Information Processing and Management 47 (2011) 647–666

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Multilingual document mining and navigation using self-organizing maps

Hsin-Chang Yang [a,*], Han-Wei Hsiao [a], Chung-Hong Lee [b]

[a] Department of Information Management, National University of Kaohsiung, Kaohsiung 811, Taiwan
[b] Department of Electrical Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan

ARTICLE INFO

Article history:
Received 30 April 2009
Received in revised form 16 November 2009
Accepted 1 December 2009
Available online 8 January 2010

Keywords:
Multilingual Web page navigation
Multilingual text mining
Self-organizing map
Hierarchy alignment

ABSTRACT

One major approach for information finding in the WWW is to navigate through some Web directories and browse them until the goal pages were found. However, such directories are generally constructed manually and may have disadvantages of narrow coverage and inconsistency. Besides, most of existing directories provide only monolingual hierarchies that organized Web pages in terms that a user may not be familiar with. In this work, we will propose an approach that could automatically arrange multilingual Web pages into a multilingual Web directory to break the language barriers in Web navigation. In this approach, a self-organizing map is constructed to train each set of monolingual Web pages and obtain two feature maps, which reveal the relationships among Web pages and thematic keywords, respectively, for such language. We then apply a hierarchy generation process on these maps to obtain the monolingual hierarchy for these Web pages. A hierarchy alignment method is then applied on these monolingual hierarchies to discover the associations between nodes in different hierarchies. Finally, a multilingual Web directory is constructed according to such associations. We applied the proposed approach on a set of Web pages and obtained interesting result that demonstrates the feasibility of our method in multilingual Web navigation.

1. Introduction

Nowadays the users of the World Wide Web (or for simplicity, the Web) try to access the huge amount of documents on the Web (or the Web pages) by searching with a search engine or browsing through hyperlinks existed within Web pages. For users who have no specific goal, browsing Web pages is often the preferred choice. However, many users have difficulty of getting start from a page which will eventually lead to their goals. Hence many portal sites emerge to provide such starting points. These sites often provide, in company with search facility, some sorts of navigating structure which organize Web pages into hierarchies, which are called Web directories henceforth. Users can then achieve a thematic navigation through such hierarchies. However, these hierarchies were generally constructed by human experts manually and were often lack of coverage, redundant, probably inconsistent, and hard to maintain. Besides, different sites often adopt different topic selection and categorization schemes which make them incapable of information exchange.

Another problem of existing Web directories comes from the monolingual nature in the construction process. Most of Web directories categorized only Web pages written in a specific language, such as English. Different Web directories have to be constructed for different native Web pages. Such monolingual interface may limit the spread of users who are unfamiliar with the used language. For example, a native Chinese may not intend to use a Web directory which provides only

* Corresponding author.
E-mail addresses: yanghc@nuk.edu.tw (H.-C. Yang), hanwei@nuk.edu.tw (H.-W. Hsiao), leechung@mail.ee.kuas.edu.tw (C.-H. Lee).
URL: http://www.im.nuk.edu.tw/yanghc (H.-C. Yang).

0306-4573/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2009.12.003

English categorization labels and Web pages. Thus, it will be convenient for users to have a Web directory providing multilingual category labels and categorizing multilingual Web pages.

There are two necessary steps in constructing a multilingual Web directory. The first step is to organize Web pages into hierarchies for easy browsing. Although other structures are possible for Web page navigation, hierarchies were most adopted since they have intrinsic categorization structures that higher-level categories represent superset of lower-level ones. Most users found this convenient since they could achieve their goals by exploiting the structures in a coarse-to-fine manner from the most general theme that meets their goals. Most popular portal sites constructed such hierarchies by human experts. Although these hierarchies have the advantages of precise and consistent, manual construction approach suffers from the enormous amount of time and labor to initiate and maintain those hierarchies and prevents it being applied on large datasets. Thus automatic approach should be more feasible for large datasets such as the Web.

The second step in constructing multilingual Web directories is to obtain the associations between different languages. One popular approach is to apply some machine translation schemes to translate terms in one language to another. Unfortunately, there is still no well recognized scheme to provide precise translation between two languages. A different approach is to match terms in different languages directly without a priori translation. This approach is also difficult since we need some kind of measurements to measure the semantic relatedness between them. Such semantic measurements are generally not able to be explicitly defined, even with human intervention. Thus we need a kind of automated process to discover the relationships between different languages. Such process is often called multilingual text mining (MLTM).

To construct Web directories, human intervention is unavoidable in present time. We need human effort in tasks such as selecting topics and revealing their relationships. Such need is acceptable only when the volume of Web pages is considerably small. However, the volume of Web pages under consideration is generally large enough to prevent manual construction. To expand the applicability of the directories, some kind of automatic process should involve during the construction of the directories. The degree of automation in such construction process may differ for different constructors with different needs. One may only need a friendly interface to automate the authoring process, and another one may try to automatically identify every component from the ground up. We recognize the importance of a Web directory not only as a navigation tool but also a desirable scheme for knowledge acquisition and representation. According to such recognition, we try to develop a scheme based on a proposed text mining approach to automatically construct Web directories. Our approach is opposite to the navigation task performed by an existing Web directory to obtain goal pages. We extract knowledge from a corpus of Web pages to construct a Web directory.

In this work, we will develop an automatic scheme to arrange multilingual Web pages into Web directories based on a multilingual text mining approach. We will first apply a machine learning algorithm (Yang & Lee, 2004) based on self-organizing maps (Kohonen, 1997) on a corpus of monolingual Web pages to identify their topics, discover the relations among them, and construct a Web directory. The construction process consists of two major tasks. The first is topic detection which identifies the major themes existed in a set of close related Web pages. This set of Web pages is called a category hereinafter. The second is the construction of a Web directory for these monolingual Web pages. A second text mining process is then applied on two monolingual directories to discover the associations between categories in different directories. The proposed method will provide users novel multilingual Web directories for easy browsing of Web pages in different languages.

Fig. 1 depicts an overview of architecture of the proposed method. When a set of monolingual Web pages is input, we will first cluster them using self-organizing map algorithm. Two feature maps are obtained to reveal the relationships among

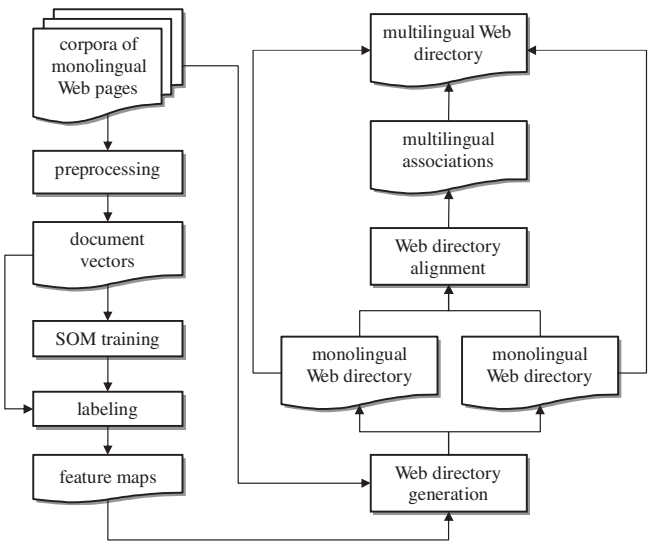

Fig. 1. The processing steps of the proposed method.