From Text to Exhibitions: A New Approach for E-Learning on Language and Literature based on Text Mining

Qiaozhu Mei
Department of Electrical Engineering and Computer Science
Vanderbilt University
Box 1679 Station B
Nashville, TN 37235 USA
qiaozhu.mei@vanderbilt.edu

Junfeng Hu
Department of Computer Science
Institute of Computational Linguistics
Peking University
100871, Beijing, China
hujf@pku.edu.cn

Abstract

Unlike many well established approaches for E-Learning on science fields, there isn't a commonly accepted approach of E-Learning on humanities fields, especially language and literature. Because the knowledge on language and literature depends too much on texts, advanced text processing has become a bottleneck for E-Learning on these domains. In traditional learning frameworks learners would easily get boring with mass pure texts. This article introduces a new approach for E-Learning on language and literature, by intelligently extracting real or virtual objects from texts and integrating them as exhibitions in a digital museum system. This article also discussed how to generate exhibitions from texts with computational linguistics methods as well as how this E-Learning framework pushes the research of computational linguistics. The discussion of E-Learning by Digital Museum is based on the design of Digital Museum of Chinese Ancient Poetry, by Peking University.

1 Introduction

Computer based Education has become a very hot and productive topic in recent years. However, most of the existing methodology and models are based on science domain. This is because the teaching and learning on science domain relies much on the ability of reasoning and computation, which directly utilizes the advantage of computer. The most important carriers of Knowledge on humanities domain, especially literature and language are textual materials. Therefore, unlike E-Learning on science and technical fields, a more intelligent way of using computer to deal with texts is required. Traditional E-Learning models on language and literature rely too much on pure text. Relevant frameworks include Digital-Archives, Digital-Libraries and Digital publications. Most of them are just "gathering mass text materials and providing them online", thus the interface between system and learners is onefold, non-interactive and lack of guidance. Learners easily get missed in excessive bald texts without a "docent [2]" to advise them how to select a well organized knowledge structure and a learning pathway. Searching and retrieving modules are provided in those models to various extents, which provide a knowledge retriever. However, it still cannot go beyond texts.

Recently, Digital Museum systems are believed to be able to provide a vivid interface which carries educational uses to participants. Teaching and learning becomes much easier from the special circumstance of learning in the presence of real objects, which inspires curiosity and creative thinking, and gives museums the potential to develop distinctive and meaningful educational experiences [5].

There are many good examples that approach E-learning on humanities fields with a system similar to a Digital Museum. The National Palace Museum system in Taiwan offers 14 courses on the cultural relics of China [3]. Digital Museums on more than 10 major fields in nature and culture have been designed along with Taiwan's nation wide Digital Museum plan. Lo, Feng-ju et' al have designed a digital museum of Chinese Ancient Literature, which provides some sub-exhibitions of poetry and fictions in formats of photocopy of the actual paper edition of ancient texts.[7] These works have been well exploring the primitive application of Digital Museum in E-Learning on Humanities Fields.

To satisfy the needs of E-Learning on Language and Literature fields, a modern digital museum should have some specific features. It should provide a mechanism to process texts, which would be able to integrate some computational linguistics methods. It should also provide a way to organize knowledge beyond the texts, and be able to provide guidance for learning. This can be achieved by generating objects out from texts and organizing them into interactive exhibitions that

can be personalized. Moreover, the digital museum framework should be reusable to different scope of background knowledge. Such a modern digital museum associating text processing mechanism is believed to be a sound approach of E-Learning on Language and Literature.

This article discussed this approach on the Digital Museum framework design, how it is associated with Computational Linguistics, and how to integrate knowledge to maximize the E-Learning efficiency. These discussions will be based on an example of the Digital Museum of Chinese Ancient Poetry Art, by Peking University 2003. [10] The following section will discuss the general framework design of digital museum. We will discuss text processing work behind the Digital Museum in Section 3, and Knowledge Processing and integration in Section 4. Some more discussion and future work will be provided at the conclusion section.

2 The Digital Museum Framework

Instead of digital library and traditional digital museum systems, which provide single function of exhibition, a modern digital museum provides multidimensional functions. Generally, a modern digital museum has three key functions, exhibition, education and research. In our design of Digital Museum for Language and Literature, the three dimansion would be: interacting theme based exhibitions from texts, E-Learning modules on language and literature, and related research on Computational Linguistics.

2.1 Digital Museum and E-Learning on Language and Literature

Digital Museum systems have gone beyond exhibitions of digital collections. Instead, they would increasingly emphasize educational uses rather than traditional exhibitions. It provides users with educational and well-motivated exhibitions [13]. UK-wide Digital Museum linked exhibitions connected by subject and theme with an integrated learning environment [6]. By 2000, the National Science Plan of Digital Museums of Taiwan has defined a specific and integrated program on how to utilize scientific technology, especially information technology, and how to digitalize the archives in both cultural and natural fields, with significant humanistic meaning. It has conducted further discussions on how to apply these kinds of digital projects and productions to education, research and industrialization, for the sake of conserving culture, promoting education, inspiring research and increment of industrialization. [3]. Knowledge on a learning topic should be organized as an exhibition theme, which is represented by a series of real or virtual objects and detailed descriptions. Exhibitions of various themes are linked together corresponding to the relativity of their themes. Learners can participate in the Digital Museum by choosing a pathway of linked exhibitions with a typical topic. Special modules will also be provided for participants to interacting with the system, which will be discussed in section 4.

2.2 General Architecture Design of a Digital Museum

The life cycle of a modern digital museum looks like a fountain model [11]. There are feedbacks from each design phase to previous phases. There are several milestones in the life cycle, each of which acts as a knowledge container and a foundation of knowledge processing on upper levels. [14]. These knowledge containers are as follows:

Milestones	***Functionality***
Information Origin Pool: (Primitive Corpus)	The mass storage of large-scale information from preliminary digitalization work.
Refined Knowledge Bases (Refined Corpus)	Database storage of useful and relevant knowledge from knowledge refining.
Metadata for Exhibitions	Metadata describing ontology, with all detailed metadata for knowledge flows, items and relations
Integrated Exhibition Base	Database for Exhibiting items, individual or integrated, for regular accessing by system.
Reusable Tool Base for Functional Modules	Tool pool for reusable module functions, individual or integrated components for various use.
Multi-functional Interface	Web-based interface for exhibitions, education and research.

Table 1: Milestones within the Digital Museum Architecture

Based on these milestones, the general architecture of a Digital Museum on Language and Literature can be represented in the following figure:

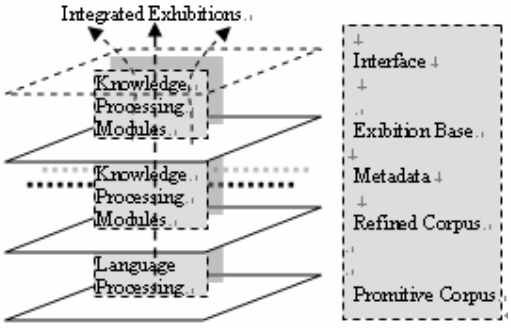

Figure1: General Architecture of a Digital Museum based on language processing

2.3 Example: Introduction to the Digital Museum of Chinese Ancient Poetry

The Digital Museum of Chinese Ancient Poetry Art [10] is a research model by Peking University, Beijing, combining E-Learning, computer assisted research on Chinese Ancient Poetry and computational linguistics. A prototype of this Digital Museum was designed in order to meet the needs of exhibition, education and research on the art of Chinese Ancient Poetry. The analysis, design and implementation of this project were on a highly abstract level.

2.3.1 Corpus, Design and Prototype System

The information origin pool and the refined knowledge base of this project were also the corpus for related computational linguistics research. It involves Chinese Ancient Poetry across 2,000 years, approximately 100,000 items [10]. Other advanced knowledge bases such as Author Information base, Image and media base, Location information base and Word lists were constructed.

In the design of this Digital Museum system, knowledge mining was divided into two types, item entity information mining and relational information mining. Item entity information was detailed to exhibiting items, characters, images, media, locations and words. Relational information reflected all aspects of relations among items. Metadata for each category of instances was defined in the design phase. Particularly, a group of items with relating meaning was structured as a virtual item class, which was also treated as a specific item.

In the prototype system, items of poetry, character, location and others were exhibited along with all related formats of knowledge. Users can leap from one item to its related items, and learn them in the context where they originally belongs. Sample exhibitions on specific themes, such as clothing, plants, food and spring were also designed.

2.3.2 E-Learning and Related research from this Digital Museum

In the dimension of learning, Digital Museum of Chinese Ancient Poetry explored the study of E-Leaning system for the language and literature features of Chinese Ancient Poetry. It enabled a way to learn a poem in its background environment, with reference to its related poetry and other related objects in multiple formats. The system also presented statistical research results of the corpus to users, such as the words usages of authors, the cooccurrence of words, the likelihood of the hidden meanings of words, which help users to be well-informed and easier to understand in learning a poem or a word.

In the dimension of research, the digital museum is closely related to specific research topics on computational linguistics, especially statistical natural language processing. We refined unknown words from the corpus though statistic methods and explored to cluster them into concepts. In this way, we studied the hidden meanings of words and poetry in context and studied the relation discovery among poems. We also conducted some research of knowledge mining and discovering from corpus, which can also inspire extended researches like Computer Assisted archaeology on Chinese Ancient Poetry.

3 Language Processing behind the Digital Museum Framework

Knowledge of humanities areas, especially language and literature, is commonly carried by texts. Therefore, the language processing, specifically the text processing will be vital for transforming pure texts and domain knowledge into abstracted exhibitions. Actually, most digital museums today haven't made good use of computational linguistics techniques. Most of them remain on organizing exhibitions manually and providing them online. Those exhibitions are relatively isolated from each other.

However, there are remarkable relations among text units and real objects and topics, which are hidden in the texts. For example, the word “willow” seems having nothing to do with “getting apart” by the semantic definitions, but in the context, “breaking a willow branch” does indicate “send-off friends”, or “seeing a friend leaving” in Chinese Ancient Poetry.

These meaningful entities and relations can be learned from the statistical analysis of large scale poetry texts. The use of computational linguistics methods here is crucial, which distinguishes it with traditional Digital Museum models. Statistical natural language processing over large scale corpus is the most significant approach we have adopted in this research.

3.1 Construction of Corpuses and Integrated Knowledge bases

The first phase of language processing is to build corpora and knowledge bases. Primitive corpora are constructed by archive digitalization. Refined corpora are constructed by applying language processors on the primitive corpus. We can use Digital Museum of Chinese Ancient Poetry for example.

For the Digital Museum of Chinese Ancient Poetry Art, the primitive corpora include texts of

poems over 1, 200, 000 lines, descriptions of 4000 authors, a name dictionary and a location dictionary. The refined corpora include a words dictionary which is thoroughly discovered from the texts, a concept base constructed by supervised word clustering and a storage of words cooccurances. Other knowledge bases include images, music, medias(reading), relics, events, and a series of expertise knowledge on Chinese Ancoent Poetry.

The general ontology of domain knowledge was carefully studied. Important entities and relations from texts and related domains were determined. Consequently, we carefully designed the metadata and chose a database system to maintain the knowledge base. This knowledge base should be expandable so that it can contain texts, entities from related domains, and relations.

The last step of this phase is to design an referencing mechanism to query and get the answer. The outcome of this phase is an integrated knowledge base, the textual part of which is the corpus for mining and knowledge discovery.

3.2 Text Mining: Extracting Objects from Texts

As soon as the corpora and knowledge bases are constructed, higher level methods of natural language processing are applied to mine in the corpus. The goal is to find objects abstracted from texts, which are organized by individual topics. Statistical natural language processing plays a very important role in this procedure, which can be described in the following three levels.

3.2.1 Extracting Direct Relevant Objects from Texts.

Textual knowledge is not "dead" in the fields of language and literature. It is interacting with knowledge in other forms, by other carrier or on other abstract level. Taking Chinese ancient poetry for example, a poem is associated to its author, its era and its writing background. The textual body of a poem also refers to certain persons, events, locations, plants, scenes, feelings and other entities, either real or virtual. In addition, there are various sources of objects relevant to the poem, such as paintings, calligraphy works, music and cultural relics, etc. All these entities above are so important to the synopsis of the poem that it is an advisable way to learn the poem with the appearance of these objects. Furthermore, relying on these directly relevant objects makes teaching and learning much more open and exciting than barely focusing on texts.

In the early phase of Digital Museum design, an integrated exhibition base is built, in which directly relevant entities of the texts are refined, stored in relational or XML databases and associated with the body of texts.

3.2.2 Discovering Hidden Entities and Relations Associated with Language Units.

As the Computer assisted research develops on these fields, we can work on the hidden knowledge of texts by means of text mining and retrieval. As language technology evolves, a computational age of language has arrived [1]. We can conduct computer assisted analytical research on language, with both linguistic and statistical approaches. In the research on the language of Chinese ancient poetry, we studied the statistical concurrences and meaningful units in the texts, extracted words from collocations and clustered words into meaningful concepts. In further research, we explored ways to study the hidden meanings of the words and collocations, especially those related to emotions of human. Consequently, expected to learn emotional characteristic of a poem, associating words, concepts and other units it refers with the similar characteristic.

On the other hand, language and texts are the most important carriers of cultural fragments. Many interesting knowledge patterns are hidden in the texts. There is a considerable proportion of Chinese ancient history and culture buried in the texts of Chinese ancient poetry, which evolutes along more than 2,000 years and involves locations all over China. By language techniques, fragments of culture can be mined from the texts, refined and stored, and finally integrated into interacting virtual scenes.

By this we can discover hidden entities and relations associated with text and expand it to analytical meaningful segments.

3.2.3 Expanding Indirect Relations.

In our framework, knowledge entities are not living alone but interacting. Both textual entities and other objects are associated to its relevant entity set. There are two kinds of relations identifying that two entities are interacting, direct relation, which have already been discussed above, and indirect relation. For instance, a poem refers to various knowledge objects, thus poems referring to the same objects are indirectly interacting with each other. These poems are involved in their relevant entity set, with "identical reference" as an indirect relation. In a more intelligent level, poems with the similar hidden meanings or relevant emotions are arranged together as a set. This set can be associated with a topic, a subject, a scene or a specific semantic cluster.

In these three approaches to expand textual knowledge into relevant objects, a former purely

textual entity has been developed as involving in the surrounding of various relevant objects, real or virtual. Thus we complete the procedure of extracting objects for exhibitions from texts. An example from poems to objects is as follows:

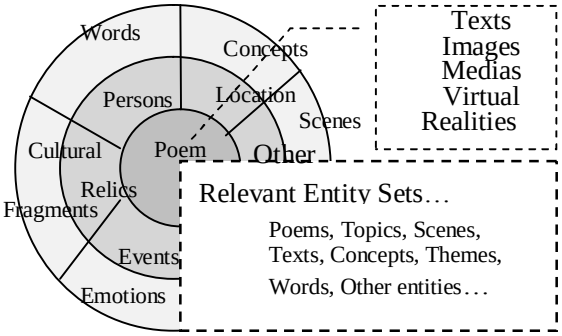

Figure2: Expanding Objects Set from a Poem Text.

3.3 Theme Driven Knowledge Discovery

From the statistical analysis on character concurrences, we applied various methods to discover unknown words from the texts. Chinese language is different from other language because there isn't natural interval from a word to another. We consider all words to be unknown in the beginning and generate a word dictionary from the filtering by mutual information value, μ-test and other statistical methods.

Upon the word dictionary, we conducted words clustering by the distance of words concurrence vectors. This procedure has abstracted concepts from words. After supervised filtering, these concepts will indicate some hidden semantic meanings.

The consecutive knowledge discovery work will be theme driven. First, a theme, or a learning topic is decided, some features and key concepts of this theme will be decided with the expert knowledge. Using statistical methods, we can find the concepts and words which are semantically similar or in some way related to this theme. Then, directly and indirectly related objects (discussed in section 3.2) will be associated with the topic. Then, reluctant units are eliminated. We will filter the most significant entities and relations, which can be represented by combinations of both concepts and words, and organize them around the theme. In this way, we can put the topic/theme back to its ancient living environment.

Further works includes rebuilding ancient scenarios where the topic belongs, and mining for relations among topics.

4 Knowledge Processing and Integration of the Digital Museum

Knowledge processing plays a very significant role in the Digital Museum framework. It is involved as a clue throughout the life cycle of the digital museum. The entire design and implementing of the digital museum is focusing on language processing, knowledge discovery and exhibition integrating. The knowledge processing procedures can be represented in the following figure:

Figure3: Knowledge Processing in this digital museum.

4.1 Knowledge Processing Hierarchy

An intelligent platform of knowledge deals with knowledge in five primary hierarchies, namely, knowledge citation, knowledge application, knowledge transmitting, knowledge learning and knowledge developing [8]. This division of knowledge hierarchies remarkably adapts the needs of an E-Learning program. In the study of this article, we make a little modification to this division and applied it to the Digital Museum system as follows:

Knowledge Citation	
Knowledge Applying	
Knowledge Learning	
	Learning and Teaching
	Knowledge Mining
Knowledge Representing	
	Knowledge Representing to Users
	Information Interacting
Knowledge Developing	

Table 1: A knowledge processing hierachy in the Digital Museum

Actually, this division is somewhat relative and not absolute. For instance, in some activities defined as knowledge representation and knowledge developing, we may also need to do knowledge citation and applying. However, this division of knowledge hierarchy would help to define the functions of Knowledge Platform and content the needs for knowledge by systems and users. [8]

The Digital Museum presents multidimensions according to the three functions of exhibition, education and research. The processing targets, procedures and emphases on Knowledge vary among dimensions.

In the dimension of exhibition, system focuses on Knowledge citation and Knowledge representing in the hierarchy above.

In the dimension of e-learning, system focuses on the hierarchy of Knowledge applying, learning and teaching, Knowledge Representing and information interaction.

In the dimension of computational linguistics research, system emphasizes the hierarchy of Knowledge Mining and Knowledge developing.

4.2 Two Types of Integration for Knowledge Objects

After discussing the generating of objects from the texts, we would be interested in how to integrate them for E-Learning.

Relating and interacting objects are extracted from texts and stored in the exhibition base. The next phase is to arrange exhibitions by selecting, dividing and integrating these objects, and construct the digital museum interface.

There are two key forms of objects integration, tutored and theme-oriented exhibitions and virtual scenarios.

In the first form, tutored theme-oriented exhibition, objects relevant to a specific subject or theme are integrated and represented in multi-modals. This interface design provides a dynamic exhibition module by grouping texts and their relevant objects in various formats together, providing docent knowledge for this topic and links to relevant topic exhibitions. Learners participate in one exhibition and go through links fitting to their needs or under instructions, thus personalized learning paths are formed.

There are two tips in tutored theme-oriented exhibitions. One is "multi-modal". Personalized exhibitions in our framework enable learning through multi channels, in forms of texts, image, music and virtual reality, etc. Also taking Chinese ancient poetry for example, we first discover the relevant scenes and hidden emotions of a poem, select objects referring to similar scenes and emotions, provide them as background materials and then integrate them with the poem. A more detailed instance is the Auto-matching poems and paintings. The other is "interactive". In our framework, a learner can add his remarks or discuss in every exhibition topic. These remarks are processed and stored as new relevant objects to this topic. Users can also provide materials or background information to an object or a topic, and can provide their own exhibition plans of new organizations of objects. The system studies the feedbacks and provides users with personalized participation paths.

The second integration form is scenarios. Knowledge objects were recorded in texts from their original living environments. By collecting and extracting relevant objects from texts and analytical researching on their relevant environmental elements such as emotions, we are able to put a textual object back to a scene representing its original living environment by rebuilding these origin scenes. Teaching and learning are made easier and more exciting with participating in the original scenes that a topic really lived. With the technology of multimedia and virtual reality, we are able to integrate objects and environmental elements surrounding a specific topic and rebuild a virtual scene, which is represented in our framework as multimedia demonstration, tests and games.

These two key integrating patterns organize various formats of objects and represent these integrated exhibitions to users in an interactive and personalized way. It maximizes the educational use of a digital museum on language and literature fields.

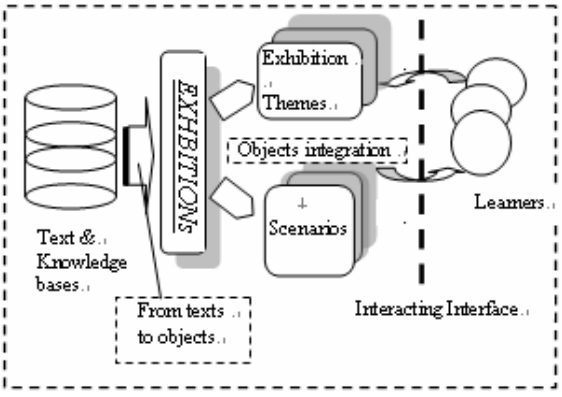

Figure3:Integrating exhibits in the Digital Museum on Chinese Ancient Poetry.

5 Conclusion

Computer-based education on language and literature has both its advantage and difficulty. On one hand it provides learners with abundant relating materials, on the other hand it's tedious

and difficult for learners to acquire knowledge in the sea of information. The approach of extracting objects from texts, and integrating them to build an interactive and vivid exhibitions enables learners both to explore in broad scope of knowledge and to enjoy exciting and comprehensible learning. Computer techniques are adopted in the framework of Digital Museums to maximize its educational use. How to make use of the methods from computational linguistics, especially statistical methods is the bottleneck or the key to success of this e-learning approach. On the other hand, the needs of e-learning and the abstracting of digital exhibitions from texts have very positive effect on pushing the research of computational linguistics. Significant techniques include unknown word discovery, clustering and other issues in text mining. Besides the conituous work on text mining, future research will focus on how to personalize the learning paths of learners, and enable in-time processing of user feedbacks. Investigations and evaluations will be made both on the e-learning system and the efficiency of text mining techniques over typical kinds of texts, like Chinese ancient poetry.

6 Acknowledgements

The authors would thank people in Institute of Computational Linguistics, Peking University, who gave great help for this research. We will especially thanks Miz. Feng-ju Lo, who has given us great help ever since the research starts.

References

1. Martin A. Nowak, Natalia L. Komarova, Partha Niyogi, Computational and Evolutionary Aspects of Language, *Nature*, VOL417, 6 June 2002
2. W.Rayward, M. Twidale, From Docent to Cyberdocent: education and Guidance in the Virtual Museum, *Archives and Museum Informatics* 13, 1999, p23-p53.
3. Ching-Chun Hsieh, Ying-Chun Hsieh et al, "Samples of Digital Archive in Taiwan National Digital Archive Program", 2003
4. Shun-tzu Tsai, Chun-ko Hsieh, Diversity and Aesthetic Appeal for a Virtual Reality World of Chinese Art, *proceeding of the Seventh International Conference on Virtual System and Multimedia*, 2001
5. "The Learning Power of Museums–A Vision for Museum Education" Published by Department for Culture, Media and Sport, United Kingdom, 2000
6. Louise Smith, "Building the Digital Museum: A National Resource for the Learning Age." *joint report of The National Museums Directors' Conference, Resource and mda*, UK, 10 August 2000
7. Feng-ju Lo, et al, Ancient Literature Museum: Design of an E-learning System for non-Chinese Major, *the 4th International Workshop on Computer, Multimedia and Education of Language*, Taiwan, 2000
8. Chuanzhong Li, Jingzhong Zhang, "Idea of Intelligent Knowledge Platform and a Rudimental Prototype", Research and Development on the World Science & Technology, Volume 23 Issue 6, 2001
9. Junfeng Hu, Shiwen Yu, Word meaning Similarity analysis in Chinese Ancient Poetry, *ICL Technical Report*, Peking University, 2001
10. Qiaozhu Mei, "A Digital Museum of Ancient Chinese Poetry Art: It's Design, Realization and Related Researches on Computational Linguistics", Thesis for Bachelor's Degree in Peking University, 2003.6
11. Krish Pillai, "The Fountain Model and Its Impact on Project Schedule", ACM SIGSOFT Software Engineering Notes, Volume 21 Issue 2, March 1996
12. Nikos Kladias, Tassos Pantazidis, Manolis Avagianos, A Virtual Reality Learning Environment Providing Access to Digital Museums, *1998 MultiMedia Modeling* October, 1998, p193
13. Jen-Shin Hong, Bai-Hsuen Chen, Jieh Hsiang, Tien-Yu Hsu, "Content Management for Digital Museum Exhibitions", Proceeding of JCDL 2001, pp.450, June 24-28, 2001
14. Qiaozhu Mei, A Knowledge Processing Oriented Life Cycle Study from a Digital Museum System., *The 42nd ACM Southeast Conference*, Huntsville, 2004