</u>
bàn phím			keyboard
	<u>bàn phím QWERTY</u>		QWERTY <u>keyboard</u>
	<u>bàn phím cảm ứng</u>		touch <u>keyboard</u>

Learning from instance.

Typically, a class name rarely co-occurs with its subclasses or its properties in a sentence. Instead, instances of that class usually appear together with its related concepts. For example:

- HTC Hero được trang bị màn hình cảm ứng điện dung 4.3 inches và bộ nhớ trong 1.5GB.

- HTC Hero is equipped with a capacitive touch screen 4.3 inches and internal memory 1.5GB.

If we already know the class of the instance, we generalize the abstraction of the instance by replacing it with its class and obtain the relation between the class and discovered concepts. In this example, mobile phone will have a “hasA” relation with screen and internal memory.

Association Rule-Based Conceptual Relation Discovery

Frequent sequential pattern mining can help in cases when lexical patterns cannot be applied. We use association rule mining to find hidden (anonymous) relations between concepts by taking into account their co-occurrence in contexts, both on a sentence level and document level.

An association rule reflects an implication between its two sides. Let $T = \{t_i \mid i = 1, 2, \dots, n\}$ denotes a set of transactions, where each transaction is a list of items. $I = \{i_1, i_2, \dots, i_m\}$ denotes a list of items. An association rule of “A implies B” states that A associates with B, where A and B belong to I, and the intersection of A and B is empty. A rule “A implies B” indicates that the appearance of A is followed by the appearance of B with an acceptable probability. The reliability of a rule is expressed by two measures *support* and *confidence*.

$$\text{support}(A \Rightarrow B) = \frac{|\{t_i \mid A \cup B \subseteq t_i\}|}{n}$$

$$\text{confidence}(A \Rightarrow B) = \frac{|\{t_i \mid A \cup B \subseteq t_i\}|}{|\{t_i \mid A \subseteq t_i\}|}$$

That is, support is the probability to see both A and B appear in the same transaction while confidence is the probability to see the consequence B when the antecedent A appears in a transaction.

At the sentence level, we aim to find concept pairs that often appear together in a sentence. We consider each sentence as a transaction where each term is an item. If a concept is discovered in a sentence, terms that belong to the concept will be merged into one item. The result of the association rule mining stage will be in a pair of concepts (C_A, C_B) with support and confidence exceeding predefined thresholds. Nevertheless, we also want to find the verb that connects these two concepts and the result is in the form of ($C_A, C_B, \text{Verb}_{\text{con}}$).

At the document level, we consider each sentence as a transaction where items are concepts that appear in it. The assumption under this mining stage is that two concepts occurring in different sentences may have a relation between them. The results are concept pairs of (C_x, C_y) that may not co-occur in one sentence. Results of association rule mining, both at the sentence level and document level, are presented to users as pairs of concepts. Users will make their own judgment on the final selection.

Experimental Results

To evaluate the performance of our work, we built a real Vietnamese ontology for the mobile phone domain as a base for comparison, which is called reference ontology, O_R . It was manually built by using the source documents from the online technical specifications of various smartphones like iPhone, HTC Hero, Motorola Milestone, Samsung Galaxy, and so on.

Applying our semi-automatic approach to build an ontology, called computer learned ontology O_C , we used Vietnamese news on many kinds of mobile phones from 2009 to 2010 as the corpus. The corpus includes about 500 news articles from Vietnamese Web sites covering such topics as the arrival of new phones, comparisons of various phones, sharing phone usage experiences, symptoms of phone problems, and their advantages and disadvantages. To obtain the input corpus, we used a crawler to download the entire subfolders of three source Web sites. Web pages published before 2009 were not included. Then we manually selected HTML pages that contain well-known phone brands such as iPhone, Nokia, HTC, Samsung, Motorola, Acer, and Sony Ericsson. The final collection of the HTML pages was used as input documents.

Firstly, we used HTMLParser to extract contents from the HTML files to generate text files. The text files were tokenized, POS tagged, and segmented using Vietnamese processing tools: vnToolkit and VLSP tools. To make the Vietnamese text documents suitable for concept extraction with GATE, we developed a convertor to convert them into annotated documents that can be used by GATE. The annotation sets contain POS labels and chunk labels of tokens.

We constructed a set of Vietnamese lexical rules using JAPE grammar for pattern-based discovery. PrefixSpan is used for mining frequent sequential patterns and association rule-based discovery. The pattern concept extraction is executed by GATE transducer based on a set of lexical patterns. Results of this stage are recorded as annotations in the annotation set of the documents to be used as input for conceptual relation extraction. The final results of concept set and conceptual relation set are proposed to the users for manual selection. Selected objects and relations are exported to OWL model by Jena toolkit.

Finally, to evaluate the performance of concept discovery algorithms, we compute the term precision and term recall scores on the comparison between the ontologies O_C and O_R . To evaluate how well relations were learned, we used the measures of taxonomic precision and taxonomic recall. These scores were acquired by comparing the concept hierarchies of the two ontologies based on their position of common concepts.

Evaluation of Concept Extraction

We adopted the most commonly used measurements for information retrieval, term precision and term recall, to measure the performance of concept extraction methods. These measures are computed based on the overlap between the set of concepts in the reference ontology O_R and the computer learned ontology O_C . Let ST be the term set of the reference

ontology and T be the term set discovered by our method. We have:

$$\text{Term Precision} = \frac{|ST \cap T|}{|T|}$$

$$\text{Term Recall} = \frac{|ST \cap T|}{|ST|}$$

We also use F_β -measure with the same weights of precision and recall to measure the overall performance, where F_β -measure is computed by:

$$F_\beta = \frac{(1 + \beta^2). \text{Precision}. \text{Recall}}{(\beta^2. \text{Precision} + \text{Recall})}$$

Here β is recall weight against precision weight. Since the purpose of our system is extracting and presenting as many materials as possible to users, we select F_2 which weights recall twice as much as precision due to the requirement on the completeness of the constructed ontology. This weight means that the more extracted concepts satisfying user requirements, the better result we will get.

$$F_2 = \frac{(1 + 2^2). \text{Precision}. \text{Recall}}{(2^2. \text{Precision} + \text{Recall})}$$

In order to examine the scalability of the system, we tested the extraction performance with various corpus sizes. We divided the set of documents into five subsets as listed in Table 3. There are one large dataset, two small datasets, and two medium datasets.

Table 3

Description of Test Document Sets

	Number of files	Total size	Number of contained concepts
Subset1	195	1.37 MB	1412
Subset2	28	249 KB	489
Subset3	34	214 KB	340
Subset4	62	768 KB	696
Subset5	60	628 KB	746

In our system, we use two parameters TF-IDF threshold b and minimum support m of frequent sequence mining to drive learning algorithms. Originally, when both two parameters are set as zero, only lexical pattern-based concept extraction algorithm is executed. By increasing the TF-IDF threshold value or minimum support score, extracted concepts will be

filtered by the corresponding parameter. These adjustments affect the number of extracted concepts. The higher the parameter values are set, the fewer the number of concepts that will be extracted. According to the size of the constructed ontology, users can adjust these parameters to increase or decrease the size of the set of concepts.

Table 4 presents our results in detail, where δ is the TF-IDF threshold, M is the minimum support, T is the number of extracted concepts, STCT is the number of relevant extracted concepts, P is the precision value, R is the recall value, and F is the F_2 -measure value.

Table 4

The Performance of Concept Extraction

	δ	M	T	STCT	P	R	F	δ	M	T	STCT	P	R	F
S1	0	1	4888	1394	0.29	0.99	0.67	0.01	1	4844	1384	0.28	0.98	0.65
S2			1270	489	0.40	1.00	0.77			1042	444	0.43	0.91	0.74
S3			819	340	0.42	1.00	0.78			767	322	0.42	0.95	0.76
S4			2005	694	0.35	1.00	0.73			1438	532	0.37	0.77	0.63
S5			1846	746	0.40	1.00	0.77			1647	680	0.41	0.91	0.73
S1	0	2	2782	932	0.34	0.66	0.56	0.01	2	2781	932	0.34	0.66	0.56
S2			696	311	0.45	0.64	0.59			646	304	0.47	0.62	0.58
S3			480	233	0.49	0.69	0.64			455	220	0.48	0.65	0.61
S4			1256	500	0.40	0.72	0.62			978	412	0.42	0.59	0.55
S5			1846	746	0.40	1.00	0.77			1130	529	0.47	0.71	0.64
S1	0	4	1239	492	0.40	0.35	0.36	0.01	4	1238	492	0.39	0.34	0.35
S2			294	151	0.51	0.31	0.34			274	145	0.53	0.30	0.33
S3			241	137	0.57	0.40	0.43			235	132	0.56	0.39	0.42
S4			597	295	0.49	0.42	0.43			507	259	0.51	0.37	0.39
S5			542	290	0.54	0.39	0.41			525	285	0.54	0.38	0.40
S1	0.005	1	4888	1394	0.29	0.99	0.67	0.02	1	3773	1110	0.30	0.79	0.60
S2			1254	489	0.39	1.00	0.76			465	216	0.46	0.44	0.44
S3			809	337	0.42	0.99	0.78			486	200	0.41	0.59	0.54
S4			1859	658	0.35	0.95	0.71			683	266	0.39	0.38	0.38
S5			1768	722	0.41	0.97	0.76			943	401	0.43	0.54	0.51
S1	0.005	2	2782	932	0.34	0.66	0.56	0.02	2	2481	862	0.34	0.61	0.53
S2			686	311	0.46	0.63	0.59			335	186	0.56	0.38	0.41
S3			478	232	0.49	0.68	0.63			289	144	0.50	0.42	0.43
S4			1197	490	0.41	0.70	0.61			494	222	0.45	0.32	0.34
S5			1209	553	0.46	0.74	0.66			728	348	0.48	0.47	0.47
S1	0.005	3	1524	586	0.39	0.42	0.41	0.02	3	1365	541	0.40	0.38	0.38
S2			363	183	0.51	0.37	0.39			198	116	0.59	0.24	0.27
S3			298	166	0.56	0.49	0.50			188	107	0.57	0.31	0.34
S4			720	336	0.47	0.48	0.48			317	166	0.52	0.24	0.27
S5			652	334	0.51	0.45	0.46			406	212	0.52	0.28	0.31

Results of extraction performance show that the lexical pattern based approach can produce high recall but relatively low precision. By increasing the TF-IDF thresholds and min-sup values, precision can be improved. Figure 5 shows the scalability of our system with respect to corpus size when $\delta = 0.005$ and $M = 2$. In our test cases, the system produced best performance when it was tested with a medium size corpus; increasing corpus size can make the extraction performance go down. Overall, after testing with various sets of input documents, the extraction performance was around 60% when appropriate thresholds were specified.

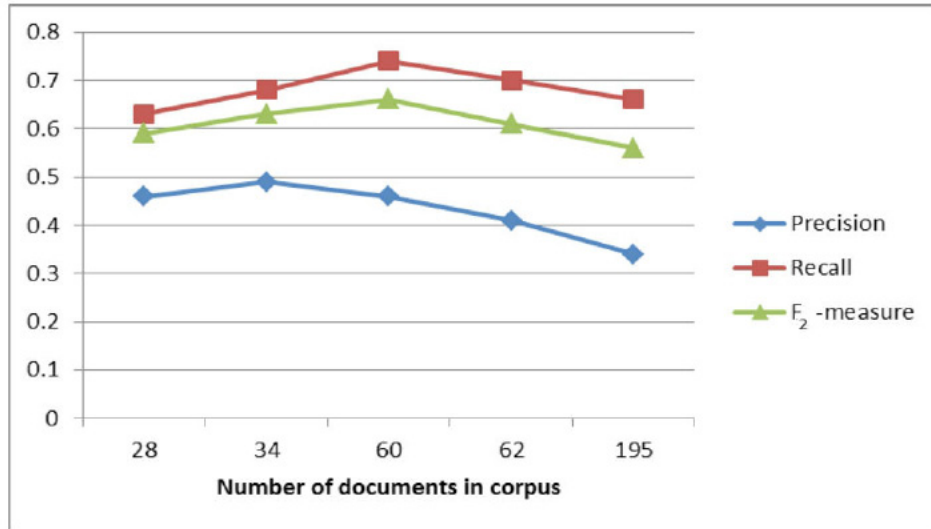


Figure 5. Extraction performance with respect to corpus size.

Evaluation of Conceptual Relation Extraction

As our purpose in the relation extraction step is to find subsumption relations, the lexical rules are built to discover patterns that contain “isA” and “hasA” relations. In the set of relations found by association rule-based extraction, we only used rules that imply “isA” and “hasA” relations. These relations are used to construct the concept hierarchy of the ontology. Some of the top extracted relations are shown in Table 5.

Table 5

Top Extracted Relations

Relations	English meaning
điện_thoại – hasA – màn_hình	phone – hasA – screen
điện_thoại – hasA – bàn_phím	phone – hasA – keyboard
màn_hình – hasA – độ_phân_giải	screen – hasA – resolution
điện_thoại – hasA – hệ_điều_hành	phone – hasA – operating_system
Android – isA – hệ_điều_hành	Android – isA – operating_system
màn_hình_cảm_ứng – isA – màn_hình	Touch_screen – isA – screen
màn_hình_cảm_ứng_điện_dung – isA – màn_hình_cảm_ứng	Capacitive_touch_screen – isA – touch_screen
bàn_phím_QWERTY – isA – bàn_phím	QWERTY_keyboard – isA – keyboard

Figure 6 is an illustration of some top extracted conceptual relations in the computer learned ontology being translated into English.

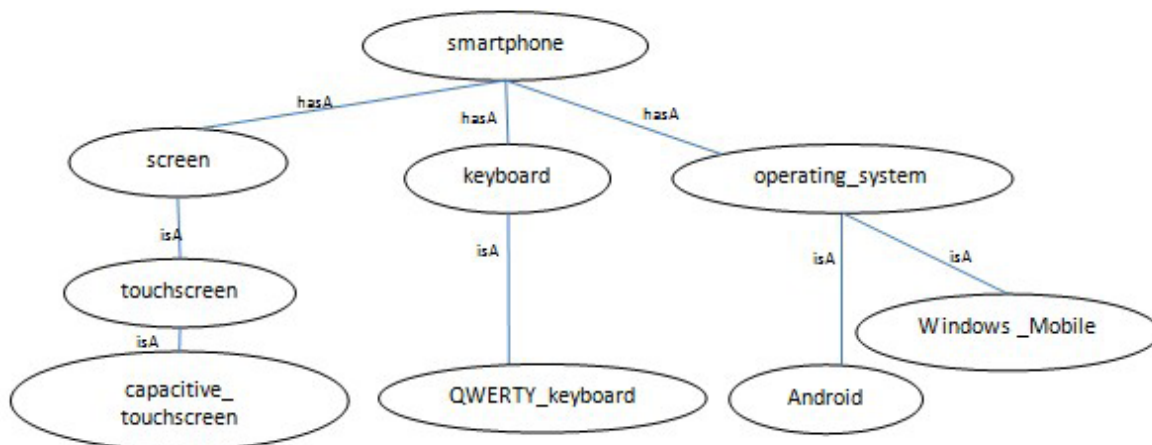


Figure 6. Some learned top level relations.

In order to evaluate how well the concept hierarchy was constructed in the ontology, we used taxonomic precision (TP) and taxonomic recall (TR) as proposed by Dellschaft and Staab (2003), in which the position of a concept in the learned hierarchy is compared with the same concept in the reference hierarchy. TP and TR are computed based on common semantic cotopy (csc) which measures the taxonomic overlap of two ontologies. The common semantic cotopy excludes all concepts which are not also available in the other ontology's concept set. Given a concept c , two ontologies O_1 and O_2 , the common semantic cotopy (csc) is defined as follows:

$$\text{csc}(c, O_1, O_2) := \{c_i | c_i \in C_1 \cap C_2 \wedge (c_i <_1 c \vee c <_1 c_i)\}$$

where C_1 and C_2 are two sets of concepts for ontologies O_1 and O_2 , respectively.

TP and TR are computed based on common semantic cotopy as follows:

$$TP_{\text{csc}}(O_1, O_2) := \frac{1}{|C_1 \cap C_2|} \sum_{c \in C_1 \cap C_2} tp_{\text{csc}}(c, c, O_1, O_2)$$

$$TR(O_1, O_2) := TR(O_2, O_1)$$

where $tp_{\text{csc}}(c, c, O_1, O_2)$ is a local precision on common semantic cotopy of the concept c and computed by:

$$tp_{\text{csc}}(O_1, O_2) = \frac{|\text{csc}(c, O_1, O_2) \cap \text{csc}(c, O_2, O_1)|}{|\text{csc}(c, O_1, O_2)|}$$

Based on TP and TR, we can compute taxonomic F-measure as:

$$TF(O_1, O_2) = \frac{2 \cdot TP(O_1, O_2) \cdot TR(O_1, O_2)}{TP(O_1, O_2) + TR(O_1, O_2)}$$

Given two ontologies O_1 and O_2 in which O_1 is the computer learned ontology and O_2 is the reference (or standard) ontology, a part of the evaluation is shown in Table 6. We only take different parts of the two ontologies into consideration, in which a concept c in O_1 has a different position as in O_2 . A number of concepts are not considered in this evaluation since they are leaf concepts linked to root nodes (things) in both ontologies. There is no need to find position difference for these concepts.

Table 6

Comparison of the Learned Ontology and Reference Ontology

O_1	O_2		$csc(c, O_1, O_2)$	$csc(c, O_2, O_1)$	
			c	$csc(c, O_1, O_2)$	$csc(c, O_2, O_1)$
			a	root, b, c	root, b, c, d
			b	root, a	root, a
			c	root, a	root, a
			d	-	root, a
			e	root	root
			f	root, h	root, h
			g	f	f
			h	f	f
			i	-	root, j, k, l, m, n
			j	root	root, i
			k	root	root, i
			k	root	root, i
			l	root	root, i
			m	root	root, i
			m	root	root, i
			o	-	root, p, q, r, s, t
			p	root	root, o
			q	root	root, o
			r	root	root, o
s	root	root, o			
t	root	root, o			

Description: a: file format; b: music file; c: video file; d: other file; e: smartphone; f: network; g: 2G network; h: 3G network; i: phone software; j: operating system; k: applications; l: music player; m: email; n: call feature; o: phone hardware; p: keyboard; q: memory; r: screen; s: battery; t: earphone

Based on the analysis shown in Table 6, we can compute $TP_{csc}(O_1, O_2) = 100\%$
 $TP_{csc}(O_1, O_2) = 100\%$, $TR_{csc}(O_1, O_2) = TP_{csc}(O_2, O_1) = 68.05\%$
 $TR_{csc}(O_1, O_2) = TP_{csc}(O_2, O_1) = 68.05\%$, and $TF_{csc}(O_1, O_2) = 0.8$
 $TF_{csc}(O_1, O_2) = 0.8$.

Usage of ViText2Onto in the Education Domain

Using ViText2Onto, we built an ontology in the education domain to show its application in real-world projects. The purpose of this project is to build a recommender system of course selection for students of the Information Technology Department at Tra Vinh University. This recommender system takes student profiles as input to output a list of recommended courses. A student profile is created based on the courses taken by the student so far. The knowledge base of this system is an ontology containing information about all courses of the bachelor program in information technology in the school.

The ontology was built based on descriptions of 50 courses in which each course description is stored in one text document. In each document, a course is presented by many units of knowledge that students must learn. Each unit corresponds to a learning objective. Each document includes course name, list of learning objectives, list of knowledge units, list of chapter titles, and the schedule of the course. These documents are available in the school's e-learning system before the beginning of each semester.

We use ViText2Onto to obtain as many concepts as there are courses and learning objective names to construct the ontology. Due to the structure of the source documents, each document only contains a plain list of learning objectives that belong to the corresponding course in which each learning objective is presented by a noun phrase, not a whole paragraph of full sentences like in news or in other types of documents. We did not use the relation extraction feature for this ontology. Consequently, after extracting concepts from documents, we put them in the ontology manually as a semi-automatic approach.

In our ontology, concepts that belong to a course form a concept tree. The concept trees of all courses in the program form the structure of the ontology. Using ViText2Onto, we were able to extract 60% of the concepts used in the ontology. For example, Figure 7 illustrates a concept tree for the course Introduction to C Programming.

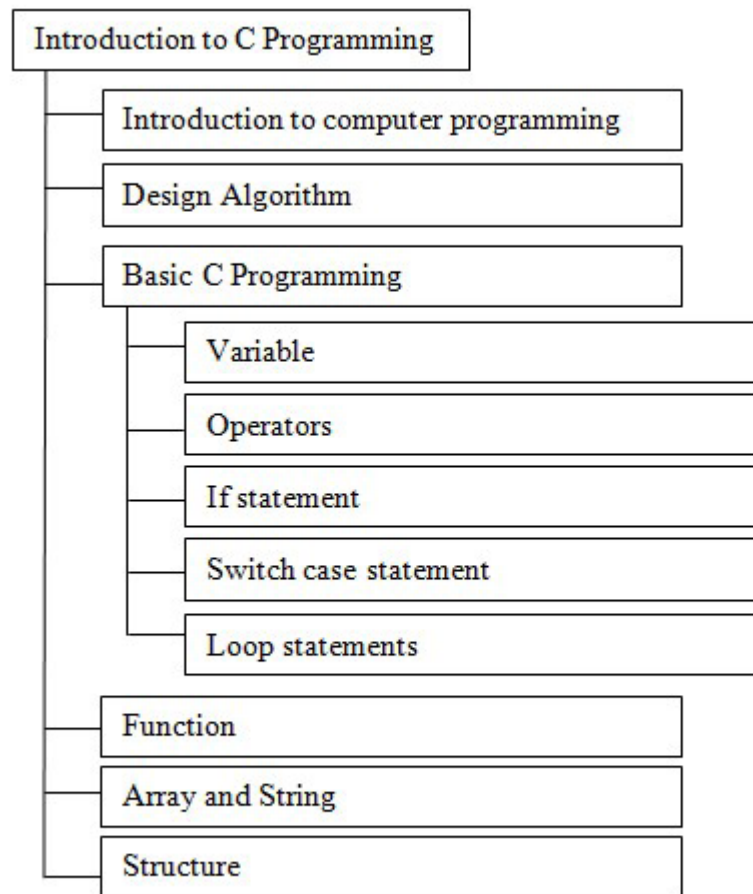


Figure 7. Concept tree for the course of Introduction to C Programming

From a general perspective, the performance of the proposed system is acceptable in supporting users to construct a Vietnamese ontology, in which the labor cost and time consumption are reduced significantly by using the semi-automatic concept extraction method. The accuracy of the system reaches above 50% with our testing datasets. More effort and further studies are on the way to boost the execution of the extraction phase. We believe the overall performance can be improved.

Conclusion and Future Research

In this research, we proposed a support system for Vietnamese ontology construction using the combination of lexical pattern-based, statistics-based, and frequent sequence pattern-based methods. The integrated approach can overcome the weaknesses of each individual method which may lead to missing concepts and relations in the discovery task. We also built a real Vietnamese ontology in the mobile phone domain using our proposed system. Then it was compared with a golden standard of manually constructed ontology. The evaluation shows that our approach has acceptable performance in concept and relation discovery.

In addition, the constructed ontology can be used as a knowledge base in many applications such as a recommendation system, text classification, and information retrieval. Based on our model many knowledge bases can be constructed easily such that more materials are available in open and distance learning.

In the near future, we would like to further automate the ontology construction by automatically learning the taxonomy part of ontology from text documents. Alternative methods of more efficient concept extraction will be considered to take the semantic aspect of documents into account.

References

- Ana, M. B., Richard, G., Marla C., Leonardo, C., & Rafael, H. B. (2009). Towards a study opportunities recommender system in ontological principles-based on semantic web environment. *WSEAS Transactions On Computers*, 2(8), 279-291.
- Blaschke, C., & Valencia, A. (2002). Automatic ontology construction from the literature. *International Conference on Genome Informatics*, 13, 201-213.
- Chen, J., Zhang, Z., Li, Q., & Li, X. (2005). A pattern-based voting approach for concept discovery on the Web. *Lecture Notes in Computer Science*, 3399, 109 -120.
- Dellschaft, K., & Staab, S. (2003). On how to perform a gold standard based evaluation of ontology learning. *International Semantic Web Conference* (pp. 228-241.)
- Jaimes, A., & Smith, J. R. (2003). *Semi-automatic, data-driven construction of multimedia ontologies*. IEEE International Conference On Multimedia and Expo.
- Lee, C. S., Kao, Y. F., Kuo, Y. H., & Wang, M. H. (2007). Automated ontology construction for unstructured text documents. *Data & Knowledge Engineering*, 60, 547-566.
- Li, P. S., & Rui, M. S. (2005). Ontology-based learning content recommendation. *International Journal of Continuing Engineering Education and Life Long Learning*, 15(3-6), 308-317.
- Maedche, A., & Staab, S. (2001). Ontology learning for the semantic web. *IEEE Intelligent Systems*, 16(2), 72-79.
- Maedche, A., & Staab, S. (2000). Semi-automatic engineering of ontology from text. *12th International Conference on Engineering and Knowledge Engineering* (pp. 231-239).
- Nguyen, C. Q., & Phan, T. T. (2009). Key phase extraction: A hybrid assignment and extraction approach. *11th International Conference on Information Integration and Web-based Applications & Services* (pp. 276-283).
- Noy, N. F., & McGuinness, D. L. (2001). *Ontology development 101: A guide to creating your first ontology* (Technical report SMI-2001-0880). Stanford Medical Informatics.
- Saman, S., Seyed, Y. B., Nor, A. M. Z., & Shahrul, A. M. N. (2012). Ontological approach in knowledge based recommender system to develop the quality of e-learning system. *Australian Journal of Basic and Applied Sciences*, 6(2), 115-123.
- Subhashin, R., & Akilandeswari, J. (2011). A survey on ontology construction methodologies. *International Journal of Enterprise Computing and Business System*, 1(1).

Zheng, Y., Dou, W., Wu, G., & Li, X. (2007). Automated Chinese domain ontology construction from text documents. Bio-inspired computational intelligence and applications. *Lecture Notes in Computer Science*, 4688, 639-648.

Zhou, J., & Wang, S. (2010). Concept mining and inner relationship discovery from text. In J. Zhang (Ed.), *New advances in machine learning*. Retrieved from Intech: <http://www.intechopen.com/books/new-advances-in-machine-learning/concept-mining-and-inner-relationship-discovery-from-text>

GATE – General Architecture for Text Engineering - <http://gate.ac.uk/>

Jena: A Semantic Web Framework for Java: <http://jena.sourceforge.net/>

OntoLT: <http://olp.dfki.de/OntoLT/OntoLT.htm>

OWL: <http://www.w3.org/TR/owl-ref/>

TextToOnto: <http://sourceforge.net/projects/texttoonto>

VLSP Project: <http://vlsp.vietlp.org:8080/demo/?page=home>

vnToolkit: <http://www.loria.fr/~lehong/tools/vnToolkit.php>

Data source Web sites:

<http://www.mainguyen.vn/tintuc/>

<http://sohoa.vnexpress.net/sh/dien-thoai/smartphone/>

<http://thegioididong.com/tin-tuc-dien-thoai,trang-chu-1.aspx>

Athabasca University 



A Qualitative Analysis Framework Using Natural Language Processing and Graph Theory



Patrick J. Tierney
Brock University, Canada

Abstract

This paper introduces a method of extending natural language-based processing of qualitative data analysis with the use of a very quantitative tool—graph theory. It is not an attempt to convert qualitative research to a positivist approach with a mathematical black box, nor is it a “graphical solution”. Rather, it is a method to help qualitative researchers, especially those with limited experience, to discover and tease out what lies within the data. A quick review of coding is followed by basic explanations of natural language processing, artificial intelligence, and graph theory to help with understanding the method. The process described herein is limited by neither the size of the data set nor the domain in which it is applied. It has the potential to substantially reduce the amount of time required to analyze qualitative data and to assist in the discovery of themes that might not have otherwise been detected.

Keywords: Qualitative analysis; graph theory; natural language processing

Qualitative research is often seen by quantitative researchers as not presenting the positivist rigor that is found in quantitative studies because of its subjectivity and its dependence on “an investigator’s own style of rigorous empirical thinking, along with the sufficient presentation of evidence and careful consideration of alternative interpretations” (Yin, 2009, p. 127). A number of software applications exist to make the often copious amount of data more manageable for the analyst, but they are mostly assistive, tracking the ascription of codes and categories to specific entities in the data, but are still dependent on the inferential capabilities of the researcher-analyst. In recent years, a new class of software applications has become available that use natural language processing (NLP) as a first step in reducing the subjective nature of qualitative analysis. Some of these applications have introduced graph theory in representing the data. This report describes an analytical process developed for a recent study that incorporated techniques from standard coding processes to include the use of graph theory. The process was developed during a recent study that explored the use of distance education (DE) technologies to support lifestyle change (Tierney, 2011). Qualitative data generated during the study were limited to the transcripts from semistructured qualitative interviews. However, the methods described herein are not limited to interview analysis, particularly in the distance-learning domain. DE researchers have long used text from discussion forums in their studies, which can represent a copious amount of information when examining threads spanning an entire semester or multiple semesters. In much the same way, social media (e.g., Facebook and Twitter) represent another vast source of data. Standard coding methods require substantial investments of time and resources, which may not always be available to researchers.

The methods herein were not used as a graphical method with deterministic outcomes per se. Rather, they assisted the qualitative researcher to visualize relationships within the data. The result was a process which may hold promise in analyzing and understanding data more quickly and in presenting a form of qualitative analysis that has some aspects of positivism, but still allows the researcher the ability to examine data and provide an understanding that can only be gained through qualitative inquiry.

Literature Review

This section provides a brief overview of the domains that form part of the new approach to qualitative data analysis, starting with the domain of quantitative text analysis. A very brief review of coding is followed by an introduction to natural language processing and its relationship to artificial intelligence and a review of the pertinent areas of graph theory. The section ends with a survey of existing software applications for computer-assisted qualitative analysis software (CAQDAS).

Quantitative Text Analysis

Evans (1996) provided a review of the state of computer-assisted analysis of text and images in the mid-1990s. At the time,

developers of software for qualitative data analysis [had]

long known that it is most appropriate to view computers as tools with which to support rather than replace human coders, especially when so few tools exist[ed] to automate sophisticated analysis procedures. (p. 271)

However, most of his discussions in that paper dealt with quantitative rather than qualitative approaches to text and image analysis. It is unclear whether his decision to not foray into the qualitative aspects of the tools and techniques reviewed was due to positivist leanings, whether his own or the research community at the time. After all, qualitative research had only taken firm hold just a few years before, following the paradigm wars of the 1980s (Denzin & Lincoln, 2011, p. 1).

Roberts (2000) provided some history on the origins of quantitative text analysis, recounting the results of a 1955 conference of Harvard University which developed the contingency analysis method:

The first step in a contingency analysis involves counting occurrences of content categories within sampled blocks of text. This produces a data matrix ... with distinct content categories (or themes) heading the columns, unique text blocks heading the rows, and counts of occurrences (of theme within block) in the cells. The analysis proceeds by computing a matrix of associations between pairs of themes. Finally, the researcher develops (usually post hoc) explanations of why some themes co-occurred and why others were disassociated (i.e., negatively associated). (pp. 260-261)

The contingency analysis method assists the researcher in keeping track of relationships between textual elements and emerging themes, but at its heart it is still a quantitative method, relying entirely on the researcher to provide the meaning within the text.

Coding Qualitative Data

Qualitative researchers often use broad ideas, themes, or concepts as tools for making generalizations. The analysis of this data can have nonvariable concepts or nominal-level variables (Neuman, 2006, p. 459). New or refined concepts are grounded in the data. Because concept formation is an integral part of qualitative data analysis and begins during data collection, “conceptualization is one way that a qualitative researcher organizes and makes sense” (p. 460). Data coding is used to assist this conceptualization process. Neuman (2006, p. 460) cites the description of codes by Miles and Huberman (1994, p. 56):

Codes are tags...for assigning units of meaning to the descriptive or inferential information compiled during a study. Codes usually are attached to “chunks” of varying size—words, phrases, sentences or whole paragraphs, connected or unconnected to a specific setting.

Neuman identifies Strauss (1987) as having defined the coding process, made up of three distinct kinds of qualitative data coding. (Neuman [2006, p. 460] also reiterates the Strauss [1987, p. 55] warning that “coding is the most difficult operation for inexperienced researchers to understand and to master.”) The three coding types are open, axial, and selective, and are sequential. In the first step, *open coding*, the researcher examines the data to condense them into preliminary analytic categories or codes (Neuman, 2006, p. 461). The next step, *axial coding*, requires the researcher to organize the codes, link them, and discover key analytic categories (p. 462). In the final step, *selective coding*, the researcher examines previous codes to identify and select specific passages that will support the conceptual coding categories that were previously developed (p. 464).

Large data sets “require considerable manual effort to analyze as researchers read and re-read the data to locate evidence to support or refute their theories” (Crowston, 2010, p. 1). However, most qualitative data analysis software packages primarily track the manual conceptualizations and categorization performed by the researcher; they do not assist with the conceptualization process itself.

Natural Language Processing

“Natural language,” as the terms suggests, is language spoken or written by humans, as opposed to a language used to program or communicate with computers. Natural language processing (NLP) falls under the rubric of artificial intelligence (AI), which is the subfield of computer science concerned with the concepts and methods of symbolic inference by computer and symbolic knowledge representation for use in making inferences. Natural language understanding by computers is one of the hardest problems of artificial intelligence due to the complexity, irregularity and diversity of human language, and the philosophical problems of meaning (natural language, n.d.). AI can be seen as an attempt to model aspects of human thought on computers (artificial intelligence, n.d.). (This report does not attempt to explain the very complex computational theories, processes, or algorithms that underpin NLP software. Instead, NLP applications are treated as a black box, with a brief description of how a qualitative researcher would use NLP software.)

In its current state of development, the setup and use of AI-based software can be tedious and time-consuming because these technologies (e.g., speech and speaker recognition, biometric-based identification such as fingerprint and face recognition, and retinal scanning) require training prior to processing input data. Crowston, Allen, and Heckman (2011, p. 2) studied two methods for training NLP software. In each approach, a portion of the data set is used for training purposes, with the remainder being used to inform the research project. The first is a *rules-based* approach which is knowledge-based, analyzing linguistic phenomena that occur within text using syntactic, semantic, and discourse information. The researcher iteratively constructs coding rules for the most abundant and obvious examples for each code. Training occurs as the rules are progressively refined for coverage and accuracy. The second approach involves *machine-learning* (ML) algorithms to “automatically learn the complex patterns underlying the extraction decisions based on statistical and se-

mantic features identified in textual data” (p. 2). Crowston et al. determined that, compared to manual rule writing, the ML process was more automatic: Training data is used to train a classifier using a machine-language algorithm that infers rules for extraction using features within the dataset itself. However, in practical terms, “human coders would still have to be used to code an initial set of data for training, but from there the trained classifier could be used to infer the code labels for the rest of the data automatically, allowing coders to shift their attention to checking the machine-coded data to further improve precision, focusing on the most important and nonautomatable job of making sense of the data.

Graph Theory

One NLP-based qualitative analysis software package uses graphs, mathematical structures used to model pair wise relations between objects from a certain collection, as one technique for representing the relationships between categories developed during analysis. Graph theory is the study of graphs and how they can be used to solve or sometimes only understand what are often very complex problems. In fact, there are graphs for which a general solution is not available, only a subset that bound by very specific conditions can be solved.

A graph refers to a collection of *vertices* or *nodes* and a collection of *edges* that connect pairs of vertices. The edges represent relationships amongst the nodes (Manber, 1989, p. 83). Any subset of a graph’s nodes and edges is called a *subgraph*. Using the simple components of nodes and edges, graphs can be used to model a large variety of both natural and human-made structures and situations.

As with all other branches of mathematics, graph theory uses very specific terminology. Graphs can be *directed*, where the relationship between the joined vertices is unidirectional and the edges are represented with an arrow. In *undirected* graphs, the relationship between the joined vertices is bi-directional, represented simply with a line (Figure 6a). An additional property of edges is that they can have a *weight* associated with them: a numerical value that quantifies the relationship of the two nodes joined by the edge (Figure 6d). Weights are sometimes referred to as *costs*. For example, the weight can represent properties such as distance, frequency, resistance, or flow.

An important aspect of graphs, which influences their analysis, is the connectedness of the vertices. A pair of vertices is *connected* if there is a path (i.e., edge) between them. In a *connected graph*, every pair of vertices are connected (Figure 1a). A *closed walk* in a graph is a path of nodes and edges between a vertex and itself (Figure 1b). A closed walk in which no edges repeat is a *circuit* (Figure 6c). A *cycle* is a circuit with no repeated vertices (Balakrishnan, 1997, p. 29; Figure 1c); an *acyclic graph* has no cycles. A specialized form of graph addressed in this report is the *tree*, a connected acyclic graph.

A lengthy discussion of the mathematics involved in studying and analyzing graphs is beyond the scope of this report. However, certain algorithms pertaining to trees are germane to the discussion because these constructs can identify relationships across entire data sets that might not be evident to all observers. A *spanning tree* of a connected, undirected graph is a tree composed of all the vertices and some (or all) of the edges in the graph (Figure

1d). Prim's algorithm (Manber, 1989, p. 208; Balakrishnan, 1997, p. 94) is used to find the *minimum weight spanning tree* within an undirected graph. Such a construct defines the acyclic subgraph that connects all the vertices of the graph at minimal cost (Figure 1e). In a subsequent section, a *maximum weight spanning tree* connects all the vertices of the graph at maximal cost (Figure 1f).

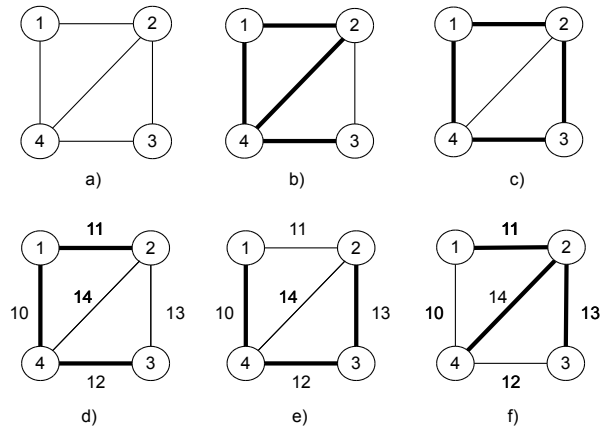


Figure 1. Graph terminology: a) Connected graph; b) closed walk (3-4-2-1-3); c) Cycle (4-1-2-3-4); d) spanning tree; e) minimal cost spanning tree; f) maximal cost spanning tree.

Computer-Assisted Qualitative Data Analysis

Many qualitative researchers use electronic spreadsheet software (e.g., OpenOffice/LibreOffice Calc, Microsoft Office Excel) to assist with managing the coding of qualitative data. Such applications aid in the capture of codes and categories, searching data associated with specific codes and categories, and the generation of histograms and other frequency representations. The ubiquity of such applications and their relatively low cost—OpenOffice/LibreOffice Calc are open source, making them available at no cost—plays a substantial role in their widespread use, particularly among student researchers.

There are several software tools specifically designed for qualitative data analysis (QDA), such as Atlas.ti and Nvivo.

[They] manage the traditional processes of manual coding and support retrieval of coded segments (Richards, 2002)...[Some] offer capabilities for automatic coding such as supporting automated searches for keywords or regular expressions but no support for semantic or higher levels of language. (Crowston, Allen, & Heckman, 2011, p. 5)

Hence, any further development of meaning that may be found in the data becomes an inductive process for which QDA tools provide no direct assistance. Some packages (e.g., Atlas.ti) are able to present network views of the data, but these views do not provide a visualization of the frequency codes or categories nor do they report the strength/weight of the relations between codes.

Open source packages offer the researcher tools for QDA. NLP libraries such as Carrot2 and Apache's OpenNLP require additional programming to form an integrated application. RapidMiner is a mature data mining tool with text processing and visualization capabilities but would require some customization or the development of ad hoc procedures to analyze qualitative research data.

The IBM SPSS Text Analytics for Surveys (TAS) package incorporates several features which bookend the categorization process. (TAS does not use the standard code and category method where categories are mutually exclusive. Instead, it uses nonexclusive categories.) Prior to the categorization step, TAS can use NLP to identify potential categories which the analyst can then keep or discard. Once categories have been developed, the visualization component of TAS offers features not found in other packages. The researcher can use the application's graphing capabilities to create category maps which show the relative strength of the relations between categories. The analyst can filter for weight ranges, further isolating specific data for closer analysis. No other packages were found to have this visualization capability.

Method

The study data for which this natural language processing/graph theory (NLP/GT) method was first used originated from a series of semistructured qualitative interviews on participants' experiences using distance education technologies for lifestyle modification/behavior change. Five interviews were conducted generating over seven hours of recordings. The purpose of this report is to document the qualitative analysis process, thus any further description of the study or how the data was collected is considered ancillary and therefore not presented here.

The qualitative data analysis for the study used the IBM SPSS Text Analytics for Surveys V4 (TAS4) tool (<http://www.spss.com/software/statistics/text-analytics-for-surveys/>) to analyze the interview data. There are two key differences between TAS4 and other qualitative data analysis tools (e.g., Atlas.ti, Nvivo). First, it does not rely solely on the reasoning and capabilities of the researcher to carry out the analysis. Instead, through the use of internal resources such as dictionaries, thesauruses, templates, and libraries, it uses NLP to identify keywords within the data, providing results that are more objective than researcher induction alone. Second, there is no defined hierarchy as would be found in codes and their categories: There are only categories. Further, the categories are not mutually exclusive as is normally the case for traditional coding using codes and categories, thus the meaning of a single data object can be placed in more than one category. When the researcher assigns categories to a single quotation or other data object, TAS4 looks for keywords from that data object in other quotes, assigning the same category when a keyword or synonym match are found.

TAS4 was selected for several reasons. Primarily, it was chosen on the recommendation of colleagues of the principal researcher on this study. In their experience, TAS4 substantially

reduced analysis time, sometimes by as much as 50% when compared to other methods/tools. TAS4's visualization capabilities, particularly the category maps, enable researchers to more quickly expose meaning and nuance, thus enabling the development of themes more quickly. Furthermore, the validity of the analysis results is augmented as a result of machine-processing, which could also result in time-savings by not having to reanalyze data based on questioning of data that can arise when analysis is completely subjective.

In order to replace the hierarchical nature normally found in codes and categories, a relationship which aids in the identification of themes, a hierarchy based on order of magnitude (i.e., powers of 10) was introduced. For this study, a *first order* category had at least 100 (10^2) shared responses, a concept that has been associated with two or more categories; with other categories; a *second order* category had from 10 to 99 (10^1) shared responses; a *third order* category had 9 (10^0) or fewer shared responses. Figure 2 illustrates these concepts. A more generalized heuristic would define the "orders" in a data set based on the category frequencies found in that data set. Orders of magnitude seemed to be appropriate, but another researcher may decide to use another scheme based on the number of categories generated and their respective frequencies. With the order of magnitude approach, first order would include those categories whose frequencies had a factor with the highest power of ten, second order would have the second highest power of ten as a factor, and so on. The maximum number of "orders" for any dataset would be the exponent of ten in the first order frequencies plus one because the factor 10^0 represents the lowest order. This can be expressed mathematically, using scientific notation as

$$\text{Category frequency} = F (10^n)$$

where F is a ratio-level measurement greater than or equal to 1 and less than 10 [i.e., $53=5.3(10^1)$, $942=9.42(10^2)$]; n is the exponent (power) of ten, also the order of magnitude. The maximum number of orders would be $n + 1$. For example, a data set with the highest category frequencies being greater than or equal to 1,000 but less than 10,000, the maximum number of orders would be 4 (i.e., $3 + 1$), first order category frequencies would have a factor of 10^3 , fourth order categories would have a factor of 10^0 .

For any category, its order, O_c , would be

$$O_c = (n_{max} + 1) - n$$

where n_{max} is the order of magnitude of the first order categories and n is the order of magnitude of the specific category.

In order to reduce the "noise" within the category graphs, the concept of *degrees of separation* was introduced. Any two categories that had a direct relationship (i.e., they had shared responses between them) were identified as having one degree of separation. Categories that had responses common with another category, but not each other, were assigned two degrees of separation. Categories separated by two categories have three degrees of separation, and so on. See Figure 2.

Results and Discussion

The steps in the analysis of the qualitative interview data collected for this study were as follows.

- 1) Transcripts were created from the audio recordings of the interviews.
- 2) The transcripts were then dissected to identify single concepts. Concepts are usually a single sentence or several sentences centered around a single idea.
- 3) The concepts were entered into a single electronic spreadsheet file, with each concept in its own cell.
- 4) Identifiers, unique codes for each interviewee, and demographic data were added for each entry.
- 5) Once completed, the spreadsheet file was imported into TAS4.
- 6) TAS4 processed the interview data using natural language processing to identify keywords within each quotation.
- 7) Categories with names meaningful to the study were created (e.g., role modeling, meta-cognition, trust, multimedia, etc.).
- 8) Categories were then assigned to individual quotes. Multiple categories were applied if

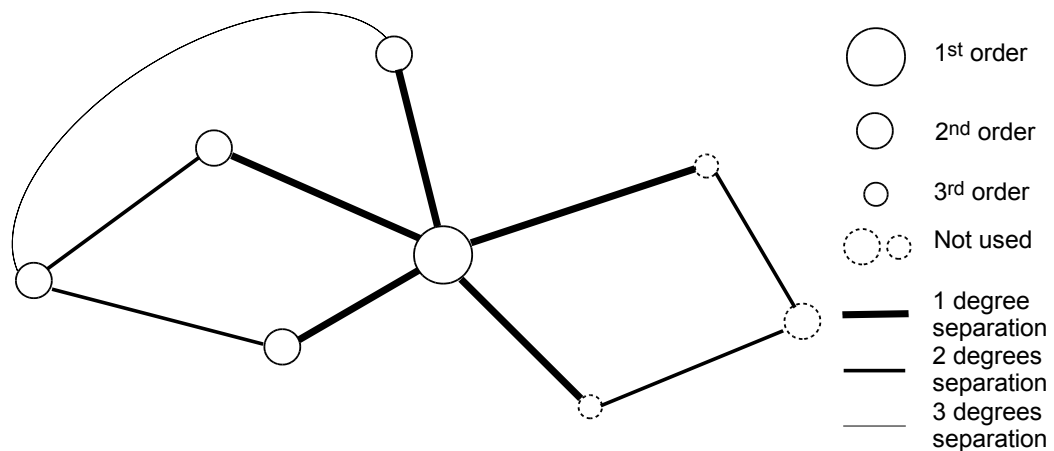


Figure 2. Code category order and degrees of separation relationships

appropriate (e.g., the quote “I am not a Facebook person” was assigned the *facebook*, *social support*, and *perception of others* categories).

- 9) If a category was improperly assigned as a result of the NLP, that quote was removed from the category. As an example, the natural language processing of the entire data set generated *people* as one of the keywords. The NLP engine uses synonyms, thus it would associate the term *person* with *people*. The quote “I am not a Facebook person” contains the

keyword person. The researcher created the category *perception of others* and linked the quote “I am not a Facebook person” to it. As a result, any data that contains the word person would automatically be associated to perception of others. The quote “I am not an ‘easy solution’ person – anything like that would not interest me” was assigned to perception of others because it contained the keyword person. However, the quote’s meaning was not about perception of others, per se, but was closer in meaning to *type of person*. The quote was removed from perception of others and linked to type of person.

10) The initial NLP generated several thousand keywords within the data. The researcher identified approximately 40 categories.

11) Five first-order categories developed. These categories were the basis of the themes found in the data analysis.

12) TAS4 created category graphs, or category webs as they are called in the software, which gave a pictorial representation of each of the first-order categories and their associated second-order categories.

The category graphs were used to develop and expand each of the themes. Most of the second order categories in a given graph had a direct relationship to the given first-order categories, hereafter referred to as having one degree of separation. Others only had an indirect relationship to the first-order category, and these were designated as having two degrees of separation. Figure 3a illustrates these concepts. Third order categories and categories with more than one degree of separation were not used in order to reduce the amount of “noise” in the graph. Figure 3b illustrates the difference between the original graph for the category *trust* and when third order categories have been removed. A third graph must be created that removes all categories and links with more than one degree of separation (Figure 3c). TAS4 was not able to isolate interview data between a given theme category and just one of its meaning categories. To overcome this limitation, an electronic spreadsheet file was constructed to allow examining only those data objects that were shared between the two categories.

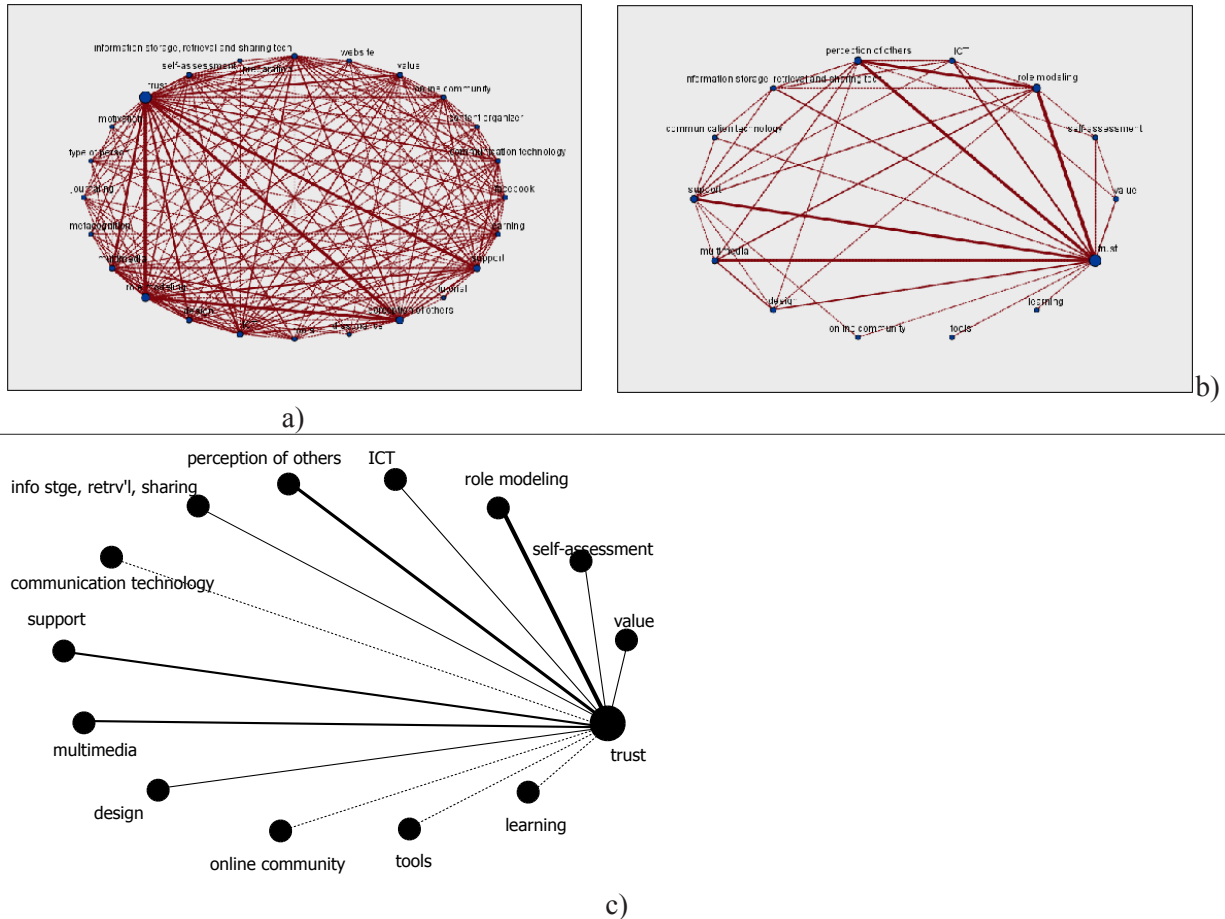


Figure 3. Shared response graphs for the “trust” category: a) all responses and categories; b) all third order categories removed; c) all remaining second degree categories removed, leaving a first order category tree.

Categories Developed From Study Data

The transcribed interview data was broken down into 505 concepts. Using the natural language processing of TAS4, 38 categories were developed. Table 1 lists those categories and the number of concepts associated with each one.

Table 1

Categories Frequencies

Category	Concept frequencies
First-order categories	
role modeling	158
multimedia	141
trust	118
design	117
support	106
Second-order categories	
self-assessment	92
ICT	90
perception of others	88
journaling	71
value	68
info storage, retrieval and sharing tech	60
learning	48
metacognition	45
tools	44
content organizer	38
preparation	26
website	22
type of person	15
online community	15
weight control beliefs	14
facebook	12
motivation	10
tutorial	10
Third-order categories	
communication technology	9
dissonance	9
self-aware	8
self-efficacy	7
resilience	7
non sequitur	6
contemplation	5
limited time	3
content	2

Theme Discovery

The use of graphs is not intended to be a graphical method of analysis with deterministic outcomes. Instead, their purpose is to aid the qualitative researcher, especially novice qualitative researchers, in discovering any structures that lie within the data and their relationships with the research question(s). To start the theme discovery within the study data set, a graph was constructed that included all the first order categories and the single degree of separation relationships between them. See Figure 4. Using the maximum spanning

tree method from graph theory, a graph was constructed that contains all of the first order categories and only the relationships amongst those activities with the highest shared response rate. See Figure 5. Upon examination of the maximum spanning tree graph, a number of category pairings were identified: design | multimedia; multimedia | role-modeling; role-modeling | trust; trust | support. These pairings were then qualitatively analyzed for meaning.

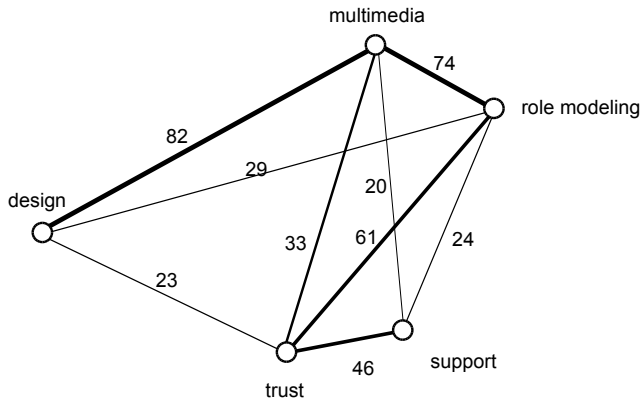


Figure 4. First-order category graph

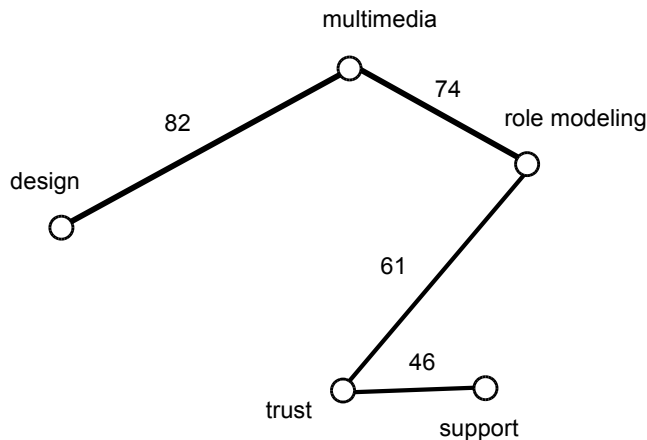


Figure 5. Maximum cost spanning tree of first order category graph.

Themes could have been developed from the maximum cost spanning tree using three subgraphs: role modeling—multimedia—design; multimedia—role modeling—trust; and, role modeling—trust—support. However, there was no obvious way, through the included edges, to relate the start and end nodes of each subpath thereby creating more cohesive themes. Therefore, in order to find such logical connections within the data, the first order category graph was revisited, removing all relationships among the first-order categories with less than approximately 24 shared responses (Figure 6). This resulted in the identification of three cycles in the graph that might help in better formulating the themes (Figure 7):

- role modeling, multimedia, and design (Figure 7a);
- role modeling, multimedia, and trust (Figure 7b);

- role modeling, trust, and support (Figure 7c).

Based on these combinations, three themes were developed. These themes and what they mean for distance-based lifestyle change are as follows: 1) Online multimedia resources for lifestyle change should be designed for role modeling; 2) Online multimedia for the purpose of

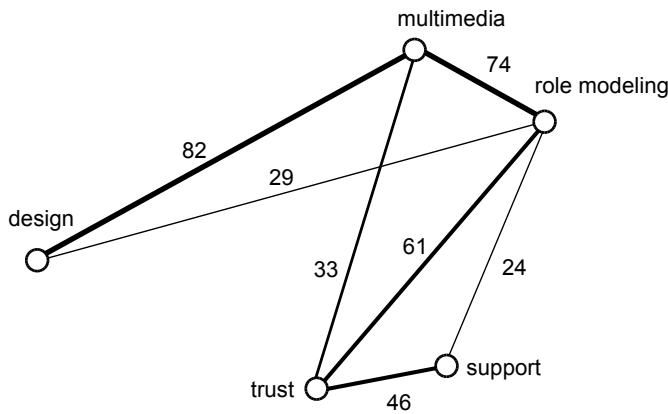


Figure 6. Reduced first order category graph

role modelling should promote a sense of safety and security, avoid risk of physical or emotional harm, and ensure that only accurate up-to-date information is provided; 3): Social support for lifestyle change should involve trusted role models.

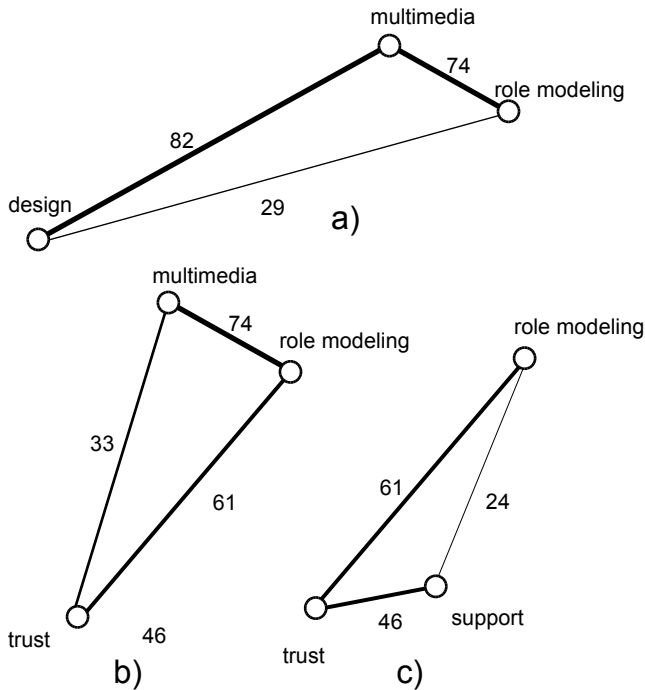


Figure 7. First-order category trinaries

Conclusion

Analyzing qualitative data can be a laborious and at times pedantic, time-consuming process. Conceptualization, especially for new researchers, can be very difficult to master. A new class of qualitative data analysis software which uses natural language processing offers tools to both simplify and accelerate the discovery of new themes and theory within that data. The concepts presented in this paper have built on natural language processing to further accelerate the discovery of what lies within the data.

This project has exposed opportunities for further research in several areas. No formal comparison was made with other tools such as Atlas.ti or Nvivo prior to embarking on the use of TAS4. Nonetheless, TAS4 does appear to have some strengths over these other software packages. Its nonexclusive use of categories uniquely allows for the generation of graphs. And with the ability to build and extend templates, the speed of analysis is almost certain to be increased when working on data sets within a given field, compared to other tools. (For the demonstration project described in this report, the categorization and theme development for 505 concepts took less than two days.) A study to identify such time savings would prove very useful in supporting the use of graph theory and natural language processing for qualitative data analysis. The cost of applications like TAS4 represent a substantial barrier to their adoption, especially in academia. The availability of open source natural language processing packages and other data/text mining software (e.g., Apache NLP, RapidMiner) may represent an avenue to the wider adoption of NLP-based qualitative data analysis.

As with all new approaches, only their use in different situations can help to debug and refine them. It is hoped that others will do exactly that to enhance the robustness of using graph theory with qualitative data natural language processing. In addition to new validation attempts of the method based on first order categories, analysis at a finer granularity than was attempted in the study for which the method was developed needs to be carried out. It is believed that when second and third order relationships are included, very subtle nuances can be discovered and teased out, further strengthening the contributions of qualitative studies.

References

- Artificial intelligence (n.d.). In *Free online dictionary of computing*. Retrieved September 1, 2011 from <http://foldoc.org/artificial+intelligence>.
- Balakrishnan, V. K. (1997). *Schaum's outline of theories and problems of graph theory*. New York, NY: McGraw-Hill.
- Crowston, K., Allen E. E., & Heckman R. (2011). Using natural language processing for qualitative data analysis. *International Journal of Social Research Methodology*. Doi:10.1080/13645579.2011.625764.
- Crowston, K. (2010). *Machine learning and rule-based automated coding of qualitative data*. Paper presented at ASIST 2010, Pittsburgh, PA.
- Denzin, N. K., & Lincoln, Y. S. (2011). *The Sage handbook of qualitative research* (4th ed.). Thousand Oaks, CA: Sage Publications, Inc.
- Manber, U. (1989). *Introduction to algorithms: A creative approach*. Reading, MA: Addison-Wesley Publishing Company.
- Miles, M. B., & Huberman, M. (1994). *Qualitative data analysis* (2nd ed.). Thousand Oaks, CA: Sage.
- Natural language (n.d.). In *Free online dictionary of computing*. Retrieved September 1, 2011 from <http://foldoc.org/natural+language>
- Neuman, W. L. (2006). *Social research methods: Quantitative and qualitative approaches* (6th ed.). Boston, MA: Pearson Education Inc.
- Roberts, C. W. (2000). A conceptual framework for quantitative text analysis: On joining probabilities and substantive inferences about texts. *Quality & Quantity*, 34(3), 259–274.
- Strauss, A. (1987). *Qualitative analysis for social scientists*. New York, NY: Cambridge University Press.
- Tierney, P. (2011). *Exploring the use of distance education technologies for lifestyle change* (Unpublished master's thesis). Athabasca University, Athabasca, AB. Retrieved from <http://hdl.handle.net/10791/12>
- Yin, R. K. (2009). *Case study research: Design and methods* (4th ed.). Thousand Oaks, CA: Sage Inc.

Athabasca University 

