Content-based Probabilistic Text Classifier for Pornorgraphic Web Filtering

งานวิจัยนี้เป็นส่วนหนึ่งในโครงงานวิจัยเรื่อง

Automated Obscenity Web Sites Filtering Systems

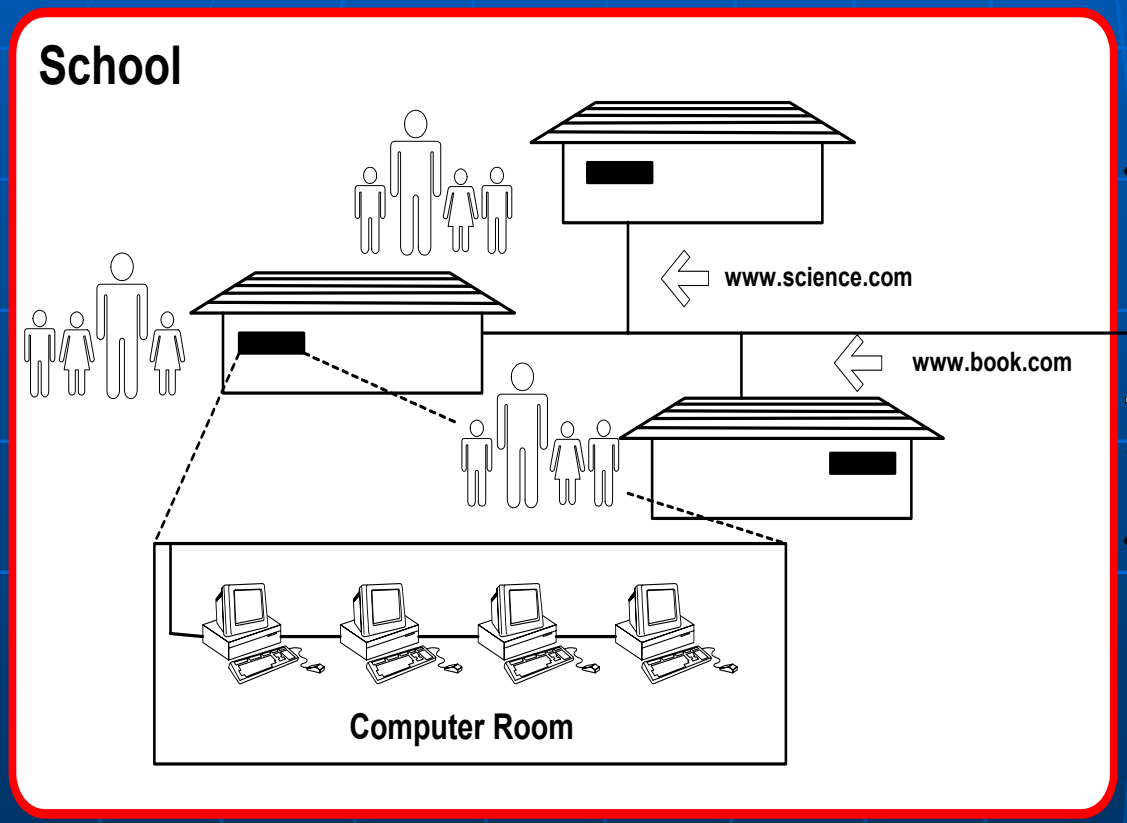
โดย

จันทิมา พลพินิจ
ชุมศักดิ์ สีบุญเรือง
รพีพร ช่ำชอง
อนิรุทธ์ โชติถนอม
สมนึก พ่วงพรพิทักษ์

คณะวิทยาการสารสนเทศ
มหาวิทยาลัยมหาสารคาม

การประชุมวิชาการสวทช. 2548 วิทยาศาสตร์และเทคโนโลยีไทยสู่เศรษฐกิจโมเลกุล วันที่ 28-30 มีนาคม 2548

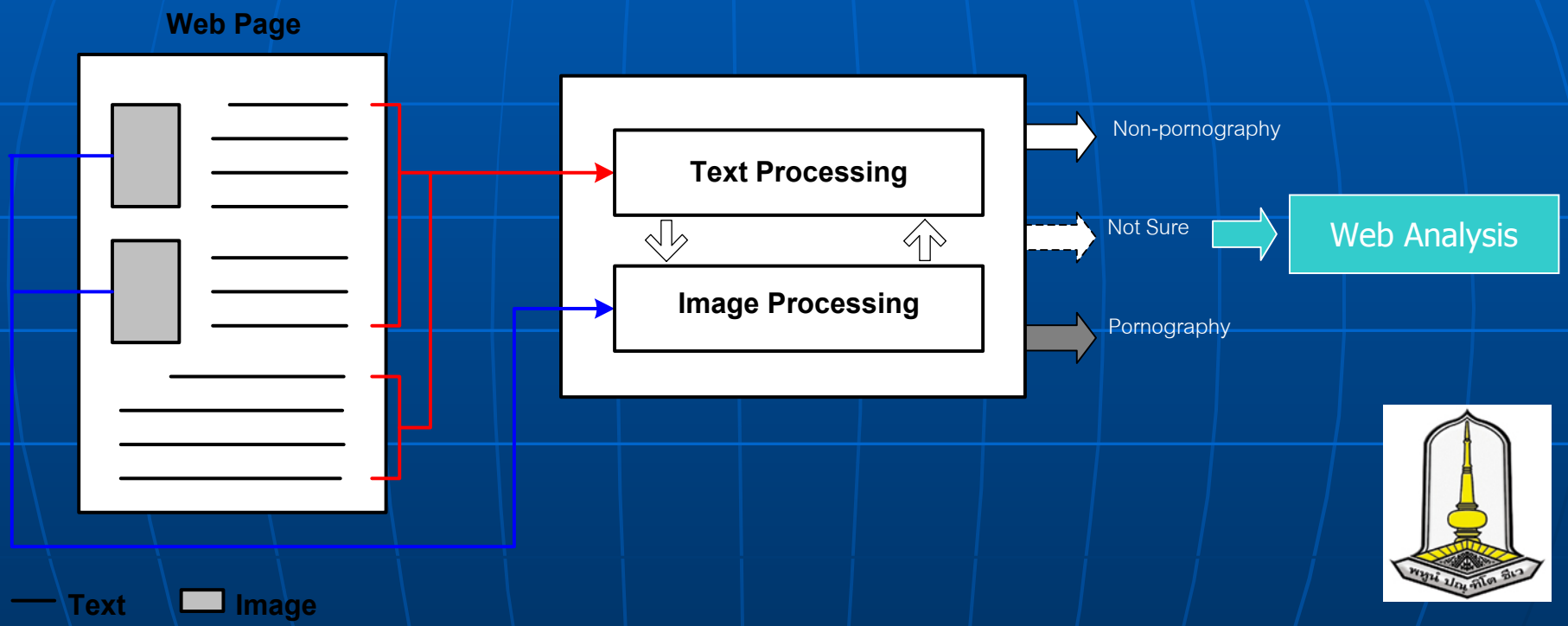
Automated Obscenity Web Sites Filtering System



การประชุมวิชาการสวทช. 2548 วิทยาศาสตร์และเทคโนโลยีไทยสู่เศรษฐกิจโมเลกุล วันที่ 28-30 มีนาคม 2548

Informatics Research Laboratory, Faculty of Informatics, Mahasarakham University

Automated Obscenity Web Sites Filtering System



หัวข้อที่นำเสนอ

- ที่มาของปัญหา (Background)
- พื้นฐานของระบบการกรองเว็บ (Web Filtering)
- งานวิจัยที่เกี่ยวข้อง (Related Work)
- กระบวนการทำงานวิจัย (Research Methodology)
- ชุดข้อมูลทดสอบและผลการวิจัย
- สรุปการวิจัย (Conclusion)

ที่มาของปัญหา (Background)

- เนื่องจากจำนวนเว็บอนาจารได้เพิ่มขึ้นอย่างมาก ดังนั้นการค้นหาเครื่องมือสำหรับป้องกันผู้คนจากเว็บที่ไม่เหมาะสมดังกล่าวจึงได้รับความสนใจมากขึ้น
- โดยทั่วไปแล้วเทคนิคการกรองเว็บ (Web Filtering) เป็นเครื่องมือที่สำคัญที่สามารถควบคุมสารสนเทศบนเว็บได้
- การประยุกต์ใช้งานของเทคนิคการกรองเว็บสามารถแบ่งได้เป็น 2 วิธีคือ เทคนิคการกรองเว็บโดยอาศัยเมตะดาต้า (Metadata) และ เทคนิคการกรองเว็บโดยอาศัยเนื้อหาของเว็บ (Content)
- งานวิจัยนี้ได้นำเสนอการกรองเว็บบนพื้นฐานของ *Content* โดยอาศัยข้อความที่ปรากฏบนหน้าเว็บเป็นปัจจัยสำคัญในการสร้างโมเดลเพื่อการกรองเว็บอนาจาร แบบอัตโนมัติ

พื้นฐานของระบบการกรองเว็บ (Web Filtering)

โดยทั่วไประบบการกรองเว็บจะแบ่งออกเป็น 2 ลักษณะคือ

1. Metadata based เป็นระบบการกรองเว็บโดยอาศัยข้อมูลหรือสารสนเทศจากภายนอกหน้าเว็บ เช่น

- การใช้ข้อมูลของผู้สร้างเว็บ (Author)
- ลิงค์ (Link)
- URL ระหว่างเอกสารหรือหน้าเว็บ
- IP Address

2. Content Based เป็นระบบการกรองเว็บที่จะใช้สารสนเทศบนหน้าเว็บ เช่น ข้อความ รูปภาพ เป็นส่วนสำคัญในการสร้างโมเดลหรือตัวกรอง (Filter) ตัวอย่างระบบที่ใช้การกรองแบบเนื้อหา เช่น

- Keyword

งานวิจัยที่เกี่ยวข้อง

- โครงการ NetProject (2001) ได้ประเมินประสิทธิภาพของระบบกรองเว็บไซต์ในเชิงพาณิชย์ที่มีใช้อยู่ในปัจจุบัน โดยใช้ 2,794 URLs ที่มี Pornorgraphic content และ 1,655 URLs ที่เป็น Normal content

- ตารางที่ 1 แสดงตัวอย่างความถูกต้องแม่นยำบางตัวอย่างของระบบกรองเว็บไซต์ในเชิงพาณิชย์ที่มีใช้อยู่ในปัจจุบันภายใต้โครงการ NetProject

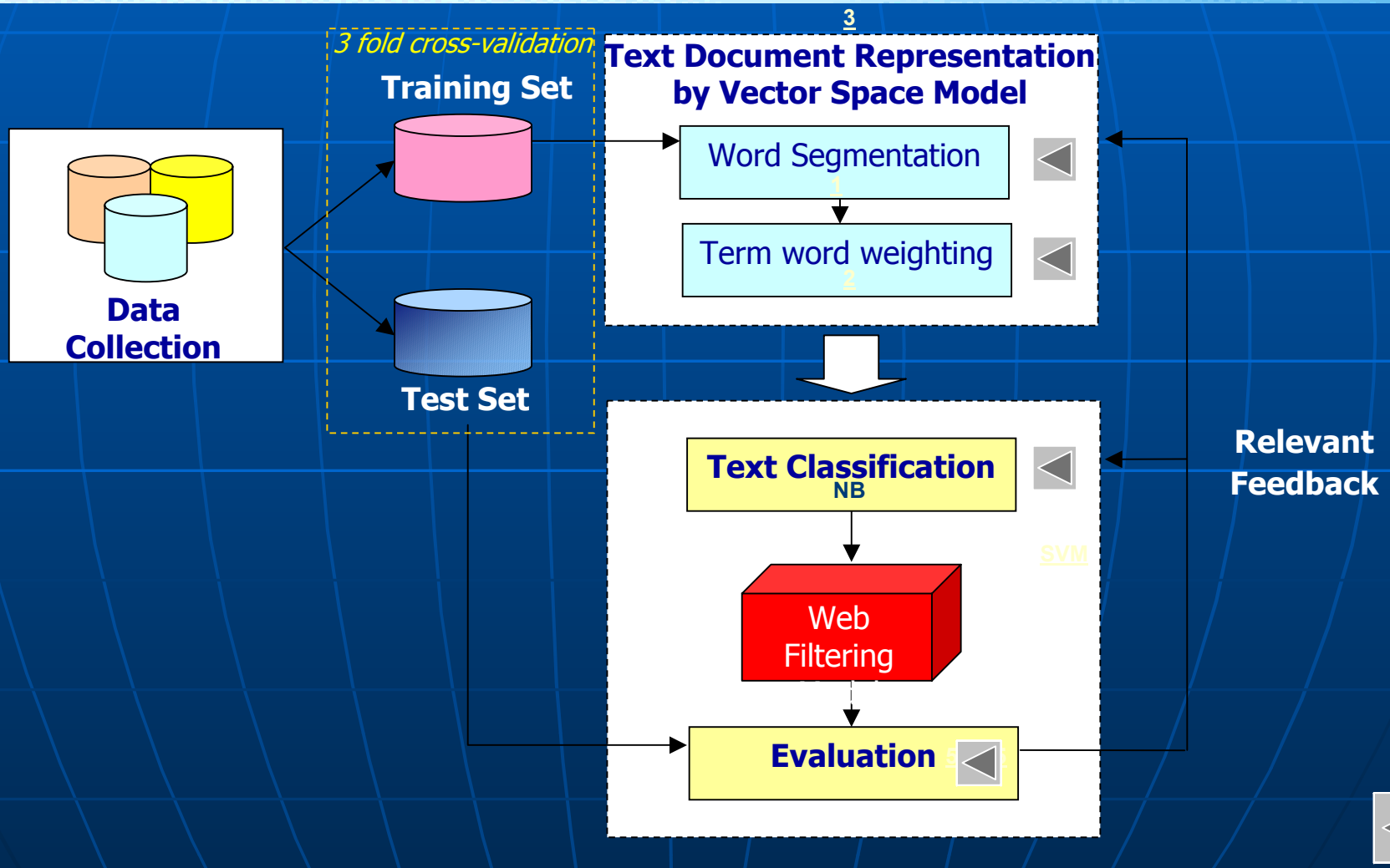
- P. Y. Lee, S. C. Hui, A. Cheuk and M. Fong. "Neural Networks for Web Content Filtering". IEEE Intelligent Systems. 2002.

- R. Du, R. Safavi-Naini and W. Susilo. "Web Filtering Using Text Classification", The 11th IEEE Int. Conf. on Network (ICON 2003), pp. 325 - 330, 2003.

ตารางที่ 1 การประเมินประสิทธิภาพของระบบกลั่นกรองเว็บไซต์ในเชิงพาณิชย์ที่มีใช้อยู่ในปัจจุบันโดยโครงการ NetProject

Filtering Tools	Blocking Effectiveness	Overblocking Rate
BizGuard	55 %	10 %
Cyber Patrol	52 %	2 %
CTBER sitter	46 %	3 %
Cyber Snoop	65 %	23 %
Internet Watcher 2000	30 %	0 %
Net Nanny	20 %	5 %
Norton Internet Security	45 %	6 %
Optenet	79 %	25 %
SurfMonkey	65 %	11 %
X-Stop	65 %	4 %

กระบวนการวิจัย



การตัดคำ (Word Segmentation)

- การตัดคำ คือ การแบ่งตัวอักขระจากข้อความ (String) เพื่อหาขอบเขตของแต่ละหน่วยคำ(Morpheme)
- การตัดคำข้อความภาษาอังกฤษจะง่ายกว่าการตัดตำข้อความภาษาไทย โดยจะใช้ช่องว่างคั่นในการแยกคำ แต่สำหรับภาษาไทย ส่วนใหญ่แล้วจะมีการเขียนในลักษณะที่ติดกันโดยไม่มีการใช้เครื่องหมายวรรคตอนใดมาคั่น
- การตัดคำข้อความภาษาไทยในงานวิจัยฉบับนี้ได้ใช้การตัดคำแบบพจนานุกรมโดย*วิธีการเทียบคำที่ยาวที่สุด* (***Longest matching***) ตัวอย่าง
- ก่อนการนำไปหาน้ำหนักคำจะตัด ***Stop words*** ก่อน

ตัวอย่างการตัดคำในภาษาไทยแบบอิงพจนานุกรม ด้วยการเทียบคำที่ยาวที่สุด

นักวิทยาศาสตร์สหรัฐฯบอกเตือนมนุษย์โลกอย่างน่าเสียวสันหลังว่า โลกกำลังก้าวไปสู่อวสานแล้ว ดีแต่ว่า กว่าชีวิตจะมลายสูญสิ้นหมด มันจะยังอยู่อีกนานถึง 500 ล้านปี

↓

Word Segmentation for Thai Text

↓

นัก / วิทยาศาสตร์ / สหรัฐ / ฯ / บอก / เตือน / มนุษย์ / โลก / อย่าง / น่า / เสียว / สันหลัง / ว่า / โลก / กำลัง / ก้าว / ไป / สู่ / อวสาน / แล้ว / ดี / แต่ว่า / กว่า / ชีวิต / จะ / มลาย / สูญ / สิ้น / หมด / มัน / จะ / ยัง / อยู่ / อีก / นาน / ถึง / 500 / ล้าน / ปี

การหาความถี่และน้ำหนักของคำ (Term Frequency and Weight)

- นำคำที่ตัดทุกคำมาหา*ความถี่* (*Term Frequency* : *TF*)
- สมการหาความถี่และน้ำหนักคำ

$$TF-IDF = TF \times IDF$$

โดย $|D|$ คือ จำนวนเอกสารทั้งหมด

DF คือ จำนวนเอกสารที่มีคำนั้น ๆ ปรากฏอยู่

IDF คือ ส่วนกลับของเอกสาร ซึ่งหาได้จาก

$$IDF = 1 + log(|D| + DF)$$

ตัวอย่างการหาความถี่และน้ำหนักของคำ
F eight

เป็นความถี่ของคำว่า "นัก" ที่ปรากฏในทุกเอกสาร

$IDF = 1 + log(|D|/DF)$
$= 1 + \log(27/3)$
$= 1.954$

			IDF	TF-IDF
นัก	3	3	1.954	5.863
วิทยาศาสตร์				.228
สหรัฐ				.228
บอก	2	2	2.114	4.228
เตือน	2	2	2.114	4.228
มนุษย์	2	2	2.114	4.228

จำนวนเอกสารที่มีคำว่า "นัก" ปรากฏ

$TF\text{-}IDF = TF \times IDF$
$= 3 \times 1.954$
$= 5.863$

นัก / วิทยาศาสตร์ / สหรัฐ / ฯ / บอก / เตือน / มนุษย์ / โลก / อย่าง / น่า / เสียว / สันหลัง / ว่า / โลก / กำลัง / ก้าว / ไป / สู่ / อวสาน / แล้ว / ดี / แต่ว่า / กว่า / ชีวิต / จะ / มลาย / สูญ / สิ้น / หมด / มัน / จะ / ยัง / อยู่ / อีก / นาน / ถึง / 500 / ล้าน / ปี

หมายเหตุ สมมติมีเอกสารทั้งหมด 27 เอกสาร

การวัดประสิทธิภาพ

ความแม่นยำ (Precision: P)

คือโอกาสของเว็บที่กรองได้ตรงกับความต้องการ

$$Precision = \frac{\#\ correct\ classes\ found}{\#\ correct\ found}$$

ความระลึก (Recall: R)

คือ เว็บที่กรองได้ตรงกับความต้องการ ส่วนด้วยเว็บที่กรองได้ถูกต้อง

$$Recall = \frac{\#\ correct\ classes\ found}{\#\ classes\ found}$$

$$\text{F-measure} = \frac{(2 \times precision \times recall)}{(precision + recall)}$$

Text Document Representation

- แสดงเอกสารภายหลังตัดคำและได้ *unique word* ซึ่งจะแสดงในรูปของ *"bag of words"* (*Vector Space Model*)

	w_1	w_2	...	w_k	...	w_v
d_1	w_{11}	w_{12}	...	w_{1k}	...	w_{1v}
d_2	w_{21}	w_{22}	...	w_{2k}	...	w_{2v}
...	...	...		...	...	...
d_N	w_{N1}	w_{N2}	...	w_{Nk}	...	w_{Nv}

กำหนดให้:

$w = (w_1, w_2, ..., w_k, ..., w_v)$

v = จำนวนคำทั้งหมดที่เป็น *unique word*

ดังนั้นเวคเตอร์ของเอกสารอนาจารคือ $d_i = (w_{i1}, w_{i2}, ..., w_{ik}, ..., w_{iv})$ เมื่อ w_{ik} เป็น ความถี่ของคำที่ *k*-th ในเอกสาร d_i

Text Classification Using Naive Bayes

- นาอีฟเบส์ (Naïve Bayes) รู้จักกันดีในงานวิจัยด้านการกระบวนการจัดกลุ่ม โดยตัวแยกประเภทจะเรียกว่า "*นาอีฟ*" (*Naive*) และเนื่องจากผลลัพธ์ที่ได้จะเป็นเกิดจากการประมวลผลค่าในเชิงสถิติบางครั้งจึงถูกเรียกว่า "*Probabilistic Classifier*"

- โดย v คือจำนวนของคำที่เป็น *unique word* ทั้งหมด

- เมื่อเอกสาร d_i จะถูกกลุ่มด้วยการจัดกลุ่มเอกสาร ของ c_j ซึ่งสามารถแสดงได้ดังสมการ

$$P(d_i \mid c_j;\theta) = (P \mid d_i \mid)\prod_{k=1}^{|d_i|} P(w_{d_{ik}} \mid c_j;\theta)^{N(w_{d_{ik}},d_i)}$$

ซึ่ง w_{dik} แทน ***k*- th** คำที่อยู่ในเอกสาร d_i และ $N(w_{dik}, d_i)$ แสดงถึงน้ำหนักของคำ w_{dik} ที่ $\theta_{w_k|c_j} = P(w_k \mid c_j;\theta)$ ปรากฏอยู่ในเอกสาร d_i และ $|d_i|$ เป็นจำนวนของคำทั้งหมดในชุดเอกสาร ดังนั้นพารามิเตอร์ของแต่ละกลุ่มเอกสารก็คือ $\sum_{k=1}^{|v|} P(w_k \mid c_j;\theta) = 1$

ผลการวิจัย

- สร้างโมเดลการกรองจาก web sites จำนวน 200 web sites (Training set)
- ทดสอบจาก 120 web sites (Test set)
- ขนาดเวคเตอร์ของคำ 6,189 คำ

Cross Validation	ชุดข้อมูลทดสอบ	P (%)	R (%)	F (%)
1-fold	เว็บทั่วไป	97	93	94.96
	เว็บอนาจาร	98	95	96.48
2-fold	เว็บทั่วไป	97	94	95.48
	เว็บอนาจาร	96	93	94.46
3-fold	เว็บทั่วไป	98	92	94.91
	เว็บอนาจาร	97	96	96.50

หมายเหตุ: ทดสอบบนพื้นฐานของ 3-fold Cross Validation

สรุปการวิจัย

- งานวิจัยนี้ประยุกต์เรื่องของการจัดกลุ่มเอกสารโดยใช้ *Probabilistic Classifier* เป็นเครื่องมือในการสร้างตัวกรองเว็บอนาจารแบบอัตโนมัติ และเป็นวิธีการกรองเว็บแบบ *Content-based*

- จากผลการทดลองพบว่าจำนวนคำที่พบในเอกสารเป็นปัจจัยที่สำคัญในการสร้างโมเดลสำหรับการกรองเว็บโดยอาศัยข้อความ เพราะฉะนั้นหากเพิ่มขนาดของ *"bag of words"* ให้มีขนาดใหญ่ขึ้นอาจหมายถึงความถูกต้องในการทดลองที่เพิ่มมากขึ้นด้วย

- กรณีที่เว็บไซต์ทั่วๆไปที่มีกลุ่มคำในเชิงอนาจาร บางครั้งก็ยากต่อการตรวจสอบ โดยเว็บไซต์ลักษณะนี้จะอยู่ในกลุ่มที่เรียกว่า "*Grey zone*" แต่กรณีนี้อาจจะปัญหาโดยการตรวจสอบและวิเคราะห์จากความสัมพันธ์ของคำ หรือความสัมพันธ์ของประโยค ในหน้าเว็บนั้นๆ

END.