

Mining the Web for Generating Thematic Metadata from Textual Data

Chien-Chung Huang
Academia Sinica, Taiwan
villars@iis.sinica.edu.tw

Shui-Lung Chuang
Academia Sinica, Taiwan
slchuang@iis.sinica.edu.tw

Lee-Feng Chien
Academia Sinica, Taiwan
lfchien@iis.sinica.edu.tw

Conventional tools for automatic metadata creation mostly extract named entities or patterns from texts and annotate them with information about persons, locations, dates, and so on. However, this kind of entity type information is often too primitive for more advanced intelligent applications such as concept-based search. In this work [1], we try to generate semantically-deep metadata with limited human intervention. The main idea behind our approach is to use Web mining and categorization techniques to create thematic metadata. For example, given the text instance “Marie Curie”(primitive metadata value) and scientists of various disciplines(thematic categories), we can categorize the former into the corresponding category “Physicists”, thus creating thematic metadata for the text instance. The problem thus we face comprises of three parts: (1) how to find proper training corpus to describe the concept of the thematic categories (e.g., Physicists); (2) how to choose a proper representational model to describe the text instances (e.g., Marie Curie); and (3) how to design the training mechanism so as to ensure high categorization accuracy.

Figure 1 depicts the overall idea of the approach, which comprises of three computational modules: Feature Extraction, HCQF (Hier-Concept Query Formulation) and Text Instance Categorization. The Feature Extraction module sends the name of text instances to Web search engines, and the returned highly-ranked search-result pages are used to describe them. Also, this module works with HCQF module and send the names of the thematic categories to search engines to acquire necessary corpora. HCQF is the core of the proposed approach, aiming at formulating effective queries to be sent to search engines and organizing the retrieved corpus so that the categories can be as well-trained as possible. HCQF specifies that (1) for each category, the query has to be formulated based on the structure of the given thematic hierarchy; (2) The training corpus for the sub-structure can enrich the corpus of the upper-level categories.

Unlike those categorization techniques relying on hand-labelled corpora, HCQF is designed with the goal of acquiring the necessary corpus to describe the “concept” of the category. The reason that we use the sub-structure corpora to help train the upper-level categories is that the con-

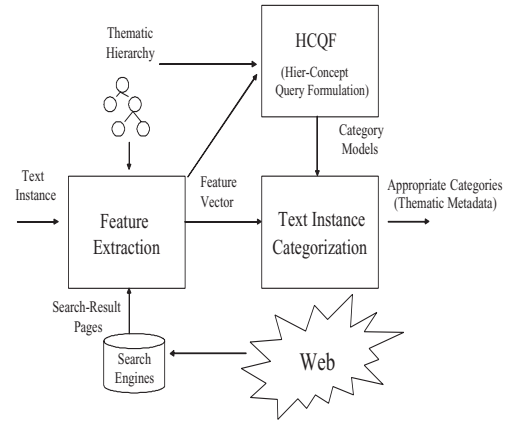


Figure 1. An abstract diagram showing the concept of the proposed approach.

cept of the lower-level categories is subsumed by the upper-level categories, therefore, the training corpora describing them can also be used to describe the upper-level category. Also, consider the case that the sub-structure is lacking, i.e., the given thematic hierarchy is a flat structure without sub-structure. We still can use some techniques to find suitable keywords and use them as pseudo sub-categories to enrich the corpus of the target categories. The category models trained by HCQF are output to Text Categorization module, which maps the input text instances into corresponding categories, generating the needed thematic metadata.

To assess the performance of our approach and explore its possible applications, we use Yahoo!’s directory as the testing bed and design a series of experiments. For example, given the 36 computer science related categories listed in Yahoo!’s directory, we categorize Computer Science terms and papers onto them; for another, we categorize the name of scientists onto a set of scientific disciplines. Overall, the accuracy achieved in these experiments is promising.

References

- [1] Full version available at <http://wkd.iis.sinica.edu.tw/publication>.