Document clustering by dynamic hierarchical algorithm based on fuzzy set type-ii from frequent item set

Saiful Bahri Musa, author

Abstrak

One of ways to facilitate process of information retrieval is by performing clustering toward collection of the existing documents. The existing text documents are often unstructured. The forms are varied and their groupings are ambiguous. This cases cause difficulty on information retrieval process. More-over, every second new documents emerge and need to be clustered. Generally, static document clus-tering method performs clustering of document after whole documents are collected. However, per-forming re-clustering toward whole documents when new document arrives causes inefficient clus-tering process. In this paper, we proposed a new method for document clustering with dynamic hierar-chy algorithm based on fuzzy set type-II from frequent item set. To achieve the goals, there are three main phases, namely: determination of keyterm, the extraction of candidates clusters and cluster hierar-chical construction. Based on the experiment, it resulted the value of F-measure 0.40 for Newsgroup, 0.62 for Classic and 0.38 for Reuters. Meanwhile, time of computation when addition of new document is lower than to the previous static method. The result shows that this method is suitable to produce so-lution of clustering with hierarchy in dynamical environment effectively and efficiently. This method also gives accurate clustering result.

Salah satu cara untuk mempermudah proses information retieval adalah dengan melakukan peng-klasteran terhadap koleksi dokumen yang ada. Dokumen teks yang ada seringkali tidak terstruktur, formatnya bervariasi, dan pengelompokannya ambigu. Hal ini menimbulkan kesulitan dalam proses information retrieval. Selain itu, setiap detik dokumen baru bartambah dan perlu untuk dikelompokkan. Pada umumnya, metode pengklasteran dokumen statis melakukan pengklasteran dokumen setelah kese-luruhan dokumen terkumpul. Namun, melakukan pengklasteran ulang terhadap keseluruhan dokumen ketika dokumen baru tiba mengakibatkan proses pengklasteran menjadi tidak efisien. Penelitian ini mengusulkan metode baru untuk pengklasteran dokumen dengan algoritma hierarki dinamis berbasis fuzzy set type-II dari frequent itemset. Untuk mencapai tujuan tersebut, terdapat 3 tahapan utama yang akan dilakukan, yaitu; ekstraksi keyterm, ekstraksi kandidat klaster dan pembangunan hirarki klaster. Berdasarkan eksperimen yang telah dilakukan diperoleh nilai F-Measure 0,40 untuk Newsgroup, 0,62 untuk Classic, dan 0,38 untuk Reuters. Sedangkan waktu komputasi pada saat penambahan dokumen dapat direduksi dibanding dengan metode statis sebelumnya. Hasil percobaan terhadap beberapa dataset koleksi dokumen menunjukkan bahwa metode ini tidak hanya sesuai untuk menghasilkan solusi peng-klasteran secara hirarki dalam lingkungan yang dinamis secara efektif dan efisien, tetapi juga membe-rikan hasil pengklasteran yang akurat.