การศึกษาเปรียบเทียบการกระจายแมทริกซ์สำหรับการค้นคืนสารสนเทศข้อความภาษาไทย

The Comparison of Matrix decomposition for Thai Text-based Information Retrieval

ทอง บุญยศ, ชลยืน หงส์ไพศาลวิวัฒน์

บทคัดย่อ

ระบบการค้นคืนสารสนเทศที่ใช้คำในข้อคำถาม (Query) ไปจับคู่กับคำหลัก (Keyword) ในเอกสารอาจไม่มีประสิทธิภาพเพียงพอ เพราะคำหลักไม่สามารถเป็นตัวแทนเอกสารได้อย่างครบถ้วน ผลของการค้นคืนจึงจะได้เอกสารที่มีคำหลักตรงกับคำในข้อคำถามของผู้ใช้เท่านั้น แต่มิใช่เป็นการค้นคืนในเชิงความหมายที่มีอยู่ในเอกสาร อีกวิธีการหนึ่งคือการใช้เทคนิคเชิงแมทริกซ์ที่มีการเปลี่ยนรูปแบบการเก็บคำสำคัญของเอกสารในรูปแบบแมทริกซ์ของคำสำคัญสำหรับเอกสาร และใช้เทคนิคการกระจายแมทริกซ์เพื่อให้ได้กลุ่มของคำหลักที่มีความสัมพันธ์กับเอกสาร งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาเปรียบเทียบกระจายแมทริกซ์สองแบบในการค้นคืนสารสนเทศข้อความภาษาไทย อันได้แก่ การกระจายแมทริกซ์แบบเซมิดีสครีต (Semidiscrete Matrix Decomposition) และการแยกค่าแบบเดี่ยว (Singular Value Decomposition) สำหรับในเอกสาร 4 สาขาวิชา ผลการทดลองพบว่า การกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตใช้เนื้อที่ในการการกระจายแมทริกซ์ (Decomposition storage) น้อยกว่าการกระจายแมทริกซ์ด้วยวิธีแยกค่าแบบเดี่ยวอย่างมีนัยสำคัญ แต่ใช้เวลาในการกระจายแมทริกซ์ (Decomposition time) มากกว่า ส่วนผลการสืบค้นสารสนเทศอันได้แก่ ความแม่นยำ (Precision) และ การเรียกกลับ (Recall) ของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตและด้วยวิธีแยกค่าแบบเดี่ยวจะขึ้นอยู่กับการปรับแต่งองค์ประกอบต่าง ๆ ในกระบวนการออกแบบและพัฒนาค้นคืนสารสนเทศ

Abstract

Information retrieval systems that use literal term matching between user query and keywords of a document can produce inefficient outputs, because the term matching does not extract knowledge or conceptual information from the document; therefore, the results of documents are retrieved based upon the matched keywords, but may be irrelevant to the concepts of documents. Another approach is the matrix oriented technique in which Latent Semantic Indexing (LSI) is applied. LSI can extract

conceptual information underlying in documents. This technique changes the keyword representation of documents into a term-by-document matrix and uses matrix decomposition techniques to refine the matrix for the better representation of term-document relationship. The main purpose of this thesis is to compare between two matrix decompositions: singular value decomposition (SVD) and semidiscrete matrix decomposition (SDD), of matrix of Thai text-based information in four domains. The results show that the SDD occupied less storage but took more time for decomposition than did the SVD significantly. Precision and recall of both matrix decomposition methods depend on various setup values during processes of design and implementation and the objectives of information retrieval systems.