วิทยานิพนธ์นี้เสนอการศึกษาเปรียบเทียบประสิทธิภาพของระบบการค้นคืนเอกสารเทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงมุม และวิธีการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียนด้วยการประยุกต์ใช้ทฤษฎีการจัดกลุ่มข้อมูลแบบ K-mean Clustering กำหนดเงื่อนไข หรือกรอบความคล้ายคลึงในการเลือกเอกสารที่เป็นคำตอบ ถ้าเอกสารใดที่มีระยะห่างกับข้อสอบถามภายใต้กรอบความคล้ายคลึงที่กำหนดจะถูกค้นคืนออกมาแสดงต่อผู้ใช้ โดยได้ทดสอบกับชุดเอกสารนิตยสารไทม์ จำนวน 425 เอกสาร และข้อสอบถามจำนวน 83 ข้อสอบถาม โดยเปรียบเทียบประสิทธิภาพของระบบการค้นคืนเอกสารทั้ง 2 รูปแบบข้างต้น ด้วยค่าความแม่นยำ, ค่าความระลึก และค่าเฉลี่ยฮาร์โมนิค

จากผลการทดลองสรุปได้ว่า ระบบการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงมุม มีค่าประสิทธิภาพทั้ง 3 ค่ามากกว่าระบบการค้นคืนเอกสารด้วยวิธีการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียน ผู้วิจัยตั้งข้อสังเกตว่าวิธีการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียนอาจจะไม่เหมาะสมสำหรับนำมาใช้ในกระบวนการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ เมื่อทดสอบด้วยชุดเอกสารนิตยสารไทม์ เนื่องจากเป็นชุดเอกสารที่มีความหลายหลายของคำสูง

ผู้วิจัยจึงได้ศึกษาว่าการประยุกต์ใช้เทคนิคการจัดกลุ่มข้อมูลแบบ K-mean Clustering บนระยะห่างเชิงมุมมากำหนดเงื่อนไขในการเลือกเอกสารที่เป็นคำตอบ จะสามารถเพิ่มประสิทธิภาพของระบบการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงมุมได้หรือไม่ ผลการทดลองแสดงให้เห็นว่าเมื่อเปรียบเทียบกับระบบการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงมุม ประสิทธิภาพของระบบค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงมุมที่ใช้เทคนิคการจัดกลุ่มข้อมูลสามารถทำให้ค่าประสิทธิภาพความแม่นยำและค่าเฉลี่ยฮาร์โมนิคดีขึ้น แต่ค่าประสิทธิภาพความระลึกต่ำลง

The thesis presents a comparison study of the efficiency between the vector space model information retrieval system using cosine angle technique and the one using Euclidean distance technique together with K-means clustering where K-means clustering is used to guide the threshold for retrieving answer documents. The experiments were conducted on the TIME Magazine collection which consists of 425 documents and 83 queries. The performance of the two information retrieval systems is compared through the use of Precision, Recall and Harmonic mean measurement.

The experimental results show that the performance of the information retrieval system using cosine angle technique is significantly better than those using Euclidean distance technique in all three measurements. It was observed that the Euclidean distance technique may be unsuitable for comparing the similarity in the TIME Magazine collection where the variation in words is extremely high.

Thus, the exploratory study was conducted to further investigate whether the use of the cosine angle technique together with K-mean clustering can improve the efficiency of the traditional cosine angle information retrieval system or not. The results show that the information retrieval system using the cosine angle together with K-mean clustering has higher Precision and Harmonic mean than those without K-mean clustering technique, but has lower Recall.