A Gateway to the Knowledge in Taiwan's Digital Archives

Tzu-Yen Hsu, Ting-Hua Chen, Chung-Hsi Hung, and Keh-Jiann Chen

Research Center for Information Technology Innovation, Academia Sinica,
Taipei 115, Taiwan
{ciyan,james,johan,kchen}@iis.sinica.edu.tw

Abstract. Taiwan's digital archives cover a broad range of cultural and natural assets. More than 2 million objects have been accumulated since the project was launched in 2002. As the number and diversity of digitised objects have increased rapidly, it has become increasingly difficult for people to gain a clear picture of the contents of the archives. To disseminate the abundant and diverse resources to the public, we are building a knowledge structure that consists of categorized keywords extracted from objects' metadata, and developing a function called "Tagging Tool" to facilitate fast and efficient mining of resources. For example, users who want to read an article enriched with archive collections can utilize the tool to identify archive specific keywords in the text automatically and annotate them with references to relevant resources. As a result, users can save a great deal of time on keyword searching, and contextualize various entities, such as historical events and people's names.

Keywords: keyword extraction, text annotation, digital library.

1 Introduction

The goal of the Taiwan e-Learning & Digital Archives Program[1] (TELDAP), which was launched in 2002, is to permanently preserve assets, such as natural resources, historical artifacts, intellectual property, and spatial data, in digitized forms. To help users retrieve archive resources through an integrated interface, a portal website called Union Catalog[2] (UC) was established in 2005. Its main functions are full-text searching, category browsing, and keyword browsing.

When archive resources grow rapidly and diversely, users tend to favor hot shared links, keywords, or social tags; however, we estimate that an increasing number of resources are never, or seldom, viewed. Motivated by Wikipedia, the Semantic Web, and ontology sharing, we propose the concept of knowledge building to enhance archive utilization. On the basis of knowledge construction, objects or documents can be classified into appropriate categories and their relationships can be mapped. Once loosely organized resources become tightly

[1] Taiwan e-Learning & Digital Archives Program, http://teldap.tw/en/
[2] Union Catalog, Taiwan, http://catalog.digitalarchives.tw/

M. Agosti et al. (Eds.): ECDL 2009, LNCS 5714, pp. 398–401, 2009.
© Springer-Verlag Berlin Heidelberg 2009

connected, it is possible to present users with more useful and meaningful resources and engage them in exploring other resources that are seldom viewed. We believe that through services, such as keyword detection and suggestions and recommendations about related resources based on knowledge construction, TELDAP's resources could be disseminated more widely and trigger more innovative add-on value. Initially, resource dissemination or knowledge sharing is performed by the Tagging Tool, which we describe in the following.

2 Knowledge Building

Our knowledge representation is inspired by the notion of ontology, which is frequently defined as an explicit specification of a conceptualization[1]. Basically, ontology is composed of "vocabulary for representing and communicating knowledge about some topic and a set of relationships that hold among the terms in that vocabulary[2]." For this reason, our knowledge model consists of categorized/hierarchical keywords, which are derived from given keywords and also extracted from the titles and descriptions of resource metadata. The steps of the knowledge building process are as follows.

- **Preprocessing:** Our data source consists of resource metadata, which describes resources based on Dublin Core's 15 elements[3] . Usually, data inside metadata is expressed as pure text; however, in our case some redundant information exists, for example, punctuation, HTML markups, and other information decoration. Therefore, we need to filter out such information in order to obtain complete sequences of Chinese characters.
- **Tokenization:** This step converts a sequence of Chinese characters into a group of tokens, which are considered meaningful units or lexicons, so as to facilitate further part-of-speech analysis and lexicon statistics. However, tokenization in Chinese does not have clear separation rules like the spaces and periods in English, and consecutive features make lexicon recognition difficult. For example, a four-character segment could be analyzed as a lexicon or recognized as two separate two-character lexicons. It is almost impossible to implement the Chinese tokenization process solely by pure logical computing. Hence, we developed a word segmentation system[4] to complete this task by using a rich lexicon database.
- **Keyword selection:** This step decides if a token should become a TELDAP keyword or if it should be abandoned. We have tried several algorithms to generate a better result at an acceptable cost. Currently, a token needs to satisfy three criteria to become a valid keyword. First, the token must be a noun type. Second, it must have a high frequency in TELDAP texts, but a low frequency in a general corpus. This means that it is not a general term, but it is relevant to TELDAP. Third, a semi-manual review decides whether the token becomes a valid keyword based on its original metadata.

[3] Dublin Core Metadata Element Set, `http://dublincore.org/documents/dces/`
[4] Word segmentation system, `http://ckipsvr.iis.sinica.edu.tw/`

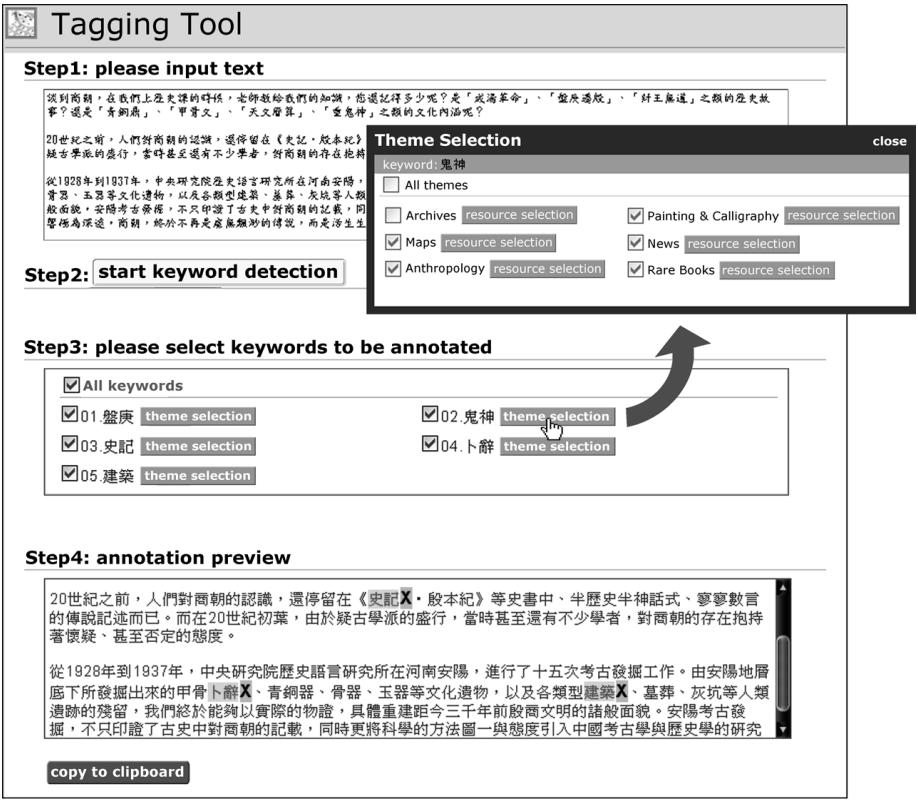

Fig. 1. The web-based Tagging Tool

There is an important task at the end of this phase. Specifically, all qualified keywords need to be added to the lexicon database to enable the subsequent tokenization process to recognize those keywords.

– **Keyword classification:** A keyword might easily relate to several themes; however, such keyword ambiguity could reduce the quality of the search results and force users to sort search lists themselves. The metadata of each resource contains at least one record of theme classification assigned by the academic institute or organization that owns the actual resource; therefore, we can categorize valid keywords into main themes based on the corresponding resources. A keyword displayed by classified resources can help users retrieve information rapidly from related categories.

3 Tagging Tool

Based on the knowledge structure, the Tagging Tool's key functions are TELDAP-specific keyword detection and selective keyword and resource annotation. The friendly step-by-step user interface is shown in Fig.1. Initially, users can input or paste any text that they want to enrich with links to TELDAP's resources,

and start the keyword detection process (Fig.1, step2), which completes text tokenization and keyword matching between text tokens and the TELDAP keyword database. Second, users can select keywords, themes, and resources relevant to the input text (Fig.1, step3) or just leave them as default settings. Finally, users can preview the annotated text and decide if each keyword is appropriate for their needs (Fig.1, step4).

The current Tagging Tool, shown in Fig.1, requires active users to decorate their articles with links to TELDAP's resources step by step. However, the user interface may not be straightforward for some users, especially those who want to get background information immediately. Therefore, it would be better if readers can have an alternative means of using Tagging Tool in a form like Google widgets or Internet browser add-ons. Obviously, the way users interact with the Tagging Tool can be developed further.

4 Future Work

The Tagging Tool is just a beginning to examine and demonstrate the potential of the knowledge structure, which still needs a more definite and transparent building process to avoid wrong calculations and incomplete thinking. After we gain more user feedback to refine keywords and the connections between keywords and resources, the TELDAP keyword repository can be shared with intended applications and institutes. It is an effective means of disseminating TELDAP's resources more widely. Moreover, a valuable keyword source linked to the rich archive collections could quite possibly inspire further innovation.

Acknowledgement

This work was supported by the National Science Council of Taiwan under Grant No. NSC98-2631-H-001-008.

References

1. Gruber, T.: A translation approach to portable ontologies. Knowledge Acquisition 5(2), 199–220 (1993)
2. Gruber, T.: What is an ontology, `http://www-ksl-svc.stanford.edu:5915/doc/frame-editor/what-is-an-ontology.html`