Chapter VI

Multilingual Web Content Mining: A User-Oriented Approach

Rowena Chau, Monash University, Australia
Chung-Hsing Yeh, Monash University, Australia

ABSTRACT

This chapter presents a novel user-oriented, concept-based approach to multilingual web content mining using self-organizing maps. The multilingual linguistic knowledge required for multilingual web content mining is made available by encoding all multilingual concept-term relationships using a multilingual concept space. With this linguistic knowledge base, a concept-based multilingual text classifier is developed. It reveals the conceptual content of multilingual web documents and forms concept categories of multilingual web documents on a concept-based browsing interface. To personalize multilingual web content mining, a concept-based user profile is generated from a user’s bookmark file to highlight the user’s topics of information interest on the browsing interface. As such, both explorative browsing and user-oriented, concept-focused information filtering in multilingual web are facilitated.
INTRODUCTION

The rapid expansion of the World Wide Web throughout the globe means electronically accessible information is now available in an ever-increasing number of languages. With the majority of this web data being unstructured text (Chakrabarti, 2000), web content mining technology capable of discovering useful knowledge from multilingual web documents, thus, holds the key to exploiting the vast human knowledge hidden beneath this largely untapped multilingual text. Moreover, users’ information interests differ. Knowledge useful to one user may not be useful to another. Mining the multilingual web content and delivering the discovered knowledge without considering the user’s information interest may not be effective.

To help each user discover knowledge specific to his domain of interest from the multilingual web, a user-oriented approach to multilingual web content mining is required. The user-oriented, concept-based, multilingual web content mining approach introduced in this chapter is such an approach. The objective of this approach is to facilitate personalized multilingual web content mining, which is important, especially when the user’s motive for information seeking is personalized global knowledge discovery.

BACKGROUND

Web content mining has attracted much research attention in recent years (Kosala & Blockeel, 2000). It has emerged as an area of text mining specific to web documents, focusing on analyzing and deriving meaning from textual collections on the Internet (Chang et al., 2001). Currently, web content mining technology is still limited to processing monolingual web documents.

The challenge of discovering knowledge from textual data which are significantly linguistically diverse has been well recognized by text mining research (Tan, 1999). In a monolingual environment, the conceptual content of documents can be discovered by directly detecting patterns of frequent features (i.e., terms) without precedential knowledge of the concept-term relationship. Documents containing an identical known term pattern, thus, share the same concept. However, in a multilingual environment, vocabulary mismatch among diverse languages implies that documents exhibiting a similar concept will not contain identical term patterns. This feature incompatibility problem, thus, makes the inference of conceptual contents using term pattern matching inapplicable.
11 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the product's webpage:

www.igi-global.com/chapter/multilingual-web-content-mining/24157?camid=4v1


www.igi-global.com/e-resources/library-recommendation/?id=1

Related Content

Service-Driven Computing: Challenges and Trends

www.igi-global.com/chapter/service-driven-computing/198648?camid=4v1a

Modeling Domain Ontology for Occupational Therapy Resources Using Natural Language Programming (NLP) Technology to Model Domain Ontology of Occupational Therapy Resources

www.igi-global.com/article/modeling-domain-ontology-for-occupational-therapy-resources-using-natural-language-programming-nlp-technology-to-model-domain-ontology-of-occupational-therapy-resources/109664?camid=4v1a
Ranking Tagged Resources Using Social Semantic Relevance
[www.igi-global.com/chapter/ranking-tagged-resources-using-social/75907?camid=4v1a](www.igi-global.com/chapter/ranking-tagged-resources-using-social/75907?camid=4v1a)

Performance Evaluation of Hierarchical SOAP Based Web Service in Load Balancing Cluster-Based and Non-Cluster-Based Web Server