At the end of 3 month of trial period, we only found 4 users with *more* than 9 items in their bookshelves along with other usage data. In the exploratory study we invited these 4 users to perform the following experimental task.

4.2 The Primary Experimental Tasks

Each user is presented with 3 sets of *abstracts* consisting of 6 abstracts in each set (after deleting title, keywords, classification and other information). Each set of 6 abstracts includes 3 abstracts taken from his or her *own* bookshelf entries, and the other 3 were randomly selected from other user's bookshelf entries. There are no overlapping items in these 3 sets of abstract presented to each user. We asked all 4 users to perform the following 3 tasks with each set of 6 abstracts on the baseline system and on the augmented system as explained below. The users were told that both servers are equipped with *Camtasia* screen recording software so that we can capture their interactions for further experimental analysis. Since the screen recording is minimally intrusive to user activities, the users should therefore perform their tasks as naturally as possible without worrying about the experimental objectives but only considering the information seeking tasks[1].

Baseline Browsing Task: For the Baseline Browsing Task, the users were asked to use the baseline system's *browsing interface* to locate the 6 target documents only through browsing. We used screen-recorder to capture user's interactions while they perform the task for further analysis. The average number of interactions each user made to locate their own 3 documents (supposedly, the familiar ones) and the other 3 documents are counted and listed in Table 1. It should be noted that the ACM CCS taxonomy consists of 4 levels (with some cross-references); and the entire DL collection consists of only about 300 items. In most cases the users succeeded to retrieve familiar documents in 4 or less interactions. However, for unfamiliar documents, the number of interactions was much higher since we consider tracking back as an additional interaction.

Baseline Search Task: For the Baseline Search Task, the users were given *another* set of 6 abstracts (3 new abstracts from their own bookshelf and the other 3 new abstracts were taken from the other's shelves). This time we requested the user to use keyword search on the baseline system to locate those documents via keyword base searching. Search results were presented as 5-documents/page (using Greenstone's default relevance ranking), and therefore, flipping a page is considered 1 interaction. Likewise, revision or editing of keywords is also considered as a new interaction. Same as before, we recorded user interactions passively for later analysis. The average number of interactions for each groups (own vs. others) are listed in Table 1. There was no regular pattern in the average number of interactions required to locate the documents regardless of familiar items (for documents from user's own bookshelf) or not (for documents selected from other's bookshelf).

[1] A warm-up session was conducted with 1 browsing and 1 searching task using a set of 2 abstracts for all 4 users.

Augmented Search Task: For the Augmented Search Task, the users were given *another* set of 6 abstracts (3 new abstracts from their own bookshelf and the other 3 new were taken from the other's shelves). This time we requested the user to use keyword search on the *augmented* system where we already have their dynamic profile and usage histories recorded and DL contents are further augmented using keyphrase extraction, ontology-driven topic inference, etc. However, the search results in the augmented system are sorted against a user's dynamic profiles (which in turn reflect their usage pattern; different from baseline search output). We asked the users to search for all 6 documents using keyword-bases search. Search results were presented as 5-documents/page (same as the baseline search task); and user interactions are recorded same as before. The average number of interactions for each groups (own vs. others) are listed in Table 1 for comparison. In most cases, the user's located their own documents with a low average number of interactions while the average number of interactions required for unfamiliar documents was higher. This is due to the fact that the search result is ranked against user's preferences and contexts and therefore flipping of pages and revision of search terms were required. The possibility of failing to use the right (combination of) keywords for unfamiliar documents can't also be fully ruled out. However, in a small-scale exploratory experimental setting, we further explore the phenomenon by interviewing the participants.

Table 1. Average number of interaction in 3 different tasks. The value shows avg # of interactions for a subgroup of 3 documents.

	Base Browsing		**Base. Search**		**Aug. Search**	
User#	*Own/3*	*Other/3*	*Own/3*	*Other/3*	*Own/3*	*Other/3*
#1	3.3	6.3	3.7	4.0	2.0	5.3
#2	4.0	6.7	4.0	4.0	3.3	4.0
#3	3.7	5.0	4.0	5.3	2.3	5.0
#4	3.0	6.7	5.0	4.7	2.7	4.3

4.3 Other Experimental Evaluations

During the interview session, we revealed our experimental methods and objectives in detail to each user. We also presented the user Dynamic Profile (Topics with currents weights using a cut-off threshold) and their initial sign-up profile side by side and requested them to rate which profile reflect their interests better. There were both strong agreements and strong disagreements. However, when we further explained how the dynamic profile takes into account of their activities (such as keywords they used and items they chose for bookshelf) and their context (such as the, interest-drift), all 4 users tend to agree that the dynamic profile is in line with the usage and activities.

In the final phase of the exploratory evaluation, we used our recommender module to extract top 5 documents for each user. We requested the user to read the title and abstracts and validate how many of these documents they would possibly read further

if they are recommended by the system on their next login. The cumulative responses of 4 users showed that 14 out of 20 recommended documents are chosen as worth reading further.

5 Conclusions

In this paper we presented a 3-Layer Digital Library Architecture which targets for intelligent and adaptive digital library services. We attempted to integrate user profile, domain ontology, usage pattern and DL content analysis together in formulating intelligent and adaptive services for digital library users. Our present experimental setup and exploratory analysis using a small-prototype show that our approach and algorithms tend to integrates several facets of DL information seeking scenario; and therefore, we plan to further enhance our algorithms and integrate them seamlessly for a large-scale collection and users.

Information seeking in the context of digital library is unique and multi-faceted. Therefore, it is desirable that DL researchers will increasingly adapt recent developments if HCI-related research and user studies in the context of digital library. We therefore, plan to integrate mixed-interactions such as, integration of browsing and search, and integration of in-turn and out-of-turn interactions [18], [19] to facilitate enhanced information seeking experience in the digital library environment.

Acknowledgment. This research was supported by Thailand Research Fund (TRF) Grant No. MRG4880112 awarded to Dr. Md Maruf Hasan. The authors also like to thank all the participants who took part in the exploratory experiment, and Ms. Yenruedee Chanwirawong who developed the system.

References

1. Chowdhury, G.G., Chowdhury, S.: Introduction to Digital Libraries. Facet Publishing, London (2003)
2. Feng, L., Jeusfeld, M.A., Hoppenbrouwers, J.: Beyond Information Searching and Browsing: Acquiring Knowledge from Digital Libraries, (Retrieved March 25) (2007), http://citeseer.ist.psu.edu/421460.html
3. Marchionini, G.: Information Seeking in Electronic Environments. Cambridge Series on Human-Computer Interaction. Cambridge University Press, Cambridge (1997)
4. Straccia, U.: Collaborative Working in the Digital Library Environment Cyclades,(Retrieved March 12) (2007), http://dlibcenter.iei.pi.cnr.it/
5. Hurley, B.J., Price-Wilkin, J., Proffitt, M., Besser, H.: The Making of America II Testbed Project: A Digital Library Service Model. The Digital Library Federation Washington DC (1999)
6. Brusilovsky, P.: Adaptive Hypermedia. User Modeling and User-Adapted Interaction 11(1-2), 87–110 (2001)
7. Crestani, F.: Application of Spreading Activation Techniques in Information Retrieval. Artificial Intelligence Review 11(6), 453–482 (1997)
8. Greenstone Digital Library Software. Project, Retrieved 2/2/2007, from http://www.greenstone.org/

9. ACM-CCS Add-on Ontology, University of Minho Web Site, (Accessed March 12, 2006), http://dspace-dev.dsi.uminho.pt:8080/en/research_about.jsp
10. Witten, I.H., Paynter, G.W., Frank, E., Gutwin, C., Nevill-Manning, C.G.: KEA: Practical Automatic Keyphrase Extraction. In: Fourth ACM Conference on Digital Libraries DL 1999, pp. 254–255. ACM, New York (1999)
11. Pitkow, J., Schütze, H., Cass, T., Cooley, R., Turnbull, D., Edmonds, A., Adar, E., Breuel, T.: Personalized Search. Communications of the ACM 45(9), 50–55 (2002)
12. Dumais, S., Cutrell, E., Chen, H.: Optimizing Search by Showing Results in Context. In: ACM Conference on Human Factors in Computing Systems (CH 2001), Seattle, WA, pp. 277–284. ACM Press, New York (2001)
13. Forecasting with Single Exponential Smoothing, NIST/SEMATECH e-Handbook of Statistical Methods. Retrieved 10/02/2007, from http://www.itl.nist.gov/div898/handbook
14. Liao, I.E., Liao, S.C., Kao, K.F., Harn, I.F.: A Personal Ontology Model for Library Recommendation System. In: Sugimoto, S., Hunter, J., Rauber, A., Morishima, A. (eds.) ICADL 2006. LNCS, vol. 4312, pp. 173–182. Springer, Heidelberg (2006)
15. Middleton, S.E., De Roure, D.C., Shadbolt, N.R.: Capturing Knowledge of User Preferences: Ontologies on Recommender Systems. In: First International Conference on Knowledge Capture (K-CAP2001), pp. 100–107 (2001)
16. Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Item-based Collaborative Filtering Recommendation Algorithms. In: 10th International World Wide Web Conference (WWW 2010), Hong Kong, pp. 285–295 (2001)
17. Ding, Y., Li, X.: Time Weight Collaborative Filtering. In: 14th ACM International Conference on Information and Knowledge Management, pp. 485–492 (2005)
18. Olston, C., Chi, E.H.: ScentTrails: Integrating Browsing and Searching on the Web. ACM Transactions on Computer-Human Interaction 10(3), 177–197 (2003)
19. Perugini, S., Ramakrishnan, N.: Personalizing Web Sites with Mixed-Initiative Interaction. IEEE IT Professional 5(2), 9–15 (2003)

From Collaborative Video Library to Annotated Learning Object Repository: Using Annotated Video Library in Personalized E-Learning

Md Maruf Hasan[1], Nophadol Jekjantuk[1], Yenruedee Chanwirawong[1], Ekawit Nantajeewarawat[2]
[1]Shinawatra University, Thailand
[2]Sirindhorn International Institute of Technology, Thammasat University, Thailand
maruf@shinawatra.ac.th; nophadol_bkk@hotmail.com; yenruedeec@isd.th.ibm.com; ekawit@siit.tu.ac.th

***Abstract* – With the proliferation of digital and video cameras, personal collection of multimedia materials such as amateur video-clips are abundant now-a-days. Most of these multimedia materials may be useful to others if they are shared and can be located easily. Semantic Web technologies hold promise to organize and re-use such non-textual information effectively. However, annotation of multimedia contents is a tedious task. Notwithstanding, as we observe growing number of community collaborations on the present Web, such content annotation can be done through online collaboration. In this research, we investigate the development of a collaborative video annotation system using open technologies and tools where people can upload, annotate and share their personal multimedia collections efficiently. We also examine how the contents acquired and annotated with this collaborative system can be transformed into (1) Reusable E-Learning Contents for personalized learning; as well as, (2) Dynamic Digital Library with exploratory search and retrieval facilities using state-of-the-art Semantic Web technologies.**

I. INTRODUCTION

Lecture materials such as those available under the MIT's Open Course Ware (OCW) initiative inherently consists of multimedia contents including video lectures and other non-textual objects [1]. Due to increased use of WWW in teaching and learning, e-learning contents have been increasingly available on the Internet. In this paper, we will use foreign language pedagogy and other intuitive (but simplified) examples to demonstrate the effectiveness of using collaborative annotation and Semantic Web technologies for sharing and re-using E-learning contents for personalized learning.

Videos play an important role in developing foreign language skills and apprehending foreign culture. Video clips for situational dialogue such as, shopping, dining, and opening a bank account etc. are highly reusable. For instance, a video-clip depicting the scenario of opening a bank account in English may easily be reused for learning French, and vice versa. Agencies involved in promotion of cultural exchange and tourism usually produce vast amount of multimedia materials. However, most of these materials are often available with a few main-stream languages. Therefore, people without any knowledge in such languages fail to make better use of such materials. By adding metadata, transcription and subtitle to video contents, we can facilitate broader use of multimedia materials. For instance, a documentary video about Thai Folk Dance is commonly available in Thai, English or some other main-stream languages. However, with metadata annotations, transcription and subtitle in other languages; such materials can be equally useful for people who are not proficient in main-stream languages.

Over the last decades, with the ubiquity of the Internet and the growing popularity of the Web, we have witnessed growing number of successful small community collaborations on the WWW [2]. The success of such collaborations largely depends on the development of a sophisticated system (a collaboration platform) with easy-to-use interface and tool-support. In this paper, we investigate the development of a *Collaborative Video Annotation System* which can facilitate collaborative learning of foreign language and promotion of cultural exchange. We argue that with the increasing availability of Internet and bandwidth, digital divide is not always due to technological barrier but often due to our inability to integrate human sharing spirit and goodwill along with the technological advances.

II. SCENARIO OF FOREIGN LANGUAGE LEARNING

Main-stream foreign language learning materials have a huge market and therefore, publishers worldwide have been actively publishing study materials for the target market. Main-stream foreign language such as English has plenty of learning materials available in many native languages. However, it is truly unfortunate (and undesirable) that a Vietnamese often needs to learn Thai using learning materials written in English, and vice versa. The market-driven phenomenon left some languages far behind from the others. Our language and culture may or may not have a feasible market but they are our unique assets, and we ought to find ways to promote them using enabling technologies. By successfully utilizing the potential of the Internet and human spirit of collaboration, it is possible to promote our language and culture in an online collaborative fashion. The development of our video annotation system is inspired by

language education and cultural exchange by making use of the potentials of ubiquitous Web and online virtual collaboration.

Unlike some technological disciplines, foreign language teaching and learning follows a typical pattern and rather static. Off-the-shelf foreign language learning kits generally consist of lessons organized into different proficiency *levels.* For instance, the *Beginner's Level* consists of lessons with basic alphabet-set, phonetics and writing system along with greetings, basic vocabulary, simple grammars and conversations. The *Intermediate Level* materials go beyond the basics and include further lessons and practices with the help of situational dialogues and further syntactic, semantic and pragmatic notes and drills. *Advanced Level* learning materials are less restrictive and often go beyond typical textbook lessons – it is not uncommon to use real newspaper articles, movies, TV news or documentaries to teach different linguistics and cultural facets to the advanced learners. In many cases, we noticed that a vast majority of these off-the-shelf software products are mere digitization of their predecessors - the video-based learning kits. However, there are some exceptions. For instance, ELLIS [3] – an English learning software tries to integrate ESL pedagogical research with computer and networking technologies. Such materials are often expensive and beyond the purchasing power of individuals.

In terms of contents, we noticed many similarities across foreign language learning materials. Unless commercially motivated, contents such as alphabet chart, pronunciation audio files, grammar flashcards, situational dialogue videos could have been reused effectively for teaching that foreign language to any native language community through some customization and annotation.

Screenplay [4], a Japanese publisher uses foreign films to help Japanese learners improve their foreign language ability. Using their Web site and multimedia products, a user can also conduct associative search to see a word in the real context. This is not only effective as endorsed by scholarly research findings [5], but also a fun activity which promotes the understanding of culture, etiquette and social practice beyond mere language learning through memorization of grammar rules, substitution drills, and the like.

III. Scenario of Cultural Exchange

Governments and agencies across the world do invest huge amount of money and efforts in producing documentaries and other multimedia materials for the promotion of culture and tourism. However, their focus remains limited to some targeted audiences (and therefore, limited to a few main-stream languages). For instance, most documentaries and videos developed in Japan are often available in Japanese, English, French, and Chinese, etc. Unless properly annotated with metadata, transcripts, subtitle or soundtrack, it is unlikely to find those materials useful for an audience who do not understand any of these languages.

With our decades of experience with other online collaborative projects, such as Project Gutenberg [6] and Aozora Bunko [7] - where copyright-free books are digitized and even translated by volunteers worldwide, we expect that such promotional video materials can equally be annotated collaboratively and free-of-charge as long as an easy-to-use collaborative system is available and online volunteers are supported with easy-to-use interface and tools.

IV. Copyright and Other Issues

It is a fact that commercially developed materials may not be available free, but it is inevitable that some promotional materials from *not-for-profit* agencies and governments may gradually become available publicly. We also rely on amateur personal videos taken by digital and video cameras as a good starting point. The Open Video Project [8] is a general digital library of publicly available videos from unrestricted domains – ranging from classroom lectures to public service or documentary videos – may also be imported and annotated in our system. However, at the moment, we target to gather and annotate video and multimedia materials in a *restricted domain* through small-community collaboration. Our primary goal is to promote cultural exchange and facilitate foreign language education while making full use of the ubiquitous connectivity of the Internet and the growing human spirit of virtual collaboration on the WWW. In doing so, we focus on developing an easy-to-use collaborative video annotation system which we have explained in this paper in detail. We also refer to our other publications on creation of sharable and reusable E-Learning materials using a component-oriented approach [9] – where we demonstrated that such annotated contents with proper metadata, and relevant structural, semantic and pedagogical information, it is possible to generate customized and personalized course materials for *any* discipline (not restricted to foreign language learning).

In any voluntary online collaboration, *motivation* plays a crucial role. The success of our initiative therefore, remains on the enthusiasm of volunteers who take pride and care about their own language and culture. The evidence of the steady growth of collaborative projects on today's Internet and WWW (e.g., the Open Video Project as mentioned earlier) is encouraging. Moreover, a collaborative project such as this one requires a *critical-mass* of volunteers and contents to become successful. It is obvious that unless we have a handful number of volunteers or a sizeable amount of contents, our initiative may not receive sufficient attention quickly.

With an initial support from *Asia-Pacific Telecommunity* (APT) and *Thailand Research Fund* (TRF), we have initiated this project and developed a fully working prototype which is currently operational at http://apt.shinawatra.ac.th/video/. Some of our international collaborators are consistently supporting us by developing or accumulating useful contents for promotion of foreign language and culture on the cyberspace. We envisage that we will continuously receive supports from governments and agencies to develop this project further.

In the rest of the paper, we will explain the implementation details and major features of this video annotation system followed by their effective application in E-Learning.

V. THE COLLABORATIVE VIDEO ANNOTATION SYSTEM

The video annotation system we developed is fully Web-based. Users interact with the system using their Java-enabled Web-browsers without having to resort to any video plug-in. We choose *MediaFrame* [10] – an open-source software for streaming video manipulation and playback. We use MediaFrame's JavaScript API, PHP scripts and MySQL database to develop our video annotation system. The technical complexities of the system remain hidden to the users since users interact with the system using an easy-to-use Web interface.

A. MediaFrame: Streaming Media Player

MediaFrame is an open-source streaming media platform in Java™ which provides a fast, easy to implement, and extremely small applet that enables web users to view audio/video contents without having to rely on external player or plug-ins. MediaFrame does not require special servers, software or programming knowledge. Integration of MediaFrame with other applications becomes easy due to its well-defined JavaScript API and its support to Mpeg-1 and Mpeg-4 video formats.

MediaFrame can deliver audio and video contents over the Internet in either streaming or progressive download mode. It is also capable of stretching media by up to 60% of its original size without a significant loss of quality. This enables us to simulate higher bit rates without the associated costs and to operate effectively in both broadband and narrowband environments. We want our system to be reachable by people in the remote area where the infrastructure is not as good, or may have to resort to satellite links for a connection. MediaFrame is capable of detecting user's connection speed in real-time and delivers media accordingly.

A full-featured *Digital Rights Management* (DRM) system is also integrated with the core of MediaFrame allowing media files to be locked to a specific domain and expired over time. This feature is crucial in disseminating copyrighted/restricted contents.

Moreover, MediaFrame is fully JavaScript controllable; making it simple to build our own control set and to integrate it with our video annotation system seamlessly. Using the JavaScript API, we could also conveniently add transcripts and subtitles to video materials. The multiple-playback-state feature also allows us to tightly integrate MediaFrame for effective video annotation and playback.

B. Metadata Annotation

For non-textual materials, such as videos, descriptive metadata plays an important role in exploratory search and browsing [11]. MPEG-7, LOM and SCROM [12] are sophisticated metadata sets for multimedia content annotation. However, we adopt a simpler subset of descriptive metadata for the system in order to avoid fears in the ordinary users. At the moment, the system allows users to annotate a video-clip with *Title, Genre, Original-Language, Keywords, Descriptions* and some other essential metadata. Whenever applicable, structural metadata such as *Video Format, Duration, Color,* etc. are extracted automatically from the video file.

C. Supplementary Annotations

The video annotation system allows a user to step through a video-clip and to add transcripts along with its original soundtracks and transcripts. User can also select transcript in an available language and translate it into a target language. During playback, users can choose from available transcription-languages to be displayed as subtitle with the video (cf. Fig 1-4).

It should be noted that for effective learning of foreign language using video, transcriptions need to be further annotated with extra information such as, pronunciation and grammatical annotations or relevant cultural notes. Such features are currently being implemented.

D. Some Representative Video-clips and Their Annotations

In this subsection, we explain a series of video-clips from our system, which summarize the effectiveness of our system in the context of language learning and cultural exchange.

Fig. 1 shows a situational dialog video-clip from ELLIS played with English (Fig. 1a) and Thai subtitle (Fig. 1b). Fig. 2 shows amateur video-clips of Kashima Jingu Shrine in Japan (Fig. 2a), and about Japanese Paper Art (Origami, Fig. 2b) – although the original soundtrack is in Japanese, English subtitles are made available through online annotation, and displayed accordingly. Fig. 3 shows a video tutorial from JEITA about *Natural Language Processing* (NLP) technology (originally with Japanese soundtrack) but is being played with English subtitles since such information has already become available through online annotation. In doing so, we made it possible for non-Japanese audience to benefit from this tutorial

For all these video-clips, descriptive metadata are also available in the system (annotated collaboratively and indexed automatically) to facilitate exploratory search and navigation [11]. Users can use keyword search (analytical strategy) or browse (partition strategy) the video collection effectively with the help of relevant metadata.

Fig. 1a. Situational dialogue in a bank (played with English subtitle).

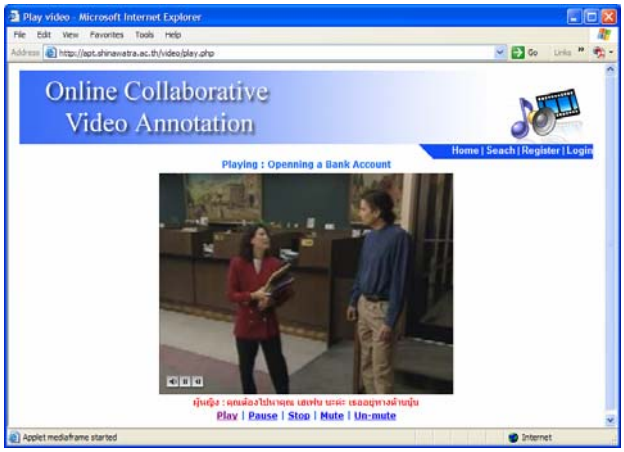

Fig. 1b. Situational dialogue in a bank (played with annotated subtitle in Thai).

Fig. 2a. Amateur video-clip: Kashima Jingu Shrine in Japan (original soundtrack in Japanese but played with annotated English subtitle).

Fig. 2b. Amateur video-clip: Japanese Origami or Paper Art (original soundtrack in Japanese but played with annotated English subtitle).

Fig. 3. A video tutorial developed by JEITA on *Natural Language Processing* – originally in Japanese, but we added English subtitles and descriptive metadata online.

Finally, Fig. 4 shows an example video annotation interface. Video is displayed with playback control, time-stamp and other information. In the process of filling in the *Time* and *Subtitle* boxes below, annotating user can choose reference subtitle from available source languages, and control the video playback with control buttons as necessary. All other annotation interfaces are similar browser-based, and therefore, easy-to use.

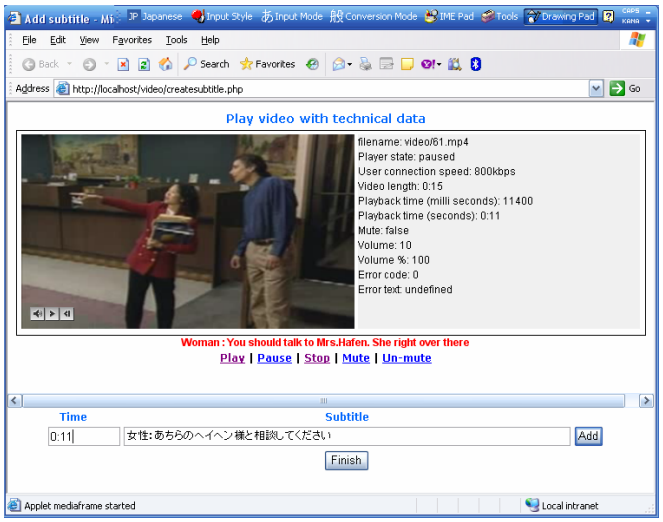

Fig. 4. Subtitle annotation interface: User can choose a source language and add subtitle in a target language. Timestamp and Playback Control are displayed as video-clip is being played.

VI. Applications of Annotated Video Library

The video annotation system serves as a *platform* for *small community collaboration* online as well as an *input channel* of contents. Multimedia contents gathered, annotated and manipulated in this way should be used (disseminated) in useful applications. The following two subsections explain

two unique applications in personalized E-Learning and dynamic Digital Library.

A. Intelligent Dissemination of E-Learning Contents

The success of gathering and disseminating sharable, reusable and customizable e-learning contents depends on developing an easy-to-use collaborative system – a system that not only hides the complexities behind a simple form-like web interface (for submission, annotation, retrieval, and the like), *but also* organizes and manipulates contents in a structured and efficient manner so that intelligent inferences can also be made. For example, content-level dependencies should be propagated and preserved at lesson or course level and so on. Semantic Web technologies including metadata and domain-ontology are used in the background so that both humans and software agents can equally effectively access and manipulate the annotated contents from the video annotation server.

We proposed a *3-Tier Architecture* for component-oriented E-Learning Content Management [9] as shown in Fig. 5. We use ontology to organize contents and ontology-based reasoning to make inferences about contents such as content dependency or other pedagogical attributes. Given that we have gathered and annotated contents with their pedagogical attributes, our current E-Learning prototype can respond to user's content needs intelligently. For instance, finding contents for *"10-hour* lessons for *introductory* Japanese focusing on *greetings and everyday conversation",* or a "*42 hours* of course materials for a *System-oriented* Database course*"* - can be created on-the-fly as long as annotated contents are available.

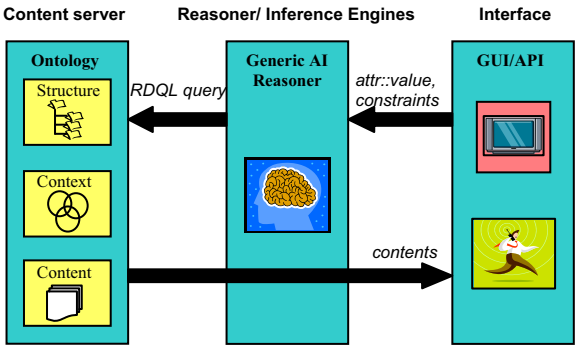

Fig. 5. *3*-Tier E-Learning Content Management Architecture.

A currently working prototype is available at http://apt.shinawatra.ac.th/ecms/. Internally, we organize the contents in an OWL-ontology based on the annotated metadata. Our current prototype uses *a Semantic Web Reasoner* (OWL-based Pellet, [13]) to identify appropriate contents from the ontology that satisfy user's criteria. Users interact with the system using Web-based interface. Fig. 6a and Fig. 6b explain how user may specify different constraints and find appropriate contents that satisfy their criteria and content dependency from a *Database* related course contents.

For example, a teacher or a learner trying to retrieve a *20-hour Foundation Course in Database* only needs to specify the criteria through the Web-based interface (Fig. 6a). User criteria and content dependencies are then verified using a G*eneric AI Reasoner* to locate the appropriate contents. The user is then presented with the appropriate contents in an organized fashion (Fig. 6b).

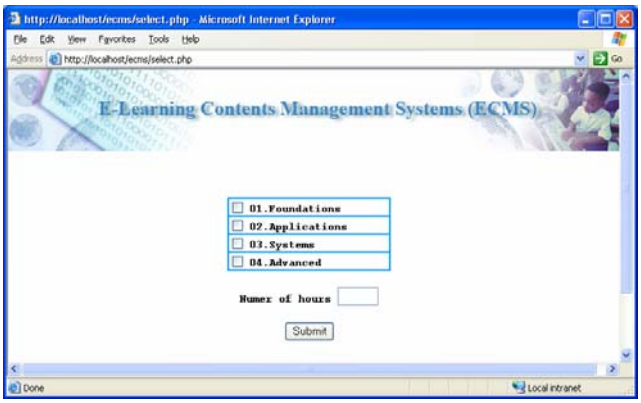

Fig. 6a. Search interface of E-Learning contents (example shows contents search for a *Database* course contents).

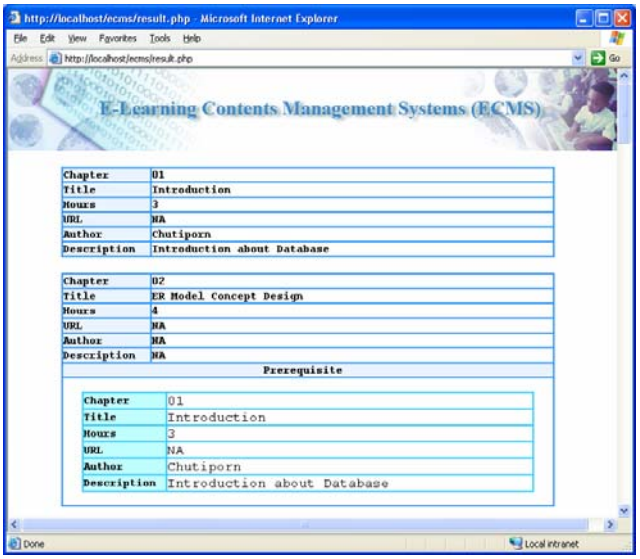

Fig. 6b. E-Learning contents retrieved and presented in an organized fashion Output reflects the criteria/constrains specified by user as well as content-level dependencies.

B. Dynamic Digital Library

Using Semantic Web technologies and exploratory search and browsing interfaces as explained in [11] and tested on Open-Video collection [8], we plan to disseminate our multimedia-contents arriving from the video annotation system (i.e., live *input channel*) in the form of *Dynamic Digital Library*. For example, a user interested in Temples and Pagodas in a particular region will define a *Library Template* by specifying the nature of contents (such as, video-clips of Temples and Pagodas in Thailand can be easily specified since our contents are annotated with descriptive metadata and organized in an ontology). Contents and their organizations in such dynamic libraries are automatically updated as new

materials are deposited or contents are updated, annotated or reorganized on the video annotation server.

VII. CONCLUSIONS

In this paper, we outlined an operational prototype of a collaborative video annotation system developed for foreign language education and cultural exchange in mind. We often ignore the fact that majority of Internet users are not proficient in foreign language and therefore, materials available on the WWW in a foreign language has little or no use to those people. Multimedia materials, such as videos are relatively easy to understand or appreciate without *complete* translation (since contextual, visual and other cues are present). Nevertheless, it is extremely difficult to locate non-textual materials on the Web using today's keyword-dependent *hit-or-miss* search engines which heavily rely on textual indexing. Metadata annotation is therefore essential and can be done effectively in small-group collaboration. We do admit that we are yet to conduct any experimental evaluation of the proposed system to justify the effectiveness of this approach. However, our future work will eventually attempt such evaluations and vigorous live trials.

We have also outlined an ontology-based *E-Learning Content Management* prototype for *personalized* E-Learning. We have plans to organize our video collections using Semantic Web technologies to support personalized foreign language education. We explained how our annotated video library with its multimedia contents can serve as (1) an *E-Learning Content Server* - storing both contents and structures (contents with annotated pedagogical attributes in an ontology) to facilitate personalized E-learning; and, (2) a *Digital Library Content Server* - storing amateur videos as well as professional multimedia (contents with annotated language and cultural attributes in an ontology) to support cultural exchange and foreign language learning tasks.

ACKNOWLEDGMENT

This research is supported by *Thailand Research Fund* (grant No. MRG4880112) and by *Asia-Pacific Telecommunity* (APT-HRD grant 2005).

REFERENCES

[1] MIT Open Course Ware Project: http://ocw.mit.edu/

[2] Johnson, C.H., "A survey of current research on online communities of practice," Internet and Higher Education Vol.4, pp.45-60.

[3] ELLIS Inc.: http://www.ellis.com/

[4] Screenplay: http://www.screenplay.co.jp/

[5] Cordilo, D.S. (1997). Using a Foreign Film to Improve Second Language Proficiency: Video vs. Interactive Multimedia, Journal of Educational Technology Systems, Vol.25 no.2 pp.169-77.

[6] Project Gutenberg: http://promo.net/pg/

[7] Aozora Bunko Project: http://www.aozora.gr.jp/

[8] Open Video Project: http://www.open-video.org/

[9] Hasan, M.M., S. Yamamoto, Y. Fujino and W. Chujo, "Towards Sharable and Reusable Contents using Semantic Web Technologies for Personalized Foreign Language Learning," IEICE General Conference 2006, Tokyo, Japan; D8: Artificial Intelligence and Knowledge Processing Track - Paper No.: D-8-1; pp.84, 2006

[10] MediaFrame: http://Mediaframe.org/

[11] Marchionini, G., "Exploratory Search: From Finding to Understanding," Communications of ACM, Vol.49, No.4, pp.41-46.

[12] Godwin-Jones R., "EMERGING TECHNOLOGIES - Learning Objects: Scorn or SCORM?" Language Learning & Technology, Vol.8, No.2, pp.7-12

[13] Pellet Reasoner: http://pellet.owldl.com/

Natural Language Processing and Semantic Web Technologies in the Digital Library

Md Maruf Hasan
School of ICT, Shinawatra University, Thailand

Ekawit Nantajeewarawat
Sirindhorn International Institute of Technology
Thammasat University, Thailand

In this article, we review the state-of-the-art of natural language processing and Semantic Web technologies, and explore their potential applications as value-added digital library services. We argue that a digital library is not merely a collection of digital or digitized objects with search and browse interfaces. Several definitions and views of digital libraries have appeared in publications over the last decade - some emphasizing the system aspect and others emphasizing the service aspect. We analyze the trends in digital libraries in an unbiased fashion, and explore how human language technology and other technologies can enhance the functions and services of future digital libraries.

Introduction

Before the 1940s, there were primitive experiments with the storage, manipulation, and retrieval of information held in mechanical form. Since World War II, computer technology has demonstrated the power of handling and processing vast amounts of electronic information with speed and efficiency. Ever since, the revolution in publishing and library work has begun and evolved through a new era known as the "eLibrary," "digital library," or "virtual library."

The American Chemical Society, through its Selective Dissemination of Information (SDI), began alerting scholars through electronic means as far back as 1962. In 1965, Chemical Biological Activities appeared both in print and on tape. Medical Literature Online (MEDLINE) and Project Gutenberg appeared in 1971. In the 1990s, CD-ROM publications seemed to be cheaper than their printed counterparts. Moreover, since the early 1990s, with the massive growth of the Internet and the popularity of the World Wide Web, things have started to change rapidly for both publishing and libraries. With ubiquitous connectivity and the proliferation of cheaper mass storage devices and powerful processors, we have discovered new means of storing, manipulating, and accessing information. Such developments have significantly changed the way we publish, store, and seek information. The result is the appearance of "digital libraries."

Easy-to-use open source digital library software has not only given rise to the creation of digital collections of plain texts or scanned images but has also facilitated the creation of music, video, and animation libraries (Witten et al., 2001a and b). As a result, nowadays we often come across rudimentary digital libraries with heterogeneous architectures offering diverse services. The ease of HTML authoring and the simplicity of the HTTP protocol gave birth to the current World Wide Web with billions of heterogeneous and semi-structured documents and links. The simplicity in developing digital collections with the help of easy-to-use tools (DL-in-a-Box from Virginia Tech, USA, Greenstone Digital Library [GSDL] from the University of Waikato, New Zealand, D-Space from MIT) is similarly giving birth to a plethora of digital libraries with information in heterogeneous structures and formats (languages, media types, and so on). The accessibility of such libraries is extended to the world via the World Wide Web. Digital libraries are inevitably facing setbacks (in the form of heterogeneity and integration issues) similar to those with the "good-old-WWW." The Semantic Web Initiative, as described by Tim Berners-Lee et al. (2001), is an attempt to overcome these setbacks with the current Web by organizing heterogeneous and semi-structured Web resources in a meaningful way. Interoperability and integration research in the context of digital libraries has also started to appear (cf. the Open Archive Initiative [OAI]).

A digital library is not merely a collection of digital (or digitized) objects with search and browse interfaces. One of the first research-

oriented definitions of "digital libraries" appeared in an article by Borgman (2000, p. 42, citing Borgman et al., 1996):

Digital libraries are a set of electronic resources and associated technical capacities for creating, searching, and using information. . . . Digital libraries are constructed — collected and organized — by [and for] a community of users and their functional capacities support the information needs and uses of that community.

The above definition certainly reflects the expectations of typical library users from a digital library in the electronic era. Several other definitions of digital libraries have continuously appeared in publications over the last decade. These definitions and views reflect various perceptions about digital libraries. In the next section, we will analyze some of those views and definitions in an unbiased fashion. In the latter sections, we will investigate what roles human language technologies (or natural language processing [NLP]) and Semantic Web technologies may play in shaping the future of the digital library.

Digital libraries - system or service or both?

Although there are various definitions and views of digital libraries, we can safely summarize them in terms of two categories: system-oriented views and service-oriented views. The following definitions are representative of such views from each category:

System-oriented definition:
(from the Greenstone Digital Library System [GSDL] Developers Team)

A digital library is a focused collection of digital objects, including text, audio, and video, along with methods for access and retrieval, and for selection, organization, and maintenance of the collection.

Service-oriented definition:
(from the Digital Library Federation)

Digital libraries are organizations that provide the resources, including the specialized staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities.

It is interesting to note that library-oriented people or entities (such as the Digital Library Federation) often emphasize the service aspects of digital libraries. On the other hand, computer-oriented people (such as the members of the GSDL Developers Team) tend to focus on the system aspects of digital libraries.

The most unbiased definition we found during the literature review is given by Gladney et al. (1994).

A digital library is an assemblage of digital computing, storage, and communications machinery together with the content and software needed to reproduce, emulate, and extend the services provided by conventional libraries based on paper and other material means of collecting, cataloguing, finding, and disseminating information. A full service digital library must accomplish all essential services of traditional libraries and also exploit the well-known advantages of digital storage, searching, and communication.

In the recent past, most articles appearing in the two special issues on digital libraries in the *Journal of the American Society for Information Science* (2000) provide excellent accounts of system-oriented digital library research, but only a few articles focus on the service aspects. On the other hand, the special issue of Information Processing and Management (1999) includes many articles focusing on the service aspects of digital libraries. Similar patterns are obvious for international conferences in the area of digital libraries and related areas such as library automation. Conferences led by computer professionals (e.g., the International Conference on Digital Libraries [ICDL] and its European and Asian counterparts, ECDL and ICADL) include mainly system-oriented presentations, but conferences led by library professionals (e.g., the Dublin Core Metadata Conference or Computers in Libraries) have inherently emphasized digital library services.

In fact, design and research issues in digital libraries are complex and multifaceted. In one of the articles in the special issue of *Information Processing and Management,* Marchionini and Fox (1999) state that "Digital library work occurs in the context of complex design space shaped by four dimensions: community, technology, services, and content." They also add that:

- The **community** dimension of digital libraries reflects social, political, legal, and cultural issues.
- **Technology** serves as the engine moving the digital library field, including technical progress in computing, networking, and more specifically information storage and retrieval, multimedia, and interface design.
- **Services** form the central focus of digital libraries and future digital libraries should facilitate digital reference services, real-time question answering, on-demand help, information literacy, and user involvement mechanisms.
- **Content** represents all possible kinds of form and genres of information (not restricted to those meant for purchase and subscription as

it is typically the case for traditional libraries): for example, print as well as digital, meant for purchase or subscription or freely available on the Web.

Therefore, we envisage that digital library researchers will eventually put the right emphasis on both the system and service aspects of digital libraries in the years to come.

Due to the fact that not every single printed (published) material will be digitized and kept in digital libraries (some published materials will never be available in a digital library), and the reality that digital experiences in a virtual space cannot possibly be materialized in physical form (some contents of digital libraries will never be available in physical form), we are heading towards a **hybrid library** era - where traditional libraries and digital libraries coexist peacefully and offer services to humanity. More precisely, a hybrid library is a library where digital and printed information resources coexist and are brought together in an integrated information service accessible locally as well as remotely (HyLife 2002).

Librarianship as a discipline has a centuries-old history and disciplinary wisdom. A major challenge in digital library research is to find out how to achieve balance during the rapid development of technologies. The relationship between traditional libraries and digital libraries are similar to those between traditional banking and online banking (including ATM machines). Online banking and ATM cash dispensers are offering value-added services to an institution that has existed for centuries.

It is foreseeable that traditional libraries and digital libraries will evolve together taking advantage of technological advancements in computing and communication technologies. With the present state of the art, many of us believe that in the foreseeable future we will be living in a world of hybrid libraries. The rational view of the hybrid library is therefore justified through a proper emphasis on both the service and system aspects of the library as a service institution taking advantage of the development of technology. Similar to the banking analogy above, traditional libraries and digital libraries will complement each other, and new innovative digital library services will enhance and surpass the services libraries have been typically offering their patrons.

Accepting the above reality, in this article we will focus on how advances in human language technologies and Semantic Web technologies may contribute to developing value-added digital services for libraries.

Evolution of electronic publishing, library automation, and the Semantic Web

Following Lancaster (1995), we summarize the evolution of electronic publishing as follows:

1. The use of computers to generate conventional paper-based publications (desktop publishing era)
2. The distribution of texts in electronic form identical to their printed counterparts (e-book era)
3. The distribution in electronic form only with eventual value-added features, such as search and manipulation (including adaptation) capabilities (value-added publishing era)
4. The generation of completely new publications that exploit the power of both computing and connectivity: e.g., hypertext and hypermedia with links, sounds, animation, and so on (new generation publishing era)

From the above trends, it is obvious that we are heading into an era where it is not always possible (or feasible) to make publications manifest in physical form. For instance, the virtual space or environment is impossible to materialize in physical form.

Similar to publishing, library automation has also evolved through many stages over the last few decades. The original card catalog has evolved through the machine readable catalog (MARC) to Dublin Core metadata. Interoperable protocols such as Z39.50 from the Library of Congress were developed and have been in use since the era of mainframe computing. The most recent, the Online Public Access Catalog (OPAC), is another example of an interoperable protocol. These protocols enable libraries to share and exchange catalog information. System-oriented digital library researchers also developed interoperable protocols and standards such as the OAI-PMH (Open Archive Initiative - Protocols for Metadata Harvesting). The trend is to move towards a union catalog (or universal catalog) - a one-stop source of information to be accessed in an organized and unified fashion locally and remotely. The trend of a union catalog is not to bring information into uniformity, but rather to bring heterogeneous information into some form of agreement (through integration and mapping). The Semantic Web approach also aims to do the same thing.

The most obvious reason behind not considering the present World Wide Web as a universal digital library is its lack of any organization. The Semantic Web initiative aims to give organization and meaning to information. Not surprisingly, similar to libraries, the Semantic Web also relies on metadata (defined as "information about information"). The oldest metadata in human history are the library's card catalog and the relational database schema. The Semantic Web uses the Resource Description Framework (RDF) and relevant schema to describe and organize information (W3C Semantic Web, 2001). It is interesting to note that the history of metadata goes

back to library science as the root of XML-based information encoding goes back to the publishing industry (XML is a subset of SGML used by publishers even before computers were invented).

Ontologies are explicit specifications of conceptualization (Gruber, 1995). In the Semantic Web context, ontologies are the formal representations of entities (concepts) and their relationships. Logical inferences are made possible on the upper layers of the Semantic Web Stack (as explained by Berners-Lee et al., 2001) due to the lower layer machine understandable representations using metadata and ontologies. We should also note that Semantic Web research is dominated by people with experience in artificial intelligence, a branch of computing mainly focused on automating information processing and mimicking human behavior in computers. There are plenty of similarities between Semantic Web initiatives and library science because both are focusing on organizing information efficiently for human (as well as machine) consumption. To some extent, many ideas in the Semantic Web are essentially borrowed from library science. Other than organizing information, Semantic Web research has also focused on interoperability, resource description, discovery, syndication, and federation, as well as personalization and adaptation issues.

We would like to conclude this section by pointing out that machine readable catalog information has been present in libraries for decades. What makes the digital library more promising to us is that in a digital library **content** is also available in machine readable (or sometimes machine understandable) form, which can be processed further. Due to such a fact, human language technology (HLT) — that is, the advances in natural (human) language processing (NLP) using computers — has a potential impact on digital library systems and services. In the next section, we summarize potential applications of human language technology and Semantic Web technologies in the context of the digital library - more appropriately, in implementing value-added services in digital libraries.

Natural language processing and Semantic Web technologies for value-added digital library services

Natural language processing technologies

NLP researchers spent decades of effort in developing fundamental algorithms and techniques in tokenization and word segmentation. Word segmentation (boundary detection) is essential for many Asian languages where words are not explicitly delimited with white spaces. Part-of-speech taggers and morphological analyzers are already available for the syntactic analysis of various languages. NLP researchers have intensively used machine learning, rule-based, and hybrid approaches to model highly ambiguous linguistic phenomena in processing human languages by computers. Parsers (including the special purpose chunk parser and shallow parsers) are also becoming available for a growing number of languages. Language independent algorithms such as word sense disambiguation tools using contextual cues are also available to disambiguate polysemous words. Although not abundant, pragmatic analysis tools and techniques such as those for anaphora resolution or the analysis of argument structures are also available for some languages. Based on fundamental research in NLP, sophisticated applications are already developed and in practical use in the following areas: named entity extraction (identifying entities such as proper names, artifacts, and so on automatically); keyword and key phrase extraction, noun-phrase extraction, and PP attachment analysis; text categorization; automatic summarization; text filtering; information retrieval; information extraction (template filling); and text filtering (similar to SDI). Similar advances have also occurred in analyzing aural and visual language processing. For instance, speech synthesis and recognition (including text-to-speech and dictation tools) and optical character recognition (such as, the recognition of hand or typewritten texts) and image/video analysis and annotation tools are also available in many languages. Integrated text analysis platforms, such as the General Architecture of Text Engineering (GATE) have already been developed for years. It is inevitable that such integrated platforms will be extended into areas beyond textual information.

Semantic Web technologies

The Semantic Web relies on XML based encoding of information (text and beyond) with RDF-schema based metadata annotations. Ontologies provide a conceptualized domain model on top of it. Reasoning and logical inferences are made over the formal representation of objects and their logical relationships. Essentially such representation and manipulation go beyond the current HTML-based encoding and HTTP-based access on the World Wide Web. To satisfy the growing demand for giving meaning to information and organizing it systematically, the Semantic Web community has resorted to new standards and protocols, such as simple object access protocols (SOAPs), Universal Description, Discovery and Integration (UDDI), and Web Service Description Languages and Web Service Flow Language (W3C Semantic Web, 2001).

It may be noted here that early library automation was inherently a straightforward problem solved with simple metadata since the catalog was its main focus. A simple

relational-schema with catalog-fields was sufficient. MARC-like modelling served the purpose and protocols such as Z39.50 were sufficient to ensure interoperability. As digital information proliferated dramatically and the expectations of information consumers (library users) increased, the complexities also grew. The eventual outcome is of course a more sophisticated (and extensible) metadata framework such as the Dublin Core. Due to the proliferation of publicly accessible archives and libraries with heterogeneous metadata architecture and content, interoperable protocols have also evolved (cf. OAI-Protocols for Metadata Harvesting). There is a similar pattern between the evolution from library automation to the digital library with the evolution of the World Wide Web to the Semantic Web.

The Semantic Web has offered two major promises: one in intelligent B2B information interchange and integration, and the other in the intelligent organization of information. Semantic Web technologies in B2B, especially in enterprise application integration and intelligent web service development, have been very successful due to the straightforward nature of the problems that they address. However, the main setback in giving meaning to and organizing information is still due to the sheer tasks of annotating information with metadata and developing ontologies for a broader domain. These tasks should be carried out by automatic (or at least semi-automatic) means. Advances in human language technologies and their successful applications in performing these tasks fairly automatically are challenging. As a motivating example, we refer to the KAON platform (KArlsruhe ONtology and Semantic Web Tool Suite), which includes tools for text to ontology conversion (Text-to-Onto). It is interesting to note that Text-to-Onto uses a parser to analyze texts to identify important noun phrases and verbs and subsequently map them into concepts and their relationships respectively. When the text collection (or corpus) is homogeneous (from a specific domain), the rudimentary ontology produced automatically is close to a rudimentary domain model.

Digital libraries may also take advantage of developments related to integration and interoperability research in the Semantic Web. Topic maps (Topic Map Consortium; Ding, 2003) are proven examples of Semantic Web content integration and interoperability. Nevertheless, personalization and adaptation issues in the context of the digital library and the Semantic Web are virtually the same. In the Semantic Web, personalization and adaptation are basically done by integrating personal profiles and activities with target objects (resources) using ontology-driven domain models and logical inferences.

Potential digital library applications of NLP and Semantic Web technologies

The machine understanding of digital content is inevitable and natural language processing technologies will continue to play an important role in this regard. Electronic content and usage patterns are useful in developing value-added digital library services. For instance, most text-based digital libraries are capable of offering full-text search with the help of automatic indexing techniques. However, for languages such as Thai, Chinese, and Japanese, where explicit word boundaries are not present, automatic indexing becomes cumbersome. Natural language processing (NLP) researchers have developed efficient algorithms and tools for word segmentation as well as other sophisticated tasks for language analysis (Cole et al., 1996). Asian language digital libraries are increasingly making use of such technological advances (Hasan et al., 2004).

Similarly, the automatic extraction of keywords (or key phrases) has been made possible for many languages using sophisticated machine learning techniques. Another newer area of research in NLP, known as "named entity extraction," has successfully demonstrated that we can automatically extract entities (such as proper name, organization, location, artifact, date, time, money, and percent) from full texts. Structural analysis tools are also developed to extract title, author, and section headings (for example) from the visual and typographical analysis of documents. Semi-automatic metadata extraction is possible by integrating and extending such techniques.

Efficient part-of-speech tagging tools, morphological analyzers, and parsers have already been developed for many languages (for examples, see NECTEC Thai Wordbreak Insertion Service; Japanese Morphological Analyzer), and they are extremely useful for text analysis (content understanding). For instance, with the help of a parser, we can easily extract noun phrases (mainly entities or concepts) from text. Tools like Text-To-Onto are able to parse a sentence to identify relationships (verbs) and concepts (noun phrases) and convert them into ontology as explained before.

Speech researchers from the NLP community have developed sophisticated tools for speech synthesis and recognition (text-to-speech and speech-to-text conversions). Such tools and techniques may offer convenient access to information for hearing impaired people. Nevertheless, for an audio or music digital library, an aural search interface is the most desirable means of search and retrieval.

Optical character recognition (OCR) tools developed by NLP researchers offer tremendous

help in extracting texts from document image collections (such as typewritten and handwritten scanned documents). Although often incorrect, such tools usually offer a manual post-editing feature. Some existing digital libraries (including the ACM Digital Library) use scanned document images with OCR-generated texts for older published items. An OCR-generated full text or summary is often used in automatic indexing.

Automatic classification based on keywords and phrases has potential in categorizing digital library collections spontaneously. By integrating user profiles with a keyword hierarchy, it is also possible to generate a personalized classification hierarchy.

The collaborative filtering technique has demonstrated great success in e-commerce by recommending to buyers items of potential interest (i.e., the "customers who bought this item also bought the following items" assumption in e-commerce maximized sales with minimal advertising cost). With the analysis of individual user usage patterns, it is possible to offer similar services in a digital library (i.e., "readers who read this item also read the following items"). There are also avenues to incorporate personal profiles and preferences into a digital library system to implement a truly personalized and adaptable digital library.

Although not entirely based on NLP research, the visualization of a digital library collection using state-of-the-art visualization techniques and keywords relationships provides a value-added interface to digital library services. Library users become accustomed to the visible and organized presentation of available items like those in a supermarket.

It should be noted that more and more digital libraries contain information in multiple languages. Cross-language information retrieval (CLIR) that can retrieve information in a query language as well as foreign languages is also a possibility in the digital library context. The common CLIR approach uses a machine-readable multilingual dictionary (machine-readable dictionaries are fundamental NLP tools!) and metadata/keyword alignment across languages. Nevertheless, there are many other approaches to CLIR (Oard, 1997).

We would like to conclude this section by pointing out that although technically feasible, most of the above-mentioned value-added features and services are still not present in many digital libraries. This is probably due to some negligence of the service-oriented and user-oriented aspects of digital library research. Nevertheless, simple features such as full-text indexing (similar to Internet search/browse engines) are mostly available with existing digital libraries. Research shows that our searching behavior in the library environment (e.g., OPAC catalog searching) and the Web environment (e.g., Google searching) is very different. Unlike Internet search engines, digital libraries have to offer a guided and systematic search interface, and are expected to produce exhaustive hits (**all** available items must be presented for a query). The traditional user-librarian relationship in the digital library era is difficult to put into practice. However, significant research effort has yet to be put into **user interface** and **interaction** with the digital library since a human librarian is not always present in the digital library environment.

Conclusion

In this article, we have tried to rediscover the digital library as a balanced blend of system and service by giving both librarians and computer scientists credit for their contributions to current developments in library automation. Although research and development in digital libraries over the past decade has been intensive due to the digital library initiatives (*DLI-I* and *DLI-II*) in the United States, *eLib* (1, 2, 3) in the United Kingdom, and similar initiatives in Europe, Japan, and elsewhere, our review and analysis show that there are still plenty of opportunities in digital library research.

This conclusion is especially true when we consider human factors along with technological advances. Obviously, the development of computing power, communication speed, and storage media has played a key role in making digital libraries successful. However, the spirit of human collaboration has silently played an equally crucial role in the advancement of digital libraries. For instance, Project Gutenberg would never have succeeded without human volunteers scanning or typing copyright free publications and depositing them into the digital library.

Developing value-added services in digital libraries is as important as building useful digital library collections. Easy-to-build digital library systems and the availability of abundant electronic content may give rise to the number of digital libraries here and there on the Web. However, they will eventually suffer the same fortune as the World Wide Web as a whole. The Semantic Web initiatives which aim to overcome the drawbacks of the current Web offer good lessons as well as enabling new technologies to develop efficient digital libraries and useful digital library services. In this article, we pointed out how natural language processing and Semantic Web technologies can contribute to developing value-added services to digital libraries. Being a service-oriented institution, libraries (digital or not) have had the sole objective for centuries to quench the human thirst for information and knowledge. As digital libraries or hybrid ones, they will eventually be

serving humanity with easy-to-use systems, and easy-to-access augmented services.

Acknowledgement

The authors would like to acknowledge support from the Thailand Research Fund in terms of a two-year research grant to investigate open digital library research issues. The project focuses on multilingual and interoperable digital libraries capable of offering some of the value-added services cited in this article.

References

Berners-Lee, T., Hendler, J., Lassila, O. 2001. The Semantic Web. Scientific American 284(5), 34-43.

Borgman, C. 2000. From Gutenberg to the Global Information Infrastructure: access to information in the networked world. ACM Press, New York.

Cole, R., et al., 1996. Survey of the state of the art in human language technology. Available at <http://cslu.cse.ori.edu/HLTsurvey/HLTsurvey.html>

Denaux, R., et al. 2005. An approach for ontology-based elicitation of user models to enable personalization on the Semantic Web. Available at <http://www2005.org/cdrom/docs/p1170.pdf>

Digital Library Federation: <http://www.diglib.org/>

Ding, H. 2003. Challenges in building semantic interoperable digital library system. Available at <http://www.idi.ntnu.no/grupper/su/courses/dif8901/Essay2003/essay2003-haoding.pdf>

General Architecture of Text Engineering: <http://gate.ac.uk>

Gladney, H., et al. 1994. Digital library: gross structures and requirements. Report from a 1994 DL Workshop. Available at <http://www.csdl.tamu.edu/DL94/paper/fox.html>

Greenstone Digital Library Suite: <http://www.greenstone.org/>

Gruber, T. 1995. Toward principles for the design of ontologies used for knowledge sharing. International Journal of Human and Computer Studies, 43(5/6), 907-928.

Hasan, M., Takeuchi, K., Isahara, H., Sornlertlamvanich, V. 2004. Digital libraries in Asian languages: a TCL initiative. Pages 365-372 in T. Sembok et al (eds.), Digital Libraries: Technology and Management of Indigenous Knowledge for Global Access. Springer's Lecture Notes in Computer Science, Vol. 2911.

HyLife, 2002. The Hybrid Library Project: <http://hylife.unn.ac.uk/>

Japanese Morphological Analyzer: ChaSen <http://chasen.aist-nara.ac.jp>

The KArlsruhe ONtology (KAON) and Semantic Web Tool Suite: <http://kaon.semanticweb.org/>

Lancaster, F. 1995. The evolution of electronic publishing. Library Trends, 43(4), 518-527.

Marchionini, G., Fox, E. 1999. Editorial: Progress towards digital libraries: augmentation through integration. Information Processing and Management, 35(3), 219-225.

NECTEC Thai Wordbreak Insertion Service: SWATH <http://ntl.nectec.or.th/services/wordbreak/>

Oard, D. 1997. Serving users in many languages: cross-language information retrieval for digital libraries. D-Lib Magazine, 3(12). Available at <http://www.dlib.org/dlib/december97/oard/12oard.html>

Topic Map Consortium: <http://www.topicmaps.org/>

W3C Semantic Web site: <http://www.w3.org/2001/sw/>

Witten, I., Bainbridge, D., Boddie, S. 2001a. Open source digital library software. D-Lib Magazine, 7(10). Available at <http://www.dlib.org/dlib/october01/witten/10witten.html>

Witten, I., Bainbridge, D., Boddie, S. 2001b. Power to the people: end-user building of digital library collections. Pages 94-203 in Proceedings of the Joint Conference on Digital Libraries, Roanoke, Virginia. Available at <http://www.acm.org/pubs/articles/proceedings/dl/379437/p94-witten/p94-witten.pdf>

Extending the Reach of Typical Library OPAC - A New Way of Community-based Resource Sharing using Distributed Cataloging and Z39.50 Protocol

Md Maruf Hasan, Kovit Bonsri[1] and Nophadol Jekjantuk[2]
Shinawatra University, Thailand
maruf@shinawatra.ac.th; kovit.b@cattelecom.com; nophadol_bkk@hotmail.com

Abstract

The Online Public Access Catalog (OPAC) is a computerized online catalog of the materials held in a library. Typically, the library OPAC databases (collections and circulation data) are centrally managed by professional library staff. In OPAC systems, the library users have a little control over the collection and circulation data. In most cases, users of a library generally belong to a community and live in close proximity (such as the students and staff of a university, or people living in the same neighborhood). The users may collectively own (and willing to share) huge amount of items of their own if there is an enabling system to help them doing so. With the advent of the Internet and WWW, people are communicating, collaborating and sharing information online in exciting new ways. In this research, we investigate a decentralized cataloging system where people in a community can create and manage shared catalogs of published materials they own. Unlike the library catalog, this community catalog is managed in a decentralized fashion, where each user is responsible to create catalog data about their own collection, and to manage circulation data as items are borrowed and returned by fellow users. We use Z39.50 protocol to integrate the centrally managed library-catalog with the decentralized community-catalog, seamlessly. Through this integration, users may access both the library-catalog and the community-catalog using a single interface.

Keywords: Digital Library; Online Public Access Catalog (OPAC); Z39.50 Protocol; Online Virtual Collaboration.

1. Introduction

For centuries, Libraries have been playing a vital role in sharing published materials among its members. The advent of Internet and WWW made it easier for people to communicate, collaborate and share information online. We are constantly witnessing exciting new applications of online virtual collaboration. In this research, we develop an online cataloging system where people in a community effectively create and manage a *shared catalog* of published materials (such as, book, DVD, or other electronic files) they own. Unlike the library catalog, this community catalog is managed in a decentralized fashion, where each user is responsible to create catalog information about their own collection, and manage circulation data as items are borrowed and returned by fellow users in the community.

The *Community Catalog System* (CCS) is developed using open source resources such as PHP scripts and MySQL databases. We borrow some ideas from DVDdb -- an open source web-based cataloging system originally developed to allow a group of users to keep track of their DVD collection and circulation. We extended the DVDdb [1] System to handle myriad of materials including books, DVD, and other electronics files. Cataloging information kept in our system is similar to those found in a library catalog, or commercial databases such as CDDB or IMDb as appropriate. In order to help users input or update catalog information effectively, we implement utilities which take unique identifier (e.g., ISBN) as input, and query an appropriate database (e.g. Library of Congress Z39.50 server) to automatically extract catalog information (e.g., title, author, publisher, etc.).

The entire system is designed to be easy to use. Users use web-based interface to interact with the system and need not to worry about the backend databases and other technical complexities. The backend catalog database is wrapped with a *Z39.50* protocol wrapper [2] to facilitate seamless integration with existing library catalogs. Therefore, when a user uses the library OPAC (Open Public Access Catalog) to search for library materials, he or she will *not only* find materials held by the library *but also* materials own by other users in the community, and vice versa. We demonstrate that the collaborative catalog maximizes resource sharing and utilization within an organization or a community. On many occasions, an item not available in the library (library-catalog) is perhaps owned by other library users (community-catalog) who are willing to share it with the fellow users in the community. Moreover, CCS is a Web-based system that helps users conveniently manage their own collections without having to setup, maintain and backup any databases. Very few people do keep record of their own collections of books, DVDs and other materials and end up losing track of some of their valuable collections

[1] Currently with CAT Telecom, Thailand
[2] Currently with Aberdeen University, UK

over time as they occasionally share them with friends and colleagues (and later forget). Nonetheless, CCS allows users to set certain items as *private* if they do not want the items to be visible to others. Other privacy and copyright-related issues need to be addressed efficiently as the system goes into wider use.

2. Relevant Research Projects and Systems

The Internet has dramatically evolved beyond simple e-mail and file transfer applications over the last few decades. The development of the WWW in recent years has offered us the opportunity to experience new and exciting collaborative applications such as, wiki, blog, various social networking and online virtual community applications. The community catalog system explained in this paper is a *new* dimension to online virtual collaboration among the members of a library (who are generally already connected people as they are affiliated to the same institute or they belong to the same community). The development of CCS is built upon the idea that people in a community tend to share their personal collections of books, music, movie and other electronic or physical media with the fellow community members provided that an easy to use cataloging and circulation management systems is available to them. The human spirit of sharing and collaboration is the key factor for this project's success. However, CCS also offers a crucial incentive to its user by offering an easy-to-use web interface for any user to create and manage his or her personal collections without having to master in database and cataloging as well as recovery and backup (which is similar to other recent web applications such as, blogging and P2P file sharing). Before introducing the CCS system in detail in the next section, in the rest of this section we will explain some relevant systems and projects which offered us valuable ideas, insights and technologies we needed for the CCS.

The DVDdb System:

The *DVDdb* system is a simple web-based movie database that allows multiple users (e.g., a group of friends) to track ownership and loans of movies. The DVDdb system is implemented using PHP and MySQL, and it is tremendously simple but extremely useful for the purpose it serves. For the development of CCS, we needed to consider *(a)* how to extend the DVDdb system to include items beyond DVD and *(b)* how to integrate CCS with existing library databases and protocols. Further information about DVDdb including a demo is available at: http://www.globalmegacorp.org/dvddb.

The isbnsearch System:

It is not practical to ask the CCS users to type in the catalog information of each and every item they own in to the system. Therefore, having a tool or add-on which allows users to type only the *unique identifier* (such as ISBN or DOI) to get the full catalog information from a database is essential. We found that the *isbnsearch* [3] virtually does similar jobs by taking advantage of Z39.50 protocol. *isbnsearch* sends ISBN query to the library Z39.50 gateway server (such as Z39.50 gateway at Library of Congress and other libraries worldwide) to create its local permanent cache. Many readers may also be aware of the existence of similar database for music and movie, namely CDDB and IMDb. Further information about *isbnsearch* is available at: http://isbnsearch.sourceforge.net/.

The ReadingList Project:

The *ReadingList* [4] project was initiated by the Loughborough University's Teaching and Learning Committee in cooperation with the university library. The motivations behind developing this service were to enable taught-course students to view reading list from a variety of sources, enable users to check availability of those materials against the Library Catalog, and avoid duplication of effort by academics, departmental administrators and the Library in maintaining separate reading lists. The ReadingList project offers an ideal example of specialized and value-added service to a *specific* user community on top of the common services a library typically offer to its general user community. Further information about the project including a demo is available at: http://lorls.lboro.ac.uk/about.html.

The ThaiLIS Initiative:

Under the auspices of the Thai Ministry of Education, the Commission on Higher Education (CHE) initiated the *ThaiLIS* [5] consortium to facilitate resource sharing among member libraries (mostly, Thai public university libraries). The *Union Catalog* service is most relevant to our CCS System since with the ThaiLIS Union Catalog users issue a *single* query and receive aggregated response about relevant items at *all* member libraries (using Z39.50 gateways and response aggregation). Online Inter-Library Loan service is also available to request items from remote libraries. ThaiLIS databases are updated and managed by library professionals at each member's site. However, in the CCS, ordinary users are responsible in updating and managing their own personal collections and loans. Unlike ThaiLIS union catalog, the greatest challenge in developing the CCS system is to enabling ordinary users to be able to share their own collections among themselves as well as their preferred libraries in a flexible way. Further information and demo about *ThaiLIS Union Catalog* is available at: http://uc.thailis.or.th/.

It should be noted that lessons from the above-mentioned projects and systems have been carefully incorporated in CCS to make it a useful value-added service for library users and community. The overview of the proposed Community Catalog System (CCS) is discussed in detail in the following two sections.

3. Overview of the Community Catalog System

As explained earlier, the *Community Catalog System* (CCS) is developed with ordinary users and simplicity in mind. Therefore, the technical complexities and details remain transparent to the users. Users simply interact with the system using web browsers. Figure 1 shows how users may use the Library OPAC Interface or Community Catalog System's own interface to search and retrieve information from *both* the library-catalog and the community-catalog using a single interface -- either via the library OPAC interface or the community-catalog interface.

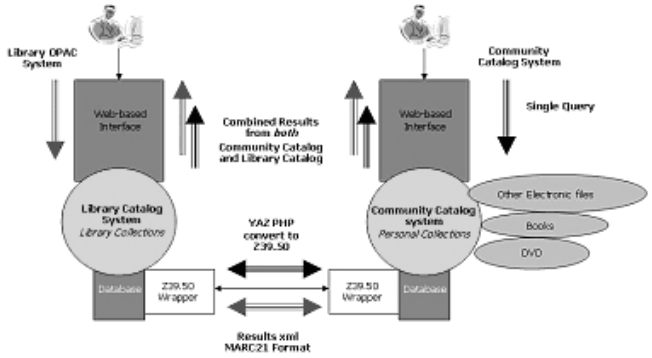

Figure 1: The Community Catalog System

The CCS also offers relevant web-based interfaces to manage personal collections and loans (e.g., catalog, circulation and user management).

Although users may access the CCS database through their library's OPAC interface with necessary scripting and adaptation, for the sake of clarity and simplicity, we will explain the CCS-oriented access scenario to the community catalog and the library catalog (via Z39.50 gateway) in the rest of this paper. We will also explain some key CCS modules and features in CCS catalog, circulation and user management.

4. Implementation and System Architecture

The Community Catalog System (CCS) is implemented as a Web-based client/server application and uses HTTP protocol between the client and the server. Interactions between backend databases (CCS databases and typical library databases) are facilitated by using Z39.50 protocol wrapper.

A. The CCS Server and Client:

We use available open-source technologies and tools to implement the *CCS Server* as outlined in Figure 2. Users use Web browsers as client to access the CCS Server. The interactions between the client and server take place in terms of simple HTTP requests and responses. The CCS Server makes extensive use of existing tools and packages [6][7][8][9] as listed in Figure 2 to seamlessly perform its desired tasks. Nevertheless, the CCS Server also implements its own modules (mainly PHP and PERL scripts) to process user inputs, interact with CCS databases and typical library databases (using Z39.50 protocol); integrate and aggregate all responses into a single output as explained later.

The CCS server also implements its own CCS catalog and circulation and user management modules so that it can be used independently (without an OPAC system).

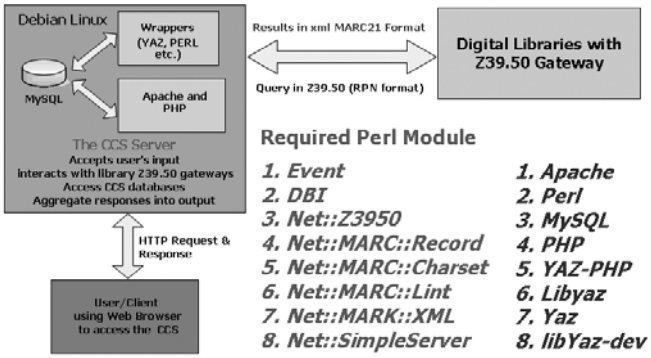

Figure 2: The CCS Client/Server Architecture

B. Use Case Diagram

The use case diagram in Figure 3 explains some of the key CCS modules such as user management; collection management and circulation management, etc. For brevity we will not elaborate every use-case scenario in this paper as they are quite obvious in the context of the community catalog system. The detail interactions with Z39.50 gateways are also omitted. Interested readers may refer to our Technical Report [10] for further details.

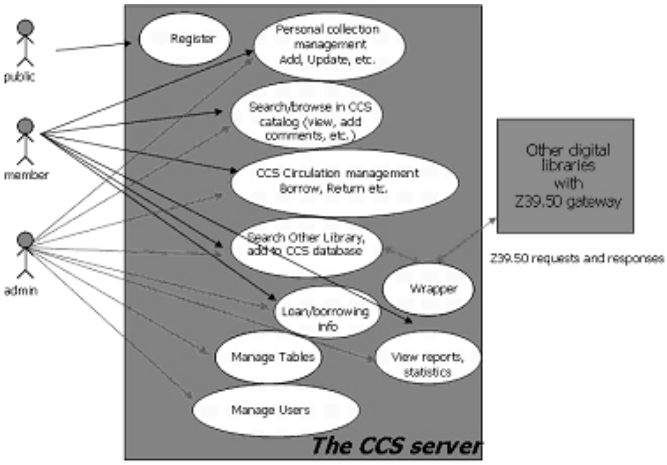

Figure 3: The CCS Use Case Diagram

C. The CCS Data Model

In Figure 4, we present a simplified data model of our CCS system. The current implementation is capable of dealing with books and articles, DVD/CDROM, electronic files, and items from the library-catalogs as shown below. Please note that we omitted certain data items (e.g.,

borrowing data) in this figure for clarity.

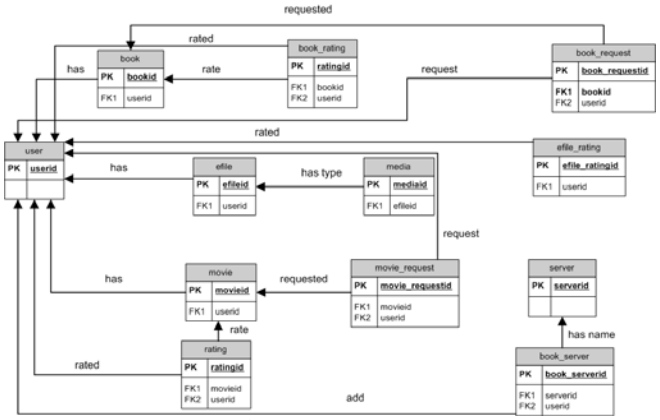

Figure 4: The CCS Data Model

D. The CCS User Interface and Operation

The CCS user interface is an easy-to-use web-based interface similar to the OPAC interface. General users (*public*) are allowed to use the system for simple tasks such as search and browsing. Upon approval from the *admin*, the *registered users* can upload and manage their own collections as well as request/borrow items from other registered users, etc. For the *admin* user, CCS offers a wide range of functionalities, such as configuration of external library and Z39.50 gateway, control and management of the entire collection and user community, etc. Some interesting features of CCS included simple but useful tool-support for users to create and mange their personal collections and circulation information. In figure 5, we explain how a user can import crucial catalog information from a Z39.50 gateway by simply typing a 10 digit ISBN number or title. Such features are attractive to CCS users as it reduces their workload drastically.

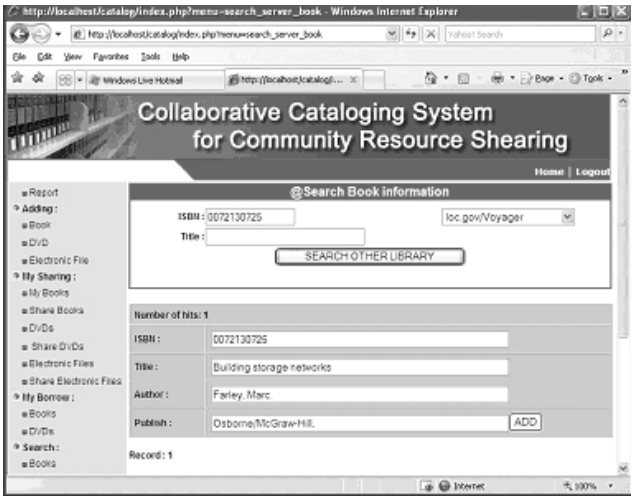

Figure 5: The CCS Interface – User can Import Catalog Info from External Database using ISBN or Title Search

Nevertheless, having been able to manage and track their personal collections without worrying about database, backup and maintenance is another rewarding incentive CCS can currently offer to its users. As the CCS system gains popularity within a big community (such as a full-fledged university), all users would benefit from the community-based resource sharing since many items they search for may be available from a friend or colleague next door. A registered user may check availability of an item in his or her library as well as all users in the community and place a request/reservation by sending e-mail or by making an online request. When the owner of the items logs in, he or she is duly notified and subsequently, may grant or deny such requests.

Other useful features, such as *(a)* New Arrival Listings alerts every CCS users about new items added by other users since they last logged in; *(b)* Separation of Personal Collection into Private (not to lend) and Public sub-collections allows users to manage their personal collection without having to compromise privacy and preference. We envisage that CCS will go through continuous further development by means of usage and user behavior analysis as well as users' survey and feedback.

5. Conclusions and Future Work

In this paper, we investigated the importance of a community catalog system that maximizes resource sharing within a community and organization. At the moment, the prototype Community Catalog System (CCS) we developed is operational in our university among small number of users which successfully integrate user's personal collections with our library collection. We have plans to further develop the system and put it in wider use for all members of our university and beyond.

The CCS features we explained in this paper undoubtedly add new dimensions in building online virtual community and facilitate community-based resource sharing. Considering the amount of electronic information everyone is accumulating these days, it goes beyond an ordinary user's ability to manage and organize them effectively without the help of an easy-to-use cataloging system like the CCS. For non-electronic (physical) items, CCS offers effective cataloging, tracking and management facilities.

We are aware of privacy, copyright and digital right related issues. Such issues need to be carefully investigated before putting the CCS into public.

Acknowledgments

This research is supported by a TRF Young Researchers' Grant (No: **MRG4880112**). The first author would also like to thank Miss Panarat Wuttiwongpakdee, his former undergraduate student for her substantial help at the early stage of this research.

References

[1] DVDdb: Source Code and Demo
http://www.globalmegacorp.org/dvddb

[2] Z39.50 Protocol: Wikipedia Article
http://en.wikipedia.org/wiki/Z39.50

[3] isbnsearch: SourceForge site
http://isbnsearch.sourceforge.net/

[4] Loughborough ReadingList Project: Source Code and Demo
http://lorls.lboro.ac.uk/about.html

[5] ThaiLIS Union Catalog: Information and Demo
http://uc.thailis.or.th/

[6] YAZ Toolkit: Source Code and Information
http://www.indexdata.dk/

[7] CPAN PERL Modules at CPAN:
http://www.cpan.org/ (search with Z39.50)

[8] Z39.50: Bib-1 Format and Attribute
http://www.loc.gov/z3950/agency/defns/bib1.html

[9] Z39.50: MARC21 Format and Attribute
http://www.loc.gov/marc/

[10] Technical Report: # SIU SS: SOT-MSIT-2007-01: *Extending the reach of library OPAC-a new way of community resource sharing using distributred catalogs and Z39.50 protocol*. Kovit Boonsri, Ekawit Nantajeewarawat, Chutiporn Anutariya and Md Maruf Hasan. MSIT Program, School of Technology, Shinawatra University, Thailand.

Towards Intelligent and Adaptive Digital Library Services: Associating Contents with User Profile and Usage Pattern in Digital Libraries

Yenruedee Chanwirawong, Md Maruf Hasan
School of Technology, Shinawatra University
Shinawatra Tower III, 1010 Viphavadi-Rangsit Road, Chatuchak, Bangkok, Thailand
yoyomimi_yen@hotmail.com, maruf@shinawatra.ac.th

Abstract- **The availability of contents, user profile and usage pattern in a digital library in machine understandable formats paves the way of processing these information further using state-of-the-art technologies to introduce intelligent digital library services. Annotating, organizing and analyzing contents based on a domain-ontology give us the ability to make topic inference through content relationships. User's inaccurate and incomplete profile can also be augmented with the help of a domain-ontology and usage pattern analysis using collaborative-filtering algorithms. Nevertheless, we can also use sophisticated mathematical models to process usage-pattern data for making content recommendations that reflects user's interest-drift. In this research, we integrate DL contents with a domain-ontology, user-profile and usage-pattern data by means of intelligent algorithms and techniques. On top of an open-source digital library system, we developed required modules to capture and manipulate necessary data with the help of efficient techniques such as ontology-driven topic inference, collaborative filtering, single exponential smoothing, etc. Our prototype can demonstrate that the retrieval results for the same query by a particular user, but on different time, may yield different result-sets (in terms of ranking) since the query, profile and contents are treated in a unified manner. We also verified that the same query from different users may yield different result sets as justified by differences in user profile and usage history. Our prototype can also recommend digital library items to users by using collaborative-filtering based usage-pattern analysis that takes user's interest-drift in consideration.**

I. Introduction

A digital library (DL) is a collection of documents in organized electronic form and accessible via search and browsing interfaces. The availability of contents, user profile and usage pattern in a digital library in machine understandable formats paves the way of processing these information further using state-of-the-art technologies to introduce intelligent services in digital library. Annotating, organizing and analyzing contents based on a domain-ontology [1] give us the ability to make topic inference through content relationships. User's inaccurate or incomplete profile can also be augmented with the help of a domain-ontology and usage pattern analysis using collaborative-filtering algorithms. Nevertheless, we can also use sophisticated mathematical models to *process* usage-pattern data for making content recommendations that reflect user's interest-drift [2]. Serving DL users with the right information which best reflect their query, profile, usage history and content relationships is only possible when we process heterogeneous information in a unified manner. In this research, we try to present such a unified approach towards developing adaptive and intelligent digital library services.

On top of an open-source digital library system, we developed required modules to capture and manipulate necessary data with the help of efficient techniques and algorithms such as ontology-driven topic inference, collaborative-filtering, single exponential smoothing, etc. Our prototype can demonstrate that the retrieval results for the same query by a particular user, but on different time, may yield different result-sets (in terms of ranking) since the query, profile and contents are treated in a unified manner. We also verified that the same query from different users may yield different result sets as justified by differences in user profile and usage history. Our prototype can also recommend digital library items to users by using collaborative-filtering based usage-pattern analysis that takes user's interest-drift in consideration.

II. Material and Methodology

A. Digital Library Services

Being organized and focused collections of information, digital libraries concentrate on a particular topic or theme; and good digital libraries will articulate the principles governing what is included there and how are they organized. Typical DL Systems provide searching and browsing interfaces. Searching implies that the user knows exactly what to look for, while browsing should assist users navigating among correlated searchable terms to look for something new or interesting. Searching interface may range from basic keyword search to field-specific advanced search, etc. Browsing interface includes categorical navigation based on certain taxonomy and meta-data such as browse by author, category and the like [3][4].

Most digital libraries provide basic search and navigation functions which maybe adequate in WWW context but are merely sufficient in the context of a digital library. In DL context, we have been continuously failing to take advantage of electronic contents using advanced content processing techniques, and also

failing to make use of user profile and usage history, etc. We argue that what makes a digital library unique is the availability of content in electronic form (which can be processed automatically and inferences can be made), and the availability of user profile and usage patterns. Unlike the WWW, Google-like Keyword Search or Yahoo-like Directory Hierarchy is certainly not adequate for harnessing information in the context of a digital library. Therefore, we tried to make use of DL content, domain-ontology, user-profile and usage patterns; and developed necessary algorithms to facilitate intelligent and adaptive services for digital library.

In a digital library, keyword search and navigation may only facilitate and entry-level access to the content. Efficient access to DL contents is only possible when we succeed to make use of heterogeneous information and facets under a unified framework. For example, (1) keyword-based search often produces enormous amount of irrelevant hits. Different users may target different information when using the same keyword to search. (2) When a DL users signs up, they may inadvertently provide incomplete or inaccurate profile information about themselves. Moreover, (3) user's profile, information-seeking behavior and information-needs, etc. change over time (a.k.a. user interest-drift). In our research, we try to augment user profile adaptively using domain-ontology and usage history. We also analyze DL contents and usage-pattern to make ontology-driven topic inference through content and usage associations. Interest-drift modeling using single exponential smoothing techniques tries to capture interest-drift in DL users adaptively.

B. The Domain Ontology

The domain-ontology helps us organizing the DL contents and representing the user profile. We used a modified version of *ACM Computing Classification System* (ACM-CCS) Ontology [5][6] in our work. The schema of the ontology is presented in Fig. 1, where each node represents a topic. It should be noted that the `hasKeyword` attribute is very useful in making topic-inference from usage-patterns and through full-text analysis. DL user's search history and bookshelf items are analyzed to extract keywords and those keywords are associated against related nodes in the domain ontology.

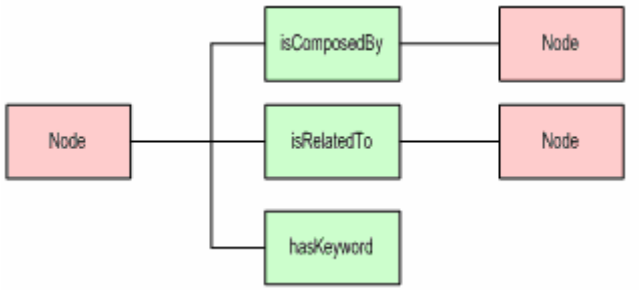

Figure 1. Schema for ACM-CCS Ontology based on CCS Taxonomy

C. Dynamic User Profile

The classes (topic nodes) defined in the ACM-CCS ontology is used to represent user profile. In the DL context, typically a user will specify their interests as they sign-up. However, such a raw profile is usually incomplete or inaccurate [7]; and requires further refinements and enhancement. We used a weight-based topic vector to represent user profile. The initial weights for each selected topic can be assigned rudimentarily. The subsequent augmented profile consists of topics and their weights as normalized and adjusted via the domain ontology by considering topic interrelationships and temporal usage history. Such a dynamic user-profile may be computed off-line in batch mode; and may more accurately reflect a user's interest and context. By keeping extra information such as user access logs including user's keyword-sets and bookshelf-items along with timestamps, we can capture user's interest-drift over time and re-weight the user-profile dynamically [2]. The mathematical formalism of interest-drift is explained below.

D. Interest-Drift Modeling

In a digital library, the user interests, information needs and information seeking behavior change over time. We adopt a policy on assigning higher weights on recent usages since topics related to recent usages reflect user's present context. We also assign exponentially weighted weights on older usage to reflect a user's general context. We use *Single Exponential Smoothing* technique (a well-known smoothing technique used in financial forecasting and engineering applications) to assign *exponentially adjusted weights* to estimate adaptive weights for each topic [8].

The smoothing is done as follows:

$$S_t = \alpha y_{t-1} + (1 - \alpha) S_{t-1} \qquad 0 < \alpha \leq 1$$

where t = any time period

S_t = the smoothed value

A = smoothing constant

y_{t-1} = the current observation

In the DL context, the frequency of a user login and user activities are directly related to the choice of the smoothing constant, α. The α values can be different for each user; and are estimated using a normalized frequency measure, fi as derived in Table 1.

$$f_i = N_i \;/\; Max_N$$

where N_i = # of transactions for $user_i$

Max_N = # of transactions for the most active user

TABLE 1
ESTIMATION OF SMOOTHING CONSTANT , α

Frequency (f_i)	Alpha (α)
frequency = 0	0
frequency < 0.3	0.05
0.3 <=frequency<0.5	0.1
0.5 <=frequency<0.7	0.2
0.7 <=frequency<0.9	0.3
frequency > = 0.9	0.4

E. Usage-Pattern Analysis using Collaborative Filtering

Collaborative Filtering (CF) is one of the key techniques for implementing a recommender system that recommends to a user a set of candidate items, which may be preferable or useful to the user based on similarity analysis with other users [9]. Collaborative Filtering techniques have been intensively used in E-commerce. The idea is easily extendable in the context of digital library. In the DL context, we can take advantage of both usage-pattern based and content-based collaborative filtering algorithms. It is interesting to note that in DL context, CF can be *bidirectional* – that is, it is not only possible to recommend new DL items for a user by analyzing profile similarity and looking at their bookshelves; but also it is possible to analyze bookshelf-item similarity and augment user-profile. We use a time-sensitive CF algorithm (similar to [2] for our purpose) to achieve adaptive effects on recommendations made.

F. Intelligent and Adaptive Digital Library Services

As a DL user undertakes new tasks or switches to new projects, an intelligent digital library should be able to adapt to such changes in the context and preference of that user. In another word, a user searching for information with the *same keyword but on different point of time* should not be served with the same retrieval results (at least the ranking of item should vary). Similarly, *two different users* searching the DL using the *same keyword at the same time* may not be served with the retrieval results. At least the ranking of the retrieval result should take into account of their profiles. We illustrated such adaptive digital library scenario in Fig. 2.

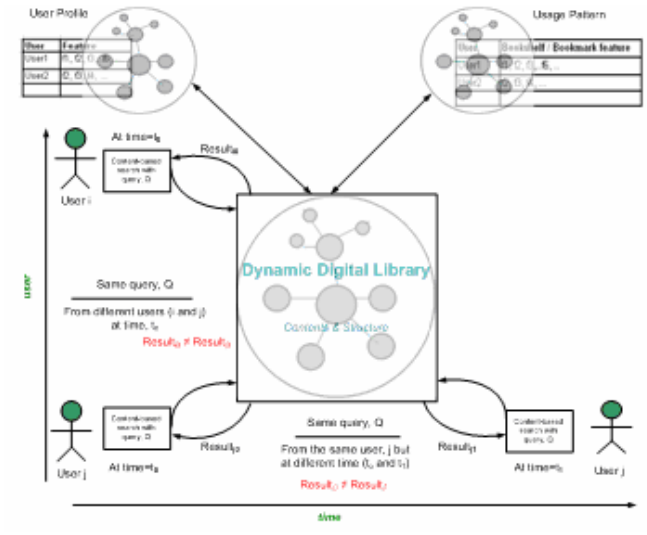

Figure 2. Intelligent and Adaptive Features in Digital Library Services
(a) Horizontal Axis illustrates that search results for a particular user using the same keyword at different time yield different retrieval results.
(b) Vertical Axis illustrates that search results for different users based on the same keyword may yield different retrieval results.

Using our prototype we try to verify that the retrieval results for the same query by a particular user, but on different time, may yield different result-sets (in terms of *ranking*) since the query; profile and contents are dynamically enhanced using intelligent algorithms based on ontology-driven topic inference, time-weighted user profile adjustment, CF-based similarity analysis and the like. We also verify that the same query from different users may yield different result sets as justified by the differences in their profiles and content relationships. Our prototype can also recommend digital library items to users by using collaborative filtering–based usage pattern analysis that reflects user's interest-drift. We test our approach using a partially synthetic datasets and analyzed the results through human judgments.

III. EXPERIMENTAL SETUP AND RESULT ANALYSIS

In order to demonstrate our proposed approach, we conducted experiments with semi-synthetic data. In this section, we will highlight some observations from those experiments.

A. User Similarity

To start with we generated random profiles for 10 users. These profiles are augmented though ontology-driven topic inference and usage pattern analysis. We use this information to find the similarity among users. Fig. 3, shows an example raw profile of users. The value '1' means a user is interested in that topic (category). The value "0" means that a user is not interested in that topic. We calculate vector similarity to find similarity among users. Fig. 4, shows user similarity. It should be noted that the profiles are continuously augmented using ontology-driven topic inference module and usage pattern analysis modules. Therefore, our CF-based recommendations reflect dynamic usage pattern of DL users and the topic associations in the DL domain.

	category																	
user	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0	0	1	1	0	0	0	1	0	1	1	0	0	1	0	0	0	0	1	1
2	0	0	0	0	0	1	0	0	1	0	1	0	0	1	1	1	1	1	1	1
3	1	1	1	0	0	1	1	0	0	1	1	1	0	0	0	1	1	1	0	1
4	0	1	0	0	1	1	0	0	0	1	1	1	1	0	0	0	0	1	1	1
5	0	0	0	1	0	0	1	1	0	1	0	0	1	1	1	0	1	0	0	0
6	0	0	0	0	1	0	0	0	1	0	1	1	0	0	1	1	0	0	0	0
7	1	1	0	1	0	0	0	1	0	1	1	1	0	0	1	1	0	1	1	0
8	1	0	0	0	0	0	1	0	1	1	1	0	1	1	0	0	0	1	1	1
9	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	1	0
10	1	1	1	0	0	0	0	0	0	0	0	1	1	0	1	0	1	0	1	1

Figure 3. Examples of initial user profiles (raw-profile)

user	1	2	3	4	5	6	7	8	9	10
1	1	0.447	0.408	0.447	0.5	0.144	0.533	0.559	0.474	0.354
2	0.447	1	0.548	0.5	0.335	0.516	0.477	0.6	0.424	0.422
3	0.408	0.548	1	0.639	0.306	0.354	0.609	0.548	0.387	0.577
4	0.447	0.5	0.639	1	0.224	0.387	0.572	0.6	0.424	0.527
5	0.5	0.335	0.306	0.224	1	0.144	0.426	0.447	0.158	0.354
6	0.144	0.516	0.354	0.387	0.144	1	0.492	0.258	0.365	0.272
7	0.533	0.477	0.609	0.572	0.426	0.492	1	0.477	0.405	0.503
8	0.559	0.6	0.548	0.6	0.447	0.258	0.477	1	0.424	0.422
9	0.474	0.424	0.387	0.424	0.158	0.365	0.405	0.424	1	0.447
10	0.354	0.422	0.577	0.527	0.354	0.272	0.503	0.422	0.447	1

Figure 4. Similarity among user s

B. Time-weight Smoothing

From DL user's access-log, we calculate α value for each user to continuously augment user's raw-profile that reflects user's context and interest-drift. If α value is small, the weight changes

faster than a bigger α value. From Fig. 5, we see that the weights are changed sharply for more active DL users (with smaller α).

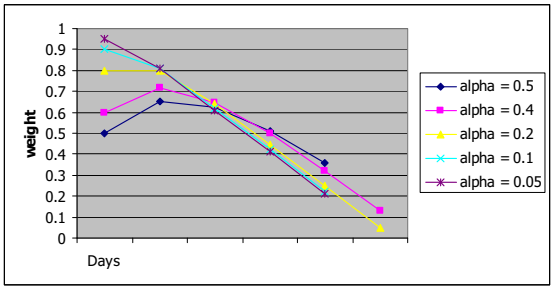

Figure 5. Graphical representations the relation between topic weight and α

In Fig. 6, we illustrate how topic-weight changes when a user starts searching the DL with new items (originally not reflected in his or her profile and usage).

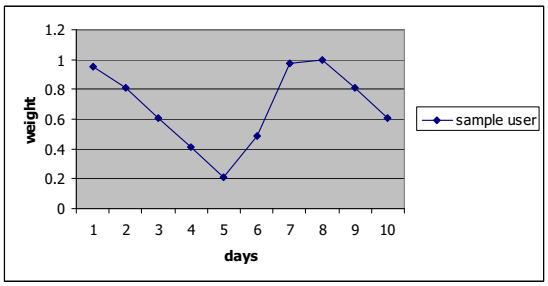

Figure 6. Weight of a topic for a particular user changing over time

IV. Conclusions and Future Work

In this paper, we presented a novel approach to associate contents with user profile and usage pattern and modeling interest-drift in Digital Libraries. Our prototype is based on a *3-Layer Architecture* as shown in Fig. 7. The core of the 3-Layer architecture is an open-source DL system surrounded by series of Add-on Modules to capture further information about contents, users and usages in Layer 2. The outer-most layer (Layer 3) consists of services developed by making intelligent use of those information with the help of efficient techniques and algorithms.

We validated our approach with preliminary experiments using synthetic data and prototype algorithms. At the moment, we are setting up a digital library with sizeable amount of contents and real users; and planning to put our prototype into practical operation. Our present experimental setup and analysis using synthetic dataset showed that our approach integrates several facets of DL information seeking scenario; and is capable of providing personalized services for digital library users. Two different users searching with the same query yield different hits. The same user searching with the same query but at different point of time, may get different hits depending on his or her recent context. For the future work, we will include systematic evaluation based on real digital library usage data.

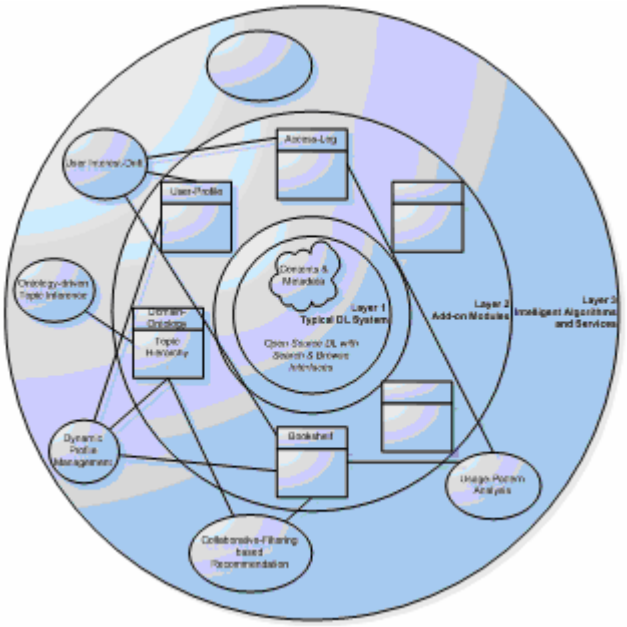

Figure 7. 3-Layer DL Architecture for Intelligent and Adaptive Digital Library Services

Acknowledgment

We would like to thank Assoc.Prof. Dr. Ekawit Nantajeewarawat and Asst.Prof. Dr. Chutiporn Anutariya for their supports and valuable comments. This research has been supported by Thailand Research Fund in terms of a TRF grant (MRG4880112) awarded to Dr. M.M. Hasan.

References

[1] S. E. Middleton, D. C. De Roure, and N. R. Shadbolt, "Capturing Knowledge of User Preferences: Ontologies on recommender systems," In Proceedings of the First International Conference on Knowledge Capture (K-CAP2001), 2001, pp. 100-107.

[2] Y. Ding and X. Li, "Time weight collaborative filtering," In Proceedings of the 14th ACM international conference on Information and knowledge management, 2005, pp. 485-492.

[3] G. G. Chowdhury and S. Chowdhury, Introduction to digital libraries, London: Facet Publishing, 2003.

[4] I. H. Witten and D. Bainbridge, How To Build a Digital Library, San Francisco, CA: Morgan Kaufman Publishers, 2003.

[5] ACM-CCS, "The ACM Computing Classification System [1998 Version]," [online]
Available: http://www.acm.org/class/1998/. [Accessed: Jan. 12, 2007]

[6] ACM-CCS Add-on Ontology, "DSpace Dev @ University of Minho," April 2003, [Online].
http://dspace-dev.dsi.uminho.pt:8080/en/research_about.jsp
[Accessed Mar. 12, 2006]

[7] I. E. Liao, S. C. Liao, K. F. Kao, and I. F. Harn, "A Personal Ontology Model for Library Recommendation System," in The 9th International Conference on Asian Digital Libraries (ICADL 2006), 2006, pp. 173-182.

[8] NIST Handbook, "NIST/SEMATECH e-Handbook of Statistical Methods," [Online]
http://www.itl.nist.gov/div898/handbook/. [Accessed: Mar. 17, 2007]

[9] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based Collaborative Filtering Recommendation Algorithms," In Proceedings of the Tenth International World Wide Web Conference (WWW10), 2001, pp. 285-295.

ThaiFone: A Robust Approach to Romanized-Thai Interface and Algorithm for Efficient Dictionary and Web Search

Md Maruf Hasan, Suchat Pupang[1]
Shinawatra University, Thailand
maruf@shinawatra.ac.th; suchat@oae.go.th

***Abstract* - To address problems related with the disparity in Thai Romanization, we proposed *ThaiFone* and *ThIME* algorithms. Although Thai Romanization Standard exists, Thai texts are often transliterated in a rudimentary fashion. Such practices are causing serious problems in efficient search and retrieval of cross-language information on today's computer applications. The *ThaiFone* algorithm is motivated by the well-established *Spell-Alike and Sound-Alike* set of algorithms widely used to cope with spelling and pronunciation discrepancies in other languages. *ThaiFone* attempts to align a set of related Romanized Thai strings and their corresponding Thai string together using heuristics mapping rules and corpus-based knowledge. *ThIME* tries to map a non-standard Romanized Thai string to its Thai and Romanized-Thai counterparts. In this paper, we outline *ThaiFone* and *ThIME* algorithms and related interfaces along with our preliminary experimental results. We highlight two practical applications of *ThaiFone* and *ThIME* namely, in – (1) cross-language Web search through Search engine API (GoogleAPI) and, (2) Thai-English Dictionary search with LEXiTRON open-source dictionary.**

I. INTRODUCTION

Cross-language Information Retrieval (CLIR) is a *subfield* of Information Retrieval (IR) dealing with retrieving information written in a language different from the language of the user's query. For example, a user may pose their query in English but retrieve relevant documents written in Thai, English or hopefully in other languages. CLIR takes advantage of multilingual dictionary, thesaurus and machine translation technologies to achieve such goals by means of partially or fully translating query or documents [1]. In today's WWW, information on a Web site or a Web page is often written in multiple languages and forms. With the help of popular keyword-based *hit-or-miss* search engines (such as Google) users are only capable of retrieving partial (incomplete) information.

In an attempt to verify the extent of incompleteness of search results from a our practical point of view, we attempt to perform search using the *original* Thai keyword *(พหลโยธิน)*, along with the variations of its common *Romanized Thai spellings* (such as Phahonyothin, *Pahonyothin*, *Paholyothin* etc.); and analyze the search results manually. The outcome of the search result depends on the search keyword (and is incomplete) as depicted in Figure 1. Internet users searching for restaurants or hospitals located at *Paholyothin* area in Bangkok often use a variety of Romanized Thai spelling or Thai strings as query. As a motivating example, let's consider a scenario where non-Thai users trying to locate restaurants or hospitals in Bangkok using a Romanized Thai query. The search engine is likely to produce a keyword-based result depending on the spelling he or she chose. However, the user may be happy to find a *Thai* "location map" of the venue which he or she can use to guide a Taxi driver to the venue. Although search engines such as Google is capable of handling query in multiple languages and forms (English, Thai as well as Romanized Thai), we can't overcome the partial retrieval phenomenon without a robust algorithm that efficiently aligns and maps non-standard Romanizations with their standard and Thai counterparts.

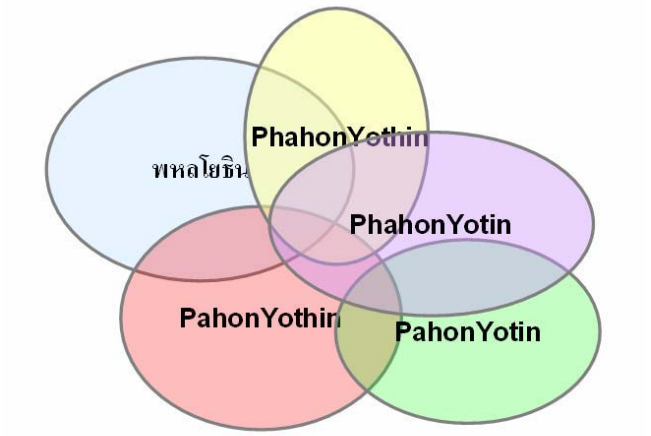

Figure 1: Partial Retrieval Phenomenon: Keyword specific non-overlapping search results (hits)

In this research, we propose *ThaiFone* and *ThIME* algorithms to address the problem related with the disparity in Thai Romanizations. We used ThaiFone algorithm and GoogleAPI [2] to demonstrate a CLIR-like effect in Web search. We also explain how *ThIME* can facilitate a robust interface to query Thai-English Dictionary without having to type Thai scripts.

II. THE THAIFONE APPROACH

ThaiFone algorithm is motivated by the well-established algorithms in dealing with *Spell-Alike and Sound-Alike* words in English and other languages. Such algorithms can be broadly categorized in *two* types: orthographic

[1] This author is also affiliated with the Office of Agricultural Economics, Royal Thai Government.

algorithms and phonological algorithms. Orthographically motivated algorithms try to quantify a distance-measure (such as *edit distance* [3]) to calculate the similarity between a misspelled word with its target (correct) spelling. Phonologically motivated algorithms (such as *Soundex, MetaPhone, etc.* [4]) try to map *raw consonant sounds* (set of *consonants*) to a smaller set of *basic sounds* (set of *phones*). Both types of algorithms are in common use in today's computer applications. Orthographic algorithms are commonly used in word processors and search engines for automatic spelling corrections and keyword suggestions, respectively. Phonological algorithms are often incorporated in relational database management systems (such as Oracle) or genealogical research to cope with phonetic variations in *Proper Names* (name of people, place, etc.). The details about these algorithms can be found in relevant literatures [5][6]. It is worthy to note that interesting applications of these basic algorithms are also experimented in medical and pharmaceutical domain - such as to help doctors and pharmacist to identify potential life-threatening mistakes involving sound-alike or spell-alike drug names. In our research we focus on development of *ThaiFone* and *ThIME* algorithms and relevant interfaces to address effective Web- and Dictionary- search in particular. However, the approach may be equally useful in other application domain.

ThaiFone tries to align a set of related Romanized Thai spellings and their corresponding Thai word together using both heuristics rules and corpus-based evidences. We consider non-standard Thai Romanizations as a kind of spelling-error; and assume that unforeseen Romanization (not yet in the dictionary) maybe a possible romanization of a Thai word based on plausible phonetic and orthographic similarity. Therefore, we needed to build **(1)** a corpus-based *Dictionary of Romanized Thai* as well as **(2)** a set of *Metaphone-like Phonetic Mapping Rules*.

A. Corpus-based Dictionary of Romanized Thai

We start by developing a corpus-based Thai Dictionary as explained in Figure 2. Standard Romanization for the Thai vocabulary is derived using a Romanization tool [7].

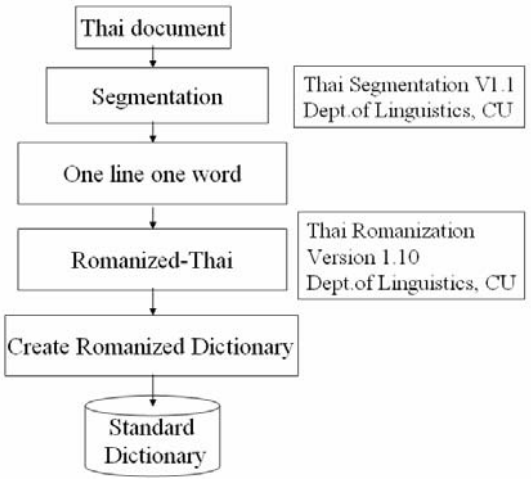

Figure 2: Creation of *Standard* Romanized-Thai Dictionary

We further augment our standard dictionary by adding non-standard *variations* of romanization using corpus-based method as explained below. First, we collect English news reports published in Thailand and Thailand related English Web pages to create a huge corpus-based dictionary (**X**) which includes both English vocabularies as well as Romanized Thai strings. We then use a large list of English vocabulary (**Y**). Finally, by subtracting **Y** from **X** (i.e., set theoretical difference, **X-Y**), we develop our initial list of Non-standard Romanized-Thai Dictionary (cf. Figure 3).

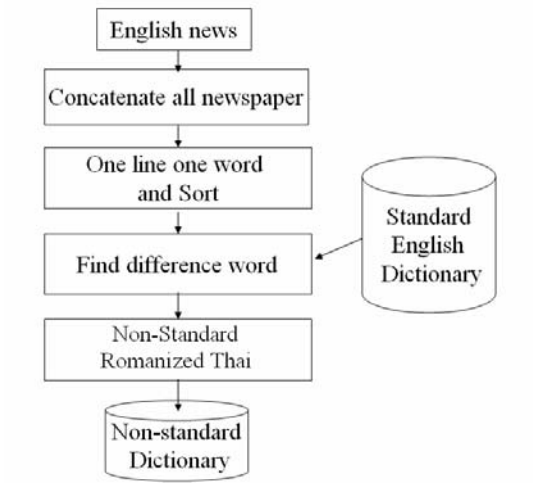

Figure 3: Creation of *Non-standard* Romanized-Thai Dictionary

In the last step (cf. Figure 4), we use *ThaiFone* similarity measures (explained later) to align and merge the standard and non-standard dictionaries together. Since an *initial* Romanized Thai Dictionary is crucial for this research, we also manually check and validate the dictionary entries to correct errors in automatic alignment.

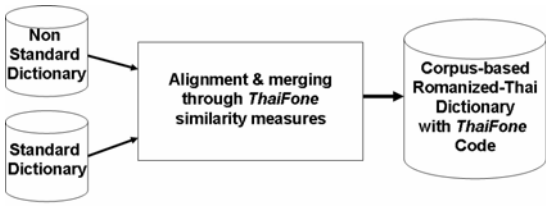

Figure 4: Creation of *Corpus-based Dictionary* of Romanized-Thai through alignment and merging of *Standard* and *Non-standard* Romanized-Thai Dictionaries.

No dictionary is complete and therefore, we propose an *incremental update* mechanism as explained in Figure 5 (lower-half of the diagram. We resort to the Google spelling-suggestions (evidence-based update) as well as *ThaiFone* similarity-measure (similarity-based alignment) to progressively update our corpus-based dictionary by adding/aligning new variations of Thai Romanizations in the initial dictionary.

B. Metaphone-like Phonetic Mapping Rules

Based on Thai Romanization Standard [8], we develop MetaPhone-like phonetic mapping rules to generate

ThaiFone codes from Romanized Thai strings. The following is the summary of techniques we used to develop the phonetic mapping rules.

- *Fuzzy algorithm* is used to map the letter that produces sound-alike for example "D" and "T" we can map to each other as described in Table 1.
- *Consonant Removal* is used to eliminate the character which produces no sound for example "TH" we can map into "T" since the letter "H" not produces any sound.
- *Duplicate Consonant Removal* is to reduce the word that has a consonant duplication for example "Gullaya" with this rules, we can reduce "ll" to "l" since the reduction of consonant letter is not affected to the pronunciation of the word.
- *Vowel Removal* is a technique to eliminate the vowel from the word as the widely used in the existing algorithms. Based on our experiments, we found that the precision of word retrieval especially for Thai language is high when all vowels are kept. Therefore, our algorithm is proposed to keep all vowels rather than eliminate.
- *Duplicate Vowel Removal* is to remove the vowel duplication for example "EE" can be mapped to "E".

Table 1: The ThaiFone Phonetic Mapping Table *(C, P)*

Sub spelling *(C)*	Sub sound *(P)*	Comments
A, AR, AE	A	
E	E	
I	I	
O	O	
ORN, OLN	ON	
U	U	
Y	I	
B	P	
CH	T	If CH last position
CH	C	Otherwise
D	T	
F	F	
G	K	
H		Remove
	H	If H is first position
J	T	If J is last position
J, JA, JO	C	
K	K	
L	L	
KUL, TOL, MOL,	N	If they are last position
KHUL, THUL,	N	If they are last position
M	M	
N	N	
NG	NG	
PH, PL, PR	P	
PHLA, PHRA	PA	
Q	K	
R	L	
R		Remove if R is after vowel
S	S	
T	D	If T is first position
TR	T	
V, VR, VW	W	
WV, WR	W	
X, XC	C	
Z	S	

ThaiFone Code Generation

ThaiFone algorithm uses a set of *M* mapping rules, each in the form of (C → P), where C is a sub-spelling and P is a sub-sound. It also uses a temporary variable, S' to store ThaiFone codes.

Input: A spelling $A = a_1.a_2...a_k$;
Output: A *ThaiFone* code S

1. $l = 0$; *length counter*
2. $S' = \{\}$; empty string
3. $w = 5$; *user defined code length*
4. for each $j, 1 \le j \le k$
   (a) If $a_j = a_{j+1}$ then $j = j + 1$; remove duplicate letters
   (b) Each $(C, P) \in M$ such that $a_1...a_j = C$, find P based on Mapping Rules.
   (c) Let S' be the concatenation of S' and P
   (d) $l = l + 1$
   (e) If $l \ge w$ End Loop
   (f) Loop
5. Output the final *ThaiFone Code*, S.

ThaiFone Similarity Measure

The *ThaiFone* similarity measure is a complex one. It makes use of longest common substring (*lcs) similarity, s* and *levenshtein distance, d* as shown in the following formula (1). We use this formula to calculate the similarity between two Romanized Words (or ThaiFone Codes).

ThaiFone Similarity:

$$S = \frac{s}{\left(\frac{(l_1 + l_2)}{2}\right)} - \frac{d}{(l_1 \times l_2)} \qquad (1)$$

$$s = lcs(t_1, t_2)$$

$$d = levenshtein(t_1, t_2)$$

where t_1 and t_2 are two words or codes with length l_1 and l_2, respectively.

III. ThaiFone Experimental Results

ThaiFone algorithm is still under development since we are trying to fine-tune the *mapping rules* and enhance our *corpus-based dictionary*. To validate our *ThaiFone* algorithm, we selected a few candidate Romanized-Thai strings and calculated the precision and recall rate. We have tried with varying ThaiFone code-lengths (between 3 and 9) and compared our retrieval results with existing algorithms. In Table 2, we summarize the precisions and recalls of ThaiFone (with code-length 7) with those of existing algorithms.

Table 2: Precision and Recall for each algorithm (ThaiFone code length is 7)

Algorithms	Recall	Precision
ThaiFone	0.70	0.92
Soundex	0.32	0.52
Metaphone	0.17	0.88
Doubble Metaphone 1	0.17	0.58
Doubble Metaphone 2	0.19	0.58

Readers should note that we have not made any changes in the existing algorithms (which were meant for English).

Therefore, the precision and recall of ThaiFone appears to be significantly higher than those of the existing algorithms. Nevertheless, we can safely conclude that the ThaiFone approach possesses the potentials of dealing with the Thai Romanization problem with further work.

In Figure 5, we show the precision of each algorithm with varying code-length. It appears that ThaiFone performs well with a longer code-length (code-length = 7) in our experiment.

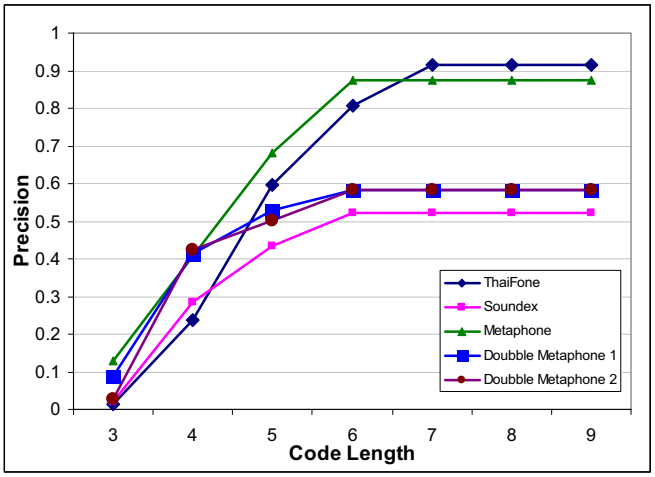

Figure 5: Precisions with varying code-lengths for each algorithm.

IV. ThaiFone Applications

A. *ThaiFone in Web Search*

ThaiFone helps us in efficient web-search by using suggested candidates as shown in Figure 6. The effect is essentially a CLIR-like effect in Web search through user feedback (selection). User may input either Thai or a Romanized-Thai string. ThaiFone returns a set of candidate words for selection. Once selection is made, we use GoogleAPI to perform the search. We also use Google spelling-suggestions to continuously update our corpus-based dictionary as shown in the lower-part of Figure 6.

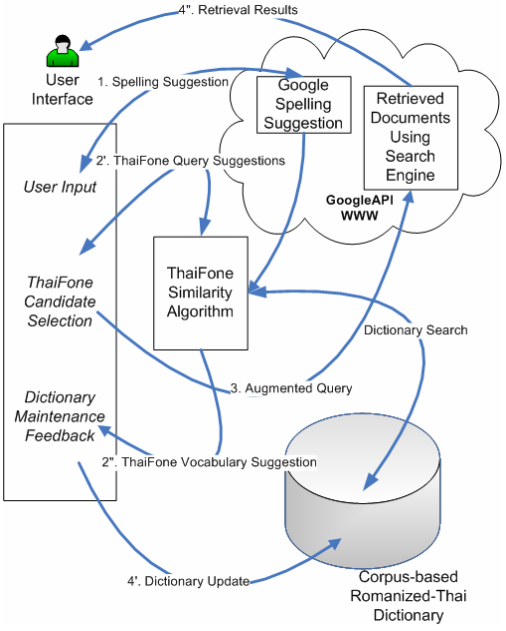

Figure 6: *ThaiFone* Interface in Web Search

B. *ThaIME in Dictionary Search*

Using *ThaiFone* we propose a Thai Input Method Editor (*ThIME*) interface to implement a dictionary search interface for Thai-English dictionary. At present, to search a Thai-English dictionary (electronic or online), users must type Thai characters which is not practical for certain users (e.g., foreigners who do not write Thai character). *ThIME* is similar to Chinese or Japanese Input Method Editor (IMEs). ThIME accepts Romanized input and makes use of *ThaiFone* to map user-input into candidate Romanized Thai and Pure Thai strings user may select before sending the query to usual dictionary search application. We integrated ThIME with the open-source Thai-English bilingual dictionary, LEXiTRON as explained in Figure 7.

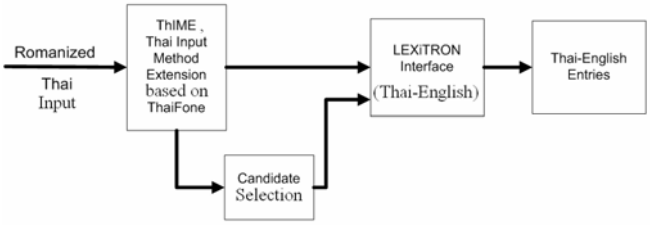

Figure 7: *ThIME* Interface in Dictionary Search

V. Conclusions and Future Work

We experimented with *ThaiFone* algorithm to deal with the problems with non-standard Thai Romanizations in information processing tasks. We tested our approach in *two* practical applications in Web and Dictionary search. Both prototypes are currently operational as stand-alone applications. In the near future, we plan to develop *Web-search Interface* using GoogleAPI and *ThaiFone*; and *Dictionary-search Interface* using LEXiTRON dictionary and *ThIME*.

Acknowledgment

This research has been supported by Thailand Research Fund in terms of a TRF grant (MRG4880112) awarded to Dr. Hasan.

References

[1] Grefenstette, G. (1998). Cross-language information retrieval. Boston, MA, Kluwer Academic Publishers.

[2] Google APIs (2007). http://code.google.com/apis/

[3] W. A. Robert, and J. F. Michael (1974). "The String-to-String Correction Problem," Journal of the ACM, 21:1, pp.168-173.

[4] Z. Justin and P. Dart (1995). "Finding Approximate Matches in Large Lexicon," Software-practice and Experience, Vol.25(3), pp. 331-345.

[5] K. Grzegorz (1999). "Alignment of Phonetic Sequences," Technical Report CSRG-402, Department of Computer Science Unversity of Toronto.

[6] K. Grzegorz (2003). "Phonetic Alignment and Similarity," Kluwer Academic Publishers.

[7] A. Wirote, and R. Wanchai (2004). "A Unified Model of Thai Romanization and Word Segmentation," PACLI, pp. 205-214.

[8] Royal Thai Institute "Royal Thai General System of Transcription," 1999.

[9] M. Kristina; and T. Robert (2002). "Pronunciation modeling for improved spelling correction," In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 144-151.

Proceedings of the Third IASTED International Conference
ADVANCES IN COMPUTER SCIENCE AND TECHNOLOGY
April 2–4, 2007, Phuket, Thailand
ISBN CD: 978-0-88986-656-0

E-LEARNING CONTENT MANAGEMENT: AN ONTOLOGY-BASED APPROACH

Nophadol Jekjantuk, Md Maruf Hasan
School of Technology, Shinawatra University
Shinawatra Tower III, 1010 Viphavadi-Rangsit Road, Chatuchak, Bangkok, Thailand
nophadol_bkk@hotmail.com, maruf@shinawatra.ac.th

ABSTRACT
Scarcity of E-Learning content being a barrier for E-Learning is no longer true on today's Internet. The current concerns are how to effectively annotate and organize available content (both textual and non-textual) to facilitate effective sharing, reusability and customization in an intelligent fashion. In this paper, we explain a component-oriented approach to organize content in an ontology. We also illustrate our *3-Tier E-Learning Content Management Architecture* and relevant Web Services and Interfaces. We use a simple yet intuitive example to successfully demonstrate the current working prototype which is capable of compiling personalized course materials on a particular topic (e.g., "Database") on-the-fly. The prototype uses the Pellet Semantic Web Reasoner as an inference engine to satisfy the constraints and criteria specified by a user (through browser based interface) or an agent (via Web Service API), and retrieves relevant content from the domain ontology in an organized fashion.

KEY WORDS
Learning Objects, E-Learning Content Management, Ontology-driven Inference, Semantic Web Services

1. Introduction

The ubiquity of the Internet and E-learning with their new educational tools and applications are rapidly changing the old way of learning. In the past, we simply distributed E-learning content on the WWW in a semi-structured fashion – with HTML tags and links. In principle, content available on the WWW is accessible ubiquitously. However, in reality, due to the limitations of keyword-oriented *hit-or-miss* search engines, we often find it hard to locate the desired content. Moreover, the absence of explicit metadata or annotations reflecting the pedagogical facets about these contents (such as content-dependency) often makes it harder to organize the discovered content in a self-sufficient courseware for *personalized* teaching and learning. Only expert users can go through the content and validate or organize heterogeneous content in a sensible manner. With the advent of Semantic Web technologies, E-Learning content annotated with proper metadata (including pedagogical attributes) and organized in an ontology may effectively facilitate efficient dissemination, discovery and reuse of content in a new way. In this new approach, not only humans but also artificial agents can discover and organize contents from heterogeneous sources and combine them into *customized courseware* that satisfies specific criteria and constraints stipulated by users or agents. Customized courseware refers to a collection of content (possibly from heterogeneous sources) where the content-related dependencies and pedagogical constraints are preserved.

In this paper, we explain a component-oriented approach to organize E-Learning content in domain ontology. This component oriented approach is inspired by the concept of *Learning Object*s (LOs) and surrounding technologies [1][2]. We also illustrate our *3-Tier E-Learning Content Management Architecture* and relevant *Semantic Web Services*. We use a simple yet intuitive example to successfully demonstrate our current working prototype which is capable of compiling personalized course materials on *"Database"* and *"Thai Language Learning"* on-the-fly. The Web-based prototype is available at http://trf.shinawatra.ac.th/ecms/. The prototype uses a generic Semantic Web Reasoner, *Pallet* [3] as an inference engine to satisfy constraints and criteria specified by a user (through browser-based interface) or an agent (via Web Services API), and retrieves relevant content from the domain ontology in an organized fashion.

2. Background

In this section, we provide background information on E-Learning, and the concept of Learning Objects (LOs). LOs are the basic foundation of a component-oriented E-Learning system. The E-Learning Content standards are also reviewed. We also review relevant state-of-the-art Semantic Web technologies and their advantages. We also refer to related work as they appear relevant.

2.1 E-Learning Content Issue

E-Learning Content is any digital resource that we can use to support learning. E-Learning Content can be divided into *two* categories: *textual* (text-based content such as, plain-text, PDF, etc.) and *non-textual*

(multimedia content such as audio visual materials and animations, etc.).

Textual content can be effectively located by using a keyword search engine, such as Google or Yahoo, producing a number of results. Only human experts can make sense of such a set of retrieved content and organize it into a customized courseware.

With the present state-of-the-art technology, *non-textual* content is still difficult to locate - even popular search engines, like Google, cannot find them if that content is given irrelevant filenames or surrounded by unassociated textual context. Ontology-based metadata annotations for non-textual content pave the way for locating and organizing non-textual content efficiently.

It should be noted that for both of the above cases (textual and non-textual content), search engines generally produce a *set* of links (URLs); and each URL subsequently points to other URLs or contents (URIs) in a nested fashion. How much of the retrieved contents are useful largely depends on the user's expertise and patience to understand and explore the property and nature of the contents.

2.2 Learning Object, Metadata and Domain Ontology

The IEEE Learning Technology Standards Committee (LTSC) describes *Learning Objects* (LO) as "any entity, digital or non-digital, which can be used, re-used or referenced during technology supported learning" [4]. Learning objects include multimedia content, textual content, instructional software and software tools, organizations. Each learning object must have a description that enables human and computer agents to search relevant contents efficiently. This means, objects must be wrapped in metadata. As the metadata is machine-readable, it must be possible for a specific system to interpret the metadata from other sources and then reuse the learning objects. However, the metadata has limitations because it is only a wrapper for the search engine to identify one object from another. However, the content interrelationship or dependencies between the LOs can not be fully described by only metadata. Therefore, a domain ontology which conceptualizes a particular domain is essential. In our research, we captured pedagogical attributes and content relationships using metadata and respective domain ontology.

2.3 E-Learning Standards

In this section, we will summarize some of the popular E-Learning standards, such as those specified by the Learning Technology Standards Committee (LTSC), the Instructional Management System (IMS), and the Advanced Distributed Learning (ADL).

2.3.1 LTSC

The LTSC was founded by the Institute of Electrical and Electronic Engineers (IEEE) which has developed many technological standards for electrical and information technologies and sciences. The purpose of LTSC is "to develop accredited technical standards, recommended practices, and guides for learning technology" [4].

2.3.2 IMS

The IMS is the most essential organization in the e-Learning community since it is a consortium of distinguished academic, corporate, and government organizations for developing and promoting *open specifications* to facilitating online distributed learning and to address interoperability issues, such as locating and reusing content, tracking learner improvement and exchanging student profiles between the systems [5].

2.3.3 ADL

The ADL developed the Sharable Courseware Object Reference Model (SCORM) [6]. SCORM is a specification for standardizing the reusability and interoperability of learning content.

SCORM focuses on two critical aspects of learning content interoperability:

1. It defines an aggregation model for packaging learning content.
2. It defines an API for enabling communications between learning content and the system that launches it.

SCORM also divides the world of learning technology into functional components. The key components are

1. Learning Management Systems (LMS)
2. Shareable Content Objects (SCOs)

SCOs are a standardized form of reusable learning object. An LMS is (for the purposes of SCORM) any system that keeps learner information, can launch and communicate with SCOs, and can interpret instructions that tell it which SCO comes next. Additional components in the SCORM model are tools that create SCOs and assemble them into larger units of learning

2.4 Semantic Web Technologies

Semantic Web is new WWW architecture that supports content with formal semantics. Therefore, the content is suitable for automated systems [7]. Such an architecture will enable automated agents to reason about Web content, and perform intelligent inferences about that content to develop customized courses delivered just in time to the user, according to their preferences and needs.

To achieve these goals we need to express the meaning (in terms of attributes) of the content by using Semantic Web technologies in several layers. The following layers are the basic ones:

- the XML layer, which represents data;
- the RDF layer, which represents the meaning of data;
- the Ontology layer, which represents the formal common agreement about the meaning of data;
- the Logic layer, which enables intelligent reasoning with meaningful data.

The Semantic Web technologies help us to develop systems that gather E-Learning content from diverse sources; process, organize and share content with other humans or artificial agents using ontology. Such an approach makes contents machine-understandable and it becomes possible to develop automated Web services with those heterogeneous contents. Three important technologies for developing the Semantic Web are eXtensible Markup Language (XML), the Resource Description Framework (RDF) and Ontology Web Language (OWL).

2.5 Relate work

In this section, we will refer to E-Learning content management approaches that are relevant to our work. Stojanovic et al. [8] examined the comprehensive benefits of E-Learning by using Semantic Web technology which provides an E-Learning portal architecture that uses ontology to address the description of learning material (content). The form of presentation (context) and the dimension of learning materials (structure) provide flexible and personalized learning materials. Tane et al.[9] developed tools called "Courseware Watchdog" which use ontology to address the different needs of teachers and students in organizing their learning materials. Dolog et al. [10] used personalized service architecture to address the gap between an adaptive educational system and a personalized functionality. It brings personalization to the semantic web to help the user find learning materials, courses or learning paths that are suitable for that user. However, there are still gaps in the semantic relationship between contents and user perception and affordance of those contents since LOs themselves are not capable of capturing individual or group perspectives. In conjunction with addressing those issues, we use domain ontology and declarative query language to represent LOs and their relationships and support navigation at the conceptual E-Leaning space. Knight et al. [11] and Amorim et al. [12] use ontologies to facilitate the representation of learning object context and expressiveness limitations found on the current XML-Schema. In addition, Agarwal et al. [13] provide the model of intelligent agents that can make E-Learning efficient since artificial agents tries to make inference across contents. In our research, we focused on a 3-Tier Architecture, where Content Tier (contents and their structures) is clearly separated from the Inference Tier as well as the Interaction Tier so that we can deal with any type of contents and user perspective using a generic inference engine as explained in subsequent Sections (cf. Figure 1).

3. System prototype

In this section, we explain the *3-Tier Content Management Architecture* after introducing the tools and techniques we used in developing our prototype. We also give a specific example of how the system works. Although, we restrict ourselves in a smaller domain (i.e. contents of a "Database Book"), our approach is equally effective for general purpose E-Learning content management as long as a domain-ontology is available to annotate contents.

3.1 Protégé, Jena2, RDQL and Pallet Reasoner

We use Protégé and OWL Editor [14] to construct domain ontology for its simplicity and popularity. We use Jena2 APIs and RDQL to interact with a generic Semantic Web Reasoner, *Pallet* (as explained below) to implement our prototype.

Jena is a Java framework for constructing Semantic Web applications and supports major ontology languages such as RDF/RDFS, DAML+OIL, and OWL (except OWL-Full). In particular, as of writing, Jena2 supports OWL-Lite, some constructs of OWL-DL and OWL-Full such as *hasValue* and partial *unionOf*. Some of the significant constructs that are not supported in Jena2 are *complementOf* and *one of* [15]. RDQL is a query language for RDF within the Jena framework. The purpose of RDQL is to extract information from RDF graphs. This means that RDQL only retrieves information stored in the model which contains a set of *N-Triples* statements. RDQL can process ontology in a number of languages including OWL. A typical RDQL query has the following form:

```
SELECT ?x
WHERE (?x shortPrefix:localName "value")
USING shortPrefix FOR <URIprefix>
```

`?x` is a variable. In the *WHERE* clause, a set of *N*-Triples define the pattern of a query. The *USING* clause defines an alias for the prefix of a URIs to simplify the URI. RDQL can also query about predicates or objects too. The limitation of RDQL is that there is no disjunction in the query. Though RDQL is relatively simple in syntax, it is efficient for most of the ontology queries.

Pellet is an open-source pure Java implementation of OWL-DL reasoner. It can be used in combination with both Jena and OWL API libraries and also provides a DIG interface. Pellet API provides functionalities validate and check consistency of ontologies, classify the taxonomy,

check entailments and answer a subset of RDQL queries (known as ABox queries in DL terminology). Pellet is an *OWL-DL Reasoner* based on the tableaux algorithms developed for expressive Description Logics. It supports the full expressivity OWL-DL including reasoning about nominals (enumerated classes). Therefore, OWL constructs *owl:oneOf* and *owl:hasValue* can be used freely. Currently, Pellet is the first and only sound and complete DL reasoner that can handle this expressivity. Pellet ensures soundness and completeness by incorporating the recently developed decision procedure for SHOIQ (the expressivity of OWL-DL plus qualified cardinality restrictions in DL terminology).

3.2 The *3-Tier* Content Management Architecture

The overview of the 3-Tier E-Learning Content Management Architecture is shown in Figure 1. The content server stores both the content and its structure in an ontology (contents may be distributes across servers and reachable with hyperlinks). The Generic AI Reasoner engine is in between the content server and user/software agent interface. The user or the agent interact with the content server through the *Reasoner* by specifying criteria (attribute-value pairs) and constraints; then the *Reasoner* locates the relevant contents (based on those conditions), and deliver the contents to the user/software agents in an *organized* fashion.

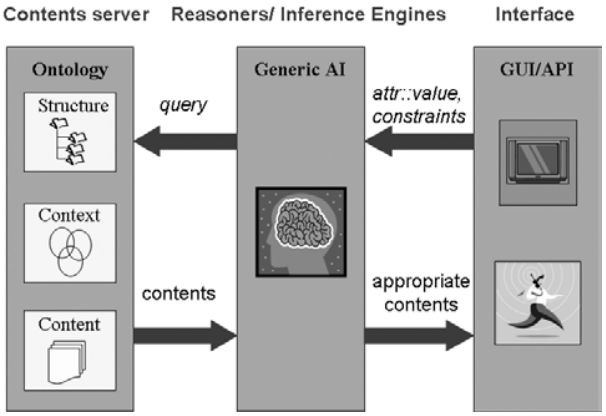

Figure 1: The 3-Tier E-Learning Content Management Architecture

a. The Content Server

We use OWL to represent and link the content on the E-Learning content server. The content dependencies and other pedagogical attributes of each object (content) are needed to be annotated in advance. Annotation may also be done in a collaborative fashion.

The domain ontology is to represent the domain knowledge. As an illustrative example, in this paper, we annotate the content of a *Database Book* in an ontology. The book consists of 4 *modules: Foundation, Applications, Systems, and Advanced*. The *Foundation* module consists of only 1 *Part*, *Applications* Module has 2 *Parts*, the *Systems* module has 3 *Parts* and the *Advanced* module has only 1 *Part*. Each *Part* consists of one or more *Chapters*. The *Chapters* across this book have direct and transitive relationships (Figure 2). We capture the content dependency and other attributes (such as *number of hours* required to deliver the content) in the domain ontology.

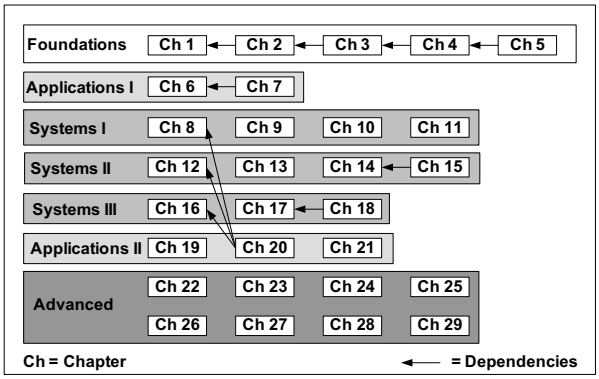

Figure 2: Organization of Database Course

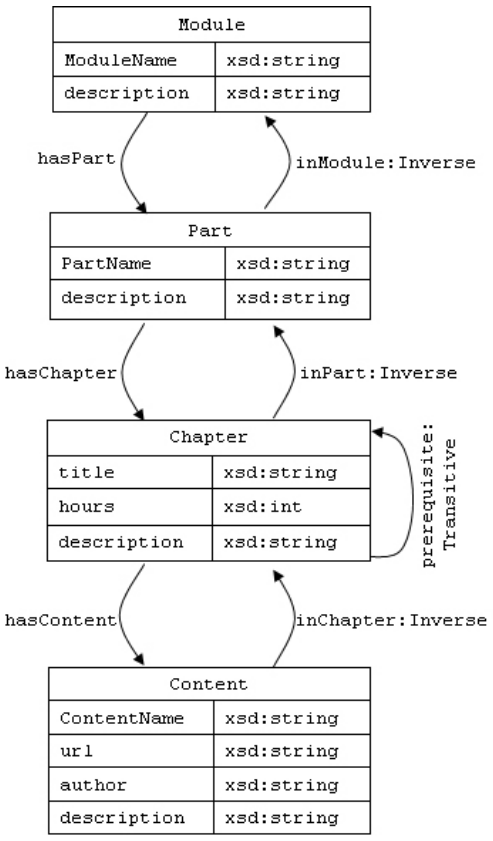

Figure 3: Content Schema for the Database Course

To illustrate the relationships or dependencies in this domain, we use 2 types of relationships, first is the direct

relationship between content such as *hasPart, hasChapter* and *hasContent* including its inverse relationship such as *inModule, inPart* and *inChapter*, respectively. Second is the indirect or transitive relationship such as *prerequisite* which means that it has inherited relationships. For example, the content of Chapter 5 is transitively related to contents of all four Chapters from Ch1 to Ch4. The content schema of the content ontology is shown in Figure 3.

b. The Generic AI Reasoner

We integrate Jena2 and Pellet OWL Reasoner and develop a Content Management Module (CMM). The CMM interacts with users or agents and make use of the Generic AI Reasoner (Jena and Pallet) to locate contents (based on user criteria) from the content server.

We use Jena2 and RDQL to query the content from the content server. The inference mechanism in Jena 2 does not support transitive relationships. Therefore, we resort to the Pellet OWL Reasoner as an *external* Reasoner. Pellet provided a Jena interface which can reduce overhead time instead of using other external Reasoner such as *Racer Pro* and *Fact* ++. The structure of Generic AI Reasoner along with CMM is explained in Figure 4.

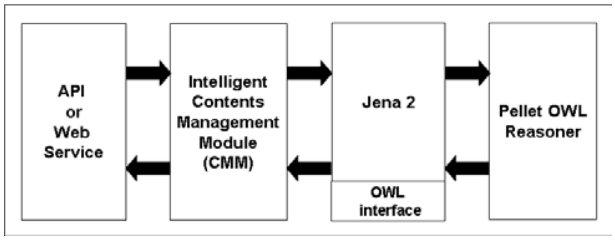

Figure 4: Organization of the Generic AI Reasoner

The system loads the ontology model in OWL from the Content Server. Jena 2 binds the model containing the original ontology. Then an RDF graph-base is ready for RDQL query. The Content Management Module (CMM) generates queries based on user's or agent's request. The CMM passes the query to Jena2 to find out the relevant relationship by using the Pellet OWL Reasoner which uses information in the base graph to generate additional entailments from the original set of statements. Finally, Jena 2 returns the content to CMM for necessary processing. The retrieved content that satisfies the query is generated in XML format via Web Service API and presented to the user in an organized fashion. Agent-based access to the content can also be done in similar fashion.

c. The Interface (API/Web Service)

The current prototype provides a *Web Browser-based Interface* for human users and a *Web Service API* for artificial agents. The web interface is written in PHP scripts. In this design we allow people to select topics and fill in the number of hour reflecting their time constraint for example. The user is then presented with content compiled in an organized courseware for personalized teaching and learning. The Web Service API defines the methods along with parameters to access content from the content server.

3.3 Operational Details of the Prototype ECMS

Content organized or annotated in ontology (OWL file) on our Content Server in this fashion can easily facilitate intelligent content dissemination to support personalized teaching and learning. In reality, teachers and learners often search specific content to satisfy their personal needs. In the context of our *Database Course* example, a teacher may already possesses partial content on *Foundational* database topics but trying to search for *Application-oriented* content to offer a 20-hour application-focused Database Course to a group of students who *do not have any foundation knowledge* in databases. Our prototype can accept such constraints and retrieve relevant content from the content server through ontology-driven reasoning as explained in the following subsections and screenshots.

a. The ECMS Server

The ECMS Server runs at a specific *TCP Port* which users can interact with by using a simple *telnet* program. The ECMS prototype is a multithread program which can support multiple users simultaneously.

The current prototype uses a *2-Mode approach* to retrieve content from the content server. In the *First Mode,* our CMM retrieves all content that is available in the OWL ontology through direct query matching and generates answers into XML form which user/software agents can easily optimize in terms of the utilization of the content; and in the *Second Mode,* constraints are further checked and validated with the help of the Generic AI Reasoner using description-logic –based inference.

b. Web-based interface

We have also implemented a Web-based interface for friendly interaction with users. The user can enter the URL of the OWL ontology in an input box and press the GO button. The Web browser then interacts with the ECMS systems via TCP/IP through a specific port. The ECMS will retrieve the collection of content which is available on the OWL ontology which it generates into XML form. Then the Web browser will convert it into HTML form and display it to the user (cf., Figure 5). The user can specify criteria and constraints based on their preferences. Responses from ECMS are generated in XML so that Web Services can be easily implemented for artificial agents in similar manner.

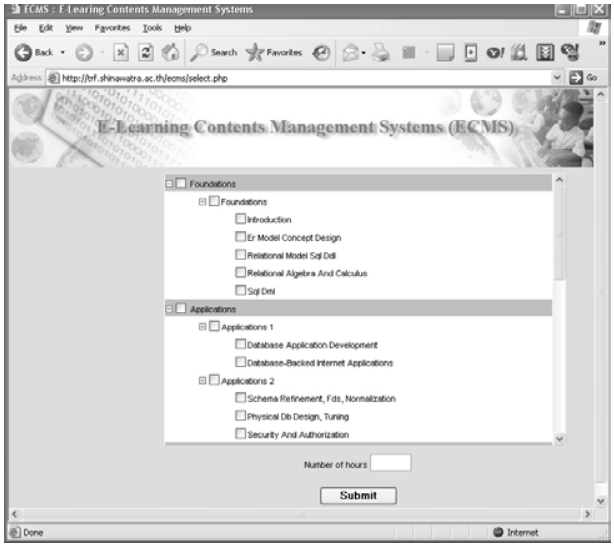

Figure 5: Web-based Input Interface for the ECMS

4. Conclusion and future work

In this paper, we explained a component-oriented approach for E-Learning Content Management using Semantic Web technologies. The prototype ECMS can efficiently organize contents for a particular domain in ontology and, therefore, with the help of a *generic* Semantic Web Reasoner, both users and software-agents can interact with the systems conveniently and can extract E-learning content efficiently. Contents do not necessarily have to be on a single server and annotation of contents can also be done collaboratively using a collaborative annotation too. We used a specific domain (Database Course Contents) and a specific example (a 20-hour Application-oriented database course for students without a Foundation in databases) to demonstrate the essence of our approach in the context of personalized E-learning. Other examples such as Contents for Foreign Language Learning are available online. We plan to apply our approach to broader domains [8][9] such as organizing the content available on a particular site such as the MIT's *Open Course Ware* initiative (http://ocw.mit.edu) or through automatic Web crawling [13]. Integration of User Profiles with E-Learning Contents as proposed by Dolog et al. [10] is also an interesting issue to explore under our 3-Tier Framework.

Acknowledgment

This research was supported by Thailand Research Fund (Grant No: **MRG4880112**) awarded to Dr. M.M. Hasan.

References

[1] D.A. Wiley, *Learning Object Design and Sequencing Theory*, Department of Instructional Psychology and Technology, Brigham Young University, 2000, pp. 142.

[2] S. Downes, Learning Objects: Resources For Distance Education Worldwide, *International Review of Research in Open and Distance Learning, 1*(2), 2001.

[3] Pellet, *Pellet - An OWL DL Reasoner*, retrieved June 23, 2006 from http://pellet.owldl.com/

[4] IEEE Learning Technologies Standards Committee *Learning Objects Meta-Data Specification*. Version 6.3, retrieved April 15, 2006 from http://ltsc.ieee.org/doc/wg12/.

[5] IMS Global Learning Consortium. *IMS Learning Design Information Model, Version 1.0 Final Specification, revision20,* retrieved April 28, 2006 from http://www.imsglobal.org/learningdesign/ldv1p0/imsld_infov1p0.html.

[6] ADL. (2005). *Advanced Distributed Learning SCORM Specification*, retrieved May 15, 2005 from http://www.adlnet.org/scorm/index.cfm.

[7] T. Berners-Lee, J. Hendler, O. Lassila, The Semantic Web**,** *Scientific American, 284*(5), 2001, 34-43.

[8] L, Stojanovic, S, Staab, R, Studer, eLearning based on the Semantic Web, *In Proceedings of the World Conference on the WWW and the Internet (WebNet 2001)*, Orlando, Florida, USA, 2001.

[9] J. Tane, C. Schmitz, G Stumme, Semantic Resource Management for the Web: An E-Learning Application, *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, New York, NY, USA, 2004, 1-10.

[10] P. Dolog, N. Henze, W. Nejdl, M. Sintek, Personalization in Distributed eLearning Environments, *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, New York, NY, USA, 2004, 170-179.

[11] C. Knight, D. Gašević, G. Richards, An Ontology-Based Framework for Bridging Learning Design and Learning Content. *Educational Technology & Society, 9*(1), 2006, 23-37.

[12] R. R. Amorim, M. Lama, E. Sánchez, A. Riera, X. A. Vila, A Learning Design Ontology based on the IMS Specification. *Educational Technology & Society, 9*(1), 2006, 38-57.

[13] R. Agarwal, A. Deo, S. Das, Intelligent agents in E-learning**,** *ACM SIGSOFT Software Engineering Notes, 29*(2), 2004, 1-1.

[14] Protégé, *Protégé - Ontology Editor and Knowledge Acquisition System*, retrieved March 2, 2006 from http://protege.stanford.edu/.

[15] Jena2, *Jena – A Semantic Web Framework for Java,* retrieved May 14, 2006 from http://jena.sourceforge.net/.

International Conference on Information and Communication Technology
ICICT 2007, 7-9 March 2007, Dhaka, Bangladesh

AUTOMATIC MUSIC CLASSIFICATION AND RETREIVAL: EXPERIMENTS WITH THAI MUSIC COLLECTION

Chakkapong Nopthaisong, Md Maruf Hasan

School of Technology, Shinawatra University
Shinawatra Tower III, 1010 Viphavadi-Rangsit Road, Chatuchak, Bangkok, Thailand
chakkapong.n@gmail.com, maruf@shinawatra.ac.th

ABSTRACT

We present the experimental results of classification and retrieval of Thai music using TreeQ (a tree-structured classifier) and LVQ (Learning Vector Quantization) algorithms in this paper. We use the HTK Toolkit in preprocessing acoustic signals including feature extraction from the Thai music collection. The training set consists of 250 songs – 50 songs from each of the 5 genres. Training is divided into *three* phases using all or some of these songs. The test set consists of 10 songs selected from 5 genres which are not included in training. We trained and tested the music classifiers using both TreeQ and LVQ algorithms with varying parameters such as, Number of Codebook (NOC) and pruning thresholds to identify the effects of different parameters and features in the Thai music classification and retrieval. We observed that TreeQ-based experiments yield faster response-times than those of LVQ; and therefore, a TreeQ-based system maybe appropriate for online (real-time) music retrieval tasks. On the other hand, LVQ-based experiments consistently yield better accuracy than those of TreeQ; and therefore, a LVQ-based system may be appropriate in the music classification task since music classification can generally be performed off-line. We also outlined a Relevance Feedback based Music Retrieval System in this paper.

Keywords: Music Classification; Music Information Retrieval; Machine Learning; Decision Tree; Self Organizing Map

1. INTRODUCTION

Music is generally classified based on its genre. Musical genres are widely used for classifying music in record stores, radio stations, in all sorts of music libraries, and presently increasingly on the Internet or on our own computers. Classifying music into a particular genre is a useful way of describing qualities that it shares with other music from the same genre, and separating it from other music. Generally, music in the same musical genre has certain similar characteristics, for example, similar types of instrument used, similar rhythmic patterns, or similar harmonic/melodic features. However, musical genres do not have clear boundaries or definitions since music is an art form that evolves constantly.

Several researchers have addressed the problem of audio classification problem. Foote [1] presented a system that can retrieve audio documents by using acoustic similarity. He measures the similarity based on statistics derived from a supervised vector quantizer, rather than matching simple pitch or spectral characteristic. His experiments show that, under different experimental settings, the Q-tree algorithm classifies audio signals with an accuracy of between 40% and 77.2%. Soltau *et al.*[2] also presented a system that can classify music genres from a multimedia database. Four categories, Rock, Pop, Techno and Classic are chosen, which are classified by ETM-NN (Explicit Time Modeling with Neural Network) and compared with the Hidden Markov Model (HMM). The result shows that the ETM-NN shows a better performance (86.1%) than those of HMMs (79.2%). Logan [3] explored the Mel Frequency Coefficient (MFCCs)-based classification; and the result shows that the Mel scale is suitable for modeling music. Tzanetakis *et al.* [4] introduced a 9-dimensional feature vector using Wavelet Transform (WT) for representing the rhythmic structure of music to test with 2-classifiers: the Guassian Mixture Model (GMM) and the Random Classification (RC). Their results show that GMM gives significantly better (86% and 74%) classification accuracy than the RC counterpart (50% and 33%) for music and voice data.

This paper describes our findings on music classification tasks using TreeQ [5] and LVQ [6] algorithms. HTK toolkit [7] is used to extract features

from Thai music. The goals of this research are: (1) to explore music classification accuracy using different techniques and tools; (2) to compare the classification performance between these techniques; (3) to make recommendations for practical applications using these techniques in music classifications and retrieval tasks.

2. APPROACHES

2.1 Training set

In our experiment, we trained different classifiers using a collection of 250 Thai musical excerpts selected from 5 different genres (50 music each), namely, Country, Folk, Oldie, Pop and Rock.

We trained each classifier in 3 phases – using 10, 25 and 50 songs from each genre (as shown in first column of Table 1) to understand the effect of the size of the training music set.

2.2 Test set

We randomly selected 10 songs not included in the training (i.e., outside test), which consisted of 1 Country, 2 Folk, 2 Oldie, 2 Pop and 3 Rock songs. The classification accuracies are reported in the rest of this paper. We admit that the training set is rather small and, therefore, we have conducted a detailed analysis of classification errors for each experimental setup.

2.3 Feature extraction

2.3.1 *Mel-scale Frequency Cepstral Coefficients (MFCCs)*

We parameterized all music files into mel-scale cepstral coefficients (MFCCs) plus energy term with the help of the HTK Toolkit.

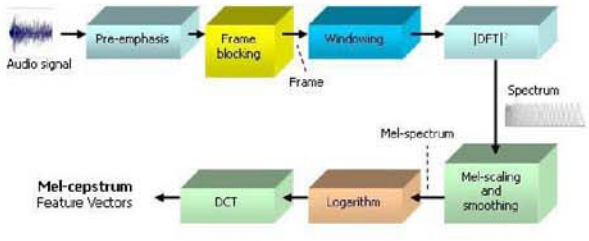

Fig. 1 Extraction of MFCC vectors

Fig. 1 shows the step of extraction of the MFCCs feature. These steps are motivated by perceptual and computational considerations. The preprocessing step involves pre-emphasizing the audio signal, dividing the signal into frames and windows. Pre-emphasis is done using a first-order finite impulse response (FIR) filter $1-0.97z^{-1}$ to increase the relative energy of the high-frequency spectrum. The aim of frame blocking is to segment the signal into statistically stationary blocks. A Hamming window is used to weight the pre-emphasized frames. Next, the Discrete Fourier Transform (DFT) is calculated for the frames. Since the human auditory system does not perceive pitch linearly, a perceptually meaningful frequency resolution is obtained by averaging the magnitude spectral components over Mel-spaced bins. This is done by using a filterbank consisting of 40 triangular filters occupying the band from 80 Hz to half the sampling rate, spaced uniformly on the Mel-scale.

2.3.2 *The HTK Toolkit*

The Hidden Markov Model Toolkit (HTK) [7] is a portable toolkit for building and manipulating hidden Markov models. HTK is primarily used for speech recognition research although it has been used for numerous other applications including research into speech synthesis, character recognition and deoxyribonucleic acid (DNA) sequencing. HTK is in use at hundreds of sites worldwide.

2.4 Classification

2.4.1 *The TreeQ package*

The TreeQ package [5] has been successfully used in speaker identification, speech and music classification, music retrieval by similarity and audio segmentation. The TreeQ package is a set of C-language libraries that implement a decision tree like an automatic machine learning algorithm. This package lets us construct a system that will learn the differences between complicated data sets. Though the system is data-driven, and thus will work on any arbitrary data, we have experimented primarily with music. The basic functions let us find the similarity, expressed as a number between two given audio files.

2.4.2 *Learning Vector Quantization (LVQ)*

LVQ [6] is a precursor of the well-known self-organizing maps (SOM, also called Kohonen feature maps). It can be seen as a special kind of artificial neural network. A neural network for learning vector quantization consists of two layers: an input and an output layer. It represents a set of reference vectors, the coordinates of which are the weights of the input or output neurons.

3. EXPERIMENTAL SETUP & RESULTS

3.1 Experimental setup for TreeQ

The music classification accuracy was validated using TreeQ against the test set which was comprised of 10 test songs. The training was divided into 3 phases using varying sizes of training music. In each phase, we used varying numbers of pruning thresholds. The experimental setup of TreeQ is shown in Fig. 2.

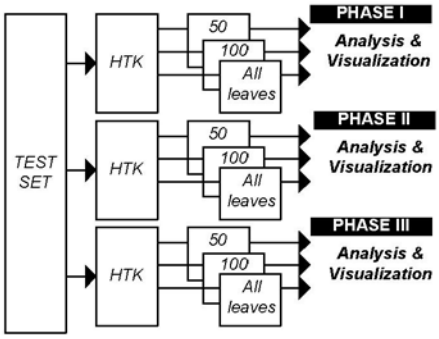

Fig. 2 Experimental Setup of TreeQ

3.2 Experimental setup for LVQ

The music classification using LVQ was conducted against 3 different training corpora of varying numbers of music. We chose varying thresholds in a number of codebooks (NOC). The experimental setup of LVQ is shown Fig. 3.

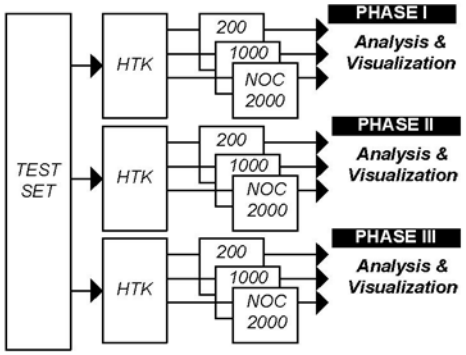

Fig. 3 Experimental Setup of LVQ

3.3 Experimental Results

3.3.1 *The Results of TreeQ*

From the overall experimental results of TreeQ, we can safely conclude that a pruning threshold of 50 performs as good results as those of 100 or more. However, TreeQ requires a bigger training set to classify music accurately. The results are summarized in Table 1.

Table 1: Classification results using TreeQ using 3 sets of training music in 3 phases

Phase	50 leaves	100 leaves	All leaves
Training Phase I (10 * 5 songs)	60%	70%	50%
Training Phase II (25 * 5 songs)	70%	60%	80%
Training Phase III (50 * 5 songs)	90%	90%	90%

3.3.2 *The Results of LVQ*

From the overall experimental results of LVQ, it appears that, with LVQ, a bigger training corpus along with a higher NOC threshold (1000 NOC or more) yield the best classification results. The results are summarized in Table 2.

Table 2: Classification results using LVQ

Phase	NOC 200	NOC 1000	NOC 2000
Phase I (10 * 5 songs)	50%	50%	70%
Phase II (25 * 5 songs)	80%	80%	80%
Phase III (50 * 5 songs)	60%	100%	100%

3.4 Comparative Analysis of Classification Results using TreeQ and LVQ

3.4.1 *Result from Phase I*

Table 3 summarizes the result which is classified by TreeQ and LVQ. This experiment was based on training with a small corpus, 10 songs * 5 Genres. The result shows that both TreeQ and LVQ present an accuracy around 50% - 70% against our test set. Only TreeQ with 100 leaves shows the best performance which is given 70% accuracy in this corpus

Table 3: Experimental results in Phase I

Phase I							
Test set		TreeQ			LVQ		
		50 Leaves	100 leaves	All leaves	NOC 200	NOC 1000	NOC 2000
Song 1	(Folk)	Folk	Folk	Folk	Folk	Folk	Folk
Song 2	(Pop)	Fok	Fok	Folk	Folk	Folk	Folk
Song 3	(Rock)	Pop	Country	Country	Country	Country	Rock
Song 4	(Folk)	Folk	Fok	Folk	Folk	Folk	Folk
Song 5	(Oldie)	Oldie	Oldie	Oldie	Rock	Rock	Oldie
Song 6	(Oldie)	Oldie	Oldie	Oldie	Oldie	Oldie	Rock
Song 7	(Pop)	Pop	Pop	Country	Country	Country	Pop
Song 8	(Rock)	Country	Rock	Country	Country	Country	Rock
Song 9	(Rock)	Country	Country	Country	Rock	Rock	Rock
Song 10	(Country)	Country	Country	Country	Country	Country	Country
Accuracy (%)		60%	70%	50%	50%	50%	70%

3.4.2 *Result from Phase II*

Table 4 shows the result which is classified by TreeQ and LVQ when training was performed against a medium corpus, 25 songs * 5 Genres. The result shows that both TreeQ and LVQ present an accuracy about 60% - 80% against our test set. TreeQ, retaining all leaves, shows the same performance as LVQ. However, TreeQ with 100 leaves and 50 leaves shows a lower performance in this phase, which is 60% and 70%. LVQ with all variation of NOC shows very good performance which is 80% accuracy against our test set.

Table 4: Experimental results in Phase II

Phase II							
Test set		TreeQ			LVQ		
		50 Leaves	100 leaves	All leaves	NOC 200	NOC 1000	NOC 2000
Song 1	(Folk)	Folk	Folk	Folk	Folk	Folk	Folk
Song 2	(Pop)	Folk	Folk	Folk	Folk	Folk	Folk
Song 3	(Rock)	Pop	Pop	Pop	Rock	Rock	Rock
Song 4	(Folk)	Folk	Folk	Folk	Folk	Folk	Folk
Song 5	(Oldie)	Oldie	Oldie	Oldie	Oldie	Oldie	Oldie
Song 6	(Oldie)	Oldie	Oldie	Oldie	Oldie	Oldie	Oldie
Song 7	(Pop)	Pop	Rock	Pop	Rock	Rock	Rock
Song 8	(Rock)	Rock	Rock	Rock	Rock	Rock	Rock
Song 9	(Rock)	Country	Country	Rock	Rock	Rock	Rock
Song 10	(Country)	Country	Country	Country	Oldie	Country	Country
Accuracy (%)		70%	60%	80%	80%	80%	80%

From this experiment, we observed that when using a bigger corpus, the system gives a better performance using both TreeQ and LVQ in all variations of number of pruned leaves and NOC. Then, we made a bigger corpus to verify our hypothesis that a bigger corpus should result in a better performance of the systems. In the next phase we experiment with the same test set against the new corpus which is built with 50 songs * 5 genres, to see whether the result of TreeQ and LVQ shows any improvement or not.

3.4.3 *Results from Phase III*

Table 5 presents all results which are classified by TreeQ and LVQ. This experiment was based on training with the biggest corpus that we made to test our system, 50 songs * 5 Genres. The result shows that TreeQ with all variations of leaves shows the same result with an accuracy of 90% against our test set while LVQ with NOC 1000 and 2000 expresses an excellent performance (100%). NOC 200 shows a lower performance at 60% accuracy against the test set.

Table 5: Experimental results in Phase III

Phase III							
Test set		TreeQ			LVQ		
		50 Leaves	100 leaves	All leaves	NOC 200	NOC 1000	NOC 2000
Song 1	(Folk)	Folk	Folk	Folk	Rock	Folk	Folk
Song 2	(Pop)	Pop	Pop	Pop	Pop	Pop	Pop
Song 3	(Rock)	Country	Pop	Pop	Rock	Rock	Rock
Song 4	(Folk)	Folk	Folk	Folk	Rock	Folk	Folk
Song 5	(Oldie)	Oldie	Oldie	Oldie	Rock	Oldie	Oldie
Song 6	(Oldie)	Oldie	Oldie	Oldie	Rock	Oldie	Oldie
Song 7	(Pop)	Pop	Pop	Pop	Pop	Pop	Pop
Song 8	(Rock)	Rock	Rock	Rock	Rock	Rock	Rock
Song 9	(Rock)	Rock	Rock	Rock	Rock	Rock	Rock
Song 10	(Country)	Country	Country	Country	Country	Country	Country
Accuracy (%)		90%	90%	90%	60%	100%	100%

4. ERROR ANALYSIS

After collecting the results from all the experiments, we analyze the performance of the system by comparing the results from the experiments in phase I, phase II and phase III which were trained by a different training corpus against the same test set. We found that the results from phase III which was trained with the biggest corpus (50 songs*5 genres) shows a better performance than those of phase I and phase II. This phenomenon can be discussed from the following angles.

In TreeQ experiments, we observed that the number of leaves such as 50 leaves, 100 leaves and all leaves gives almost the same result, taking less than 2 seconds. However, a tree with 50 leaves shows a closer result than those of the others. For this reason, we can summarize that the music classification is suitable to use with a tree with 50 leaves.

Based on the LVQ experiment, we observed that a small number of NOC express a good result when testing against a small or medium sized corpus, while with a bigger corpus the higher number of NOC shows a better performance. Not only the number of NOC but also task fulfillment in terms of time consumed should be considered in terms of cost. Experiments show that with NOC 200, 1000 and 2000, 2, 6 and 12 seconds respectively are spent to measure the similarity of each song. This means the best compromise using this LVQ is to use NOC 1000, which performs the best in terms of cost and time.

Comparing TreeQ and LVQ, we observed that the improvement of the result will increase when we make a bigger corpus for the system. The results in phase III of the experiment show that both TreeQ and LVQ show the best performance; TreeQ in all variation of pruning shows 90% accuracy against the test set while LVQ with NOC 1000 and NOC 2000 presents 100% accuracy. We also noticed that LVQ with NOC 200 shows a lower performance by being