

**METASEARCH ENGINE USING FUZZY LOGIC AND  
CONTENT-BASED CLUSTERING APPROACHES**

**WERASAK YUOKOOLBODEE**

**A THESIS SUBMITTED IN PARTIAL FULFILLMENT  
OF THE REQUIREMENTS FOR  
THE DEGREE OF MASTER OF SCIENCE  
(COMPUTER SCIENCE)  
FACULTY OF GRADUATE STUDIES  
MAHIDOL UNIVERSITY  
2004**

**ISBN 974-04-5096-2  
COPYRIGHT OF MAHIDOL UNIVERSITY**

**METASEARCH ENGINE USING FUZZY LOGIC AND CONTENT-BASED CLUSTERING APPROACHES**

WERASAK YUOKOOLBODEE 4337420 SCCS/M

M.Sc.(COMPUTER SCIENCE)

THESIS ADVISORS : CHOMTIP PORNPANOMCHAI Ph.D., DAMRAS  
WONGSAWANG Ph.D.**ABSTRACT**

Due to the enormous data on the internet it is very difficult for the web users to find the information they want. Search engines have been created to solve the problem by matching the query with the entry data in the database and retrieving the information for web users. However, the amount of data on the internet is still increasing rapidly. Gathering all available data on the internet and keeping it up to date is very hard for a single search engine. Therefore, web developers have developed a system that can send a query to several search engines at one time and integrate the results into one list. The system is called a metasearch engine. Its features enhance time saving, increase searching coverage and help users to find relevant information more quickly than using several single search engines.

Normally, in searching data, not all terms are equally important in a query. Thus, it is important to allow users to indicate the relative importance of various terms by weighting them. It is noteworthy that weighting may have an impact on the entire set of retrieved document. This thesis proposes the use of a fuzzy logic ranking. The fuzzy logic ranking will retrieve the information followed by terms that web users can weight by themselves. Furthermore, the ranking mechanism works well when web users can formulate a well-defined query for searches. However, a lot of users formulate very short queries and are unfamiliar with the topics they are looking for, therefore, most results satisfy their needs. In solving such a problem, a content-based clustering (CBC) technique is considered very helpful as it will divide the results into several groups, which may be matched to the areas of the users interest.

CBC aims to yield the clusters that can fulfill a precision of results. In addition, a prototype of a metasearch engine using fuzzy logic and content-based clustering was developed as proof of the proposed concept. Although the system takes several minutes to retrieve all of the results pages, the experiments show that this system outperforms other systems.

**KEY WORDS : METASEARCH ENGINE / FUZZY LOGIC /CLUSTERING/  
INFORMATION RETRIEVAL/**

197 pp. ISBN 974-04-5096-2

เครื่องมือค้นหาแบบเมตาเสิร์ช โดยใช้ตรรกศาสตร์คลุมเครือและการจัดกลุ่มข้อมูลด้วยเนื้อหาของข้อมูล (METASEARCH ENGINE USING FUZZY LOGIC AND CONTENT-BASED CLUSTERING APPROACHES)

วีรศักดิ์ เข่ากุลบดี 4337420 SCCS/M

วท.ม. (วิทยาการคอมพิวเตอร์)

คณะกรรมการควบคุมวิทยานิพนธ์ : ชมทิพ พรพนมชัย, Ph.D., คำรัส วงศ์สว่าง, Ph.D.

#### บทคัดย่อ

เนื่องจากจำนวนข้อมูลที่มีมากบนอินเทอร์เน็ตมันจึงเป็นเรื่องยากสำหรับผู้ใช้ในการค้นหาข้อมูลที่ต้องการ Search Engine จึงถูกสร้างขึ้นมาเพื่อแก้ปัญหาโดยทำเปรียบเทียบระหว่างคำที่ต้องการค้นหา กับข้อมูลทั้งหมดที่เก็บในฐานข้อมูลและส่งผลที่ได้กลับไปยังผู้ใช้ อย่างไรก็ตามจากจำนวนข้อมูลบนอินเทอร์เน็ตที่เพิ่มขึ้นอย่างรวดเร็วส่งผลให้การเก็บรวบรวมข้อมูลบนอินเทอร์เน็ตและปรับปรุงให้ทันสมัยอยู่เสมอเป็นเรื่องยากเมื่อใช้ Search Engine เพียงตัวเดียว ดังนั้นนักพัฒนาเว็บจึงพัฒนาระบบที่สามารถส่งคำที่ต้องการค้นหาไปยังหลายๆ Search Engine เพียงครั้งเดียวและรวบรวมผลที่ได้ทั้งหมดเป็นรูปแบบเดียวกัน ระบบนี้เรียกว่า Metasearch Engine ซึ่งจุดเด่นของระบบนี้คือลดเวลาในการรับส่งข้อมูล, ครอบคลุมการค้นหาข้อมูล และช่วยผู้ใช้ในการค้นหาข้อมูลที่เกี่ยวข้องได้ดีกว่าการใช้ Search Engine หลายๆตัว

โดยทั่วไปในการค้นหาข้อมูล คำทุกคำที่ใช้ในการค้นหานั้นจะมีความสำคัญไม่เท่าเทียมกัน ดังนั้นการอนุญาตให้ผู้ใช้กำหนดน้ำหนักของคำที่ต้องการหาจึงเป็นเรื่องสำคัญ เพราะน้ำหนักที่กำหนดจะมีผลกระทบกับผลลัพธ์ทั้งหมดที่จะได้ ในวิทยานิพนธ์นี้นำเสนอการจัดลำดับข้อมูลโดยใช้ตรรกศาสตร์คลุมเครือ การจัดลำดับข้อมูลจะทำการเรียกค้นข้อมูลตามน้ำหนักของแต่ละคำที่ผู้ใช้กำหนดด้วยตนเอง นอกจากนี้กระบวนการการจัดลำดับของข้อมูลจะทำได้ก็ต่อเมื่อผู้ใช้กำหนดคำที่ต้องการค้นหาได้ดี อย่างไรก็ตามผู้ใช้ส่วนใหญ่จะใช้คำสั้นๆในการค้นหาและก็ไม่คุ้นเคยกับหัวข้อที่ต้องการค้นหาทำให้ผลลัพธ์ส่วนใหญ่ที่ได้ไม่ตรงกับความต้องการ ในการแก้ไขปัญหาวิธีการจัดกลุ่มข้อมูลด้วยเนื้อหาของข้อมูลจะช่วยให้ได้มาก โดยจะทำการจัดแบ่งข้อมูลออกเป็นหลายๆกลุ่ม ซึ่งทำให้ตรงกับความสนใจของผู้ใช้

การจัดกลุ่มข้อมูลด้วยเนื้อหาของข้อมูลมีจุดมุ่งหมายเพื่อให้ได้ผลลัพธ์ที่ถูกต้อง นอกจากนี้ยังได้สร้างต้นแบบเครื่องมือค้นหาแบบเมตาเสิร์ช โดยใช้ตรรกศาสตร์คลุมเครือและการจัดกลุ่มข้อมูลด้วยเนื้อหาของข้อมูลเพื่อพิสูจน์แนวความคิดของวิทยานิพนธ์นี้ด้วย แม้ว่าระบบนี้จะใช้เวลาหลายนาที่ในการเรียกค้นข้อมูลทั้งหมด แต่จากผลการทดลองแสดงให้เห็นว่าระบบนี้ทำงานได้ดี