

190941

ห้องสมุดงานวิจัย สำนักงานคณะกรรมการการวิจัยแห่งชาติ



190941

รายงานการวิจัย

การศึกษาและวิจัยการทำงานของระบบค้นคืนสารสนเทศโดยใช้อัลกอริทึม VIPS  
A Study of Information Retrieval System Using Vision based Pages  
Segmentation (VIPS) Algorithms

นางสาวสุธีรา พันธุ์ธีรานุรักษ์

ได้รับทุนสนับสนุนงานวิจัยจากเงินรายได้ ประจำปีงบประมาณ 2554

คณะวิศวกรรมศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

6 00 256047

ห้องสมุดงานวิจัย สำนักงานคณะกรรมการการวิจัยแห่งชาติ



190941

## รายงานการวิจัย

การศึกษาและวิจัยการทำงานของระบบค้นคืนสารสนเทศโดยใช้อัลกอริทึม VIPS  
A Study of Information Retrieval System Using Vision based Pages  
Segmentation (VIPS) Algorithms



นางสาวสุธีรา พันธุ์ธีรานุรักษ์

ได้รับทุนสนับสนุนงานวิจัยจากเงินรายได้ ประจำปีงบประมาณ 2554

คณะวิศวกรรมศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

## กิตติกรรมประกาศ

โครงการวิจัยนี้สำเร็จลุล่วงได้เป็นอย่างดีโดยได้รับความช่วยเหลือจาก คุณเกศินี ทองตันไตรย์ คุณชญานี จารุพัฒนะสิริกุล และ คุณชนิกตา ธรรมสร่างกูร อดีตนักศึกษาของภาควิชาวิศวกรรมสารสนเทศ คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ที่ช่วยทำการทดลองต่าง ๆ

นางสาวสุธีรา พันธุ์ธำรงค์

ชื่อโครงการ (ภาษาไทย) การศึกษาและวิจัยการทำงานของระบบค้นคืนสารสนเทศโดยใช้อัลกอริทึม VIPS

ชื่อโครงการ (ภาษาอังกฤษ) A Study of Information Retrieval System Using Vision based Pages Segmentation (VIPS) Algorithms

แหล่งเงิน ..... เงินรายได้คณะวิศวกรรมศาสตร์

ประจำปีงบประมาณ ..... 2554 ..... จำนวนเงินที่ได้รับการสนับสนุน ..... 64,900 ..... บาท

ระยะเวลาทำการวิจัย ..... 1 ..... ปี ตั้งแต่ 1 ตุลาคม 2553 ถึง 30 กันยายน 2554

ชื่อ-สกุล หัวหน้าโครงการ และผู้ร่วมโครงการวิจัย พร้อมระบุ หน่วยงานต้นสังกัดและ อีเมล

..... นางสาวสุธีรา พันธุ์ธีรานุรักษ์

..... สาขาวิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์

..... สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง kpsuthee@kmitl.ac.th

คำสำคัญ (Keywords) Information Retrieval System, Web-based Retrieval Systems, Relevant Feedback Algorithms

### บทคัดย่อ

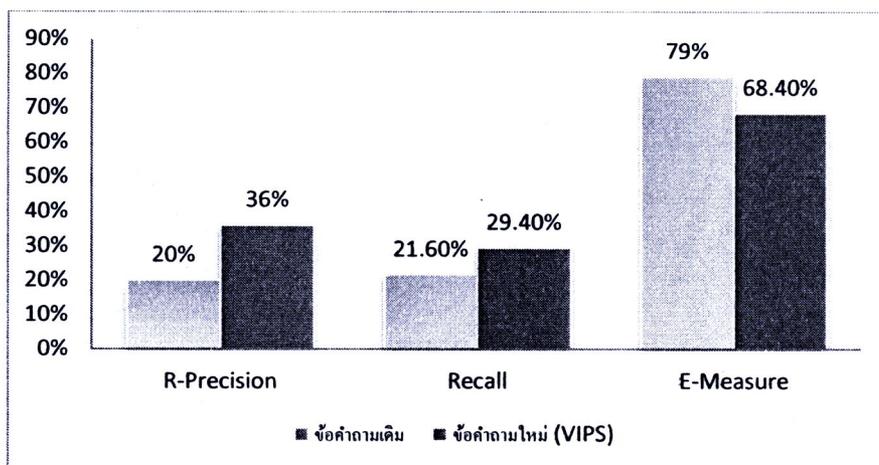
190941

โครงการวิจัยนี้เป็นการศึกษาการทำงานของระบบค้นคืนสารสนเทศบนเว็บ โดยมีการเก็บการค้นคืนย้อนกลับจากผู้ใช้และการใช้อัลกอริทึม Vision Base Page Segmentation (VIPS) เพื่อทำการแบ่งเว็บเพจที่ได้จากการค้นคืนออกเป็นบล็อก และหาคำที่จะนำมาเพิ่มในข้อความเดิมเพื่อสร้างเป็นข้อความใหม่ โดยได้ทำการจำลองระบบการค้นคืนสารสนเทศบนเว็บเพื่อทำการเก็บการค้นคืนย้อนกลับจากผู้ใช้ แล้วนำเว็บเพจที่ผู้ใช้เลือกมาหาข้อความใหม่เปรียบเทียบกับการค้นคืนย้อนกลับจากผู้ใช้แล้วนำเว็บที่ผู้ใช้เลือกไปแบ่งเป็นบล็อกโดยใช้อัลกอริทึมวีไอพีเอส จากนั้นให้ผู้ใช้เลือกบล็อกที่แบ่งได้ซึ่งผู้ใช้เห็นว่าเกี่ยวข้องกับความต้องการอีกครั้งหนึ่ง จากผลการทดลองพบว่าข้อความใหม่ที่เกิดจากการเพิ่มคำ โดยนำอัลกอริทึมวีไอพีเอสมาใช้นั้นทำให้ได้ผลการค้นคืนที่เกี่ยวข้องกับสิ่งที่ผู้ใช้ต้องการมากขึ้น

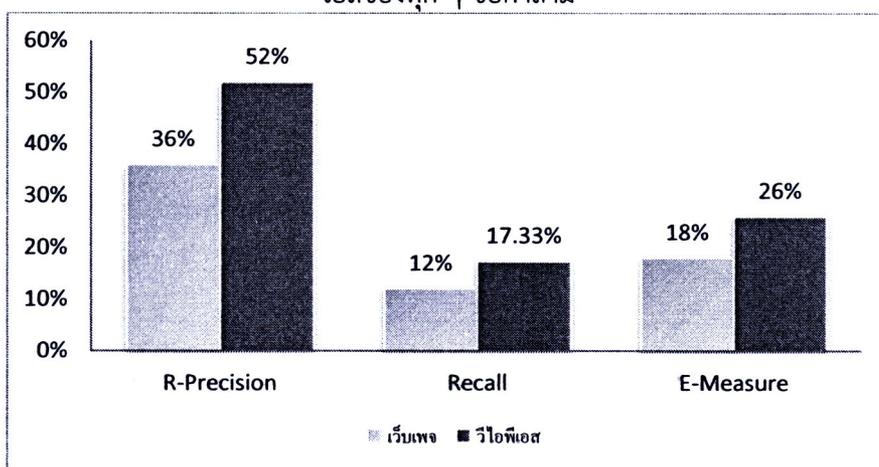
### ABSTRACT

This research project aims to study web search system using relevant feedback and Vision-based Page Segmentation (VIPS) algorithm. Web information retrieval using relevant feedback was simulated to compare results of two experiments. One is result from web search system using relevant feedback. Another is web search system using relevant feedback and VIPS algorithm. In our experiments, we can show that the expansion terms from using VIPS algorithm is meet user requirements more than web search using relevant feedback.

### รูปผลงานวิจัย



รูปที่ 1 กราฟแสดงการเปรียบเทียบประสิทธิภาพผลการค้นคืนย้อนกลับจากข้อคำถามเดิมและการใช้อัลกอริทึมวีไอพีเอสของทุก ๆ ข้อคำถาม



รูปที่ 2 กราฟแสดงการเปรียบเทียบประสิทธิภาพผลการค้นคืนย้อนกลับจากการเพิ่มข้อคำถามที่มาจากทั้งหน้าเว็บเพจและการใช้อัลกอริทึมวีไอพีเอสของทุกๆข้อคำถาม

จากรูปกราฟที่ 1 และรูปกราฟที่ 2 จะเห็นว่าผลการเปรียบเทียบการประเมินประสิทธิภาพของการค้นคืนย้อนกลับจากการเพิ่มข้อคำถามที่มาจากหน้าเว็บเพจนั้นมีค่าน้อยกว่าการใช้อัลกอริทึมวีไอพีเอสในทุกๆค่า ดังนั้นอัลกอริทึมวีไอพีเอสมีส่วนช่วยเพิ่มประสิทธิภาพในการค้นคืนให้ดีขึ้นและได้เว็บเพจที่เกี่ยวข้องกับสิ่งที่ผู้ใช้งานต้องการมากยิ่งขึ้น

# สารบัญ

	หน้า
กิตติกรรมประกาศ .....	I
บทคัดย่อภาษาไทย .....	III
บทคัดย่อภาษาอังกฤษ .....	III
รูปผลงานวิจัย .....	IV
สารบัญ .....	V
สารบัญตาราง .....	VII
สารบัญรูป .....	VIII
บทที่ 1 บทนำ .....	1
1.1 ความสำคัญและที่มาของปัญหา .....	1
1.2 วัตถุประสงค์ .....	2
1.3 ขอบเขตของโครงการวิจัย .....	2
1.4 สมมติฐานของการศึกษา .....	2
1.5 วิธีการดำเนินการวิจัย .....	3
1.6 ประโยชน์ที่คาดว่าจะได้รับ .....	3
บทที่ 2 วิธีการดำเนินการวิจัย .....	4
2.1 ลูซีน .....	4
2.1.1 การประยุกต์ใช้ลูซีนสำหรับการพัฒนาระบบค้นคืนสารสนเทศ (Lucene for IR Application) ..	4
2.1.2 โครงสร้างทางสถาปัตยกรรมของลูซีน (lucene API) .....	5
2.2 การเลือกเทอมของเวกเตอร์โมเดล .....	10
2.2.1 การเลือกเทอม (Sort order) .....	10
2.3 วีไอพีเอส (Vision Based Pages Segmentation) .....	11
2.3.1 โครงสร้างเนื้อหา (Vision Based Content Structure : VIPS) .....	11
2.3.2 ขั้นตอนการทำงานของอัลกอริทึมวีไอพีเอส .....	13
บทที่ 3 ผลการวิจัย .....	20
3.1 ส่วนประกอบของระบบ .....	20

## สารบัญ (ต่อ)

	หน้า
3.1.1 การค้นคืนสารสนเทศ .....	21
3.1.2 การแบ่งเว็บเพจออกเป็นบล็อก .....	23
3.1.3 การหาข้อความใหม่ .....	23
3.2 การเตรียมข้อมูล .....	24
3.3 การทำงานของระบบค้นคืนสารสนเทศบนเว็บโดยใช้การค้นคืนย้อนกลับ .....	26
3.3.1 ขั้นตอนการทำงานของระบบค้นคืนสารสนเทศบนเว็บ .....	27
3.3.2 ขั้นตอนการทำงานของระบบค้นคืนย้อนกลับ .....	28
3.3.3 ขั้นตอนการทำงานของระบบหาข้อความใหม่ .....	29
3.4 การทำงานของระบบค้นคืนสารสนเทศบนเว็บที่ใช้อัลกอริทึมวีไอพีเอสและการค้นคืนย้อนกลับ .....	30
3.4.1 ขั้นตอนการทำงานของโปรแกรมวีไอพีเอส .....	30
3.4.2 ขั้นตอนการทำงานของโปรแกรมบันทึกไฟล์ของแต่ละบล็อก .....	32
3.4.3 ขั้นตอนการทำงานของระบบหาข้อความใหม่ .....	33
3.5 การสร้างไฟล์ดัชนี .....	35
บทที่ 4 ผลการทดสอบระบบ .....	37
4.1 ผลการทดลอง .....	37
4.1.1 การทดลองที่ 1 เปรียบเทียบข้อความใหม่ที่มาจากทั้งหน้าเว็บเพจและข้อความใหม่ที่มาจากการใช้อัลกอริทึมวีไอพีเอส .....	37
4.1.2 การทดลองที่ 2 วัดประสิทธิภาพของระบบจากค่า R-Precision, Recall และ E-Measure .....	38
4.2 สรุปผลการทดลอง .....	40
บทที่ 5 สรุปผลการวิจัยและข้อเสนอแนะ .....	43
5.1 สรุปผลการวิจัย .....	43
5.2 ข้อเสนอแนะ .....	43
5.3 แนวทางในการพัฒนาต่อ .....	43

## สารบัญ (ต่อ)

	หน้า
เอกสารอ้างอิง.....	44

## สารบัญตาราง

ตารางที่	หน้า
2.1 ตารางระบุความหมายของฟิลด์อินเด็กซ์.....	6
2.2 ตารางระบุความหมายของฟิลด์สไตร์.....	6
2.3 กฎในการแบ่งเว็บเพจออกเป็นบล็อก.....	14
2.4 กฎที่แตกต่างกันของแท็กที่แตกต่างกัน.....	15
4.1 ข้อคำถามใหม่และค่าคะแนนของคำว่า “panda”.....	37
4.2 ข้อคำถามใหม่และค่าคะแนนของคำว่า “aids”.....	37
4.3 ข้อคำถามใหม่และค่าคะแนนของคำว่า “java”.....	37
4.4 ข้อคำถามใหม่และค่าคะแนนของคำว่า “sushi”.....	38
4.5 ข้อคำถามใหม่และค่าคะแนนของคำว่า “titanic”.....	38
4.6 การเปรียบเทียบผลการประเมินของคำว่า “panda”.....	38
4.7 การเปรียบเทียบผลการประเมินของคำว่า “aids”.....	39
4.8 การเปรียบเทียบผลการประเมินของคำว่า “java”.....	39
4.9 การเปรียบเทียบผลการประเมินของคำว่า “sushi”.....	39
4.10 การเปรียบเทียบผลการประเมินของคำว่า “titanic”.....	39
4.11 การประเมินผลการค้นคืนย้อนกลับของคำว่า “panda”.....	39
4.12 การประเมินผลการค้นคืนย้อนกลับของคำว่า “aids”.....	39
4.13 การประเมินผลการค้นคืนย้อนกลับของคำว่า “java”.....	40
4.14 การประเมินผลการค้นคืนย้อนกลับของคำว่า “sushi”.....	40
4.15 การประเมินผลการค้นคืนย้อนกลับของคำว่า “titanic”.....	40
4.16 ตารางแสดงผลการเปรียบเทียบประสิทธิภาพผลการค้นคืนย้อนกลับจากข้อคำถามเดิมและการใช้อัลกอริทึมวีไอพีเอสของทุก ๆ ข้อคำถาม.....	40
4.17 ตารางแสดงผลการเปรียบเทียบประสิทธิภาพผลการค้นคืนย้อนกลับจากการจากเพิ่มข้อคำถามที่มาจากทั้งหน้าเว็บเพจและการใช้อัลกอริทึมวีไอพีเอสของทุก ๆ ข้อคำถาม.....	41

# สารบัญรูป

รูปที่	หน้า
1.1 ระบบค้นคืนสารสนเทศ .....	1
1.2 ระบบค้นคืนสารสนเทศที่มีการค้นคืนย้อนกลับจากผู้ใช้ (Relevance Feedback).....	3
2.1 ภาพแสดงการนำลูซินไปประยุกต์ใช้ในระบบค้นคืนสารสนเทศ .....	4
2.2 การทำงานระหว่างสถาปัตยกรรมหลักของลูซิน.....	5
2.3 การเก็บข้อมูลแบบเพิ่มข้อมูลผกผัน .....	7
2.4 ขั้นตอนการค้นคืน .....	8
2.5 ตัวอย่างหน้าเว็บของ Yahoo! Shopping Auctions และโครงสร้างของหน้าเว็บเพจที่ถูกแบ่ง .....	12
2.6 โครงสร้างเนื้อหาของเว็บ Yahoo! Shopping Auctions .....	13
2.7 ขั้นตอนการทำงานของอัลกอริทึมวีไอพีเอส .....	13
2.8 แสดงตัวอย่างการตัดออกเป็นบล็อกของตัวอย่างเว็บ .....	16
2.9 แสดงตัวอย่างการหาขั้นตอนการหาตัวแบ่ง.....	17
2.10 (a) แสดงคอมทรี (b) แสดงส่วนย่อยของหน้าเพจ (c) แสดงตัวแบ่งและค่าน้ำหนักระหว่างแต่ละบล็อก .....	18
2.11 แสดงตัวอย่างของการสร้างโครงสร้างเนื้อหา .....	19
3.1 แผนภาพทำงานของระบบค้นคืนสารสนเทศบนเว็บโดยใช้การค้นคืนย้อนกลับ .....	20
3.2 แผนภาพทำงานของระบบค้นคืนสารสนเทศบนเว็บโดยใช้ อัลกอริทึมวีไอพีเอสและการค้นคืนย้อนกลับ.....	21
3.3 การค้นคืนสารสนเทศโดยลูซิน .....	21
3.4 แสดงโปรแกรมเว็บสฟิงซ์.....	24
3.5 กำหนดการใช้งานของแทป Pages .....	25
3.6 แสดงการทำงานของเว็บสฟิงซ์ .....	25
3.7 หน้าเว็บเพจแรกของระบบจำลอง.....	26
3.8 หน้าเว็บเพจรับข้อความจากผู้ใช้.....	26
3.9 ผลการค้นคืนของข้อความ.....	27
3.10 หน้าเว็บเพจให้ผู้ใช้เลือกเว็บเพจที่เกี่ยวข้องกับสิ่งที่ผู้ใช้ต้องการ.....	27
3.11 หน้าเว็บด้านล่างสุดของเพจที่ทำการค้นคืนได้ .....	28
3.12 ผลจากการกดปุ่ม Relevance Page .....	28

## สารบัญรูป (ต่อ)

รูปที่	หน้า
3.13 แสดงยูอาร์แอลที่ผู้ใช้เลือกและการใส่ค่าไดเรกทอรีเพื่อหาข้อความใหม่ .....	29
3.14 ข้อความใหม่จากเว็บเพจที่ผู้ใช้เลือกที่เกี่ยวข้องกับสิ่งที่ผู้ใช้ต้องการ .....	29
3.15 หน้าหลักของระบบจำลอง.....	30
3.16 แสดงการลิงก์ไปยังโปรแกรมวีไอพีเอสและโปรแกรมการบันทึกไฟล์ของแต่ละบล็อก.....	31
3.17 ส่วนติดต่อกับผู้ใช้ของโปรแกรมวีไอพีเอส .....	31
3.18 แสดงการใส่ค่ายูอาร์แอลของเว็บเพจที่ต้องการแบ่งเป็นบล็อก.....	32
3.19 ผลของการแบ่งเว็บเพจเป็นบล็อกจากโปรแกรมวีไอพีเอส.....	32
3.20 โปรแกรมการบันทึกไฟล์ของแต่ละบล็อกโดยใส่ชื่อไฟล์ที่ต้องการแยกไฟล์ของแต่ละบล็อก.....	33
3.21 การทำงานของโปรแกรมการบันทึกไฟล์ของแต่ละบล็อก .....	33
3.22 การใส่ไดเรกทอรีของไฟล์ดัชนีเว็บเพจที่แบ่งเป็นบล็อกแล้ว .....	34
3.23 บล็อกที่เกี่ยวข้องกับข้อความตั้งต้น .....	34
3.24 หน้าเพจหลังจากกดปุ่ม Relevance Box จะแสดงยูอาร์แอลที่ผู้ใช้เลือกและการหาข้อความใหม่ .....	35
3.25 ผลการหาข้อความใหม่จากบล็อกที่ผู้ใช้เลือกที่เกี่ยวข้อง.....	35
3.26 การเข้าไปในไดเรกทอรีของ webapps ของ Apache Tomcat 6.0.....	35
3.27 การทำไฟล์ดัชนีของข้อมูลภายในโปรแกรม Databox และนำไปเก็บใน โปรแกรม indexdata.....	36
4.1 กราฟแสดงการเปรียบเทียบประสิทธิภาพผลการค้นคืนย้อนกลับจากข้อความเดิมและการใช้... อัลกอริทึมวีไอพีเอสของทุก ๆ ข้อความ .....	41
4.2 กราฟแสดงการเปรียบเทียบประสิทธิภาพผลการค้นคืนย้อนกลับจากการเพิ่มข้อความที่มาจากทั้งหน้าเว็บเพจและการใช้อัลกอริทึมวีไอพีเอสของทุก ๆ ข้อความ.....	42