Fast and Effective Cluster-based Information Retrieval using Frequent Closed Itemsets

Yousef Djenouri\textsuperscript{a}, Asma Belhadi\textsuperscript{b}, Philippe Fournier-Viger\textsuperscript{c}, Jerry Chun-Wei Lin\textsuperscript{d}

\textsuperscript{a}IMADA, Southern Denmark University, Odense, Denmark
\textsuperscript{b}RIMA Lab, USTHB, Algiers, Algeria
\textsuperscript{c}School of Humanities and Social Sciences, Harbin Institute of Technology (Shenzhen), China
\textsuperscript{d}School of Computer Science, Harbin Institute of Technology (Shenzhen), China

Abstract

Document Information retrieval consists of finding the documents in a collection of documents that are the most relevant to a user query. Information retrieval techniques are widely-used by organizations to facilitate the search for information. However, applying traditional information retrieval techniques is time consuming for large document collections. Recently, cluster-based information retrieval approaches have been developed. Although these approaches are often much faster than traditional approaches for processing large document collections, the quality of the documents retrieved by cluster-based approaches is often less than that of traditional approaches. To address this drawback of cluster-based approaches, and improve the performance of information retrieval both in terms of runtime and quality of retrieved documents, this paper proposes a new cluster-based information retrieval approach named ICIR (Intelligent Cluster-based Information Retrieval). The proposed approach combines k-means clustering with frequent closed itemset mining to extract clusters of documents and find frequent terms in each cluster. Patterns discovered in each cluster are then used to select the most relevant document clusters to answer each user query. Four alternative heuristics are proposed to select the most relevant clusters, and two alternative heuristics for choosing documents in the selected clusters. Thus, eight versions of the proposed approach are obtained. To validate the proposed approach, extensive experiments have been carried out on well-known document collections. Results show that the designed approach outperforms traditional and cluster-based information retrieval approaches both in terms of execution time and quality of the returned documents.

Keywords: Document information retrieval, Data mining, Big collections,
دریافت فوری
متن کامل مقاله

امکان دانلود نسخه تمام متن مقالات انگلیسی
امکان دانلود نسخه ترجمه شده مقالات
پذیرش سفارش ترجمه تخصصی
امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
امکان دانلود رایگان ۲ صفحه اول هر مقاله
امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
دانلود فوری مقاله پس از پرداخت آنلاین
پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات