THE INFORMATION RETRIEVAL SYSTEM
FOR
THE REVENUE DEPARTMENT INTRANET

KANJANA SRISOONTORN

A RESEARCH PROJECT SUBMITTED IN PARTIAL
FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE
(COMPUTER SCIENCE)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2008

ระบบค้นคืนสารสนเทศสำหรับเครือข่ายภายในกรมสรรพากร (THE INFORMATION RETRIEVAL SYSTEM FOR THE REVENUE DEPARTMENT INTRANET)

กาญจนา ศรีสุนทร 4837928 SCCS/M

วท.ม. (วิทยาการคอมพิวเตอร์)

คณะกรรมการควบคุมโครงงานวิจัย : ดำรัส วงศ์สว่าง, Ph.D
ชมทิพ พรพนมชัย, Ph.D

บทคัดย่อ

โครงงานวิจัยฉบับนี้พัฒนาระบบค้นคืนสารสนเทศของเครือข่ายภายในกรมสรรพากร โดยได้นำเทคนิคใหม่ของ Information Retrieval มาประยุกต์ใช้กับระบบเดิม กระบวนการพัฒนาประกอบด้วย 2 ส่วน: 1) Text operations และ 2) Searching และ Retrieval Text operations มีขั้นตอนในการประมวลผล 4 ขั้นตอน: 1) วิเคราะห์คำศัพท์ 2) ลบคำ stopwords 3) เลือกคำ keywords และ 4) สร้าง index ให้กับ keywords สำหรับส่วนของการค้นมีการค้นอยู่ 4 แบบ คือ 1) การค้นด้วยภาษาธรรมชาติ 2) การค้นแบบกำหนดเงื่อนไข 3) การค้นที่ให้ความสำคัญของคำที่ค้นและ 4) การค้นจาก Text โดยตรงหรือการค้นคำที่มีการคีย์ผิด ในการทดลองได้ประมวลผลข้อมูลเอกสารจำนวน 600 เอกสาร ประกอบด้วย แนวปฏิบัติ 41, คำสั่ง 284, ระเบียบ 38, และประกาศ 237 เอกสาร

ผลการทดลอง แสดงให้เห็นว่าโครงงานวิจัยนี้ สามารถค้นคืนเอกสารที่เกี่ยวข้องและเรียงเอกสารที่เกี่ยวข้องจากมากไปหาน้อย ที่ similarity threshold 90 เปอร์เซ็นต์ ได้ precision เฉลี่ยเท่ากับ 76.55 เปอร์เซ็นต์และ recall เฉลี่ยเท่ากับ 25.52 เปอร์เซ็นต์ คาดหวังว่าระบบที่พัฒนาขึ้นนี้จะทำให้ระบบค้นคืนสารสนเทศสำหรับเครือข่ายภายในกรมสรรพากรมีประสิทธิภาพ

98 หน้า

THE INFORMATION RETRIEVAL SYSTEM FOR THE REVENUE DEPARTMENT INTRANET

KANJANA SRISOONTORN 4837928 SCCS/M

M.Sc.(COMPUTER SCIENCE)

RESEARCH PROJECT ADVISORS: DAMRAS WONGSAWANG, Ph.D,
CHOMTIP PORNPANOMCHAI, Ph.D

ABSTRACT

This project proposes an information retrieval system for the Revenue Department intranet(IRRD). The proposed system employs modern information retrieval techniques to improve the legacy system. The process is divided into 2 parts: 1) Text operations, 2) Searching and retrieval. Text operations consist of four steps: a) lexical analysis, b) elimination of stopwords, c) index terms selection, and d) creating index. Searching and retrieval consist of four types: 1) Natural language search, 2) Boolean model search, 3) Vector model search, and 4) Exact string matching or approximate string matching. Experiments are performed by using 600 HTML file, which is composed of 41 Policies, 284 Instructions, 38 Regulations, and 237 Notifications.

The experimental results showed that the proposed system can search and retrieve the relevant documents and rank them by their relevancies from the most to the least. We get an average precision of 76.55%, and average recall of 25.52% at a similarity threshold of 90%. It is expected that the system developed will contribute to effective searching for the information retrieval system of the Revenue Department Intranet.

KEY WORDS: INFORMATION RETRIEVAL / TEXT OPERATION /
BOOLEAN MODEL / VECTOR MODEL /
EXACT STRING MATCHING / APPROXIMATE STRING MATCHING

98 pp.