A Graph-Based Biomedical Literature Clustering Approach Utilizing Term’s Global and Local Importance Information

Xiaodan Zhang, Drexel University, USA
Xiaohua Hu, Drexel University, USA and Jiangxi University of Finance and Economics, China
Jiali Xia, Jiangxi University of Finance and Economics, China
Xiaohua Zhou, Drexel University, USA
Palakorn Achananuparp, Drexel University, USA

ABSTRACT

In this article, we present a graph-based knowledge representation for biomedical digital library literature clustering. An efficient clustering method is developed to identify the ontology-enriched k-highest density term subgraphs that capture the core semantic relationship information about each document cluster. The distance between each document and the k term graph clusters is calculated. A document is then assigned to the closest term cluster. The extensive experimental results on two PubMed document sets (Disease10 and OHSUMED23) show that our approach is comparable to spherical k-means. The contributions of our approach are the following: (1) we provide two corpus-level graph representations to improve document clustering, a term co-occurrence graph and an abstract-title graph; (2) we develop an efficient and effective document clustering algorithm by identifying k distinguishable class-specific core term subgraphs using terms’ global and local importance information; and (3) the identified term clusters give a meaningful explanation for the document clustering results.

Keywords: global ranking; graph-based document clustering; pagerank

INTRODUCTION

PubMed (www.ncbi.nlm.nih.gov/PubMed) is a service of the U.S. National Library of Medicine that includes more than 17 million citations from MEDLINE and other life science journals for biomedical articles back to the 1950s. As a large digital archive, PubMed is widely used by biomedical domain scientists. For such a large
dataset, a nonspecific query can easily return more than 100,000 search results. Automatically clustering the search results and providing topic-specific core terms for each cluster can help domain experts concentrate on the relevant articles and disregard the nonrelevant ones. Therefore, we present an efficient and effective clustering method for this purpose.

Proper document representation is very important for document clustering. Conventional document clustering tends to represent a document as a bag of words and then cluster documents using vector cosine similarity or other similarity measures. One main limitation of these document clustering methods is that they are usually black box clustering, invisible to end users and lacking the ability to interpret clustering results.

Steinbach, et al. (2000) argue that each document class has a “core” vocabulary of words, and the remaining “general” words may have similar distributions in different classes. Thus, two documents from different classes can share many general words (i.e., stop words) and thus be treated similarly in terms of vector cosine similarity. The ideal situation is that we use only the distinguishable terms to cluster documents in a much lower dimensionality to improve accuracy and efficiency. However, to discover these distinguishable core terms is not a trivial problem when we have no knowledge about the document class in advance. If these class-specific core terms are successfully identified, they can be used not only for document clustering but also for the interpretation of document clustering results.

Motivated by this discussion, we develop an approach to represent a collection of documents as a term co-occurrence graph or an abstract-title term graph. Global ranking methods such as PageRank (Page, et al., 1998) and HITS (Kleinberg, 1999) are applied to the graph to detect the class-specific core terms. Then an efficient algorithm is designed to grow some semantic-related and graphically connected core term clusters from these top-ranked terms. A document is then assigned to its closest term cluster.

PageRank and HITS-based algorithms have been very popular for improving Web document retrieval (a directed hyperlink graph) and text summarization such as LexRank (Erkan & Radev, 2004) (an undirected sentence similarity graph). Take PageRank as an example. If a Web page has more in-links from the Web pages with a higher PageRank score, the Web page gets a higher PageRank score. However, when such algorithms are used to identify the globally important terms within an undirected term co-occurrence graph for a document collection, they face the problem of term noises. The top-ranked terms can be either class-specific core terms or class-unspecific general terms, because those general terms have very dense connections with other terms. Moreover, an undirected graph contains no reference information (i.e., who cites whom). Ideally, we should keep the links between the class-specific core terms and remove the links that are from the class-unspecific general terms. Therefore, we can remove the “noise” of the general terms and let the ranking of the class-unspecific general terms go down and the class-specific core terms go up. Then we simply pick up a set of the top-ranked terms to initialize the “core” of a cluster.

In this study, two types of term graph representations are presented to discount the effects of the general terms and to strengthen the impacts of the class-specific core terms. One approach is to construct an undirected corpus-level term co-occurrence graph, where each term is a vertex and each edge represents the frequency of co-occurrence. Statistical filtering and semantic filtering are applied sequentially to remove the “noisy” co-occurrence edges. The edges between the two terms are thus filtered out if they are not statistically dependent on each other and semantically related. The other approach is to construct a directed corpus-level abstract-title term graph with the abstract terms pointing to the title terms. Title terms are usually
Fusion Cubes: Towards Self-Service Business Intelligence
www.igi-global.com/article/fusion-cubes-towards-self-service/78287?camid=4v1a