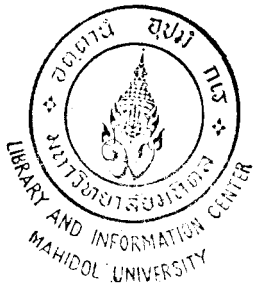

DOCUMENT ALIGNMENT WITH THESAURUS-LIKE DICTIONARY WORDLIST

PRACHYA YODPRASIT

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE
(COMPUTER SCIENCE)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2000

ISBN 974-664-488-2

3737107 SCCS/M : MAJOR : COMPUTER SCIENCE : M.Sc. (COMPUTER SCIENCE)
KEY WORDS : DOCUMENT / RETRIEVE / ALIGNMENT / RELEVANT / ENGLISH
PRACHYA YODPRASIT : DOCUMENT ALIGNMENT WITH THESAURUS-LIKE DICTIONARY WORDLIST. THESIS ADVISOR : DAMRAS WONGSAWANG, Ph.D., SUKANYA PHONGSUPHAP, Ph.D. 64p. ISBN 974-664-488-2

Document Alignment is one of the most useful searching tools in information retrieval system. Part or whole of a document can be used as an input text enquiry key to perform searching for relevant documents in the collection instead of user specified key words. Neither manually extracting significant keywords from a document nor looking for other keywords having the same semantics is needed to form a query. Document clustering is one of the most important applications of such a kind of searching. The main problem of document alignment is the automatic significant keyword extraction. Many approaches and methods have been currently proposed to get as many keywords as possible and as much relevancy as possible. However, the efficiency of the retrieval in terms of precision and recall is still not good enough for some special types of documents. In this research we proposed the keywords extracting and filtering scheme, called DAEDW (Document Alignment with English Dictionary Wordlist), for document alignment by using a thesaurus-like dictionary. After stopwords were filtered out, they were put into two groups; dictionary words and non-dictionary words. The non-dictionary words are usually appeared to be some specific names which had a higher significance of keywords. We also took advantage of the thesaurus-like dictionary to get synonymous keywords. A collection of documents consisted of news and articles which were simulated to test with the proposed alignment scheme. We found that DAEDW could increase the efficiency of retrieval in both precision and recall. Furthermore, the document ranking output was also improved. This thesis presented DAEDW in details including analysis and the computer implementation. Experimental results are discussed and future developments are also suggested.

3737107 SCCS/M : สาขาวิชา : วิทยาการคอมพิวเตอร์ : วท.ม. (วิทยาการคอมพิวเตอร์)

ปรัชญา ยอดประสิทธิ์ : การค้นหาเอกสารที่มีเนื้อความคล้ายกันโดยใช้ตารางคำอภิธาน (DOCUMENT ALIGNMENT WITH THESAURUS-LIKE DICTIONARY WORDLIST). คณะกรรมการควบคุมวิทยานิพนธ์ : ดำรัส วงศ์สว่าง, Ph.D. , สุกัญญา พงษ์สุภาพ, Ph.D. 64 หน้า, ISBN 974-664-488-2

Document Alignment เป็นเครื่องมือในการค้นหาข้อมูล อยู่ในแขนงหนึ่งของวิชา Information Retrieval โดยการนำเอกสาร(Document) ที่ผู้ใช้สนใจ ซึ่งอาจจะใช้แค่ส่วนหนึ่งหรือตัวเอกสารทั้งหมดนำมาใช้เป็นโจทย์ (Keyword) ในการค้นหา และค้นคืนเอาเอกสารที่น่าจะเกี่ยวข้อง มาแสดงให้ผู้ใช้เลือกดู ซึ่งการใช้ตัวเอกสารที่สนใจเป็น Keyword ในการค้นหาเอกสารที่ต้องการเช่นนี้ ทำให้ผู้ใช้ไม่ต้องกำหนด Keyword หรือค้นหา Keyword จากเอกสารด้วยตนเอง อีกทั้งยังไม่ต้องป้อน Keyword ที่พ้องความหมายในการสร้างคำสั่งค้นหา Document Alignment มีส่วนเกี่ยวข้องกับ Document Clustering ซึ่งเป็นการจัดหมวดหมู่ของเอกสารเพื่อใช้ในการจัดเก็บและค้นหาข้อมูล ปัญหาสำคัญที่พบร่วมกันคือการค้นหาและเลือกเอา Keyword ที่มีนัยยะสำคัญต่อเอกสารออกมาโดยอัตโนมัติอย่างไร ซึ่งปัจจุบันมีการนำเสนอแนวทางและวิธีการต่าง ๆ ที่จะเลือกเอา Keyword ที่มีนัยยะสำคัญออกมาจากเอกสารหลายวิธี แต่อย่างไรก็ตามเมื่อมีการวัดประสิทธิผลที่ได้โดยใช้ Precision/Recall มาเป็นเครื่องมือวัดก็ยังได้ผลไม่ดีเพียงพอ สำหรับเอกสารบางประเภท และในงานวิจัยฉบับนี้ได้นำเสนอวิธีการนำ Thesaurus-Like Dictionary เรียกว่า DAEDW (Document Alignment with English Dictionary Wordlist) เข้ามาช่วยในการเลือก Keyword ที่มีนัยยะสำคัญออกมาจากเอกสาร หลังจากได้มีการกลั่นกรองเอาคำหยุด (Stopword) ออกไปแล้ว คำที่เหลือก็จะถูกแบ่งออกเป็นสองกลุ่ม กลุ่มแรกคือกลุ่มคำที่อยู่ในแฟ้มข้อมูล Thesaurus-Like Dictionary และกลุ่มที่สองคือคำที่ไม่ปรากฏอยู่แฟ้มข้อมูล ซึ่งมักจะเป็นชื่อเฉพาะและมักจะเป็น Keyword ที่มีนัยยะสำคัญสูงในเอกสาร นอกจากนั้นแฟ้มข้อมูล Thesaurus-Like Dictionary ยังสามารถเก็บรวบคำศัพท์ที่มีความหมายเดียวกันได้ และจากการที่ทดสอบด้วยเอกสารตัวอย่างเช่นข่าวและบทความ พบว่า DAEDW สามารถเพิ่มประสิทธิผลในการค้นคืนเอกสารที่เกี่ยวข้อง เมื่อวัดด้วย Precision/Recall ได้ดี ทำให้ผลที่ได้คือเอกสารที่ได้มีความเกี่ยวข้องกับเอกสารที่ผู้ใช้ใช้เป็นโจทย์มากขึ้น วิทยานิพนธ์ฉบับนี้ได้กล่าวถึงรายละเอียดของ DAEDW รวมทั้งการวิเคราะห์การทำให้เป็นผลด้วยคอมพิวเตอร์ ผลการทดลอง การอภิปรายผล รวมทั้งแนวทางในการพัฒนาต่อไป