

การจัดเรียงลำดับคำดัชนีถือว่าเป็นขั้นตอนสำคัญสำหรับการค้นคืนเอกสาร ในวิทยานิพนธ์นี้เราได้วิจัยเพื่อค้นหาวิธีที่ดีในการจัดเรียงลำดับคำดัชนีในเอกสาร XML [1] ตามแบบเวกเตอร์โมเดล [2] ซึ่งตามวิธีเวกเตอร์มาตรฐานจะจัดเรียงลำดับคำดัชนีตามความถี่ที่ปรากฏในเอกสารเพียงอย่างเดียว [3] งานวิจัยนี้ได้ปรับปรุงวิธีการจัดเรียงลำดับคำดัชนีแบบเวกเตอร์สำหรับเอกสาร XML โดยมีการเพิ่มค่าน้ำหนักที่แตกต่างกันให้กับคำที่ปรากฏในแท็ก XML ที่ต่างกัน โดยแท็กที่มีความสำคัญมากกว่าจะกำหนดให้มีค่าน้ำหนักมากกว่า นอกจากนี้ยังใช้จำนวนของคำในแท็กมาเข้าร่วมคำนวณเพื่อเป็นค่าน้ำหนักด้วย โดยการคิดค่าน้ำหนัก 2 วิธีดังกล่าวรวมกับวิธีการคำนวณกับความถี่ของคำแล้ว ทำให้เราแยกแยะพิจารณาการคำนวณค่าน้ำหนักคำดัชนีได้เป็น 4 วิธี คือ 1. วิธีรวมค่าน้ำหนักแท็กที่พบ 2. วิธีรวมค่าน้ำหนักแท็กกับความถี่ที่พบ 3. วิธีรวมค่าน้ำหนักแท็กที่พบและค่าจำนวนคำในแท็ก 4. วิธีรวมค่าน้ำหนักแท็กกับความถี่ที่พบและค่าจำนวนคำในแท็ก จากผลการทดลองหาวิธีการกำหนดค่าน้ำหนักแท็กสำหรับแต่ละวิธีคำนวณค่าน้ำหนักคำดัชนี โดยใช้ค่า Precision เฉลี่ยและค่า R-Precision เฉลี่ยเป็นเครื่องมือวัดประสิทธิภาพ พบว่าการให้ค่าน้ำหนักแท็กโดยใช้เงินดิกอัลกอริทึมเป็นการให้ค่าน้ำหนักแท็กที่ดีที่สุด และผลการทดลองหาวิธีคำนวณค่าน้ำหนักคำดัชนีที่ดีที่สุดสำหรับ 4 วิธีข้างต้น โดยใช้ค่า Precision เฉลี่ยและค่า R-Precision เฉลี่ยเป็นเครื่องมือวัดประสิทธิภาพ และใช้การให้ค่าน้ำหนักแท็กโดยผู้ใช้เป็นค่าน้ำหนักแท็กที่นำมาทดลองพบว่าวิธีที่ 4 มีประสิทธิภาพดีกว่าวิธีการคำนวณอื่น ๆ รองลงมาคือวิธีที่ 3 , วิธีที่ 1 และวิธีที่ 2 ตามลำดับ

Ranking index terms is an important step for efficient information retrieval. In this thesis we have investigated in order to find a good method for ranking index terms based on a vector model for XML documents [1]. According to this model [2], indexes from documents are ranked according to only the frequency of their appearance in the documents [3]. In this thesis we have improved the conventional vector model by adding different weights to words appearing in different tags of XML documents. The more important the tag, the more weight is assigned to it. In addition, the number of words appearing (a word count) in a tag is also taken into account for determining weights for the index terms. By considering the 3 factors, i.e. tag weights, the word count in a tag, and word frequency (according to the conventional vector model), we have identified 4 ways (methods) for ranking index terms in XML documents: (1) using tag weights only (2) using tag weights and word frequency (3) using tag weights and tags' word counts (4) using tag weights, word frequency, and tags' word counts altogether. In the experimentation we use average Precision and average R-Precision to compare each method with vary tag weight. The result has shown that defining tag weight by genetic algorithm is the best method for every method and for each method the fourth method is the best method follow by the third, the first and the second method.