Available online at www.sciencedirect.com

ScienceDirect

Decision Support Systems 45 (2008) 413 – 428

Decision Support Systems

www.elsevier.com/locate/dss

A collaborative filtering-based approach to personalized document clustering

Chih-Ping Wei [a,*], Chin-Sheng Yang [b], Han-Wei Hsiao [c]

[a] *Institute of Technology Management, College of Technology Management, National Tsing Hua University, Hsinchu, Taiwan, ROC*
[b] *Department of Information Management, College of Management, National Sun Yat-sen University, Kaohsiung, Taiwan, ROC*
[c] *Department of Information Management, College of Economics and Management, National University of Kaohsiung, Kaohsiung, Taiwan, ROC*

Available online 18 May 2007

Abstract

Document clustering is an intentional act that reflects individual preferences with regard to the semantic coherency and relevant categorization of documents. Hence, effective document clustering must consider individual preferences and needs to support personalization in document categorization. Most existing document-clustering techniques, generally anchoring in pure content-based analysis, generate a single set of clusters for all individuals without tailoring to individuals' preferences and thus are unable to support personalization. The partial-clustering-based personalized document-clustering approach, incorporating a target individual's partial clustering into the document-clustering process, has been proposed to facilitate personalized document clustering. However, given a collection of documents to be clustered, the individual might have categorized only a small subset of the collection into his or her personal folders. In this case, the small partial clustering would degrade the effectiveness of the existing personalized document-clustering approach for this particular individual. In response, we extend this approach and propose the collaborative-filtering-based personalized document-clustering (CFC) technique that expands the size of an individual's partial clustering by considering those of other users with similar categorization preferences. Our empirical evaluation results suggest that when given a small-sized partial clustering established by an individual, the proposed CFC technique generally achieves better clustering effectiveness for the individual than does the partial-clustering-based personalized document-clustering technique.

Keywords: Document clustering; Personalization; Collaborative filtering; Hierarchical agglomerative clustering (HAC); Text mining

1. Introduction

With the advances and proliferation of the Internet, available information sources have grown tremendously in number and sheer volume, primarily as a result of global connectivity and ease of publishing. To manage this ever-increasing volume of documents, organizations and individuals typically organize documents into categories (or category hierarchies) to facilitate their document management and support subsequent document retrieval and access. In turn, the development of an effective document-clustering mechanism has become essential for the efficient and effective document management of organizations and individuals.

Document clustering entails the automatic organization of a large document collection into distinct groups of similar documents that reflect general themes hidden within the corpus [21,22,33]. However, according to the context theory of classification, document-clustering

[*] Corresponding author. Tel.: +886 3 574 2219; fax: +886 3 574 5310.
E-mail address: cwei@mis.nsysu.edu.tw (C.-P. Wei).

0167-9236/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2007.05.008

behaviors of individuals not only involve the attributes (including contents) of documents but also depend on who is performing the task and in what context [3,11,26,28]. Therefore, document clustering is an intentional act that reflects individuals' unique preferences with regard to the semantic coherency or relevant categorization of documents [40]. For example, given a set of research articles related to "data mining," researchers engaged in developing novel data mining techniques may prefer organizing the articles according to underlying techniques (e.g., classification analysis, clustering analysis, association rules, sequential patterns). In contrast, researchers who are applying data mining techniques to solve business questions generally would prefer categories based on application domains (e.g., banking, manufacturing, health care, telecommunications). Furthermore, even when they use the same categorization scheme, different researchers may vary in the granularity of the categories they employ. Some researchers, for example, might use a single category for all articles related to classification analysis, whereas others may employ a set of increasingly specific categories (e.g., decision tree induction, neural network, Bayes classification) for the same collection. Effective document clustering therefore must consider individual preferences and needs in order to support personalized document categorization [17,46].

Traditional document-clustering techniques generally have been anchored in pure content-based analysis. As a consequence, most existing document-clustering techniques are not tailored to individuals' preferences and therefore are unable to facilitate personalization. In other words, the categorization scheme exhibited in such non-personalized clusters may not conform to a user's expectation and perception. However, a user's document search typically is guided by his or her categorization scheme [14,38]. Thus, when searching documents with a one-for-all categorization scheme, a user generally undertakes a semantic internalization process [34] to comprehend the target categorization scheme or experiences a coadaptation process that adjusts his or her own categorization scheme and, at the same time, reinterprets and adapts the target categorization scheme to his or her needs [31,32]. The semantic internalization and coadaptation processes unnecessarily increase the user's cognitive load. Consequently, he or she likely spends more time or has difficulty locating documents of interest because of the discrepancy between the one-for-all categorization scheme and his or her expectation [46]. The described inefficiency or ineffectiveness of document retrieval and access may adversely affect the efficiency, quality, and satisfaction of decision making that requires references to various documents relevant to the target decision context.

Several prior studies have sought to incorporate noncontent information into the document-clustering process to improve clustering effectiveness. Because the noncontent often is associated with individual users, these document-clustering techniques represent possible approaches to achieve personalized document clustering. For example, the adaptive [50] and user-oriented [13,35] document-clustering techniques take the documents' relevance to user queries into account when clustering a collection of documents. To support personalized document clustering for a target user, all queries should be made by that user. However, both techniques assume that all documents in the collection D to be clustered must appear in at least one set of relevant documents for a query. If a document is considered irrelevant to all queries, its similarity with other documents in D cannot be estimated. As the size of D expands, the number of documents considered irrelevant to all queries increases, creating a serious problem to both techniques. Furthermore, document relevance to queries often is associated with query contexts. Therefore, heterogeneity in query contexts may constrain the effectiveness of these techniques for facilitating personalized document clustering. Kim and Lee [21] propose a semi-supervised document-clustering technique, a hybrid of content- and noncontent-based document-clustering approaches, that considers not only document content similarity but also users' perceptions of document similarity using a relevance feedback mechanism. Relevance feedback offers a means to achieve personalized document clustering, but such real-time feedback is prohibitively tedious and unpractical in terms of both time and cognitive efforts. Moreover, relevance feedback is impractical in many document-clustering applications (e.g., cluster-based browsing by digital libraries and search engines), thereby making it an unfeasible solution for personalized document clustering.

To address the limitations inherent to these techniques, Wei et al. [46] propose a partial-clustering-based personalized document-clustering approach that incorporates a target user's partial clustering into the document-clustering process. Let the set of documents to be clustered be D. A partial clustering denotes a user's categorization of a subset of documents in D. In some application environments, the partial clustering of a user is readily available. For example, some digital libraries and online information providers offer personal bookshelves (e.g., "my bookshelf," "my favorites," "my eNews") so that users can organize documents of interest into their personal folders. When a set of documents is retrieved for clustering for a specific user, some documents in the set