Chapter 6
The Clustering of Large Scale E-Learning Resources

Fei Wu
Zhejiang University, China

Wenhua Wang
Zhejiang University, China

Hanwang Zhang
Zhejiang University, China

Yueting Zhuang
Zhejiang University, China

ABSTRACT

E-learning resources increase vastly with the pervasion of the Internet. Thus, the retrieval of e-learning resources becomes more and more important. This chapter introduces an approach to retrieve e-learning resources from large-scale dataset. The basic idea behind that method is, the authors cluster the whole resources into topics first, and only search from those clusters which are the most tightly relevant to the query. To make the clustering feasible to large-scale dataset, the authors adapt affinity propagation in MapReduce framework and therefore the so called parallel affinity propagation is proposed. The proposed approach could improve the retrieval of e-learning resources by understanding users’ underlying intentions.

INTRODUCTION AND RELATED WORK

With the pervasion of the Internet, e-learning is more and more popular, which provides a brand new way for people to learn without attending face-to-face class. With e-learning, the student and the teacher use online technology to interact, which profits from a combination of techniques including computer networks, multimedia, content portals, digital libraries, search engines, etc.. The worldwide e-learning industry is estimated to be worth over 38 billion euros according to conservative estimates. With the prevalence of e-learning, the amount of learning resources also grows exponentially, which makes it not feasible to access them only by clicking links. Thereby, an effective
mechanism is needed to locate the resources, with which people could find the e-learning materials they want with facility. To accurately locate the e-learning materials a user is seeking for, system has to guess the user’s underlying intentions from the text typed in, rather than merely return the results from literally matching, particularly when the user is not familiar with the terminologies of the field which he/she is trying to learning. Thus, leveraging the data mining technology to locate the resources semantically related to the querying text becomes meaningful. Nowadays, the e-learning resources comprise texts, images, videos, audios and materials in other modalities, however, text materials are the best choice to be analyzed and understood, in that texts account for the biggest part and only the text resources reflect the information most directly. Besides, taking efficiency and the expensive mining cost into account, it’s reasonable to focus only on the text materials and to neglect materials in other modalities. Therefore, e-learning materials and e-learning resources in this chapter mostly mean e-learning text documents.

In data mining technology, clustering is to partition a data set into subsets, so that the data in each subset share some common trait. See (Jain et al.,1999; Xu et al., 2005) for details. Therefore, clustering is an effective method to discover clues when little is known about the data. Besides, e-learning resources are intrinsically appropriate to be clustered, in that fields of materials concerning likely overlap fields of others, and materials concerning similar fields probably use the same words, particularly the same terminologies. For example, two physics books will likely use words such as ‘energy’, ‘force’, ‘mass’, and ‘charge’ repeatedly, which consequently strengthens the correlation between the books. So we adopt the clustering method to preprocess the e-learning resources to mine the correlations among the materials.

Traditionally, measures of text similarity have been used for a long time in applications in natural language processing and related areas (Corley & Mihalcea, 2005). One of the earliest applications of text similarity is perhaps the vector model in information retrieval, where the document most relevant to a user’s query is determined by ranking documents in a collection in reversed order of their similarity to the given query (Salton & Lest, 1968). In the vector space model, a document is represented by a vector indexed by the terms of the corpus, so two documents that use semantically related but distinct words will therefore show no similarity (Kandola et al., 2002). Many methods were proposed to explicitly or implicitly discover the similarity between different terms, such as (Landauer, Foltz & Laham, 1998; Corley & Mihalcea, 2005; Kandola et al., 2002), as well as other WordNet based methods (Budanitsky & Hirst, 2001). Those information retrieval technologies work well with toy data. However, because of memory limitations of stand-alone computers and the expensive computation cost of matrix operations such as singular value decomposition, the situation becomes intractable when they are applied to large-scale data, taking an unendurably long time or even being interrupted due to out of memory. To tackle the unavoidable problem, usually two categories of methods are used. The first kind is to use matrix factorization and to merge the results using a mathematic method, such as (Pauca et al., 2004; Xu et al., 2003). The other kind is to parallelize the learning procedure and to compute each part on distributed computers in parallel, such as (Graf et al., 2005; Collobert et al., 2004). In this paper, we take advantage of MapReduce framework (Dean & Ghemawat. 2004), which helps to run a specially designed program on distributed computers. After the original large dataset is clustered, we then construct semantic spaces on the resultant relatively small-scale datasets to carry out semantic retrieval on e-learning materials.

In this paper, we proposed a method to manage the e-learning resources and to retrieve the semantically related materials according to users’ underlying intentions. To tackle the problem in-
Related Content

Cross-Institutional Blended Learning in Teacher Education: A Case Study
www.igi-global.com/chapter/cross-institutional-blended-learning-teacher/62143?camid=4v1a

Pop Lyrics and Mobile Language Learning: Prospects and Challenges
www.igi-global.com/article/pop-lyrics-and-mobile-language-learning/188411?camid=4v1a

Unearthing Invisible Buildings: Device Focus and Device Sharing in a Collaborative Mobile Learning Activity
www.igi-global.com/article/unearthing-invisible-buildings/60136?camid=4v1a

Sustaining Mobile Learning and its Institutions
www.igi-global.com/chapter/sustaining-mobile-learning-its-institutions/62131?camid=4v1a