EFFICIENT DOCUMENT CLUSTERING USING SUFFIX ARRAY

KOVIT KARAWATREEDECH

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE
(COMPUTER SCIENCE)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2005

ISBN 974-04-5658-8

การจัดกลุ่มเอกสารอย่างมีประสิทธิภาพ (EFFICIENT DOCUMENT CLUSTERING USING SUFFIX ARRAY)

โกวิท คารวะตรีเดช 4237680 SCCS/M

วท.ม. (วิทยาการคอมพิวเตอร์)

คณะกรรมการควบคุมวิทยานิพนธ์ : ดำรัส วงศ์สว่าง, Ph.D., สุดสงวน งามสุริยโรจน์, Ph.D., ชมทิพ พรพนมชัย, Ph.D.

บทคัดย่อ

ปัจจุบันการใช้เครื่องมือค้นหาข้อมูลบน Internet โดยใช้ Search Engine ได้รับความนิยมเป็นอย่างมาก แต่ทว่าผลลัพธ์ที่ได้รับจาก Search engine บางครั้งก็ไม่ตรงกับที่ผู้ใช้ต้องการหรือบางครั้งข้อมูลที่ต้องการค้นหาอาจจะอยู่ใน Page หลัง ๆ ไม่ได้อยู่ใน Page แรก ๆ ของเอกสาร ทำให้ผู้ใช้ต้องเสียเวลาในการเข้าถึงข้อมูลที่ต้องการค้นหา จากปัญหาที่เกิดขึ้น โดยการหาวิธีที่จะทำการจัด กลุ่มของเอกสารที่มีเนื้อหาเหมือนกันหรือใกล้เคียงกันจัดให้อยู่ในกลุ่มเดียวกัน Suffix Tree Clustering เป็นเทคนิคที่ดีเป็นที่นิยมใช้ กันในปัจจุบันและมีประสิทธิภาพ โดยวิธีการนี้ใช้ Suffix Tree Algorithm ในการหา String Matching แต่เนื่องจากวิธีนี้ใช้ Memory ค่อนข้างสูงในการทำงานรวมไปถึงเป็นวิธี ที่ค่อนข้างซับซ้อนในการ Implement ซึ่งอาจจะส่งผลให้มีการทำงานผิดพลาดได้ ใน Thesis นี้จึงพยายามที่แก้ปัญหาข้างต้นโดยการใช้เทคนิคของ Suffix Array แทนที่ Suffix Tree

Suffix Tree Algorithm เป็นเทคนิคที่ใช้หา String Matching เช่นเดียวกับเทคนิคของ Suffix Tree clustering จากการศึกษาค้นคว้าของผู้จัดทำพบว่า ข้อดีหลักๆของ Suffix Array Algorithm ก็คือ ง่ายต่อการพัฒนาและยังต้องการขนาดของหน่วยความจำที่น้อยกว่าวิธี Suffix Tree clustering นอกเหนือจากนั้นทางผู้จัดทำยังได้มีการเปรียบเทียบเวลาที่ใช้การประมวลผลของทั้งสองวิธี พบว่า Suffix Tree Clustering นั้นประมวลผลได้เร็วกว่าในกรณีที่ข้อมูลมีขนาดเล็ก แต่จะช้ากว่าในกรณีที่ข้อมูลมีขนาดใหญ่ สาเหตุหลักก็คือ Suffix Tree Clustering ต้องการขนาดของหน่วยความจำที่มากกว่า Suffix Array Algorithm ในการเก็บข้อมูลโครงสร้าง ดังนั้นเมื่อขนาดของข้อมูลใหญ่ขึ้น ขนาดของ Tree ก็จะขยายขึ้นอย่างรวดเร็วกว่า Array ซึ่งก็จะส่งผลให้ใช้เวลามากขึ้นในการรักษาสภาพโครงสร้าง เมื่อโครงสร้างของ Tree มีการใช้เนื้อที่มากกว่าความจุของหน่วยความจำ เนื่องจากเหตุผลดังกล่าว Suffix Array Algorithm จึงเหมาะกับข้อมูลที่มีขนาดใหญ่ หรือ ทำงานบนเครื่องที่มีขนาดของหน่วยความจำที่จำกัด

182 หน้า. ISBN 974-04-5658-8

EFFICIENT DOCUMENT CLUSTERING USING SUFFIX ARRAY

KOVIT KARAWATREEDECH 4237680 SCCS/M

M.Sc. (COMPUTER SCIENCE)

THESIS ADVISORS: DAMRAS WONGSAWANG, Ph.D., SUDSANGUAN NGAMSURIYAROJ, Ph.D., CHOMTIP PORNPANOMCHAI, Ph.D.

ABSTRACT

Nowadays the Search Engine is popular among users in searching the information on the Internet. However, the results are sometimes unmatched with users' need or placed in the latter pages of the document, and the searching wastes their time to arrive at. Such problems have been solved by an idea of clustering that is grouping documents having the same or similar content together. A Suffix Tree Clustering (STC), one of the most popular clustering techniques currently in use, is an efficient technique, which employs the Suffix Tree Algorithm in performing a string matching. The shortfall of the technique is that it requires an extensive memory to execute and due to the complexity of the implementation, some errors may occur. This thesis solves these technical problems by using Suffix Array instead of Suffix Tree.

Similar to the Suffix Tree Clustering technique, the Suffix Array Clustering technique employs a Suffix Array Algorithm in doing a string matching. From our intensive study and program development for the Suffix Array Algorithm and Suffix Tree Algorithm, we found that the major advantages of the Suffix Array Algorithm over the Suffix Tree Algorithm were that the Suffix Array Algorithm is easier to implement and generally requires less memory. In addition, we compared the speed of execution of the two techniques and found that the Suffix Tree Clustering is faster when applied to a small document collection, but slower than, the Suffix Array Clustering when applied to a large document collection. The rationale is that the STC requires more memory space than SAC in maintaining its structure. Thus in the case of collection grows larger, the size of tree grows up faster then that of array. This results in more time taken in maintaining the structure when the tree structure goes beyond the capacity of memory. Hence, SAC is suitable for a large document collection or suitable to run on the computer system with very limited memory resource.

KEY WORDS: DOCUMENT CLUSTERING / SUFFIX TREE / SUFFIX ARRAY

182 P. ISBN 974-04-5658-8