Future Generation Computer Systems 32 (2014) 222–231

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

A framework for automated construction of resource space based on background knowledge

Xu Yu[a], Li Peng[a], Zhixing Huang[a,b,*], Hai Zhuge[a,b,c,d,**]

[a] Semantic Grid Lab, School of Computer and Information Science, Southwest University, 400715, Chongqing, China
[b] Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, 100190, Beijing, China
[c] Nanjing University of Posts and Telecommunications, 210003, Nanjing, China
[d] Aston University, B4 7ET, Birmingham, UK

HIGHLIGHTS

- Our framework combines statistical topic model with human background knowledge.
- Candidate axes are generated by utilizing category hierarchy of Wikipedia.
- Three different ranking strategies are proposed for picking out final axes.
- Resources are mapped to different axes and important ones can be found out.

ARTICLE INFO

Article history:
Received 24 January 2013
Received in revised form
2 July 2013
Accepted 17 July 2013
Available online 5 August 2013

Keywords:
Resource space model
Semantic graph
Latent Dirichlet allocation
Wikipedia

ABSTRACT

Resource Space Model is a kind of data model which can effectively and flexibly manage the digital resources in cyber-physical system from multidimensional and hierarchical perspectives. This paper focuses on constructing resource space automatically. We propose a framework that organizes a set of digital resources according to different semantic dimensions combining human background knowledge in WordNet and Wikipedia. The construction process includes four steps: extracting candidate keywords, building semantic graphs, detecting semantic communities and generating resource space. An unsupervised statistical language topic model (i.e., Latent Dirichlet Allocation) is applied to extract candidate keywords of the facets. To better interpret meanings of the facets found by LDA, we map the keywords to Wikipedia concepts, calculate word relatedness using WordNet's noun synsets and construct corresponding semantic graphs. Moreover, semantic communities are identified by GN algorithm. After extracting candidate axes based on Wikipedia concept hierarchy, the final axes of resource space are sorted and picked out through three different ranking strategies. The experimental results demonstrate that the proposed framework can organize resources automatically and effectively.

1. Introduction

The physical space, socio space, mental space and cyber space will cooperate with each other to form the cyber-physical society [1,2]. In the cyber space, digital resources are valuable assets which are stored in personal computers, accumulated in organizational networks or distributed in different websites. With the increasing amount of digital resources, one pressing task is to organize them automatically based on some meaningful semantic relationships. To efficiently and effectively manage those resources, we can classify and organize the digital resources from multiple perspectives (also called as multiple facets [3] or multiple dimensions). In each facet, the resources can be further described by using a form of the concept hierarchy or taxonomy tree.

Resource Space Model (RSM) is such a data model that can effectively organize and flexibly manage resources by normalizing classification semantics [2,4,5]. It enables users to operate resources spaces in the cyber space according to the classification in their mental spaces. RSM classifies objects into categories at different granularity levels, establishes links between known objects and discovers clues between known or unknown objects which are essential for new-generation semantic data models. Furthermore, some models based on RSM have been proposed to manage various resources in different spaces of cyber-physical society or modeled the mental space combined with the self-organized semantic link network [2,6]. Although the model is very powerful and

[*] Corresponding author at: Semantic Grid Lab, School of Computer and Information Science, Southwest University, 400715, Chongqing, China. Tel.: +86 023 68253262.

[**] Correspondence to: School of Computer, Nanjing University of Posts and Telecommunications, 210003, Nanjing, China and Computer Science Group, School of Engineering and Applied Science, Aston University, B4 7ET, Birmingham, UK.

E-mail addresses: huangzx@swu.edu.cn (Z. Huang), zhuge@ict.ac.cn (H. Zhuge).

0167-739X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.future.2013.07.017

useful, how to automatically construct the resource space remains a challenge.

Statistical topic model such as Latent Dirichlet Allocation (LDA) [7,8] can map large amount of texts into multiple topics robustly. Each topic can be regarded as a semantic facet. A list of words is presented under each topic and these words come from the documents in the corpus. Weights of the generated topics represent their relatedness to each facet. However, the topics generated by LDA are often hard to understand and the facet number of the model should be predefined. The model lacks a mechanism for incorporating human knowledge. On the other side, human knowledge recorded in WordNet and Wikipedia has fruitful concepts and rich category information. Especially, researchers have shown that the Wikipedia entries are reliable identifiers for conceptual entities and can be used for annotating Web resources and knowledge assets [9–11].

RSM can manage various types of resources, for simplicity, in this paper we aim at the text resource. We proposed a framework of automatically constructing multidimensional resource space, combining Latent Dirichlet Allocation and human background knowledge. Topics from corpus are extracted by utilizing the statistical topic model. Keywords of documents generated on the basis of those topics construct semantic graphs which their semantic relatedness is computed from WordNet. By means of a community detection algorithm used in complex network analysis, semantic communities in the graph are discovered. Axis information is generated by using our algorithms after mapping the keywords to appropriate Wikipedia entities. At last, the axes generated are used for building the resource space. With the proposed method, the results of topic model are better explained and resource space is constructed automatically.

The rest of the paper is organized as follows: in Section 2, we briefly discuss the related work about RSM, topic model and human background knowledge. In Section 3, we describe the framework and the detailed processes of resource space construction. In Section 4, the experimental results are presented. Finally, we come to a conclusion and point out the future work in Section 5.

2. Preliminaries and related work

The most basic method or the natural solution way for organizing various resources in the cyber-physical society is classification. Usual classification is single dimension. There are two major reasons to use multidimensional classifications: (1) increasing or reducing dimension is an effective way to specialize or generalize knowledge in mind; (2) human need to explore large-scale resources sets from multiple dimensions [2,12].

Resource Space Model (RSM) [2,4,5] is proposed to manage the resources based on the normalization of the classification. A resource space is a multi-dimensional classification semantic space. Each dimension of resource space specifies a type of classification method. The normal forms and operations based on orthogonal classification semantics are also proposed to manage deterministic classification semantics. The basic semantic elements of the Resource Space Model are resource space, axis, coordinate, point and resource [12–14]. Resource Space Model, OWL and databases can be integrated to form a powerful semantic platform that enables different semantic models to enhance each other [15]. More detailed discussions about RSM are in the book [5].

Statistical topic models [8,16–18] provide a general data-driven framework for automated discovery of high-level knowledge from large collections of text documents. It allows each document to be represented by multiple topics. The basic concept underlying topic modeling is that each document is composed of a probability distribution over topics, where each topic represents a probability distribution over words. Although topic models can potentially discover a broad range of themes in a dataset, the interpretability of the learned topics is not ideal [19]. Combining human-defined background knowledge with the topic model is a solution for this problem.

As we known, most concepts of human knowledge can be found in WordNet or Wikipedia. WordNet is a large lexical database of English developed at the Cognitive Science Laboratory of Princeton University, similar to a traditional thesaurus but with a richer structure. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), with each expressing a distinct concept. Synsets are interlinked by means of conceptual–semantic and lexical relations. It is one of the most important and widely used lexical resources for natural language processing tasks [20], such as automatic text classification [21], information retrieval [22], text summarization [23,24] and word sense disambiguation [25,26]. The relationships in WordNet are hypernyms and hyponyms, part-meronyms, antonyms and entails. In our method, noun network in WordNet is used to extract semantic relatedness for word pairs.

Another widely used human knowledge repository is Wikipedia. It is a multilingual, web-based encyclopedia representing the outcome of a continuous collaborative effort of a large number of volunteer contributors. Wikipedia has become one of the largest online repositories of encyclopedic knowledge, with millions of entries in a large number of languages. The basic entry in Wikipedia is an entity (or page), which defines a concept or an event and consists of a hypertext document with hyperlinks to other pages within or outside Wikipedia website. Wikipedia are widely used for keyword extraction [27], topic identification [10,28], information extension [29,30] and topic evolution [31]. Wikipedia is also used for improvement on managing and exploring resource effectively. In the work of [32], they use Wikipedia to help understand a user's query intent. An unsupervised technique for automated extraction of facts useful for browsing text databases is presented in [33].

From the above works, the fruitful structure of links and categories in Wikipedia is a reliable resource for text processing and resource management. However, these methods cannot build a multidimensional, hierarchical space for managing resources and their functionality can be further improved. As mentioned before, statistical topic model can easily discover the multiple topics from large collection of text documents. A natural combining statistical topic model and human background knowledge becomes an approach of resource space construction.

3. The framework of resource space construction

This section presents the framework of automated resource space construction. We begin with a brief introduction of the Latent Dirichlet Allocation.

3.1. Latent Dirichlet allocation

The Latent Dirichlet Allocation (LDA) model is a typical statistical language topic model introduced by Blei et al. [7] for extracting a set of topics that describe a collection of documents. The basic idea of LDA is that each document is composed of a probability distribution over topics, where each topic represents a probability distribution over words. The model can be represented as a three hierarchical Bayesian model. Given a corpus consisting of M documents, the generative process for the document $d \in \{1, \dots, M\}$ is defined as follows:

1. select a distribution over topics $\theta|\alpha \sim \mathit{Dirichlet}(\alpha)$
2. for each word $n \in 1, 2, \dots, N$:
   (a) select a topic $z_n|\theta \sim \mathit{Discrete}(\theta)$