Revealing Modern History of Japanese Philosophy Using Natural Language Processing and Visualization

Hideki Mima, Katsuya Masuda, Susumu Ota, and Shunya Yoshimi
Center for Knowledge Structuring, University of Tokyo

Type of presentation: Poster
Keywords: Natural language processing, visualization, Japanese philosophy, thoughts, knowledge structuring
Contact email address: mima@t-adm.t.u-tokyo.ac.jp
Postal address: 7-3-1 Hongou Bunkyou-ku Tokyo 113-8656, Japan

Abstract

The purpose of this study was to reveal the modern history of Japanese philosophy using natural language processing (NLP) and visualization. Knowledge[1] has been increasing at an exponential rate with advances in science and technology in recent years resulting in massive amounts of knowledge that have been extremely difficult to process manually. Thus, it is important to utilize information technologies (IT) to support new discoveries of knowledge from large numbers of resources, such as literature. To implement the study, we have developed:

1) A corpus representing a modern history of Japanese philosophy,

2) A computational model for extracting ontology[2] from the corpus, and

3) An interactive user interface (UI) to support new discoveries of knowledge.

We chose "Shisou" (thoughts) by the Japanese publisher Iwanami Shoten for the target corpus, which is one of the most representative journals of philosophy in Japan that has an almost 90 year history from 1921 to the present-day. It is comprised of about 8,600 papers and more than 160,000 pages of textual data. The first step in this study was to develop a technology to digitize such large amounts of textual data from physical books (semi-) automatically. Because the target was too huge to digitize manually (i.e.,

[1] Although the definition of knowledge is domain-specific, our definition of knowledge here is the particles represented by ontology, which is the (hierarchical) collection and classification of (technical) terms used to recognize their semantic relevance.

[2] Although the definition of ontology is also domain-specific, our definition of ontology here is, as previously mentioned, the (hierarchical) collection and classification of (technical) terms used to recognize their semantic relevance.

by typing), a rapid, accurate and low-cost approach was required. Thus, we developed an Optical Character Reader (OCR) based (semi-) automatic book-digitizing system, in which we integrated three processes:

i) Book scanning
ii) OCR
iii) Automatic document style recognition

The input for the system were physical books and the output was a full-text corpus with meta-data, i.e. titles, authors, page numbers, and dates.

We propose a knowledge structuring (KS) system[1] to integrate NLP and the visualization-based interactive UI for the model of ontology extraction and UI. The system architecture is modular, and it integrates five components (Fig. 1): a) information (ontology) extraction, b) a corpus database, c) information retrieval, d) similarity calculations, and e) visualization.

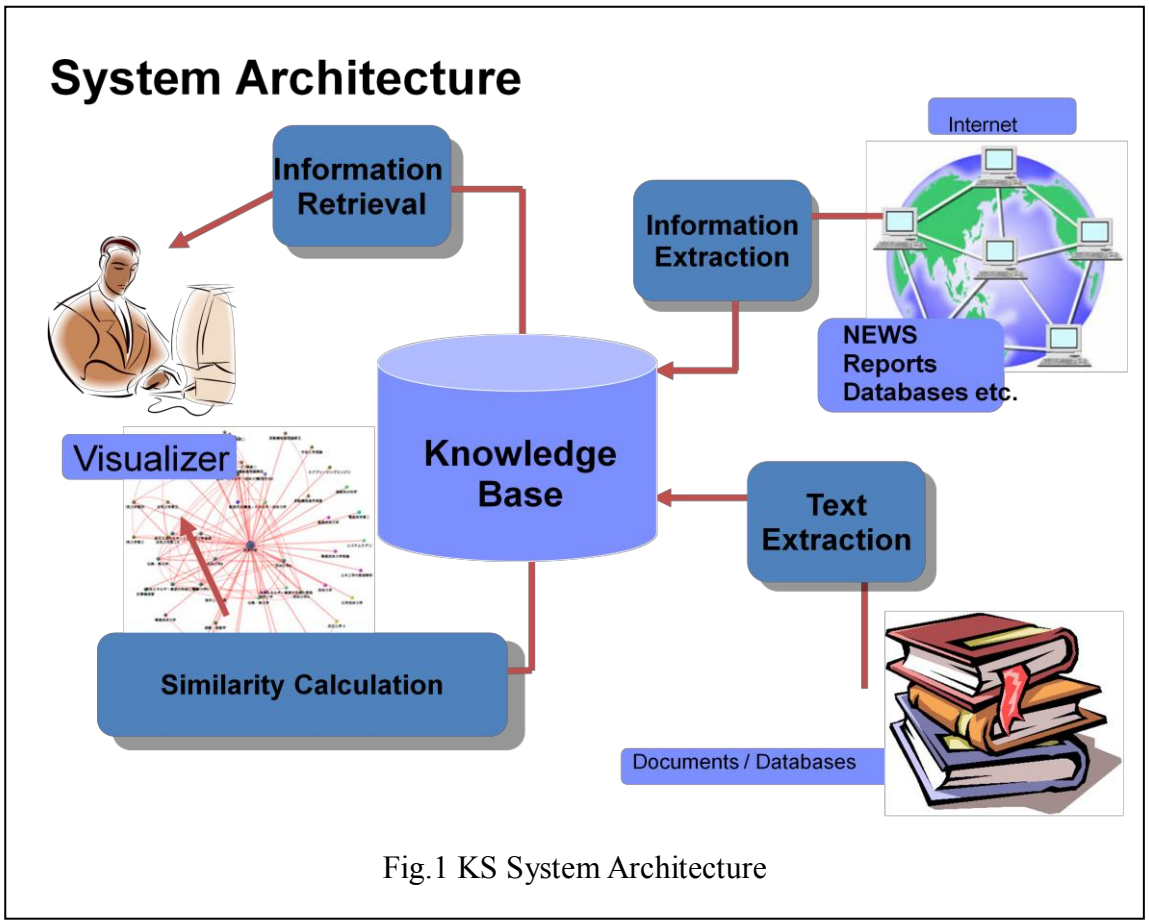

Fig.1 KS System Architecture

The main objective of the system was to facilitate knowledge acquisition from documents and generate ideas through terminology-based real-time calculations of document similarities and their visualization with an interactive UI. Fig. 2 outlines the visualization of knowledge structures for *shisou* papers relevant to the keyword "shisou (thoughts)" in the 1930s. The system constructs a graph to structure knowledge in which

the nodes (dots) reflect relevant papers with the keyword, and the links between the nodes reflect semantic similarities that are calculated based on terminological information in the papers. Additionally, the locations of all nodes are calculated and optimized when the graph is drawn. The distance between each node depends on how close they are in meaning. Cluster recognition is also carried out based on the detection of groups of papers in which every combination of papers that are included is strongly linked (i.e., their similarity exceeds a threshold). As seen in Fig. 2, several clusters are automatically recognized and category names such as “Marxism”, “socialism” and “right-wing thoughts” are also automatically assigned to clusters to facilitate an overview of thoughts discussed in these papers.

We have currently finished digitizing and creating a “Shisou” textual database of the 20 years from 1940 to 1959 and installed it in the KS system. Several experiments on text digitization were conducted to evaluate the OCR and style recognition process to improve accuracy. We obtained more than 98% accuracy in OCR, about 90% accuracy in style recognition according to the latest evaluation.

We expect to discover new knowledge on the historical flow of Japanese thinking during one of its most important eras from before World War II to the present-day by digitizing and analyzing huge amounts of historical textual data with the system.

References

[1] Mima, H. and Ananiadou, S. "An application and evaluation of the C/NC-value approach for the automatic term recognition of multi-word units in Japanese." *International Journal on Terminology*, 6 (2), pp. 175–194, 2000.

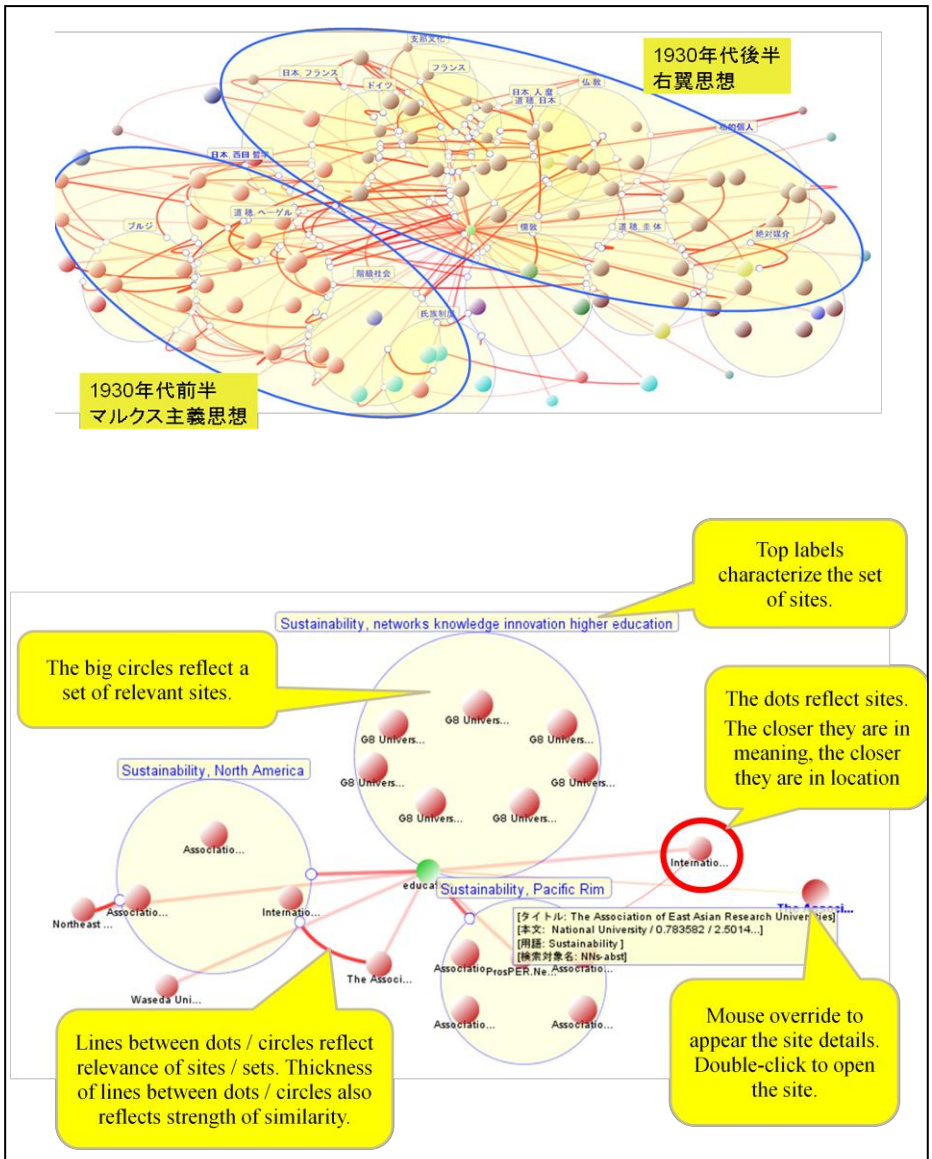

Fig.2 Visualization of knowledge structures