

การพัฒนาภาษาที่ใช้สอบถามสำหรับการสืบค้น

สารสนเทศของพระไตรปิฎก



นางสาวจิรากร เกียรติไพบูลย์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

ภาควิชาวิศวกรรมคอมพิวเตอร์

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

พ.ศ. 2533

ISBN 974-577-387-5

ลิขสิทธิ์ของบัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

016706

117419631

QUERY LANGUAGE DEVELOPMENT FOR TRI PITAKA

INFORMATION RETRIEVAL SYSTEM

Miss Jiraporn Kiatpaiboon

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science

Department of Computer Engineering

Graduate School

Chulalongkorn University


1990

ISBN 974-577-387-5

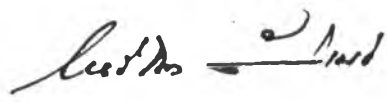
หัวข้อวิทยานิพนธ์ การพัฒนาภาษาที่ใช้สอบถามสำหรับการสืบค้นสารสนเทศของพระไตรปิฎก
โดย นางสาวจิรากร เกียรติไพบูลย์
ภาควิชา วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษา รองศาสตราจารย์ ดร. ศุภชัย ตั้งวงศ์คานต์

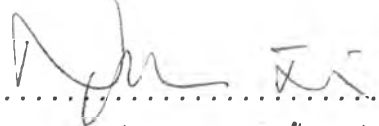


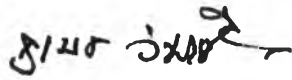
บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้บัณฑิตวิทยาลัยนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรบัณฑิต

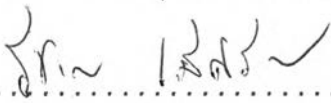
.....  คณบดีบัณฑิตวิทยาลัย
(ศาสตราจารย์ ดร. ถาวร วิชาภัย)


คณะกรรมการสอบวิทยานิพนธ์

.....  ประธานกรรมการ
(รองศาสตราจารย์ ไกรวิชิต ตันติเมธ)

.....  อาจารย์ที่ปรึกษา
(รองศาสตราจารย์ ดร. ศุภชัย ตั้งวงศ์คานต์)

.....  อาจารย์ที่ปรึกษาร่วม
(ผู้ช่วยศาสตราจารย์ สุมเมธ วิชาชัยสุรพล)

.....  กรรมการ
(ผู้ช่วยศาสตราจารย์ วิชาญ เลิศวิภาตระกูล)

.....  กรรมการ
(อาจารย์ จารุมาทร ปิ่นทอง)



จรรยา เกียรติไพบูลย์ : การพัฒนาภาษาที่ใช้สอบถามสำหรับการสืบค้นสารสนเทศของพระไตรปิฎก (QUERY LANGUAGE DEVELOPMENT FOR TRI PITAKA INFORMATION RETRIEVAL SYSTEM) อ.ที่ปรึกษา รศ.ดร.ศุภชัย ตั้งวงศ์ศานต์, ผศ.สุเมธ วัชรชัยสุรพล, 116 หน้า. ISBN 974-577-387-5

พระไตรปิฎกเป็นที่ประมวลไว้ซึ่งพระธรรมวินัย ที่เป็นหลักการใหญ่ของพระพุทธศาสนา พระสงฆ์สาวกและพุทธศาสนิกชนที่เคร่งครัดจะต้องใส่ใจศึกษาและปฏิบัติตามคำสอนในพระไตรปิฎก นอกจากนี้ นักวิชาการหลายสาขาได้ใช้พระไตรปิฎกเป็นแหล่งค้นคว้าหาความรู้ในแขนงวิชาที่สนใจได้ด้วย แต่ในความพยายามที่จะค้นหาเรื่องราวใด จากพระไตรปิฎกนั้น นอกจากอุปสรรคในเรื่องของภาษาบาลีแล้ว ยังมีอุปสรรคด้านปริมาณข้อมูลจำนวนมากถึง 24.23 ล้านตัวอักษรใน 45 เล่มพิมพ์ จึงเป็นการยากที่จะค้นหาและรวบรวมเรื่องราวใดจากพระไตรปิฎกให้ได้อย่างครบถ้วนภายในเวลาอันสั้น

ในปี พ.ศ.2531 มหาวิทยาลัยมหิดลได้จัดทำ พระไตรปิฎกฉบับคอมพิวเตอร์ ขึ้นเป็นครั้งแรก โดยได้พัฒนาบนเครื่องไมโครคอมพิวเตอร์ประเภทไอบีเอ็มพีซี และเรียกชื่อว่า BUDSIR ในลำดับต่อมาได้พัฒนา BUDSIR-II ซึ่งเป็นฉบับอักษรโรมันได้สำเร็จ ด้วยชุดซอฟต์แวร์ทั้งสอง ผู้ใช้จะสามารถค้นหา คำวลี หรือ ข้อความใด ๆ ในพระไตรปิฎกมาได้อย่างรวดเร็ว ถูกต้อง และครบถ้วนสมบูรณ์

แม้ว่า BUDSIR และ BUDSIR-II ใช้ในการสืบค้น คำ วลี ในพระไตรปิฎกได้อย่างมีประสิทธิภาพก็ตาม แต่ยังมีข้อจำกัดในการสืบค้น ที่ไม่อาจสนองความต้องการค้นในระดับที่ซับซ้อนของผู้ใช้ได้ จึงเป็นปัญหาที่ต้องศึกษาหาแนวทางแก้ไขปรับปรุงที่เหมาะสม และทำการพัฒนาระบบใหม่

งานวิจัยนี้ ได้พัฒนาชุดซอฟต์แวร์ชื่อ BUDSIR-III ซึ่งได้ขยายความสามารถด้านการสืบค้นออกไปให้ใช้ค้นพระไตรปิฎกได้อย่างกว้างขวางมากกว่าชุดซอฟต์แวร์รุ่นก่อน และในการค้นพระไตรปิฎกนั้น BUDSIR-III ได้จัดให้ผู้ใช้งานโต้ตอบกับระบบด้วยตัวเอง โดยใช้ชุดคำสั่งสอบถาม ซึ่งคำสั่งสอบถามเบื้องต้น คือ ปฏิบัติการแบบบูล ได้แก่ AND, OR, NOT นอกจากนี้ ยังมี ADJ หรือปฏิบัติการค้นหาคำประชิด และชุดคำสั่งอรรถประโยชน์ต่าง ๆ ยิ่งกว่านั้น ด้วยหลักการพื้นฐานของ BUDSIR-III ที่มีการเก็บบันทึกสารสนเทศผลลัพธ์ไว้ในแฟ้ม F-Set จึงทำให้สามารถขยายขอบข่ายของการค้นออกไปได้อย่างกว้างขวาง ซึ่งจะตอบสนองความต้องการค้นของผู้ใช้ในระดับที่ซับซ้อนได้ดี รวมทั้งมีลักษณะที่ยืดหยุ่น และสามารถพลิกแพลงให้ใช้เก็บผลลัพธ์ในรูปแบบลักษณะต่าง ๆ ได้ดีเช่นกัน

ในการวิจัย ได้ยกตัวอย่างการสืบค้นในลักษณะที่ซับซ้อนไว้หลายรูปแบบ ซึ่งแสดงให้เห็นถึงความสามารถ ความยืดหยุ่น ตลอดจนประสิทธิภาพในด้านการสืบค้นของ BUDSIR-III ได้อย่างชัดเจน

ภาควิชาวิศวกรรมคอมพิวเตอร์.....
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์.....
ปีการศึกษา2532.....

ลายมือชื่อนิสิตจรรยา เกียรติไพบูลย์.....
ลายมือชื่ออาจารย์ที่ปรึกษาสุเมธ วัชรชัยสุรพล.....
ลายมือชื่ออาจารย์ที่ปรึกษาร่วมสุเมธ วัชรชัยสุรพล.....



JIRAPORN KIATPAIBOON : QUERY LANGUAGE DEVELOPMENT FOR TRI PITAKA INFORMATION RETRIEVAL SYSTEM. THESIS ADVISOR : ASSOC.PROF. SUPACHAI TANGWONGSAN, PH.D., ASST.PROF. SUMET VACHARACHAISURAPOL, 116 PP. ISBN 974-577-387-5

The Tripitaka or Pali Canon is the collection of the principal teachings in Buddhism which Buddhist monks and followers are required to study and practice in accordance with the Doctrine and the Discipline. Moreover, scholars in various fields are also enthusiastic about learning the Pali Canon for their research interests. In the endeavour to pursue a particular subject in the Tripitaka, however, not only does one have to overcome the barrier of the Pali language, but one must also encounter overwhelming amount of information consisting of 24,23 million characters in 45 printed volumes which make it extremely difficult to retrieve the information in question effectively.

The first attempt to computerize the entire Tripitaka was initiated by Mahidol University in 1988. The computerized Tripitaka, together with the software named BUDSIR were implemented on IBM personal computer. With BUDSIR, and the later version BUDSIR-II in Romanized Pali, users are able to search whatever words, phrases or even themes in this huge database quickly, accurately and exhaustively.

Although BUDSIR and BUDSIR-II greatly facilitate the search process in Tripitaka, however, they cannot serve for complex queries. There are problems that require further investigation and system development.

This research presents the development of the software, called BUDSIR-III, which greatly enhances the search features and capability, compared to the previous versions. BUDSIR-III enables the user to search Tripitaka through a most effective dialogue with the system via a set of query commands. Basic query commands are Boolean operators such as AND, OR, NOT. Others are ADJ for word adjacency search, and utility commands. Moreover the introduction of F-Set files in BUDSIR-III provides users a broad spectrum of search capabilities as well as flexibility and various output options.

Examples of complex search queries are presented to demonstrate the BUDSIR-III capability, flexibility and performance as well.

ภาควิชา วิศวกรรมคอมพิวเตอร์
สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์
ปีการศึกษา 2532

ลายมือชื่อนิติต ศาสตราจารย์ ดร. สุพจน์ วัชรราชสุรพอล
ลายมือชื่ออาจารย์ที่ปรึกษา
ลายมือชื่ออาจารย์ที่ปรึกษาช่วย
.....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้ ได้สำเร็จลุล่วงไปได้ด้วยความช่วยเหลือให้คำแนะนำอย่างดียิ่งของรองศาสตราจารย์ ดร. ศุภชัย ตั้งวงศ์ศานต์ และ ผู้ช่วยศาสตราจารย์ สุเมธ วัชรชัยสุรพล อาจารย์ที่ปรึกษาวิทยานิพนธ์ โดยเฉพาะท่านรองศาสตราจารย์ ดร. ศุภชัย ตั้งวงศ์ศานต์ ได้กรุณาให้คำแนะนำหลักการเขียนวิทยานิพนธ์ รวมทั้งภาษาที่ใช้ในการเขียนอย่างละเอียด ช่างเจ้ารู้สึกซาบซึ้งในความกรุณาของท่านเป็นอย่างยิ่ง

การวิจัยครั้งนี้ได้ใช้ฐานข้อมูลพระไตรปิฎก ของมหาวิทยาลัยมหิดล และใช้อุปกรณ์บริษัทของสำนักคอมพิวเตอร์ มหาวิทยาลัยมหิดล โดยเจ้าหน้าที่ของสำนักคอมพิวเตอร์ได้อำนวยความสะดวกแก่ช่างเจ้าเป็นอย่างดี จึงขอขอบพระคุณอย่างสูงมา ณ ที่นี้

จิรากร เกียรติไพบูลย์



สารบัญ

	หน้า
บทคัดย่อภาษาไทย	ง
บทคัดย่อภาษาอังกฤษ	จ
กิตติกรรมประกาศ	ฉ
สารบัญตาราง	ญ
สารบัญภาพและผังงาน	ฎ
บทที่	
1 บทนำ	1
1.1 การพัฒนาบูตเซอ์	1
1.2 ข้อจำกัดในระบบสืบค้นสารสนเทศของบูตเซอ์และแนวทางใหม่	3
1.3 การออกแบบและพัฒนาระบบสืบค้นสารสนเทศพระไตรปิฎกระบบใหม่	5
1.4 ขอบเขตของการวิจัยและเวลาที่ใช้	8
2 พระไตรปิฎกแบบคอมพิวเตอร์ (BUDSIR)	10
2.1 ประวัติความเป็นมา	10
2.2 โครงสร้างของภาษาบาลีและการจัดเก็บเข้าคอมพิวเตอร์	11
2.2.1 โครงสร้างของภาษาบาลี	11
2.2.2 การจัดเก็บข้อมูลภาษาบาลีเข้าคอมพิวเตอร์	12
2.3 บูตเซอ์ (BUDSIR)	15
2.3.1 โครงสร้างของบูตเซอ์	15
2.3.2 ระบบสืบค้นสารสนเทศของบูตเซอ์	17
2.4 ปัญหาและขีดจำกัดในการค้นของบูตเซอ์	19
3 หลักการและแนวความคิดเกี่ยวกับระบบสืบค้นสารสนเทศ	23
3.1 หลักการของระบบสืบค้นสารสนเทศ	23
3.2 ฐานข้อมูลและขอบข่ายการค้นของระบบสืบค้นสารสนเทศทั่วไป	24
3.2.1 ระบบ DIALOG	24
3.2.2 ระบบ STAIRS	25

	หน้า
3.2.3 ระบบ MEDLARS	26
3.3 หลักการของขบวนการสืบค้นสารสนเทศกับความต้องการค้นหาของผู้ใช้ ..	27
3.3.1 ความต้องการค้นในระดับพื้นฐาน	27
3.3.2 ความต้องการค้นในระดับที่ซับซ้อน	27
3.3.3 แนวทางสำคัญในการจัดทำขบวนการสืบค้นสารสนเทศ	31
4 การออกแบบระบบสืบค้นสารสนเทศพระไตรปิฎกระบบใหม่	33
4.1 หลักการของระบบสืบค้นสารสนเทศพระไตรปิฎกระบบใหม่	33
4.2 โครงสร้างฐานข้อมูล	34
4.2.1 โครงสร้างฐานข้อมูลพระไตรปิฎก	34
4.2.2 โครงสร้างกลุ่มแฟ้มข้อมูล F-Set	35
4.3 โครงสร้างของขบวนการสืบค้นสารสนเทศพระไตรปิฎกในระบบใหม่	41
4.3.1 ปฏิบัติการแบบบูล	41
4.3.2 ปฏิบัติการประชิด	42
4.3.3 การกำหนดลำดับก่อนหลังของการดำเนินการปฏิบัติการ	48
4.3.4 ขบวนการสืบค้นสารสนเทศพระไตรปิฎก	48
- การแปลงยอคำสั่ง	49
- การค้นคำในโครงสร้างแบบทรี ของแฟ้มพจนานุกรม	50
- การค้นส่วนของคำ (ประเภท 1) ในแฟ้มพจนานุกรม	51
- การค้นส่วนของคำ (ประเภท 2) ในแฟ้มพจนานุกรม	54
- พังงานของขบวนการสืบค้นสารสนเทศของ BUDSIR-III ..	56
5 ผลของการพัฒนา BUDSIR-III	59
5.1 ระบบสอบถามของ BUDSIR-III	59
5.1.1 เมนูหลักของระบบสอบถาม	59
5.1.2 คำสั่งต่าง ๆ ของระบบสอบถาม	62
- คำสั่ง HELP	63
- คำสั่ง DISPLAY	63
- คำสั่ง EXPAND	66
- คำสั่ง FLUSH	70
- คำสั่ง PRINT	70

	หน้า
- คำสั่ง READ	70
- คำสั่ง SDISPLAY	76
- คำสั่ง SEARCH	83
- คำสั่ง PHRASE	84
5.2 ตัวอย่างการใช้งานระบบสอบถาม	91
5.3 คุณลักษณะเฉพาะของระบบ BUDSIR-III	96
5.3.1 กลุ่มโปรแกรมระบบสอบถามของ BUDSIR-III	96
5.3.2 คุณสมบัติเฉพาะของคอมพิวเตอร์ที่ใช้งาน	96
5.4 ความเร็วในการค้นของ BUDSIR-III	96
6 สรุปผลการวิจัยและข้อเสนอแนะ	98
6.1 สรุปผลการวิจัย	98
6.1.1 ขอบข่ายการค้นของ BUDSIR-III	98
6.1.2 ความเร็วในการค้นของ BUDSIR-III	99
6.1.3 ความถูกต้องในการค้นของ BUDSIR-III	99
6.1.4 การใช้เนื้อที่เก็บสารสนเทศผลลัพธ์	100
6.1.5 ฝั่งชั้นช่วยงานของ BUDSIR-III	100
6.2 ข้อเสนอแนะ	101
6.2.1 การอัดข้อมูล	101
6.2.2 การแปลภาษาโปรแกรมด้วยเครื่อง	102
6.2.3 ระบบผู้เชี่ยวชาญ	102
6.2.4 พระไตรปิฎกภาษาบาลีฉบับนานาชาติ	103
บรรณานุกรม	104
ประวัติผู้เขียน	105



สารบัญตาราง

	หน้า
รูปที่ 2.1 ตาราง ASCII สำหรับตัวอักษรไทย-บาลี-อังกฤษ	13
รูปที่ 5.17 ตารางแสดงความเร็วในการค้นด้วยรูปแบบต่าง ๆ ของ BUDSIR-III ..	97



สารบัญภาพและแผนผังงาน

	หน้า
รูปที่ 2.2 ตัวอย่างข้อความภาษาบาลีอักษรไทยที่เก็บเข้าคอมพิวเตอร์	14
รูปที่ 2.3 โครงสร้างฐานข้อมูลพระไตรปิฎก	16
รูปที่ 2.4 ผังงานของขบวนการสืบค้นสารสนเทศของ BUDSIR	18
รูปที่ 3.1 โครงสร้างฐานข้อมูลของระบบ STAIRS	25
รูปที่ 3.2 โครงสร้างฐานข้อมูลของระบบ MEDARS	26
รูปที่ 4.1 แผนผังแสดงหลักการของ BUDSIR-III	33
รูปที่ 4.2 ตัวอย่างการเชื่อมโยงของคำศัพท์ในโครงสร้างแบบทรี	34
รูปที่ 4.3 โครงสร้างของกลุ่มแฟ้มข้อมูล F-Set และข้อมูลที่เก็บ	38
รูปที่ 4.4 ภาพแสดงการเพิ่มระเบียบเข้าแฟ้มไดเรกทอรีของ F-Set	39
รูปที่ 4.5 ภาพแสดงการลบระเบียบออกจากแฟ้มไดเรกทอรีของ F-Set	40
รูปที่ 4.6 ผังงานของปฏิบัติการ AND ระหว่าง F-Set #21 และ #22	43
รูปที่ 4.7 ผังงานของปฏิบัติการ OR ระหว่าง F-Set #21 และ #22	44
รูปที่ 4.8 ผังงานของปฏิบัติการ NOT ระหว่าง F-Set #21 และ #22	45
รูปที่ 4.9 ผังงานของปฏิบัติการ ADJ ระหว่าง F-Set #21 และ #22	46-47
รูปที่ 4.10 ภาพจำลองวิธีการไล่ค้นหาคำศัพท์ในโครงสร้างแบบทรี	51
รูปที่ 4.11 แผนผังของ KMP ที่สร้างจากสายอักษร "กุลกมา"	52
รูปที่ 4.12 แสดงการกำหนดค่า FLINK ในกรณีมีกลุ่มอักษรซ้ำกันในสายอักษร	53
รูปที่ 4.13 แผนผังของ KMP แบบปรับปรุง ที่สร้างจากสายอักษร "กุลก?มา"	55
รูปที่ 4.14 ผังงานของขบวนการสืบค้นสารสนเทศพระไตรปิฎกในระบบใหม่	57-58
รูปที่ 5.1 จอภาพของเมนูหลัก	60
รูปที่ 5.2 ผังงานของโปรแกรมหลัก	61
รูปที่ 5.3 ผังงานของคำสั่ง HELP	64
รูปที่ 5.4 จอภาพของคำสั่ง HELP	65
รูปที่ 5.5 ผังงานของคำสั่ง DISPLAY	67
รูปที่ 5.6 จอภาพต่าง ๆ ของคำสั่ง DISPLAY	68-69
รูปที่ 5.7 ผังงานของคำสั่ง EXPAND	71
รูปที่ 5.8 จอภาพต่าง ๆ ของคำสั่ง EXPAND	72-73

	หน้า
รูปที่ 5.9	ผังงานของคำสั่ง FLUSH 74
รูปที่ 5.10	ผังงานของคำสั่ง PRINT 75
รูปที่ 5.11	ผังงานของคำสั่ง READ 77-78
รูปที่ 5.12	จอภาพต่าง ๆ ของคำสั่ง READ 79-81
รูปที่ 5.13	ผังงานของคำสั่ง SETS-DISPLAY 82
รูปที่ 5.14	ผังงานของคำสั่ง PHRASE 85
รูปที่ 5.15	จอภาพต่าง ๆ ของคำสั่ง PHRASE 86-90
รูปที่ 5.16	แสดงการใช้คำสั่งในระบบสอบถามของ BUDSIR-III และผลลัพธ์ ... 92-95