บทที่ 4
การทดสอบโปรแกรม

การทดสอบโปรแกรมจัดกลุ่มบทคัดย่อทางวิชาการด้วยอัลกอริทึมเคมีเดียนส์ ได้เสนอแนวทางการทดสอบโปรแกรมเพื่อดูความถูกต้อง ผลจากการทดสอบ และสรุปผลการทดสอบ ดังมีรายละเอียดดังต่อไปนี้

4.1 แผนการทดสอบโปรแกรม

แนวทางในการทดสอบโปรแกรม เพื่อดูความถูกต้องของการจัดกลุ่ม โดยทดลองรันโปรแกรมกับบทคัดย่อที่มีจำนวนหลากหลาย เช่น 100 บทคัดย่อ 500 บทคัดย่อ และ 1,000 บทคัดย่อ แต่จากผลการจัดกลุ่มพบว่าจำนวนบทคัดย่อ ไม่ได้มีผลกระทบกับประสิทธิภาพของโปรแกรม ดังนั้นเพื่อให้ง่ายต่อการวัดผลได้ชัดเจน จึงขอนำเสนอการจัดกลุ่มบทคัดย่อโดยใช้ชุดข้อมูลทดสอบเป็นแฟ้มข้อมูลเอกสารบทคัดย่อทางวิชาการที่เกี่ยวกับฐานข้อมูลจำนวนทั้งหมด 100 บทคัดย่อ ใช้แฟ้มข้อมูลคำสำคัญที่เกี่ยวกับฐานข้อมูล โดยมีคำสำคัญจำนวนทั้งหมด 80 คำ และระบุจำนวนกลุ่มที่ต้องการแบ่งเป็น 10 กลุ่ม ซึ่งผลการทดสอบจะอธิบายเป็นส่วน ๆ ที่สัมพันธ์กัน ดังรายละเอียดต่อไปนี้

โปรแกรมจัดกลุ่มบทคัดย่อทางวิชาการด้วยอัลกอริทึมเคมีเดียนส์ ได้ทำการทดสอบด้วยแฟ้มข้อมูลคำสำคัญ ซึ่งประกอบด้วยคำสำคัญ ดังตารางที่ 4.1

ตาราง 4.1 คำสำคัญที่ใช้ในการทดสอบโปรแกรม

ลำดับ	คำสำคัญ	ลำดับ	คำสำคัญ	ลำดับ	คำสำคัญ
1	access	28	index	55	redundant
2	algorithm	29	insertion	56	relation
3	ambiguity	30	integrate	57	reliability
4	anomalies	31	integration	58	retrieval
5	architecture	32	integrity	59	retrieve
6	audio	33	interface	60	reusable
7	availability	34	internet	61	reuse
8	classification	35	metadata	62	robust
9	classifier	36	meta-data	63	scalability

ตาราง 4.1 (ต่อ) คำสำคัญที่ใช้ในการทดสอบโปรแกรม

ลำดับ	คำสำคัญ	ลำดับ	คำสำคัญ	ลำดับ	คำสำคัญ
10	classify	37	method	64	scalable
11	cluster	38	methodology	65	schema
12	consistency	39	mining	66	search
13	constraint	40	model	67	secure
14	correctness	41	multidatabase	68	security
15	database management system	42	multimedia	69	technique
16	dbms	43	multi-media	70	test
17	deletion	44	object-orientation	71	tool
18	design	45	object-oriented	72	traditional
19	distribute	46	on-line	73	transparency
20	document database	47	optimization	74	unambiguity
21	dynamic	48	optimizing	75	usability
22	efficiency	49	parallel	76	validation
23	fault	50	performance	77	verification
24	image	51	queries	78	video
25	implement	52	query	79	warehouse
26	improve	53	real-time	80	web
27	inconsistent	54	recovery		

ข้อมูลบทคัดย่อทางวิชาการที่เป็นแฟ้มข้อมูลเอกสารที่จะใช้ในการแบ่งกลุ่มนั้น จะต้องถูกแปลงให้อยู่ในรูปของจุดข้อมูล เพื่อจะนำมาใช้กับอัลกอริทึมเคมีเดียนส์ โดยจะทำการนับจำนวนคำสำคัญในแต่ละบทคัดย่อตามคำสำคัญที่ได้กำหนดไว้ในแฟ้มข้อมูลคำสำคัญ ถ้าคำในบทคัดย่อตรงกับคำสำคัญใดก็ทำการนับจำนวนแล้วเก็บไว้ในตารางคำสำคัญ ซึ่งจะขอแสดงเนื้อหาของบทคัดย่อที่ใช้ทดสอบโปรแกรมบางส่วน ดังตารางที่ 4.2 โดยที่ตัวเข้ม หมายถึงคำสำคัญ

ตาราง 4.2 ตัวอย่างเนื้อหาของบทคัดย่อ

ลำดับ	ชื่อบทคัดย่อ	เนื้อหา
1	Data mining problems in medicine	The principle of any retrospective on patient data-based investigation is **search**ing the patients by problem or sign, but not by name. With a proper problem-encoded archival database, the data **mining** process would be easy. One would only need to input the request and obtain the proper data in a short time. Medical archives are frequently based on paper records only, with the patient name as the entry key. To find the proper record in such an archive, a detection strategy is needed. The process continues with collecting the usually enormous amount of papers, finding the appropriate records within them, and finally encoding and arranging them in a table. The whole process can be separated into patients, paper and data **mining**. Because of their slowness, these phases can be the most time-consuming part of a medical data-based investigation. The author describes his data **mining** experience.
2	Query folding	**Query** folding refers to the activity of determining if and how a **query** can be answered using a given set of resources, which might be materialized views, cached results of previous **queries**, or **queries** answerable by other databases. We investigate **query** folding in the context where **queries** and resources are conjunctive **queries**. We develop an exponential time **algorithm** that finds all complete or partial foldings, and a polynomial time **algorithm** for the subclass of acyclic conjunctive **queries**. Our results can be applied to **query optimization** in centralized databases, to **query** processing in **distribute**d databases, and to **query** answering in federated databases.
3	Dynamic query re-optimization	Very long running **queries** in database systems are not uncommon in non **traditional** application domains such as **image** processing or data warehousing analysis. **Query optimization**, therefore, is important. However, estimates of the **query** characteristics before **query** execution are usually inaccurate. Further, system configuration and resource **availability** may change during long evaluation period. As a result, **queries** are often evaluated with sub-optimal plan configurations. To remedy this situation, we have **design**ed a novel approach to re-optimize suboptimal **query** plan configurations on-the-fly with Conquest, an extensible and **distribute**d **query** processing system. A **dynamic** optimizer considers reconfiguration cost as well as execution cost in determining the best **query** plan configuration.

เมื่อทำการนับคำสำคัญในตัวอย่างบทคัดย่อทางวิชาการข้างต้น จะได้ตารางคำสำคัญ โดยหนึ่งบทคัดย่อจะเทียบได้กับหนึ่งจุดข้อมูลในอัลกอริทึมเคมีเดียนส์ ดังตารางที่ 4.3 โดยในที่นี้จะขอแสดงตัวอย่างจำนวนคำสำคัญเพียงบางส่วนเท่านั้น

ตาราง 4.3 ตัวอย่างตารางคำสำคัญ

คำสำคัญ / บทคัดย่อ	search	mining	query	image	distribute	dynamic	algorithm
1	1	3	0	0	0	0	0
2	0	0	6	0	1	0	2
3	0	0	6	1	1	1	0

จากนั้นนำจุดข้อมูลทั้งหมดที่ได้จากตารางคำสำคัญมาทำตามอัลกอริทึมเคมีเดียนส์ เพื่อจัดกลุ่มข้อมูล ซึ่งผลลัพธ์จะได้กลุ่มของบทคัดย่อทางวิชาการที่มีความคล้ายคลึงกันตามจำนวนกลุ่มที่ต้องการ

4.2 ผลการทดสอบโปรแกรม

หลังจากที่นำแฟ้มข้อมูลบทคัดย่อทางวิชาการที่เกี่ยวกับฐานข้อมูล จำนวน 100 บทคัดย่อ มาทำการทดสอบกับโปรแกรมจัดกลุ่มบทคัดย่อทางวิชาการด้วยอัลกอริทึมเคมีเดียนส์ โดยใช้แฟ้มคำสำคัญข้างต้น และระบุจำนวนกลุ่มที่ต้องการแบ่งเป็น 10 กลุ่ม ได้ผลการทดสอบดังนี้

ค่าผิดพลาดยกกำลังสอง

519.389526367188

458.905151367188

446.585540771484

446.585540771484

ส่วนผลการจัดกลุ่มแสดงดังตารางที่ 4.4 ซึ่งมีรายละเอียดของรายชื่อบทคัดย่อในแต่ละกลุ่ม และคำสำคัญที่ปรากฏในบทคัดย่อของกลุ่มนั้น ๆ ที่มีความถี่มากที่สุดสามอันดับแรก พร้อมทั้งจำนวนของคำสำคัญนั้นด้วย โดยชื่อบทคัดย่อที่มีเครื่องหมาย *** หมายถึงบทคัดย่อที่เป็นศูนย์กลางของกลุ่ม และในที่นี้ขอแสดงผลการทดสอบเพียงแค่บางส่วน ในรายละเอียดของผลการจัดกลุ่มและเนื้อหาของบทคัดย่อที่นำมาทดสอบสามารถดูได้ที่ภาคผนวก

ตาราง 4.4 ผลการจัดกลุ่มบทคัดย่อบางส่วน

กลุ่ม	ชื่อบทคัดย่อ	คำสำคัญ		
1		method	object-oriented	schema
	1. An extensible object-oriented database testbed	0	5	6
	2. An object-oriented prototype for a geophysical database	0	5	3
	3. Disk management for object-oriented databases	0	6	0
	4. Method-induced partitioning schemes for object-oriented databases	5	4	0
	5. Serializability in object-oriented database systems***	2	6	0
2		access	mining	web
	1. A Tool For World-Wide Web Access Log Analysis***	5	0	5
	2. Mobile agents for World Wide Web distributed database access	4	0	4
	3. Data mining for Web intelligence	0	6	11
	4. Performance modelling and metrics of database-backed Web sites	0	1	8
	5. Supporting dynamic interactions among Web-based information sources	5	0	5
	6. Web Database and Its Applications in Teaching Database	0	0	7
3		model	secure	security
	1. Data and applications security developments and directions	0	1	6
	2. Identifying and representing the security semantics of an application	1	1	7
	3. Multilevel database security using information clouding***	1	0	5
	4. Multilevel secure databases a new approach	0	2	5
	5. Proceedings of the 1988 IEEE Symposium on Security and Privacy	3	2	4
	6. Providing security in a phone book database using triggers	0	1	3
	7. Security model consistency in secure object-oriented systems	10	2	7

4.3 สรุปผลการทดสอบโปรแกรม

จากผลการทดสอบข้างต้นจะเห็นว่าค่าผิดพลาดยกกำลังสองมีค่าลดลงเรื่อย ๆ ในแต่ละรอบของการจัดกลุ่ม จนกระทั่งไม่มีการเปลี่ยนแปลงค่าผิดพลาดยกกำลังสอง ซึ่งถือว่าเป็นค่าผิดพลาดยกกำลังสองที่น้อยที่สุดแล้ว โปรแกรมจึงหยุดการทำงาน และแสดงผลการจัดกลุ่มบทคัดย่อออกมา ซึ่งจากผลการจัดกลุ่มที่แสดงออกมา จะเห็นว่าบทคัดย่อที่อยู่ในกลุ่มเดียวกันจะมีความคล้ายคลึงกันมากกว่ากลุ่มอื่น ซึ่งในการตรวจสอบว่าบทคัดย่อภายในกลุ่มมีความคล้ายคลึงกันหรือไม่ อาจพิจารณาที่ชื่อของบทคัดย่อ และจำนวนคำสำคัญที่ปรากฏของบทคัดย่อ ตัวอย่างเช่น บทคัดย่อในกลุ่มที่ 1 จากชื่อของบทคัดย่อจะเห็นว่าทุกบทคัดย่อในกลุ่มที่ 1 เกี่ยวกับเรื่องของ object-oriented database ทั้งหมด โดยมีบทคัดย่อที่ 5 เป็นศูนย์กลางของกลุ่ม และจากจำนวนคำสำคัญสามอันดับแรกที่มีความถี่มากที่สุดของกลุ่มก็แสดงให้เห็นว่าบทคัดย่อในกลุ่มที่ 1 มีคำสำคัญที่คล้ายคลึงกันมาก ทำให้สรุปได้ว่า อัลกอริทึมเคมีเดียนส์สามารถนำมาประยุกต์ใช้กับการจัดกลุ่มบทคัดย่อทางวิชาการได้เป็นอย่างดี และโปรแกรมจัดกลุ่มบทคัดย่อทางวิชาการด้วยอัลกอริทึมเคมีเดียนส์ก็สามารถทำการจัดกลุ่มบทคัดย่อได้อย่างถูกต้อง