บทที่ 4

ผลการวิจัย (Results)

หลังจากดำเนินการตามขั้นตอนในบทที่ 3 แล้วในบทนี้จะเป็นการบรรยายถึงผลการทดลองดังรายละเอียดต่อไปนี้

4.1 ข้อมูลของชุดเอกสารที่ใช้ในการวิจัย

ชุดเอกสารที่ใช้ในการวิจัยเป็นบทคัดย่อของงานวิจัย 4 สาขา กล่าวคือ วิทยาการคอมพิวเตอร์ (Computer science corpus – COM) ประวัติศาสตร์ (History corpus – HIS) บรรณารักษศาสตร์ (Library science corpus – LIB) และ รัฐศาสตร์ (Political science corpus – POL) ทั้งหมดถูกเก็บไว้ในรูปแบบของแฟ้มเอกสาร (Text file format) มีจำนวนเอกสาร 429 426 478 และ 505 ฉบับตามลำดับ ค่าเฉลี่ยของขนาดเอกสารในแต่ละสาขาเท่ากับ 1,283.00 2,008.93 1,903.04 และ 1,442.54 ไบต์ (Byte) ตามลำดับ ขนาดของชุดเอกสารทั้งหมดโดยเฉลี่ยก่อนการตัดคำเท่ากับ 508.31 835.75 888.33 และ 711.41 กิโลไบต์ (KB) ตามลำดับ ในขณะที่ขนาดของชุดเอกสารทั้งหมดโดยเฉลี่ยหลังการตัดคำเท่ากับ 603.14 1,036.32 1,098.65 และ 878.13 กิโลไบต์ (KB) ตามลำดับ สาเหตุที่มีขนาดของเอกสารมีขนาดใหญ่ขึ้นกว่าเอกสารก่อนการตัดคำเนื่องจากโปรแกรมได้แทรกตัวคั่น (delimiter) คั่นระหว่างคำเพื่อแยกคำให้ชัดเจน เพื่อสะดวกในการสร้างแมทริกซ์คำ-เอกสาร (Term-by-document matrix) ในขั้นตอนของการวิจัย ดูรายละเอียดในตารางที่ 4.1

ชุดเอกสารที่ผ่านการตัดคำแล้วถูกนำมาสร้างเป็นแมทริกซ์คำ-เอกสาร (Term-by-document matrix) โดยที่แต่ละสดมภ์ของแมทริกซ์แทนเอกสารแต่ละฉบับในชุดเอกสาร ส่วนแถวของแมทริกซ์แทนจำนวนคำดรรชนี (Index term) หรือ คำ (Term) ในชุดเอกสาร สำหรับสมาชิกของแมทริกซ์แสดงจำนวนคำที่ปรากฏในชุดเอกสารแต่ละฉบับ ดังที่กล่าวแล้วในบทที่ 3 แมทริกซ์คำ-เอกสารที่สร้างได้มีคำดรรชนีจำนวน 4,886 6,724 5,284 และ 5,880 คำตามลำดับ มีค่าลำดับชั้น (Rank) ของแมทริกซ์เท่ากับ 429 421 478 และ 505 ตามลำดับ ความหนาแน่นของแมทริกซ์ (Density) เท่ากับร้อยละ 1.58 2.07 1.88 และ 1.74 ตามลำดับ ซึ่งแสดงว่าสมาชิกของแมทริกซ์ส่วนใหญ่เท่ากับศูนย์ ดังแสดงในตารางที่ 4.2

ตารางที่ 4.1

ข้อมูลของชุดเอกสารที่ใช้ในการวิจัย

ข้อมูล	COM	HIS	LIB	POL
จำนวนเอกสาร	429	426	478	505
จำนวนคำ	104,710	183,087	175,791	148,339
ขนาดเอกสารเล็กสุด (Byte)	456	550	546	394
ขนาดเอกสารใหญ่สุด (Byte)	4,086	5,161	5891	3971
ขนาดโดยเฉลี่ยของชุดเอกสารทั้งหมด (Byte)	1,213.00	2,008.93	1903.04	1442.54
ขนาดเอกสารทั้งหมดโดยเฉลี่ย (KB)				
ก่อนการตัดคำ	508.31	835.75	888.33	711.41
หลังการตัดคำ	630.14	1,036.32	1,098.65	878.13

ตารางที่ 4.2

ข้อมูลของแมทริกซ์คำ-เอกสาร

ข้อมูล	COM	HIS	LIB	POL
จำนวนเอกสาร	429	426	478	505
จำนวนคำดรรชนี	4,886	6,724	5,284	5,080
ค่าเฉลี่ยของจำนวนคำดรรชนี	163.25	285.87	254.36	190.56
จำนวนคำดรรชนีที่น้อยที่สุด	49	77	73	47
จำนวนคำดรรชนีที่มากที่สุด	738	791	803	519
ค่าลำดับชั้น (Rank)	429	421	478	504
ความหนาแน่น (%)	1.58	2.07	1.88	1.74

งานวิจัยนี้ใช้การเปลี่ยนแปลงสัมพัทธ์ (Relative change) ที่ร้อยละ 0 10 20 30 40 เพื่อเปรียบเทียบการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตกับวิธีแยกค่าแบบเดี่ยว โดยใช้ค่าลำดับชั้นของแมทริกซ์ (Rank) ที่การเปลี่ยนแปลงสัมพัทธ์ (Relative change) ของวิธีแยกค่าแบบเดี่ยวเป็นค่าลำดับชั้นในการเปรียบเทียบผลการทดลอง ดังแสดงในตารางที่ 4.3

ตารางที่ 4.3
ค่าลำดับชั้นที่การเปลี่ยนแปลงสัมพัทธ์ต่างๆ

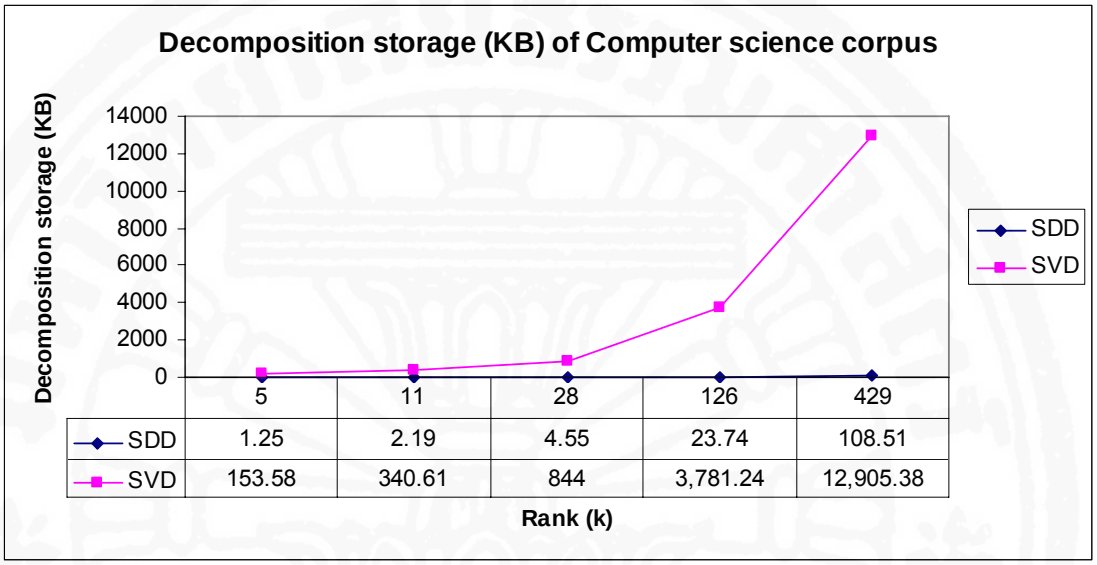
ชุดเอกสารสาขา	การเปลี่ยนแปลงสัมพัทธ์ (ร้อยละ)				
	0	10	20	30	40
วิทยาการคอมพิวเตอร์(COM)	429	126	28	11	5
ประวัติศาสตร์ (HIS)	421	141	35	11	3
บรรณารักษศาสตร์ (LIB)	478	79	18	8	3
รัฐศาสตร์ (POL)	505	149	34	10	2

4.2 ขนาดพื้นที่ประมวลผลการกระจายแมทริกซ์ (Decomposition storage)

เมื่อเราทำการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีต (Semidiscrete Matrix Decomposition – SDD) และวิธีแยกค่าแบบเดี่ยว (Singular Value Decomposition – SVD) จำเป็นต้องใช้พื้นที่หน่วยความจำและพื้นที่ดิสก์ในการประมวลผลซึ่งปรากฏผลดังนี้

ภาพที่ 4.1

ขนาดพื้นที่ประมวลผลของชุดเอกสารสาขาวิทยาการคอมพิวเตอร์

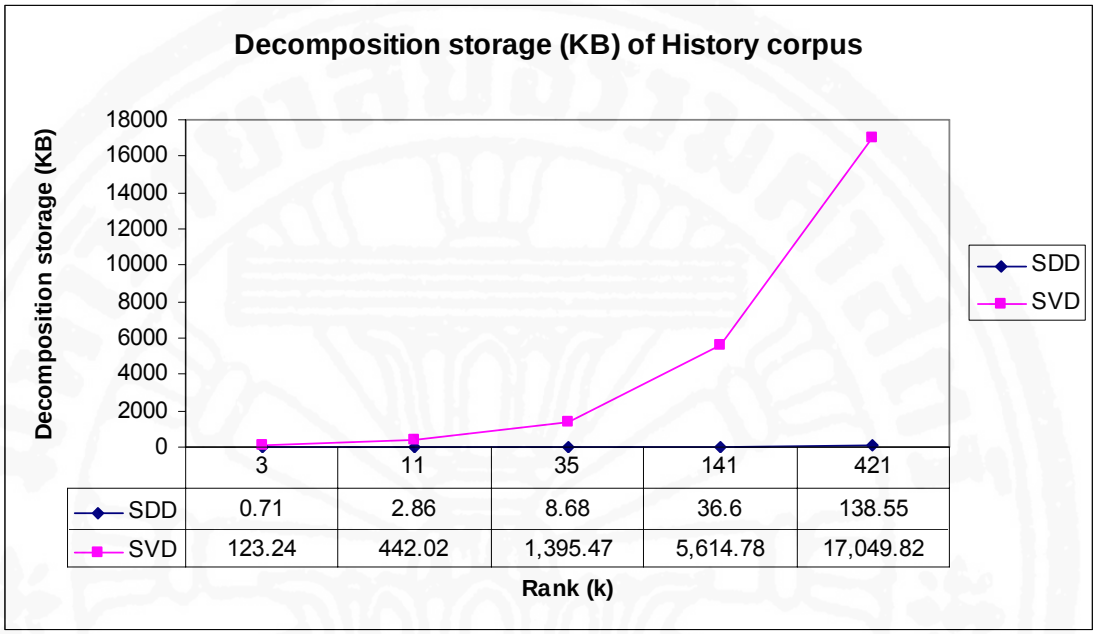


	5	11	28	126	429
SDD	1.25	2.19	4.55	23.74	108.51
SVD	153.58	340.61	844	3,781.24	12,905.38

สาขาวิทยาการคอมพิวเตอร์ (Computer science corpus)

ขนาดพื้นที่ในการประมวลผลโดยเฉลี่ยของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตเท่ากับ 28.06 กิโลไบต์ (KB) ส่วนของวิธีแยกค่าแบบเดี่ยวเท่ากับ 3,604.96 กิโลไบต์ (KB) ซึ่งใหญ่กว่าวิธีเซมิดีสครีตถึง 128.53 เท่า และขนาดพื้นที่ในการกระจายแมทริกซ์คำ-เอกสารสูงขึ้นเมื่อค่าลำดับชั้น (Rank) ของแมทริกซ์มากขึ้นทั้งสองวิธี ดูภาพที่ 4.1 และผนวก ฉ

ภาพที่ 4.2
ขนาดพื้นที่ประมวลผลของชุดเอกสารสาขาประวัติศาสตร์



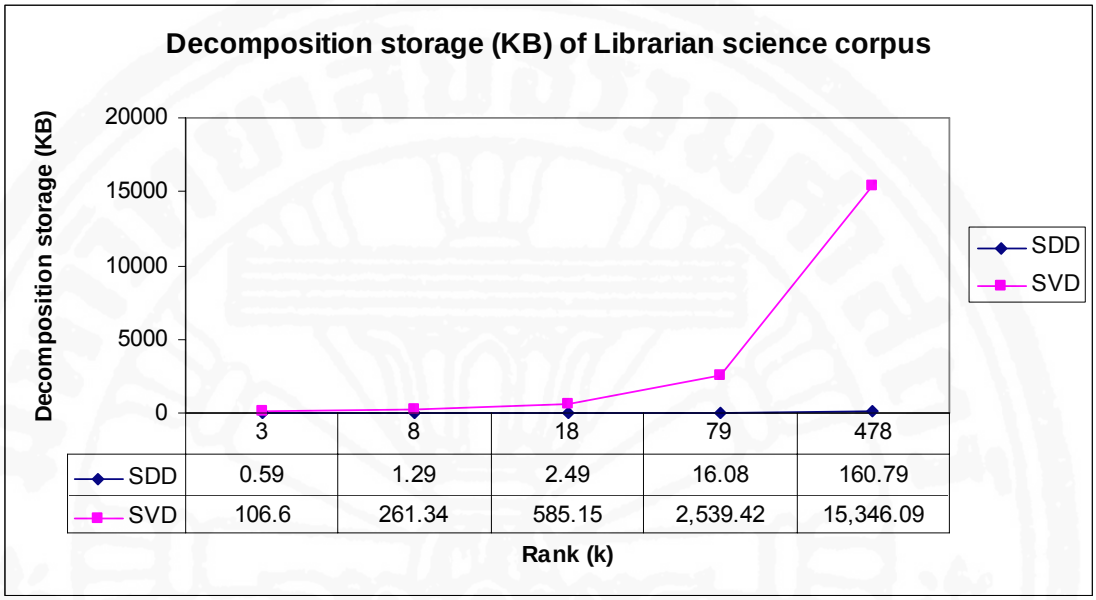
	3	11	35	141	421
SDD	0.71	2.86	8.68	36.6	138.55
SVD	123.24	442.02	1,395.47	5,614.78	17,049.82

สาขาประวัติศาสตร์ (History corpus)

ขนาดพื้นที่ในการประมวลผลโดยเฉลี่ยของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตเท่ากับ 37.48 กิโลไบต์ (KB) ส่วนของวิธีแยกค่าแบบเดี่ยวเท่ากับ 4,925.07 กิโลไบต์ (KB) ซึ่งใหญ่กว่าวิธีเซมิดีสครีตถึง 131.41 เท่า และขนาดพื้นที่ในการกระจายแมทริกซ์คำ-เอกสารสูงขึ้นเมื่อค่าลำดับชั้น (Rank) ของแมทริกซ์มากขึ้นทั้งสองวิธี ดูภาพที่ 4.2 และผนวก ฉ

ภาพที่ 4.3

ขนาดพื้นที่ประมวลผลของชุดเอกสารสาขาบรรณารักษศาสตร์



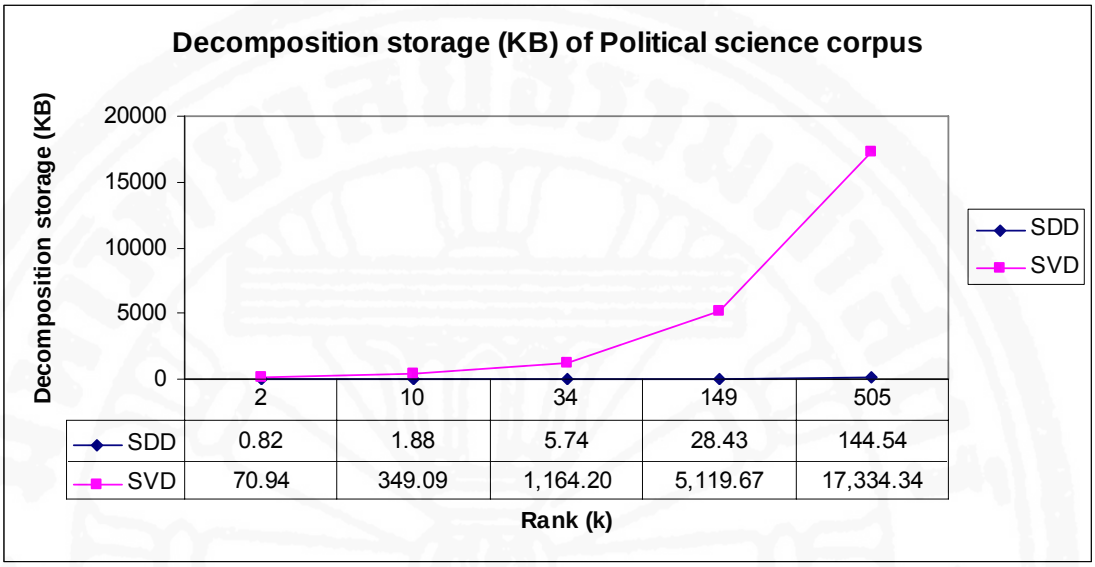
	3	8	18	79	478
SDD	0.59	1.29	2.49	16.08	160.79
SVD	106.6	261.34	585.15	2,539.42	15,346.09

สาขาบรรณารักษศาสตร์ (Librarian science corpus)

ขนาดพื้นที่ในการประมวลผลโดยเฉลี่ยของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตเท่ากับ 36.25 กิโลไบต์ (KB) ส่วนของวิธีแยกค่าแบบเดี่ยวเท่ากับ 3,767.72 กิโลไบต์ (KB) ซึ่งใหญ่กว่าวิธีเซมิดีสครีตถึง 103.94 เท่า และขนาดพื้นที่ในการกระจายแมทริกซ์คำ-เอกสารสูงขึ้นเมื่อค่าลำดับชั้น (Rank) ของแมทริกซ์มากขึ้นทั้งสองวิธี ดูภาพที่ 4.3 และผนวก ฉ

ภาพที่ 4.4

ขนาดพื้นที่ประมวลผลของชุดเอกสารสาขารัฐศาสตร์



	2	10	34	149	505
SDD	0.82	1.88	5.74	28.43	144.54
SVD	70.94	349.09	1,164.20	5,119.67	17,334.34

สาขารัฐศาสตร์ (Political science corpus)

ขนาดพื้นที่ในการประมวลผลโดยเฉลี่ยของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตเท่ากับ 36.28 กิโลไบต์ (KB) ส่วนของวิธีแยกค่าแบบเดี่ยวเท่ากับ 4,807.65 กิโลไบต์ (KB) ซึ่งใหญ่กว่าวิธีเซมิดีสครีตถึง 132.51 เท่า และขนาดพื้นที่ในการกระจายแมทริกซ์คำ-เอกสารสูงขึ้นเมื่อค่าลำดับชั้น (Rank) ของแมทริกซ์มากขึ้นทั้งสองวิธี ดูภาพที่ 4.4 และผนวก ฉ

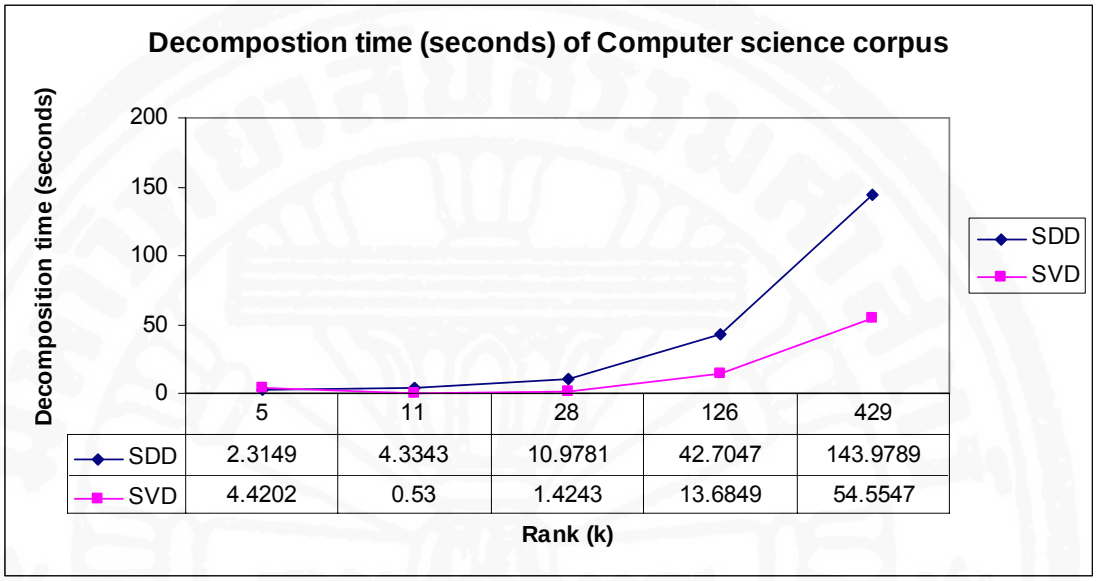
กล่าวโดยสรุปการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตใช้พื้นที่ในการประมวลผลน้อยกว่าวิธีแยกค่าแบบเดี่ยวอย่างมีนัยสำคัญทางสถิติ (Paired-Sample T-Test ที่ระดับความเชื่อมั่น 95 %) และขนาดพื้นที่ในการกระจายแมทริกซ์คำ-เอกสารสูงขึ้นเมื่อค่าลำดับชั้น (Rank) ของแมทริกซ์มากขึ้นทั้งสองวิธีในทุกสาขาวิชา

4.3 **เวลาที่ใช้ในการประมวลผลการกระจายแมทริกซ์** (Decomposition time)

เวลาที่ใช้ในการประมวลผลการกระจายแมทริกซ์หมายถึงเวลาที่ใช้ไปในการประมวลผล โดยคิดเป็นวินาที ซึ่งปรากฏผลการทดลองดังนี้

ภาพที่ 4.5

เวลาที่ใช้ในการประมวลผลของชุดเอกสารสาขาวิทยาการคอมพิวเตอร์



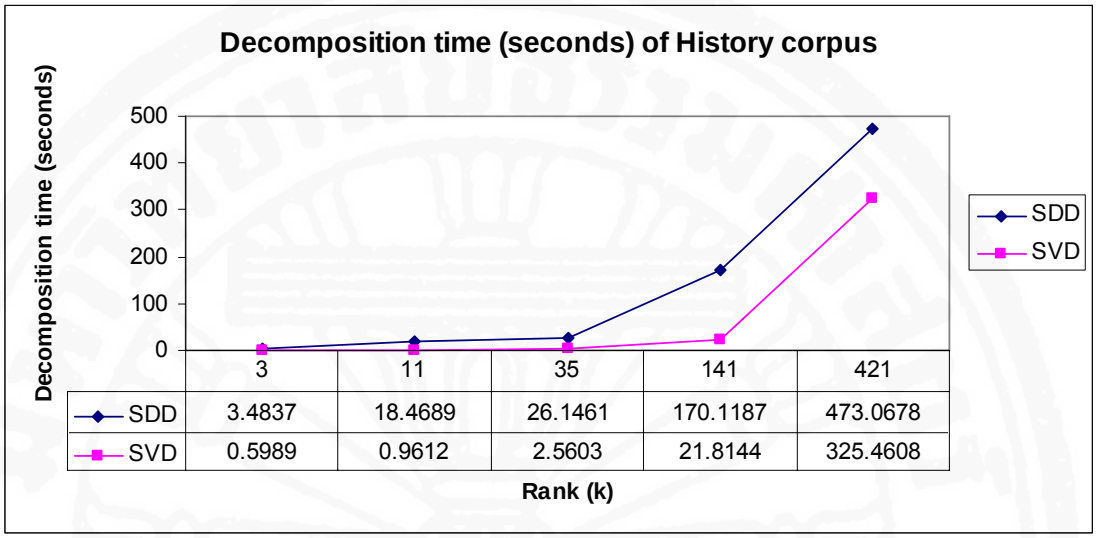
	5	11	28	126	429
SDD	2.3149	4.3343	10.9781	42.7047	143.9789
SVD	4.4202	0.53	1.4243	13.6849	54.5547

สาขาวิทยาการคอมพิวเตอร์ (Computer science corpus)

เวลาที่ใช้ในการประมวลผลโดยเฉลี่ยของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตเท่ากับ 40.86 วินาที (seconds) ส่วนของวิธีแยกค่าแบบเดี่ยวเท่ากับ 14.92 วินาที (seconds) ซึ่งน้อยกว่าวิธีเซมิดีสครีตถึง 2.74 เท่า และเวลาที่ใช้ในการกระจายแมทริกซ์คำ-เอกสารสูงขึ้นเมื่อค่าลำดับชั้น (Rank) ของแมทริกซ์มากขึ้นทั้งสองวิธี ดูภาพที่ 4.5 และผนวก ช

ภาพที่ 4.6

เวลาที่ใช้ในการประมวลผลของชุดเอกสารสาขาประวัติศาสตร์



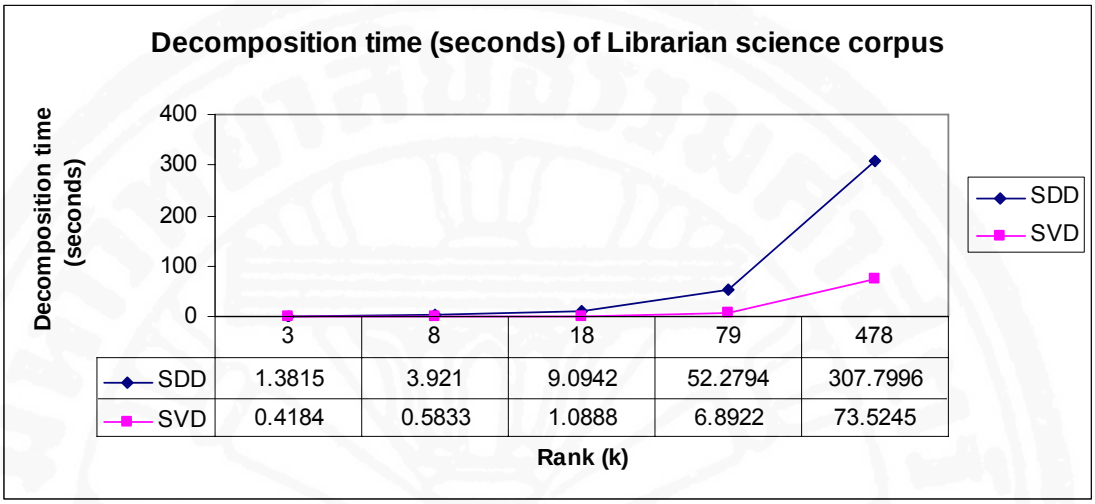
	3	11	35	141	421
SDD	3.4837	18.4689	26.1461	170.1187	473.0678
SVD	0.5989	0.9612	2.5603	21.8144	325.4608

สาขาประวัติศาสตร์ (History corpus)

เวลาที่ใช้ในการประมวลผลโดยเฉลี่ยของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตเท่ากับ 138.26 วินาที (seconds) ส่วนของวิธีแยกค่าแบบเดี่ยวเท่ากับ 70.28 วินาที (seconds) ซึ่งน้อยกว่าวิธีเซมิดีสครีตถึง 1.97 เท่า และเวลาที่ใช้ในการกระจายแมทริกซ์คำ-เอกสารสูงขึ้นเมื่อค่าลำดับชั้น (Rank) ของแมทริกซ์มากขึ้นทั้งสองวิธี ดูภาพที่ 4.6 และผนวก ช

ภาพที่ 4.7

เวลาที่ใช้ในการประมวลผลของชุดเอกสารสาขาบรรณารักษศาสตร์



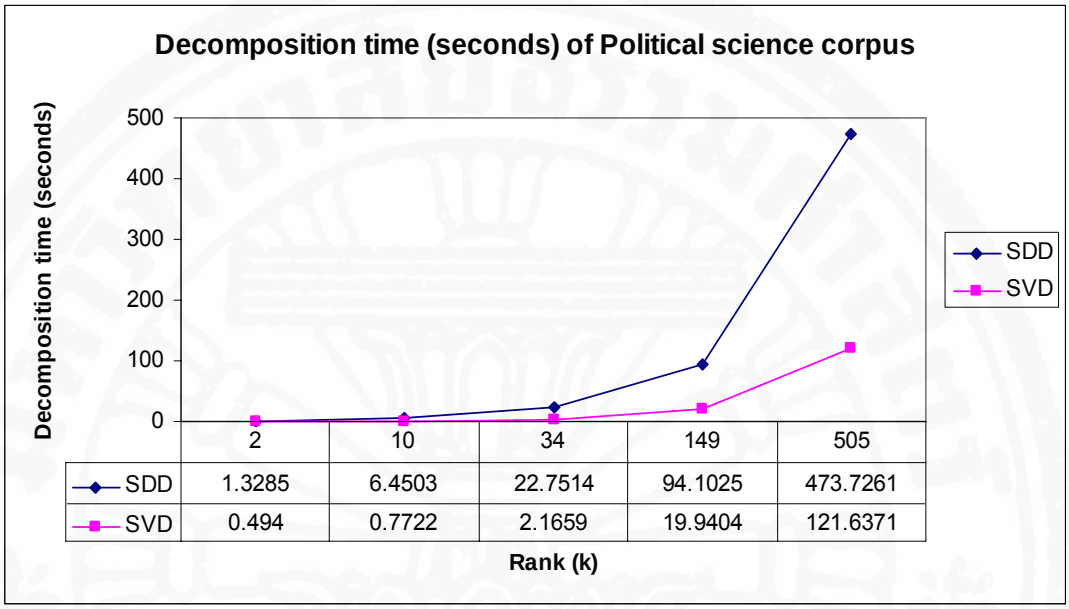
	3	8	18	79	478
SDD	1.3815	3.921	9.0942	52.2794	307.7996
SVD	0.4184	0.5833	1.0888	6.8922	73.5245

สาขาบรรณารักษศาสตร์ (Librarian science corpus)

เวลาที่ใช้ในการประมวลผลโดยเฉลี่ยของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตเท่ากับ 74.90 วินาที (seconds) ส่วนของวิธีแยกค่าแบบเดี่ยวเท่ากับ 16.50 วินาที (seconds) ซึ่งน้อยกว่าวิธีเซมิดีสครีตถึง 4.54 เท่า และเวลาที่ใช้ในการกระจายแมทริกซ์คำ-เอกสารสูงขึ้นเมื่อค่าลำดับชั้น (Rank) ของแมทริกซ์มากขึ้นทั้งสองวิธี ดูภาพที่ 4.7 และผนวก ช

ภาพที่ 4.8

เวลาที่ใช้ในการประมวลผลของชุดเอกสารสาขารัฐศาสตร์



	2	10	34	149	505
SDD	1.3285	6.4503	22.7514	94.1025	473.7261
SVD	0.494	0.7722	2.1659	19.9404	121.6371

สาขารัฐศาสตร์ (Political science corpus)

เวลาที่ใช้ในการประมวลผลโดยเฉลี่ยของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตเท่ากับ 119.67 วินาที (seconds) ส่วนของวิธีแยกค่าแบบเดี่ยวเท่ากับ 29.00 วินาที (seconds) ซึ่งน้อยกว่าวิธีเซมิดีสครีตถึง 4.13 เท่า และเวลาที่ใช้ในการกระจายแมทริกซ์คำ-เอกสารสูงขึ้นเมื่อค่าลำดับชั้น (Rank) ของแมทริกซ์มากขึ้นทั้งสองวิธี ดูภาพที่ 4.8 และผนวก ช

กล่าวโดยสรุปการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตใช้เวลาในการประมวลผลมากกว่ากว่าวิธีแยกค่าแบบเดี่ยวอย่างมีนัยสำคัญทางสถิติ (Paired-Sample T-Test ที่ระดับความเชื่อมั่นที่ 95 %) และเวลาที่ใช้ในการกระจายแมทริกซ์คำ-เอกสารสูงขึ้นเมื่อค่าลำดับชั้น (Rank) ของแมทริกซ์มากขึ้นทั้งสองวิธีในทุกสาขาวิชา

4.4 ประสิทธิภาพของการค้นคืนสารสนเทศ (Retrieval performance)

ประสิทธิภาพของการค้นคืนระบบค้นคืนสารสนเทศพิจารณาจากความสามารถของระบบในการตอบสนองเอกสารที่เกี่ยวข้องกับข้อคำถามที่ผู้ใช้ป้อนเข้าไปสู่ระบบ ในการวิจัยนี้ใช้วัด

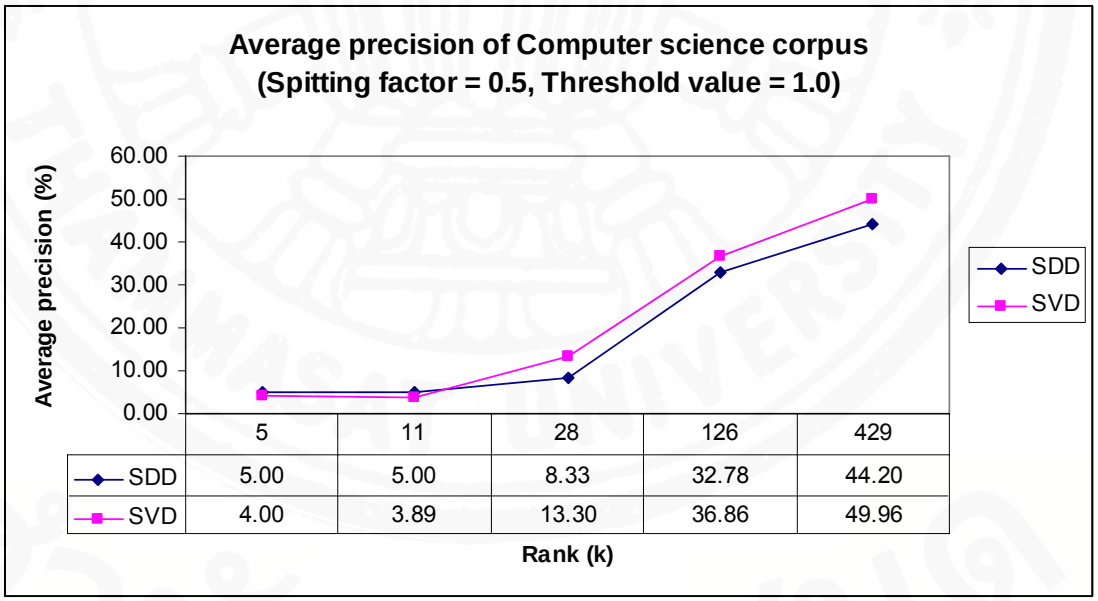
ประสิทธิภาพของระบบโดยใช้ค่าเฉลี่ยของความแม่นยำ (Precision) การเรียกกลับ (Recall) และเวลาที่ใช้ในการประมวลผลข้อคำถาม (Query time) ตามความหมายที่กล่าวถึงในบทที่ 3

4.4.1 **ผลของปัจจัยแยกแมทริกซ์** (Splitting factor)

ถ้าพิจารณาจากสมการที่ 3.7 และ 3.8 ปัจจัยแยกแมทริกซ์ (Splitting factor) อาจมีผลต่อการค้นคืนสารสนเทศ จึงได้ทดลองโดยใช้ค่าปัจจัยแยกแมทริกซ์ (Splitting factor) ตั้งแต่ -1 -0.5 0 0.5 และ 1 และให้ค่าขีดแบ่งความเหมือนของเอกสาร (Threshold value) เท่ากับ 1 ปรากฏว่า ปัจจัยแยกแมทริกซ์ (Splitting factor) ไม่มีผลต่อการค้นคืนสารสนเทศในชุดเอกสารทั้งสี่สาขา กล่าวคือได้ผลลัพธ์เหมือนกันทั้งความแม่นยำ (Precision) และการเรียกกลับ (Recall) รายละเอียดแสดงไว้ภาคผนวก ซ (ตารางที่ ซ.1 ซ.2 ซ.3 และ ซ. 4)

ภาพที่ 4.9

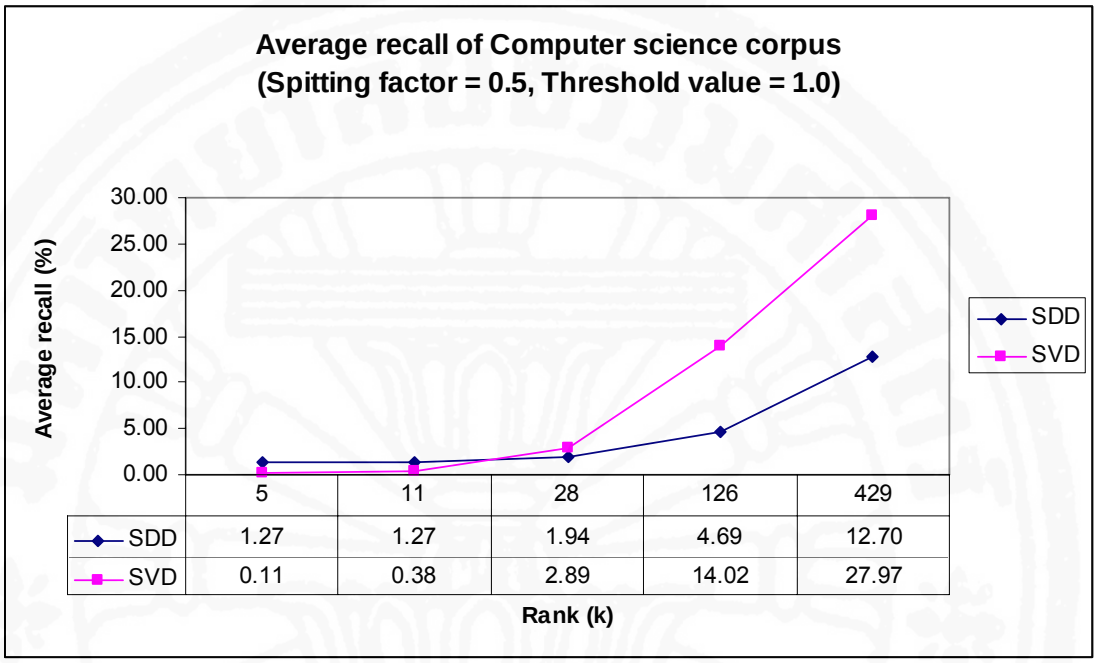
ค่าความแม่นยำเฉลี่ยของการกระจายแมทริกซ์ชุดเอกสารสาขาวิทยาการคอมพิวเตอร์



	5	11	28	126	429
SDD	5.00	5.00	8.33	32.78	44.20
SVD	4.00	3.89	13.30	36.86	49.96

ภาพที่ 4.10

ค่าการเรียกกลับเฉลี่ยของการกระจายแมทริกซ์ชุดเอกสารสาขาวิทยาการคอมพิวเตอร์



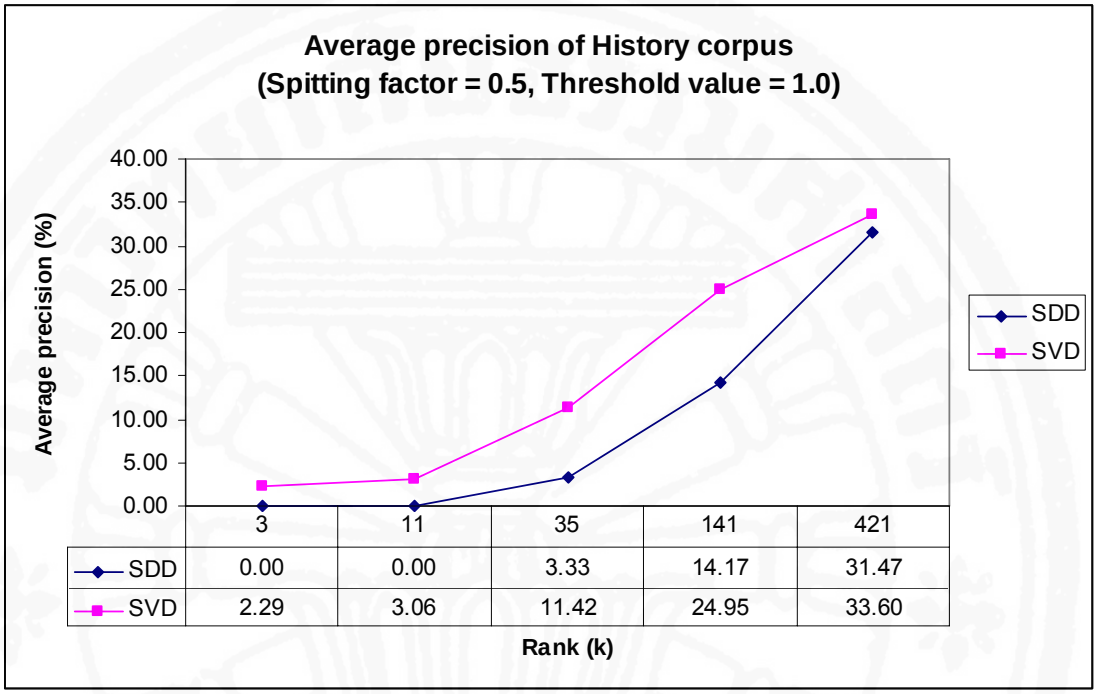
	5	11	28	126	429
SDD	1.27	1.27	1.94	4.69	12.70
SVD	0.11	0.38	2.89	14.02	27.97

<u>สาขาวิทยาการคอมพิวเตอร์ (Computer science corpus)</u>

ที่ปัจจัยแยกแมทริกซ์เท่ากับ 0.5 และค่าขีดแบ่งเท่ากับ 1.0 ค่าความแม่นยำเฉลี่ย (Average precision) ของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตต่ำกว่าวิธีการแยกค่าแบบเดี่ยวอย่างมีนัยสำคัญทางสถิติ และมีแนวโน้มสูงขึ้นเมื่อค่าลำดับชั้น (Rank) ของแมทริกซ์มากขึ้น ส่วนค่าการเรียกกลับเฉลี่ย (Average recall) ของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตสูงกว่าวิธีการแยกค่าแบบเดี่ยวเล็กน้อยเมื่อค่าลำดับชั้น (Rank) ของแมทริกซ์มีค่าน้อย แต่เมื่อค่าลำดับชั้น (Rank) ของแมทริกซ์มีค่ามากขึ้น ค่าการเรียกกลับเฉลี่ย (Average recall) ของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตน้อยกว่าวิธีการแยกค่าแบบเดี่ยว ดูตารางที่ ซ.1 และภาพที่ 4.9 และภาพที่ 4.10

ภาพที่ 4.11

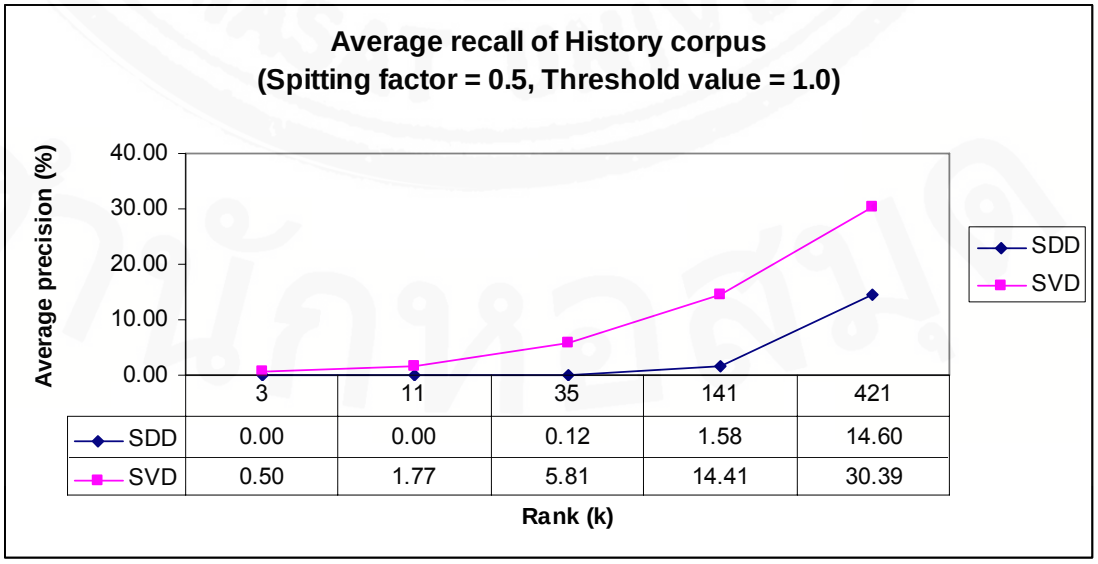
ค่าความแม่นยำเฉลี่ยของการกระจายแมทริกซ์ชุดเอกสารสาขาประวัติศาสตร์



	3	11	35	141	421
SDD	0.00	0.00	3.33	14.17	31.47
SVD	2.29	3.06	11.42	24.95	33.60

ภาพที่ 4.12

ค่าการเรียกกลับเฉลี่ยของการกระจายแมทริกซ์ชุดเอกสารสาขาประวัติศาสตร์



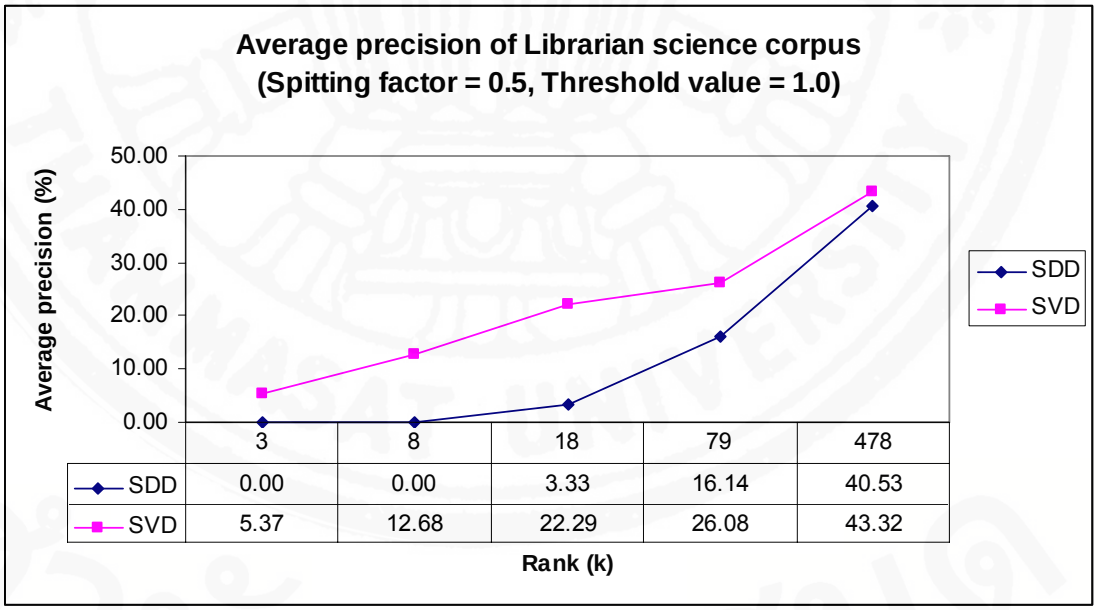
	3	11	35	141	421
SDD	0.00	0.00	0.12	1.58	14.60
SVD	0.50	1.77	5.81	14.41	30.39

สาขาประวัติศาสตร์ (History corpus)

ที่ปัจจัยแยกแมทริกซ์เท่ากับ 0.5 และค่าขีดแบ่งเท่ากับ 1.0 ค่าความแม่นยำเฉลี่ย (Average precision) ของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตต่ำกว่าวิธีการแยกค่าแบบเดี่ยว และมีแนวโน้มสูงขึ้นเมื่อค่าลำดับชั้น (Rank) ของแมทริกซ์มากขึ้นทั้งสองวิธี เช่นเดียวกับค่าการเรียกกลับเฉลี่ย (Average recall) ของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตน้อยกว่าวิธีการแยกค่าแบบเดี่ยวและมีแนวโน้มสูงขึ้นเมื่อค่าลำดับชั้น (Rank) ของแมทริกซ์มากขึ้นทั้งสองวิธีเช่นเดียวกัน และค่าความแม่นยำเฉลี่ย (Average precision)ของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตมีแนวโน้มมีค่าใกล้เคียงกับวิธีการแยกค่าแบบเดี่ยวมากขึ้นเมื่อค่าลำดับชั้น (Rank) ของแมทริกซ์มากขึ้นดูตารางที่ ซ.3 และภาพที่ 4.11 และภาพที่ 4.12

ภาพที่ 4.13

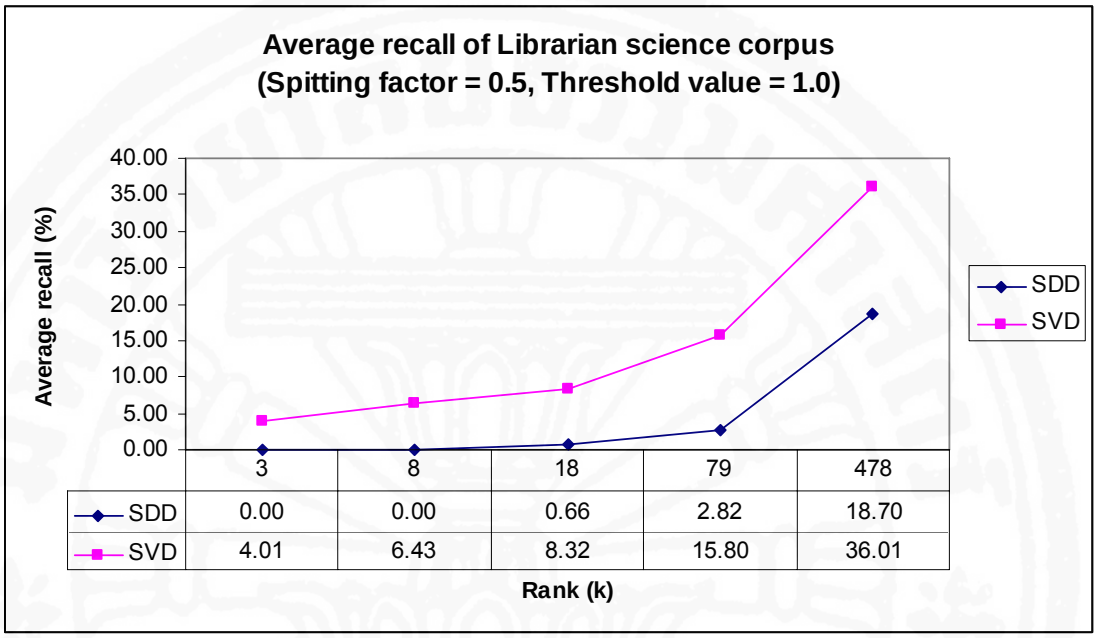
ค่าความแม่นยำเฉลี่ยของการกระจายแมทริกซ์ชุดเอกสารสาขาบรรณารักษศาสตร์



	3	8	18	79	478
SDD	0.00	0.00	3.33	16.14	40.53
SVD	5.37	12.68	22.29	26.08	43.32

ภาพที่ 4.14

ค่าการเรียกกลับเฉลี่ยของการกระจายแมทริกซ์ชุดเอกสารสาขาบรรณารักษศาสตร์



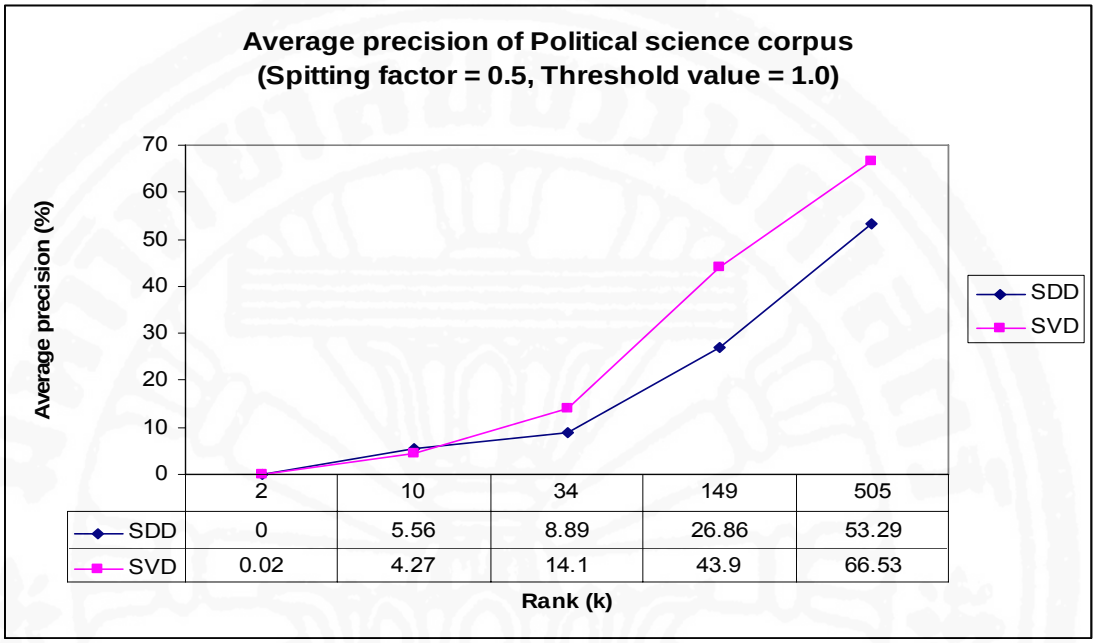
	3	8	18	79	478
SDD	0.00	0.00	0.66	2.82	18.70
SVD	4.01	6.43	8.32	15.80	36.01

สาขาบรรณารักษศาสตร์ (Librarian science corpus)

ที่ปัจจัยแยกแมทริกซ์เท่ากับ 0.5 และค่าขีดแบ่งเท่ากับ 1.0 ค่าความแม่นยำเฉลี่ย (Average precision) ของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตต่ำกว่าวิธีการแยกค่าแบบเดี่ยว และมีแนวโน้มสูงขึ้นเมื่อค่าลำดับชั้น (Rank) ของแมทริกซ์มากขึ้นทั้งสองวิธี เช่นเดียวกับค่าการเรียกกลับเฉลี่ย (Average recall) ของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตน้อยกว่าวิธีการแยกค่าแบบเดี่ยวและมีแนวโน้มสูงขึ้นเมื่อค่าลำดับชั้น (Rank) ของแมทริกซ์มากขึ้นทั้งสองวิธีเช่นเดียวกัน และค่าความแม่นยำเฉลี่ย (Average precision)ของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตมีแนวโน้มมีค่าใกล้เคียงกับวิธีการแยกค่าแบบเดี่ยวมากขึ้นเมื่อค่าลำดับชั้น (Rank) ของแมทริกซ์มากขึ้น ดูตารางที่ ซ.3 และภาพที่ 4.13 และภาพที่ 4.14

ภาพที่ 4.15

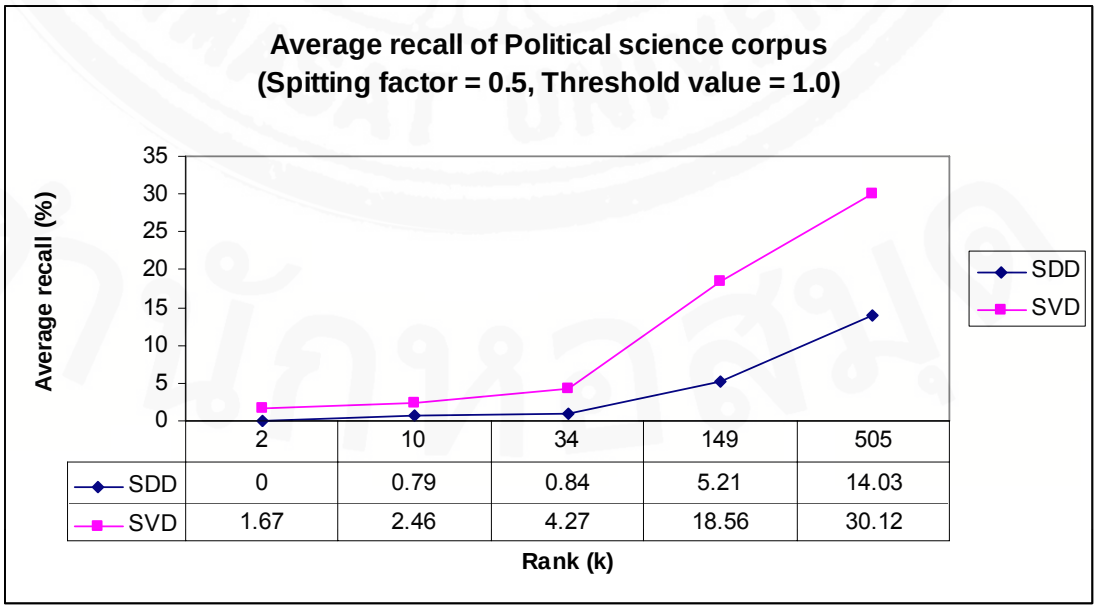
ค่าความแม่นยำเฉลี่ยของการกระจายแมทริกซ์ชุดเอกสารสาขารัฐศาสตร์



	2	10	34	149	505
SDD	0	5.56	8.89	26.86	53.29
SVD	0.02	4.27	14.1	43.9	66.53

ภาพที่ 4.16

ค่าการเรียกกลับเฉลี่ยของการกระจายแมทริกซ์ชุดเอกสารสาขารัฐศาสตร์



	2	10	34	149	505
SDD	0	0.79	0.84	5.21	14.03
SVD	1.67	2.46	4.27	18.56	30.12

สาขารัฐศาสตร์ (Political science corpus)

ที่ปัจจัยแยกแมทริกซ์เท่ากับ 0.5 และค่าขีดแบ่งเท่ากับ 1.0 ค่าความแม่นยำเฉลี่ย (Average precision) ของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตต่ำกว่าวิธีการแยกค่าแบบเดี่ยว และมีแนวโน้มสูงขึ้นเมื่อค่าลำดับชั้น (Rank) ของแมทริกซ์มากขึ้นทั้งสองวิธี เช่นเดียวกับค่าการเรียกกลับเฉลี่ย (Average recall) ของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตน้อยกว่าวิธีการแยกค่าแบบเดี่ยวและมีแนวโน้มสูงขึ้นเมื่อค่าลำดับชั้น (Rank) ของแมทริกซ์มากขึ้นทั้งสองวิธีเช่นเดียวกัน ดูตารางที่ ซ.3 และภาพที่ 4.15 และภาพที่ 4.16

กล่าวโดยสรุป ที่ปัจจัยแยกแมทริกซ์เท่ากับ 0.5 และค่าขีดแบ่งเท่ากับ 1.0 ค่าความแม่นยำเฉลี่ย (Average precision) ของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตต่ำกว่าวิธีการแยกค่าแบบเดี่ยวอย่างมีนัยสำคัญทางสถิติ และมีแนวโน้มสูงขึ้นเมื่อค่าลำดับชั้น (Rank) ของแมทริกซ์มากขึ้นทั้งสองวิธี เช่นเดียวกับค่าการเรียกกลับเฉลี่ย (Average recall) ของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตน้อยกว่าวิธีการแยกค่าแบบเดี่ยวอย่างมีนัยสำคัญทางสถิติและมีแนวโน้มสูงขึ้นเมื่อค่าลำดับชั้น (Rank) ของแมทริกซ์มากขึ้นทั้งสองวิธีเช่นเดียวกันในทุกสาขาวิชา ดูภาคผนวก ซ

4.4.2 ผลของค่าขีดแบ่งความเหมือน (Similarity threshold value)

ค่าขีดแบ่งความเหมือน (Similarity threshold value) หมายถึงค่าที่กำหนดขึ้นเพื่อระบุว่า เอกสารใดมีความเกี่ยวข้องกับข้อคำถามที่เราป้อนเข้าสู่ระบบ โดยพิจารณาจากค่าความเหมือนของเอกสาร หากค่าความเหมือนของเอกสารมีค่ามากกว่าค่าขีดแบ่งความเหมือนที่กำหนดขึ้นหมายความว่าเอกสารนั้นมีความเกี่ยวข้อกับข้อคำถาม แต่ถ้าน้อยกว่าจะถือว่าไม่เกี่ยวข้องกับข้อคำถาม เพื่อดูผลของค่าขีดแบ่งความเหมือนที่มีต่อการค้นคืนสารสนเทศ ได้กำหนดค่าขีดแบ่งความเหมือนแตกต่างกันเป็น 5 ระดับกล่าวคือ 0.5 1.0 1.5 2.0 และ 2.5 ตามลำดับ เนื่องจากปัจจัยแยกแมทริกซ์ไม่มีผลต่อประสิทธิภาพการค้นคืนสารสนเทศ การวิจัยนี้จึงเลือกใช้ค่าปัจจัยแยกแมทริกซ์ (Splitting factor) เท่ากับ 0.5 เช่นเดียวกับงานของ (Kolda and O'Leary, 1998, p. 322-346) ผลการทดลองปรากฏว่า ค่าขีดแบ่งความเหมือนมีผลต่อความแม่นยำ (Precision) และการเรียกกลับ (Recall) ดังนี้

สาขาวิทยาการคอมพิวเตอร์ (Computer science corpus)

ค่าความแม่นยำเฉลี่ย (Average precision) ของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตสูงขึ้น เมื่อค่าขีดแบ่งความเหมือนมากขึ้น ในขณะที่ค่าการเรียกกลับเฉลี่ย (Average recall) ลดลงเมื่อค่าขีดแบ่งความเหมือนมากขึ้น

ส่วนค่าความแม่นยำเฉลี่ย (Average precision) และค่าการเรียกกลับเฉลี่ย (Average recall) ของการกระจายแมทริกซ์ด้วยวิธีแยกค่าแบบเดี่ยวมีแนวโน้มเป็นเช่นเดียวกับการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีต ดูตารางที่ ซ.5

สาขาประวัติศาสตร์ (History corpus)

ค่าความแม่นยำเฉลี่ย (Average precision) ของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตมีค่ามากที่สุดที่ค่าขีดแบ่งความเหมือนเท่ากับ 1.50 ในขณะที่ค่าการเรียกกลับเฉลี่ย (Average recall) ลดลงเมื่อค่าขีดแบ่งความเหมือนมากขึ้น

ค่าความแม่นยำเฉลี่ย (Average precision) ของการกระจายแมทริกซ์ด้วยวิธีแยกค่าแบบเดี่ยวมีค่ามากที่สุดที่ค่าขีดแบ่งความเหมือนเท่ากับ 1.50 เช่นกัน ในขณะที่ค่าการเรียกกลับเฉลี่ย (Average recall) ลดลงเมื่อค่าขีดแบ่งความเหมือนมากขึ้น ดูตารางที่ ซ.6

สาขาบรรณารักษศาสตร์ (Librarian science corpus)

ค่าความแม่นยำเฉลี่ย (Average precision) ของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตสูงขึ้น เมื่อค่าขีดแบ่งความเหมือนมากขึ้น ในขณะที่ค่าการเรียกกลับเฉลี่ย (Average recall) ลดลงเมื่อค่าขีดแบ่งความเหมือนมากขึ้น

ค่าความแม่นยำเฉลี่ย (Average precision) ของการกระจายแมทริกซ์ด้วยวิธีแยกค่าแบบเดี่ยวมีค่ามากที่สุดที่ค่าขีดแบ่งความเหมือนเท่ากับ 1.0 ในขณะที่ค่าการเรียกกลับเฉลี่ย (Average recall) ลดลงเมื่อค่าขีดแบ่งความเหมือนมากขึ้น ดูตารางที่ ซ.7

สาขารัฐศาสตร์ (Political science corpus)

ค่าความแม่นยำเฉลี่ย (Average precision) ของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตสูงขึ้น เมื่อค่าขีดแบ่งความเหมือนมากขึ้น ในขณะที่ค่าการเรียกกลับเฉลี่ย (Average recall) ลดลงเมื่อค่าขีดแบ่งความเหมือนมากขึ้น

ค่าความแม่นยำเฉลี่ย (Average precision) ของการกระจายแมทริกซ์ด้วยวิธีแยกค่าแบบเดี่ยวมีค่ามากที่สุดที่ค่าขีดแบ่งความเหมือนเท่ากับ 1.5 ในขณะที่ค่าการเรียกกลับเฉลี่ย (Average recall) ลดลงเมื่อค่าขีดแบ่งความเหมือนมากขึ้น ดูตารางที่ ซ.8

กล่าวโดยสรุป ค่าความแม่นยำเฉลี่ย (Average precision) ของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตสูงขึ้น เมื่อค่าขีดแบ่งความเหมือนมากขึ้น ในขณะที่ค่าการเรียกกลับเฉลี่ย (Average recall) ลดลงเมื่อค่าขีดแบ่งความเหมือนมากขึ้น ในขณะที่ค่าความแม่นยำเฉลี่ย (Average precision) ของการกระจายแมทริกซ์ด้วยวิธีแยกค่าแบบเดี่ยวมีค่ามากที่สุดที่ค่าขีดแบ่ง

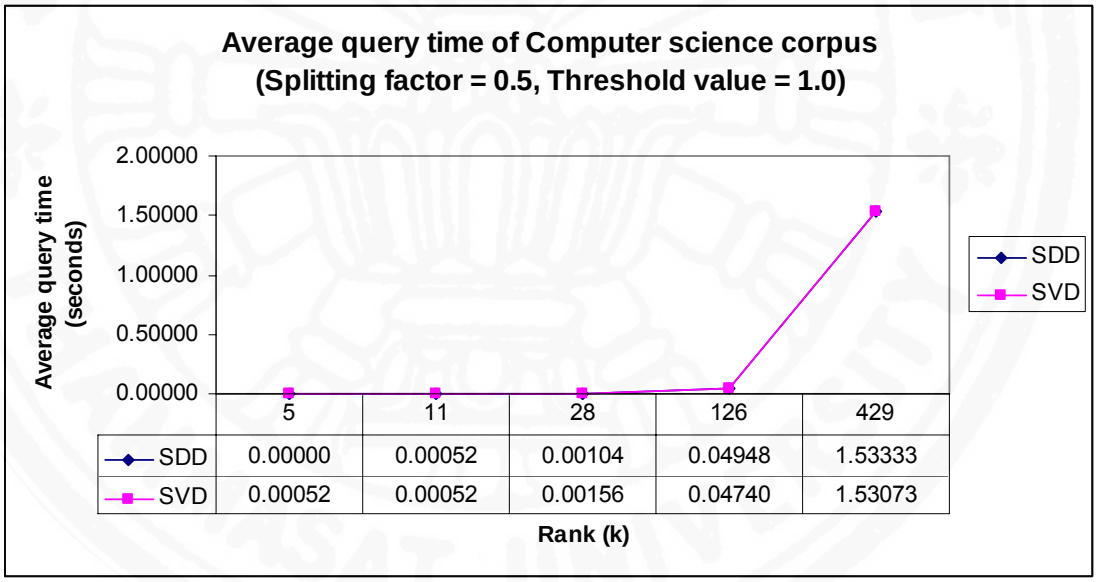
ความเหมือนเท่ากับ 1.5 ในขณะที่ค่าการเรียกกลับเฉลี่ย (Average recall) ลดลงเมื่อค่าขีดแบ่งความเหมือนมากขึ้น ดูภาคผนวก ซ

4.4.3 **เวลาที่ใช้ในการประมวลผลข้อคำถาม** (Query time)

เวลาที่ใช้ในการประมวลผลข้อคำถาม (Query time) หมายถึงเวลาที่ใช้ไปเป็นวินาทีในการประมวลผลข้อคำถามระหว่างเวกเตอร์ของข้อคำถามกับแมทริกซ์คำ-เอกสารเพื่อหาค่าความเหมือนระหว่างข้อคำถามกับเอกสาร โดยสมการ 3.7 และ 3.8

ภาพที่ 4.17

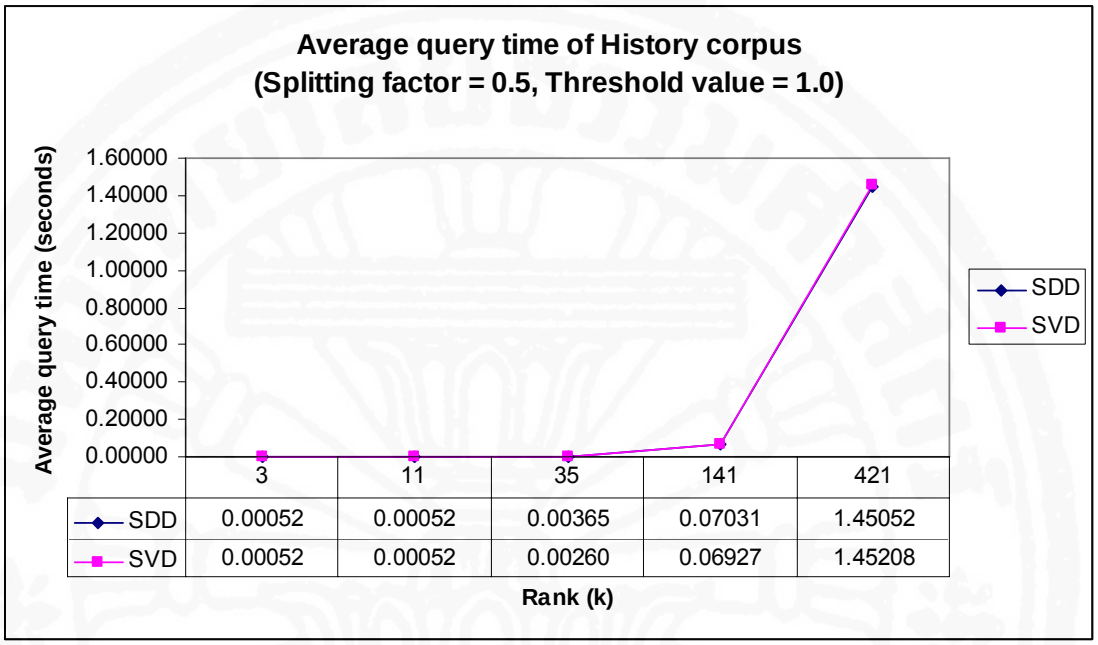
เวลาที่ใช้ในการประมวลผลข้อคำถามเฉลี่ยของชุดเอกสารสาขาวิทยาการคอมพิวเตอร์



	5	11	28	126	429
SDD	0.00000	0.00052	0.00104	0.04948	1.53333
SVD	0.00052	0.00052	0.00156	0.04740	1.53073

ภาพที่ 4.18

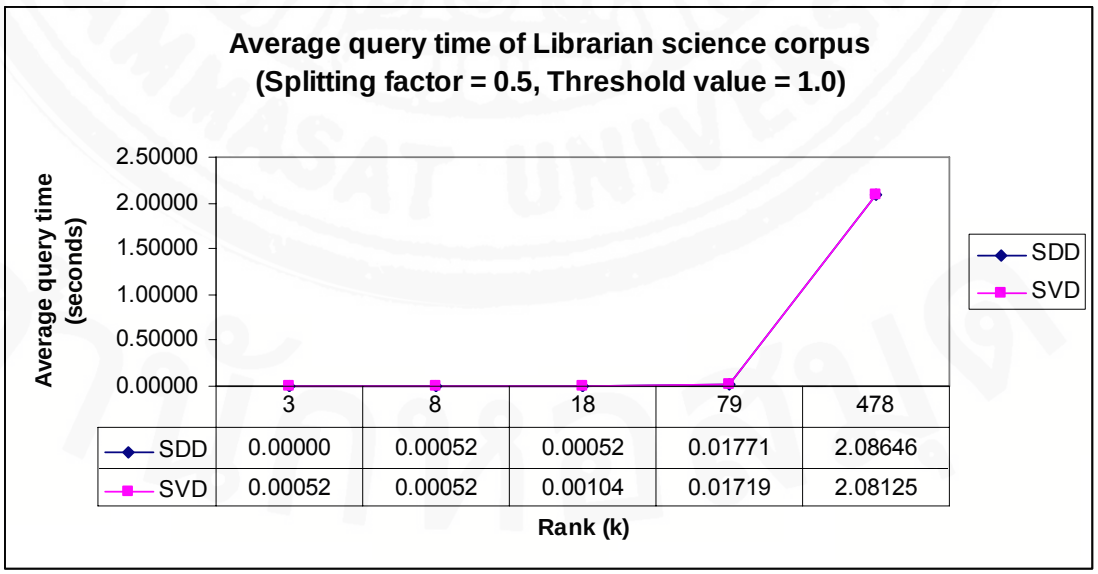
เวลาที่ใช้ในการประมวลผลข้อคำถามเฉลี่ยของชุดเอกสารสาขาประวัติศาสตร์



	3	11	35	141	421
SDD	0.00052	0.00052	0.00365	0.07031	1.45052
SVD	0.00052	0.00052	0.00260	0.06927	1.45208

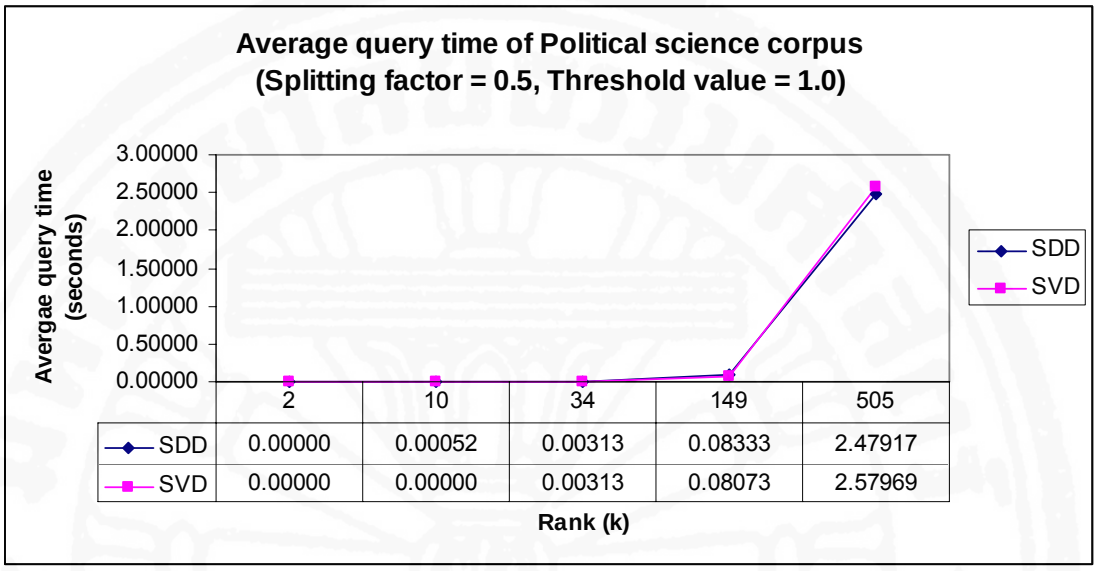
ภาพที่ 4.19

เวลาที่ใช้ในการประมวลผลข้อคำถามเฉลี่ยของชุดเอกสารสาขาบรรณารักษศาสตร์



	3	8	18	79	478
SDD	0.00000	0.00052	0.00052	0.01771	2.08646
SVD	0.00052	0.00052	0.00104	0.01719	2.08125

ภาพที่ 4.20
เวลาที่ใช้ในการประมวลผลข้อคำถามเฉลี่ยของชุดเอกสารสาขารัฐศาสตร์



	2	10	34	149	505
SDD	0.00000	0.00052	0.00313	0.08333	2.47917
SVD	0.00000	0.00000	0.00313	0.08073	2.57969

ผลการทดลองปรากฏว่า เวลาที่ใช้ในการประมวลผลข้อคำถาม (Query time) โดยเฉลี่ยของชุดข้อคำถามของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตเปรียบเทียบกับการกระจายแมทริกซ์ด้วยวิธีแยกค่าแบบเดี่ยว โดยใช้ค่าปัจจัยแยกแมทริกซ์เท่ากับ 0.5 และ ค่าขีดแบ่งความเหมือนเท่ากับ 1.0 ของชุดเอกสารในทุกสาขามีความใกล้เคียงกันมาก ไม่แตกต่างกันอย่างมีนัยสำคัญทางสถิติ (Paired-Sample T-Test ที่ระดับความเชื่อมั่น 95%) ดังภาพที่ 4.17 ถึงภาพที่ 4.20

อนึ่ง เมื่อเปรียบเทียบเวลาเฉลี่ยที่ใช้ในการประมวลผลข้อคำถามของการกระจายแมทริกซ์โดยสมการที่ 3.8 ที่ใช้ในงานวิจัยน้อยกว่าเวลาเฉลี่ยที่ใช้ในการประมวลผลข้อคำถามเฉลี่ยโดยสมการของ Kolda และ คณะ (Kolda and O'Leary, 1998, p. 322-346) ตามสมการ 3.9 [2]อย่างมีนัยสำคัญทางสถิติ (Paired-Sample T-Test ที่ระดับความเชื่อมั่น 95%)

[2] รายละเอียดของการเปรียบเทียบระหว่างสมการที่ 3.8 กับสมการที่ 3.9 ด้วยวิธีการพิสูจน์ทางคณิตศาสตร์และการทดลองกับชุดข้อมูลได้แสดงไว้ในภาคผนวก ญ

กล่าวโดยภาพรวมสำหรับการค้นคืนสารสนเทศในชุดเอกสารทุกสาขาวิชา ค่าความแม่นยำเฉลี่ย (Average precision) ของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตจะต่ำกว่าวิธีการแยกค่าแบบเดี่ยวอย่างมีนัยสำคัญทางสถิติ และมีแนวโน้มสูงขึ้นเมื่อค่าลำดับชั้น (Rank) ของแมทริกซ์มากขึ้นทั้งสองวิธี เช่นเดียวกับค่าการเรียกกลับเฉลี่ย (Average recall) ของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตน้อยกว่าวิธีการแยกค่าแบบเดี่ยวอย่างมีนัยสำคัญทางสถิติและมีแนวโน้มสูงขึ้นเมื่อค่าลำดับชั้น (Rank) ของแมทริกซ์มากขึ้นทั้งสองวิธีเช่นเดียวกัน

4.5 สรุปผลการทดลอง

ผลการทดลองสามารถสรุปได้ ดังนี้

(1) พื้นที่ที่ใช้ในการประมวลผลการกระจายแมทริกซ์ (Decomposition storage) ด้วยวิธีเซมิดีสครีตใช้น้อยกว่าวิธีแยกค่าแบบเดี่ยวอย่างมีนัยสำคัญทางสถิติ ซึ่งสอดคล้องกับสมมุติฐานของงานวิจัย และงานของ Kolda และคณะ (Kolda and O'Leary, 1998, p. 322-346)

(2) เวลาที่ใช้ในการประมวลผลการกระจายแมทริกซ์ (Decomposition time) ด้วยวิธีเซมิดีสครีตใช้น้อยกว่าวิธีแยกค่าแบบเดี่ยวอย่างมีนัยสำคัญทางสถิติ ซึ่งสอดคล้องกับสมมุติฐานของงานวิจัย และงานของ Kolda และคณะ (Kolda and O'Leary, 1998, p. 322-346)

(3) ความแม่นยำ (Precision) และการเรียกกลับ (Recall) ของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตต่ำกว่าวิธีแยกค่าแบบเดี่ยว อย่างมีนัยสำคัญทางสถิติในชุดเอกสารทุกสาขา ซึ่งขัดกับสมมุติฐานของงานวิจัย โดยปัจจัยแยกแมทริกซ์ (Splitting factor) ไม่มีผลต่อความแม่นยำ (Precision) และการเรียกกลับ (Recall) ส่วนค่าขีดแบ่งความเหมือนของเอกสาร (Similarity threshold value) มีผลต่อความแม่นยำ (Precision) และการเรียกกลับ (Recall)

(4) เวลาที่ใช้ในการประมวลผลข้อคำถาม (Query time) ของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตไม่แตกต่างกับวิธีแยกค่าแบบเดี่ยว อย่างมีนัยสำคัญทางสถิติ ซึ่งขัดกับสมมุติฐานของงานวิจัย และงานของ Kolda และคณะ (Kolda and O'Leary, 1998, p. 322-346) ซึ่งรายงานว่าเวลาที่ใช้ในการประมวลผลข้อคำถาม (Query time) ของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตใช้เวลาน้อยกว่าวิธีแยกค่าแบบเดี่ยว