บทที่ 6
การประเมินผลการค้นคืนข้อมูล

วัตถุประสงค์

1. เพื่อให้นักศึกษาทราบถึงวิธีการประเมินผลการค้นคืนข้อมูล
2. เพื่อให้นักศึกษาทราบถึงส่วนประกอบของระบบที่เกี่ยวข้องกับการประเมินผล
3. เพื่อให้นักศึกษาทราบถึงหลักเกณฑ์การประเมินผลเกี่ยวกับผู้ใช้
4. เพื่อให้นักศึกษาทราบถึงปัญหาของวิธีการวัด Recall และ Precision
5. เพื่อให้นักศึกษาทราบถึงตัววัดที่เรียกว่า Fallout

สารบัญ

สำหรับในบทนี้จะกล่าวถึงการประเมินผลการค้นคืนข้อมูล

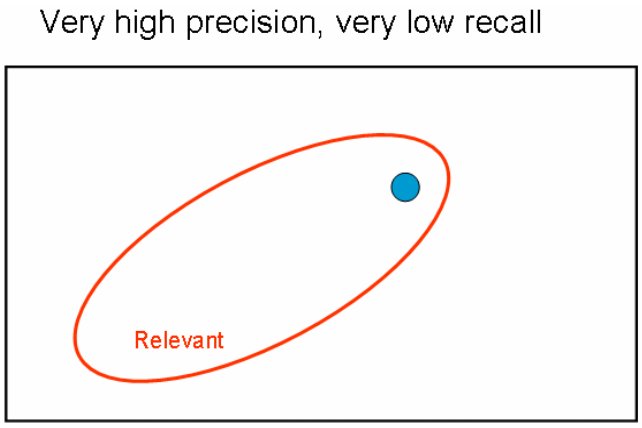
6.1 การประเมินผล

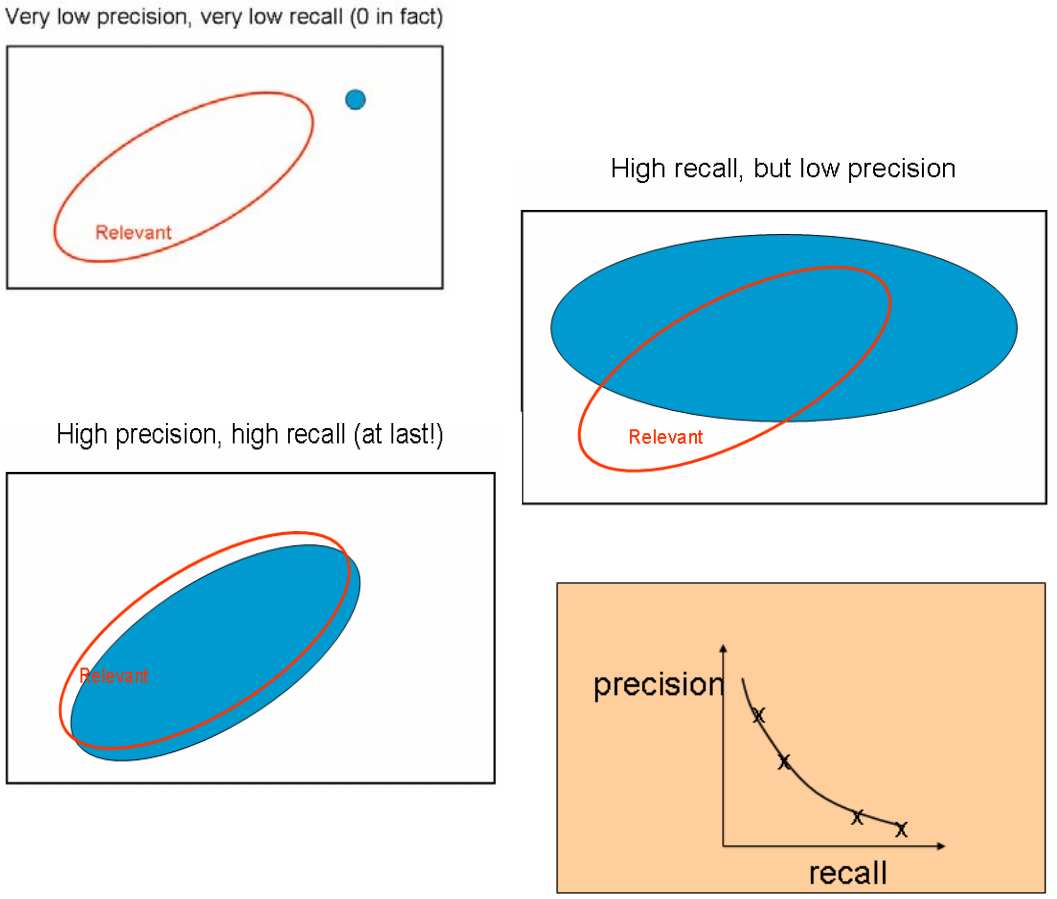
ในระบบใหญ่ๆ มีหลายๆแง่มุมที่เราจะนำมาใช้ในการประเมินผลได้ ในที่นี้ได้เน้นถึงการประเมินผลในทัศนคติของผู้ใช้ และตรวจสอบส่วนประกอบต่างๆ ของระบบที่ถูกนำเข้ามาใช้ประเมินผล

ระบบสืบคืนสารสนเทศจะเปิดโอกาสให้ผู้ใช้หลาย ๆ คน สามารถที่จะดึงรายการเอกสารที่ต้องการจากกลุ่มของเอกสารที่ถูกบันทึกเก็บไว้ในระบบโดยเร็วและประหยัดที่สุด คุณค่าของระบบจะขึ้นอยู่กับความสามารถของระบบในการที่จะดึงสารสนเทศได้ใกล้เคียงกับสิ่งที่ตนเองต้องการมากที่สุดและได้อย่างรวดเร็ว เป็นระบบที่ใช้ง่าย โต้ตอบได้อย่างรวดเร็วและไม่เสียค่าใช้จ่ายมากนักในการนำมาใช้

การที่จะประเมินผลระบบค้นคืนสารสนเทศนั้น เราอาจจะแบ่งประเด็นการประเมินผลออกเป็น 2 ประเด็น คือ

1. ในแง่ของประสิทธิผลของการค้นคืนข้อมูล (Retrieval Effectiveness) ซึ่งระบบที่ดีควรที่จะสามารถดึงเอกสารที่เกี่ยวข้องส่วนใหญ่ได้และขจัดเอกสารที่ไม่เกี่ยวข้องกับข้อ คำถามของผู้ใช้ออกไปให้มากที่สุด เราสามารถวัดประสิทธิผลด้วยค่า Recall และPrecision ของผลลัพธ์ที่ได้จากการค้นหา หรืออาจจะกล่าวได้ว่าประสิทธิผลของระบบ คือความสามารถของระบบในการที่จะให้บริการสารสนเทศตามที่ผู้ใช้ต้องการ ดังรูปที่ 6.1





รูปที่ 6.1 : แสดง Retrieved vs. Relevant Documents

2. ในแง่ของประสิทธิภาพ (Efficiency) ซึ่งอาจจะวัดได้จากทรัพยากรของระบบคอมพิวเตอร์ที่ใช้ เช่น เนื้อที่เก็บความจำและ CPU Time เป็นต้น หรือวัดได้จากค่าใช้จ่ายในการปฏิบัติการของระบบ, ความพยายามของผู้ใช้ (User Effort) ในการที่จะทำให้การค้นการเป็นไปอย่างมีประสิทธิภาพ เป็นต้น

การประเมินผลที่สมบูรณ์แบบจะต้องเกี่ยวข้องทั้งประสิทธิผลและประสิทธิภาพของระบบค้นคืนสารสนเทศ เมื่อเปรียบเทียบระหว่างประสิทธิภาพและประสิทธิผลแล้ว

6.2 สาเหตุของการประเมินผล

เหตุผลที่สำคัญในการประเมินผลระบบค้นคืนสารสนเทศ ได้แก่

1. เราอาจจะต้องการที่จะเปรียบเทียบความสามารถของระบบที่มีอยู่กับระบบค้นคืนสารสนเทศอื่นๆ
2. เราอาจจะต้องการที่จะทราบว่า เมื่อส่วนประกอบบางส่วนของระบบได้เปลี่ยนไป System Performance จะเปลี่ยนแปลงไปมากน้อยแค่ไหน ยกตัวอย่าง เช่น เราอาจจะต้องการทราบว่า การทำงานของระบบเปลี่ยนแปลงไปมานน้องเพียงไร เมื่อมีการเปลี่ยนชนิดของข้อเรียกร้อง หรือเมื่อข้อมูลที่ถูกนำมาเก็บนั้นได้เปลี่ยนแปลงไป
3. ในกรณีที่จะนำเอาส่วนประกอบใหม่ของระบบเข้ามารวมอยู่กับระบบที่มีอยู่เดิม ซึ่งเราจะต้องจำลอง (Simulate) การปฏิบัติการของระบบใหม่ก่อนที่จะมีการสร้างระบบจริงขึ้นมา

6.3 การทดสอบระบบ

การที่จะประเมินผลระบบนั้น จำเป็นต้องเกี่ยวข้องกับการทดสอบระบบ ซึ่งในการทดสอบระบบนั้น จะต้องใช้ส่วนประกอบต่อไปนี้ คือ

1. รายละเอียดเกี่ยวกับระบบและส่วนประกอบต่างๆ ของระบบหรือตัวแบบของระบบที่จะถูกตรวจสอบ
2. ข้อสมมุติฐาน (Hypothesis) ชุดหนึ่งที่จะถูกทดสอบ หรือต้นแบบ (Prototype) เกี่ยวกับตัวแบบที่จะถูกวัด
3. หลักเกณฑ์ที่จะแสดงถึงจุดประสงค์ของการทำงานของระบบและวิธีการที่จะวัดเกณฑ์ของการทำงานของระบบเป็นปริมาณตัวเลข
4. วิธีการที่จะได้มาซึ่งข้อมูลและวิธีการประเมินผล

สมมุติว่ามีผู้ใช้คนหนึ่งที่ต้องการจะตรวจดู Dialog System และกำหนดว่า ความเร็วในการค้นหาข้อมูลนั้นเร็วพอที่จะค้นหาในฐานข้อมูลหรือไม่ หลักเกณฑ์ที่ใช้ก็คือ กำหนดให้เวลาที่ยาวนานที่สุดในการที่จะส่งคำตอบกลับมา (Maximum Allowable Response Time)

การที่จะวัดความสามารถการทำงานของระบบ (System Performance) นั้นจำเป็นที่จะต้องใช้หลักเกณฑ์การวัดที่เรียกว่า Objective Quantitative Criteria ซึ่งสามารถวัดเป็นปริมาณตัวเลขเห็นได้ชัด Objective Measurements นั้นค่อนข้างที่จะง่ายในการทำความเข้าใจและไม่มีอคติลำเอียงจากผู้ทำการประเมินผล Objective Measurements อาจจะหาได้โดยการบันทึก การสังเกตโดยตรงซึ่งจะใช้แบบสอบถาม และเทคนิคในการสัมภาษณ์อีกวิธีหนึ่ง, อาจจะใช้วิธีการทางเครื่องกลในการรวบรวมข้อมูลที่ต้องการ แต่ไม่ว่าในกรณีใดๆ ที่กล่าวมานั้นจะต้องมีการเลือกพารามิเตอร์ (ซึ่งเป็นค่าคงที่ที่จะแสดงถึงประโยชน์หรือคุณค่าของระบบ)ที่สำคัญต่อวัตถุประสงค์ของการประเมินผล, สามารถถูกวัดได้โดยง่ายและมีความสัมพันธ์เกี่ยวเนื่องกัน บางพารามิเตอร์อาจจะถูกระบุได้โดยง่าย ยกตัวอย่าง ขนาดของกลุ่มเอกสารและ System Response Time มีหลายๆพารามิเตอร์ที่สำคัญอื่นๆ ที่มีความเกี่ยวข้องซึ่งกันและกัน ความสามารถของระบบที่จะดึงเอกสารที่เกี่ยวข้องนั้นจะขึ้นอยู่กับการแสดงรูปแบบ (Representation) ของเอกสาร, วิธีการค้นหาและคุณสมบัติของผู้ใช้ นอกไปจากนี้บางพารามิเตอร์อาจจะไม่สามารถระบุได้ ความสามารถที่จะดึงเอกสารที่เกี่ยวข้อง ไม่สามารถจะวัดได้ถ้าในกลุ่มเอกสารนั้น ไม่มีเอกสารที่เกี่ยวข้องกับเรื่องที่เราต้องการ

ในบางสถานการณ์นั้น ค่าของพารามิเตอร์ไม่อาจจะกำหนดค่าที่แน่นอนถูกต้องได้ หรือค่อนข้างจะลำบากในการที่จะหาค่านั้น เช่น เราอาจจะไม่ทราบจำนวนทั้งหมดของเอกสารที่เกี่ยวกับเรื่องใดเรื่องหนึ่ง แต่เราอาจจะเพียงคาดคะเนด้วยการใช้ Sampling Techniques มักจะมีการใช้ Probabilistic Models เนื่องจากมีหลายๆ พารามิเตอร์ที่จะกลายเป็นค่าคงที่เมื่อมีการสังเกตทดลองหลายๆ ครั้ง เช่น จำนวนเฉลี่ยของเอกสารที่เกี่ยวข้องกับข้อคำถามหรือจำนวนเฉลี่ยของเอกสารที่เกี่ยวข้องที่ถูกดึงโดยระบบ อาจจะถูกประมาณค่า ในขณะที่มีหลายๆ เอกสารที่ Match กับ Single Request หรือโดยการจัดการเกี่ยวกับเอกสารจำนวนหนึ่งในการค้นหาที่แตกต่างกันหลายๆ ครั้ง

6.4 ส่วนประกอบของระบบที่เกี่ยวข้องกับการประเมินผล

ก่อนที่จะมีการตรวจสอบเกี่ยวกับการประเมินผลพารามิเตอร์ต่างๆ จำเป็นที่จะต้องมีการพิจารณาส่วนประกอบของระบบสารสนเทศและสิ่งแวดล้อมของระบบ เพื่อที่จะกำหนดว่า System Performance นั้น ได้รับอิทธิพลจากสิ่งแวดล้อมของระบบและการทำงานของระบบ

เราจะอธิบายรายละเอียดเกี่ยวกับส่วนประกอบต่าง ๆ เหล่านี้ ได้ดังนี้

1. พารามิเตอร์ที่เกี่ยวข้องกับนโยบายของอินพุท ซึ่งรวมทั้งอัตราความผิดพลาดและการเสียเวลา (Time Delays) ในการที่จะใส่เอกสารใหม่ในกลุ่มเอกสาร, ความล่าช้า(Time Lag) ตั้งแต่การรับเอกสารหนึ่งที่กำหนดให้มาจนกว่าเอกสารนั้นจะปรากฏในแฟ้มข้อมูล และ Collection Coverage ซึ่งเป็นอัตราส่วนของเอกสารที่เกี่ยวข้องที่ถูกบรรจุจริง ๆ ในแฟ้มข้อมูล

2. รูปแบบทางกายภาพของอินพุท (Physical Input Form)ได้แก่ รูปแบบของเอกสารและความหมายของเอกสาร รูปแบบของเอกสารนั้น อาจจะอยู่ในรูปแบบต่อไปนี้ ชื่อเรื่อง (Title) , บทย่อ (Abstract) , บทสรุป (Summary) , หรือข้อความเต็ม (Full Text) ซึ่งจะมีผลต่อการสร้างดัชนีและการค้นหาเอกสาร นอกจากนี้ยังมีผลต่อค่าใช้จ่ายของระบบด้วย

3. โครงสร้างของแฟ้มข้อมูลที่ใช้ในการค้นหา (Organization of Search Files) ซึ่งจะมีผลต่อขั้นตอนของการค้นหาข้อมูล, Response Time, ความพยายามของผู้ควบคุมระบบ และอาจรวมทั้งประสิทธิภาพของระบบดึงข้อมูล

4. ภาษาดัชนี (Indexing Language)ภาษาดัชนีประกอบด้วยเซตของดัชนีและกฎต่าง ๆ ที่ถูกใช้เพื่อกำหนดดัชนีเหล่านี้ ให้กับเอกสารและข้อคำถามในการค้นหาข้อมูล ในระหว่างขั้นตอนของการสร้างดัชนี ค่าที่เหมาะสมกับการแสดงเนื้อหาของเอกสารจะถูกเลือกจากภาษาดัชนีและกำหนดให้กับรายการเอกสารตามกฎที่กำหนดขึ้นมา พารามิเตอร์ที่สำคัญเกี่ยวกับเรื่องนี้ก็คือ Exhaustivity และ Specificity ของภาษาดัชนี

Exhaustive Indexing Language จะบรรจุดัชนีต่าง ๆ ที่ให้ความหมายครอบคลุมไปทุก ๆ เรื่องราวสาขาวิชา (Subject Areas) ที่มีอยู่ในกลุ่มของเอกสารทั้งหมด ในทำนองเดียวกัน อาจกล่าวได้ว่า Exhaustive Indexing Product จะแสดงว่า ดัชนีต่าง ๆ ที่กำหนดให้กับเอกสารเหล่านั้นจะสะท้อนถึงเนื้อหาของเรื่องราวของเอกสารได้อย่างถูกต้อง แต่ Specific Indexing Language มักจะใช้ดัชนีที่มีความหมายเจาะจงในแคบและถูกต้อง

โดยปกติแล้ว Retrieval System Performance มักจะถูกวัดโดยการใช้ค่า Recall และ

Precision ซึ่ง Recall จะวัดความสามารถของระบบในการที่จะดึงเอกสารที่เกี่ยวข้องออกมา ในขณะที่ Precision จะวัดความสามารถในการที่จะขจัดเอกสารที่ไม่เกี่ยวข้องออกไป Indexing Exhaustivity ที่อยู่ในระดับสูงจะให้ Recall ที่สูง ซึ่งทำให้สามารถที่จะดึงเอกสารที่เกี่ยวข้องได้มาก ในขณะเดียวกัน เมื่อมีการใช้ดัชนีที่มีความเจาะจงสูง Precision ก็จะมีแนวโน้มว่าจะสูงด้วย เนื่องจากเอกสารที่ถูกดึงออกมาส่วนใหญ่ มักจะเป็นเอกสารที่เกี่ยวข้อง หรืออาจจะกล่าวได้ว่า การที่จะให้ Recall สูงนั้น ควรใช้ Exhaustive Indexing Language เพราะจะได้เอกสารต่าง ๆ ที่ครอบคลุมเนื้อหาของเรื่องราวที่กำหนดมา แต่ถ้าต้องการจะให้ได้ Precision ที่สูง ควรที่จะใช้ภาษาดัชนีที่มีความเจาะจงสูง และดัชนีควรที่จะมีตัวบ่งชี้ เนื้อหา (Content Indicators) เพิ่มเติม เช่น น้ำหนักของคำ หรือความสัมพันธ์ระหว่างคำนี้กับดัชนีอื่น ๆ เป็นต้น

5. วิธีการสร้างดัชนี (Indexing Operation) ถ้าการสร้างดัชนีถูกกระทำด้วยมือคนที่ได้รับการอบรมมาอย่างดี ตัวแปรที่จะมีผลกระทบต่อวิธีการสร้างดัชนีนั้น จะมีความสัมพันธ์ไม่เพียงแต่กับ Exhaustivity ของการ Indexing, และ Specificity ของดัชนีที่ถูกกำหนดมาเท่านั้น ยังมีความสัมพันธ์กับอิทธิพลของประสบการณ์ของผู้สร้างดัชนีต่อ Performance และความถูกต้องของ Assigner Terms

6. การวิเคราะห์คำถาม (Question Analysis)เป็นการยากในการกำหนดดัชนีให้กับ Information Requests การสร้าง Boolean Statements ที่มีความหมาย และการเปรียบเทียบข้อเรียกร้องที่ถูกวิเคราะห์แล้วกับข่าวสารที่ถูกเก็บบันทึกนั้น ทั้งหมดเป็นงานที่ค่อนข้างยุ่งยาก โดยหลักการแล้ว วิธีการวิเคราะห์แยกแยะเนื้อหาของเอกสารและข้อเรียกร้องนั้นเหมือนกัน ความคิดในเรื่อง Exhaustivity และ Specificity นั้น สามารถนำไปใช้กับข้อคำถามได้เช่นเดียวกับเอกสาร ดังนั้น Exhausive Query Indexing ที่ใช้คำที่มีความเจาะจงสูงสร้าง Search Recall และ Precision สูงที่สุด ในทางปฏิบัติ , การประมวลผลข้อคำถามค่อนข้างที่จะแตกต่างจากการสร้างดัชนีของเอกสาร เนื่องจากผู้ใช้จำเป็นที่จะต้องเกี่ยวข้องกับการประมวลผลข้อเรียกร้องโดยตรง มากกว่าที่จะเกี่ยวข้องกับการสร้างดัชนีให้กับเอกสาร

7. วิธีการค้นหา (Search Operations) เป็นการยากที่จะวัดวิธีการค้นหาโดยใช้ Objective Parameters เนื่องจากผู้ใช้ระบบไม่ค่อยจะระบุ Recall หรือ Precision Requirements, หรือไม่ค่อยจะประเมินผลเกี่ยวกับผลลัพธ์ที่ได้จากระบบ อย่างไรก็ตาม กลวิธีในการค้นหาข้อมูลนั้น ยังจำเป็นที่จะต้องมีการทบทวน เพื่อที่จะสนองตอบต่อ Specific Recall และ Precision Requirements ของผู้ใช้ ในบรรดาคุณลักษณะต่าง ๆ ที่ควรที่จะถูกรวมอยู่ในการวัด Search Performance ได้แก่ ชนิดของโครงสร้างแฟ้มข้อมูลที่ถูกนำมาใช้ ,การเปรียบเทียบข้อคำถามกับเอกสาร ,อิทธิพลของกลวิธีการค้นหาข้อมูลต่อ Response Time,

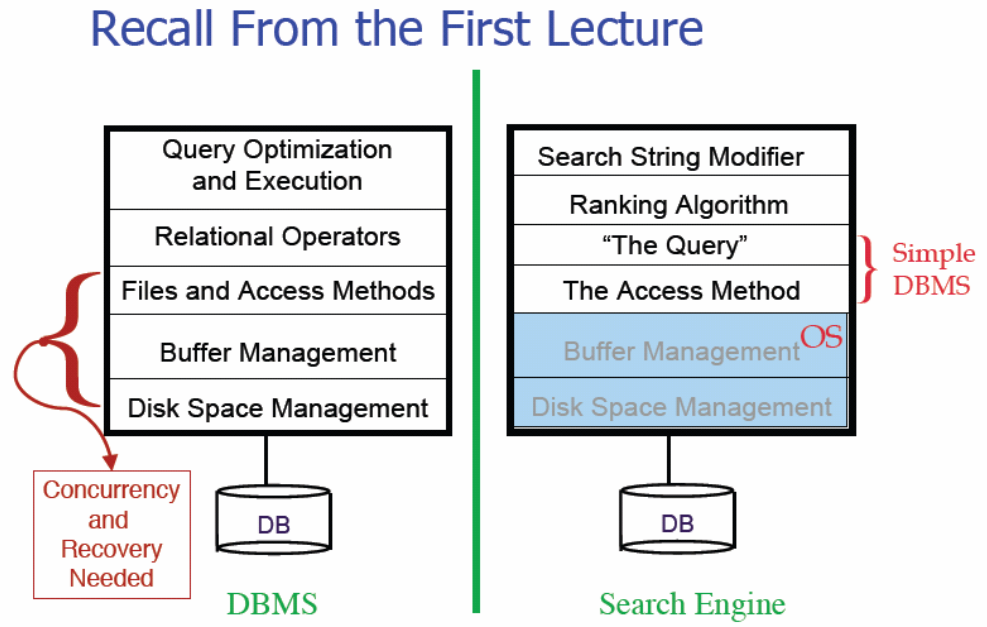
และบน Search Performance และมาตรฐานความเกี่ยวข้องผู้ใช้ระบบ (หมายถึง มาตรฐานที่ผู้ใช้ระบบใช้ในการพิจารณาว่าผลลัพธ์ที่ได้จากการค้นหานั้น เกี่ยวข้องกับเรื่องที่ผู้ใช้ต้องการมากน้อยแค่ไหน)

8. รูปแบบของการแสดงผลลัพธ์ เป็นการแสดงรูปแบบทางกายภาพของเอกสารที่ถูกค้นพบโดยระบบ เพื่อที่จะสนองตอบต่อข้อคำถามของผู้ใช้ รูปแบบผลลัพธ์ที่ปรากฏออกมาจะมีผลต่อปริมาณความพยายามของผู้ใช้ที่จะดูผลลัพธ์ที่ได้จากค้นหาข้อมูลและความพอใจครั้งสุดท้ายที่ได้รับจากการค้นหาข้อมูลในแต่ละครั้ง รูปแบบของผลลัพธ์นั้นยิ่งมีความสมบูรณ์มากขึ้นเท่าไร, งานการประเมินความเกี่ยวข้อง (Relevance Assessment Task) สำหรับผู้ใช้ก็จะยิ่งง่ายขึ้น ในทางกลับกัน ในขณะที่ผลลัพธ์ได้ถูกขยายจากหมายเลขเอกสารง่าย ๆ ไปยัง Full Document Texts ซึ่งจะทำให้เวลาที่ต้องการในการที่จะตรวจสอบ Search Results ก็จะเพิ่มไปด้วย

จากที่กล่าวมาแล้วข้างต้นนี้ จะเห็นว่าส่วนประกอบต่าง ๆ และพารามิเตอร์ของระบบค้นคืนสารสนเทศจะมีผลต่อการทำงานของระบบ และผลลัพธ์ของการประเมินผล แต่ละส่วนประกอบสามารถถูกตรวจสอบแยกต่างหากกัน, หรือสามารถที่จะเปรียบเทียบกับระบบอื่น แต่ละพารามิเตอร์จะมีผลต่อระบบ และความสำคัญของแต่ละพารามิเตอร์จะไม่สามารถที่จะประเมินค่าได้ ถ้าไม่นำเอาวัตถุประสงค์ในการใช้ระบบมาพิจารณา

6.5 หลักเกณฑ์การประเมินผลเกี่ยวกับผู้ใช้

ระบบสารสนเทศอาจจะถูกตรวจสอบทั้งจากแง่คิดของผู้ใช้หรือจากความเห็นของผู้จัดการระบบ ในปัจจุบัน ผู้จัดการระบบอาจจะถูกสมมุติให้รวมถึงทุก ๆ คนที่มีผลต่อนโยบายหรือการเงินของระบบ หรือผู้ที่รับผิดชอบเกี่ยวกับการปฏิบัติงานของระบบ เนื่องจากว่า เป็นการสมควรที่จะสมมุมิว่า ระบบสารสนเทศควรที่จะสนองตอบต่อความต้องการของผู้ใช้ระบบหลาย ๆ ประเภท หลักเกณฑ์ประสิทธิผลของประโยชน์ต่อผู้จัดการระบบนั้นเหมือนกับผลประโยชน์ของผู้ใช้ระบบ ระบบควรที่จะสนองต่อความต้องการของผู้ใช้ ข้อผิดพลาดในการดึงเอกสารที่เกี่ยวข้อง ชนิดของผลลัพธ์ที่ได้จากการค้นหาข้อมูล (เช่น เป็นชื่อเรื่อง, บทย่อ หรือ ข้อความเต็ม)เป็นต้น



รูปที่ 6.2 : สถาปัตยกรรมของ DBMS และ Search Engine

Collection Coverage ซึ่งเป็นปริมาณที่ทุก ๆ เอกสารที่เกี่ยวข้องถูกครอบคลุมอยู่ในระบบ หรืออาจจะกล่าวได้ว่า เป็นอัตราส่วนของจำนวนเอกสารที่เกี่ยวข้อง ต่อจำนวนเอกสารทั้งหมดในระบบ พารามิเตอร์ที่เกี่ยวข้อง ได้แก่ ชนิดของอุปกรณ์นำข้อมูลเข้า, ชนิดและขนาดของอุปกรณ์เก็บความจำ ,ความยากง่ายของการวิเคราะห์เนื้อหา เป็นต้น

จากหลักเกณฑ์ต่าง ๆ ที่กล่าวมาแล้วข้างต้น ยกเว้น Recall และ Precision นอกนั้นค่อนข้างที่จะง่ายในการวัด ความพยายามของผู้ใช้อาจจะวัดจากเวลาที่ใช้ในการสร้างข้อคำถาม , การโต้ตอบกับระบบและการตรวจสอบผลลัพธ์ของระบบ Response Time สามารถถูกวัดได้โดยตรง Collection Coverage อาจจะวัดได้ยากลำบากบ้าง ถ้าหากไม่ทราบจำนวนของเอกสารที่เกี่ยวข้องในเรื่องราวนั้น ๆ

การวัด Recall และ Precision นั้นค่อนข้างจะทำได้ยากแม้ในทางทฤษฎีและทั้งทางปฏิบัติ ปัญหาในการกำหนด Recall และ Precision ก็คือการตีความหมายของความเกี่ยวข้อง มีนิยามเกี่ยวกับความเกี่ยวข้องที่เป็นไปได้ ซึ่งมีความว่า

"ความเกี่ยวข้องนั้นเป็นความสอดคล้องระหว่างข้อคำถามกับเอกสาร, กล่าวคือ, ปริมาณที่เอกสารต่าง ๆ จะครอบคลุมเอกสารที่เหมาะสมหรือเกี่ยวเนื่องกับข้อคำถาม"

นิยามนั้นจะทำให้เป็นไปได้ในการที่จะพิจารณาความเกี่ยวข้องเป็นคุณสมบัติทางด้านตรรกศาสตร์ ระหว่างรายการเอกสารคู่หนึ่ง ซึ่งจะวัดโดยระดับที่เอกสารหนึ่งเกี่ยวเนื่องกับเรื่องราวของสารสนเทศที่ต้องการของผู้ใช้ อาจจะถือว่าเอกสารนั้น “เกี่ยวข้อง” แม้ว่าผู้ใช้อาจจะคุ้นเคยกับเอกสารนั้นก่อนที่จะมีการสร้างข้อคำถาม หรืออาจจะมาคุ้นเคยกับเอกสารในระหว่างความพยายามในการค้นหาครั้งก่อน ๆ ที่กล่าวมานี้เป็นความเกี่ยวข้องในแง่ของ Objective View ซึ่งจะพิจารณาความเกี่ยวข้องระหว่างข้อคำถามกับเอกสารเพียงเท่านั้น แต่ความเกี่ยวข้องในแง่ Subjective View นั้น นอกจากจะพิจารณาจากเนื้อหาของเอกสารแล้ว ยังเกี่ยวข้องกับสถานภาพของความรู้ของผู้ใช้ในขณะที่กำลังทำการค้นหาข้อมูล และเอกสารอื่น ๆ ที่ถูกดึงออกมา ซึ่งผู้ใช้รู้เกี่ยวกับเอกสารนั้น ๆ ดังนั้น, นิยามความเกี่ยวข้องจะขึ้นอยู่กับการใช้แต่ละเอกสารของผู้ใช้ เอกสารที่ตรงกับความต้องการของผู้ใช้ อาจจะถูกระบุเป็นส่วนหนึ่งของเอกสารที่ถูกเก็บบันทึกในระบบ และตรงกับความต้องการสารสนเทศของผู้ใช้ที่ได้ในช่วงเวลาของการดึงข้อมูล เอกสารหนึ่งอาจจะถือว่าเกี่ยวข้อง ถ้าการเอกสารนั้นเกี่ยวข้องกับหัวเรื่องที่ต้องการ การวัดประสิทธิภาพของการดึงข้อมูลอาจจะทำได้โดยง่าย ถ้าใช้ความเกี่ยวข้องในแง่ Objective View มากกว่า Subjective View แต่ก็อาจจะมีความยุ่งยากในการประเมินความเกี่ยวข้องของเอกสารต่อข้อคำถาม เราจะกำหนดระดับความเกี่ยวข้องไว้ที่จุดใด และจะประเมินความเกี่ยวข้องได้อย่างไร เรื่องนี้ยังเป็นปัญหาที่ยังตกลงกันได้ยาก ด้วยเหตุนี้ จึงทำให้มีการเสนอความเกี่ยวข้องในรูปแบบของ Probabilistic Terms ซึ่งในกรณีนี้, ความเกี่ยวข้องก็คือฟังก์ชันของ Probability ที่ความคล้ายคลึงกันระหว่างข้อคำถามกับเอกสารจะนำผู้ใช้ไปสู่จุดที่ยอกรับเอกสารที่ได้มาจากค้นหา ในการตอบสนองต่อข้อเรียกร้องของผู้ใช้

ในทางปฏิบัติ, จำเป็นที่จะต้องมีการสมมุติว่า เราสามารถประเมินความเกี่ยวข้องในแง่ของ Objective View ซึ่งจะประเมินความเกี่ยวข้องของเอกสารต่อข้อคำถามโดยพิจารณาจากแหล่งข้อมูลภายนอก ดังนั้น ระบบหนึ่งอาจจะถูกพิจารณาว่า มีประสิทธิผล ถ้าผลลัพธ์ที่ได้จากการประเมินผลนั้นเป็นที่น่าพอใจ

6.6 การคำนวณ Recall และ Precision

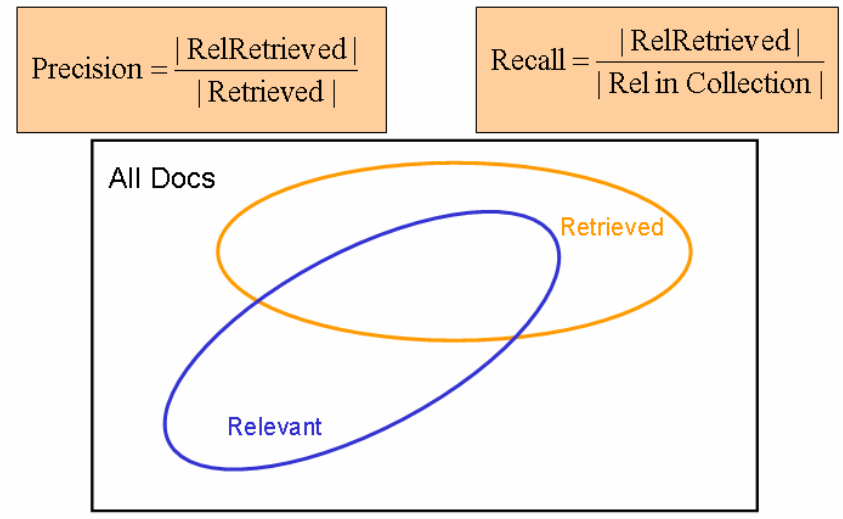
Recall นั้นถูกกำหนดให้เป็นอัตราส่วนของเอกสารที่เกี่ยวข้องที่ถูกดึงออกมาจากจำนวนเอกสารที่เกี่ยวข้องทั้งหมด ในขณะที่ Precision เป็นอัตราส่วนของเอกสารที่ถูกดึงออกมาแล้วข้อความ จากจำนวนเอกสารที่ถูกดึงออกมาทั้งหมด ในทางปฏิบัติ, สารสนเทศที่ต้องการของผู้ใช้แต่ละคนนั้นย่อมแตกต่างกันไป ซึ่งผู้ใช้บางคนอาจจะต้องการ Recall ที่สูง, กล่าวคือ, ทุกสิ่งที่ถูกดึงออกมาเป็นเรื่องที่น่าสนใจ ในขณะที่อีกคนหนึ่งอาจจะต้องการ

ถ้าหากมีเส้นตัดผ่าน Document Collection เพื่อที่จะแยกแยะรายการเอกสารที่ถูกดึงออกมา ให้ออกจากรายการเอกสารที่ไม่ได้ถูกดึงออกมา ดังแสดงในรูป 6.3 และถ้าหากเราสามารถมีวิธีการที่จะแยกแยะเอกสารที่เกี่ยวข้องออกจากเอกสารที่ไม่เกี่ยวข้อง Recall มาตรฐาน R และ Precision มาตรฐาน P อาจจะกำหนดได้โดย

$$\text{Recall} = \frac{\text{จำนวนของเอกสารที่เกียวข้องทั้งหมดและถูกดึงออกมา}}{\text{จำนวนเอกสารที่เกี่ยวข้องทั้งหมด}}$$

และ

$$\text{Procision} = \frac{\text{จำนวนของเอกสารที่เกียวข้องและถูกดึงออกมา}}{\text{จำนวนเอกสารที่ถูกดึงออกมาทั้งหมด}}$$



รูปที่ 6.3 Partition of Collection

ถ้าเราสามารถที่จะพิจารณาความเกี่ยวข้องระหว่าง ข้อกำหนดกับแต่ละเอกสารในกลุ่มเอกสาร และถ้าเราสามารถแยกแยะระหว่างเอกสารที่ถูกดึงออกมากับเอกสารที่ไม่ได้ถูกดึงออกมา เราจะสามารถคำนวณตัววัดเหล่านี้ได้โดยตรง

ในระบบดึงข้อมูลแบบดั้งเดิม ข้อคำถามจะถูกแสดงเป็น Boolean Combinations ของดัชนีที่ใช้ในการค้นหา เอกสารที่ถูกดึงออกมานั้น จะประกอบด้วย ทุก ๆ เอกสารที่ถูกดึงออกมาของคีย์เวิร์ดต่าง ๆ ที่ถูกระบุในข้อคำถามนั้นรวมกัน แต่ละข้อคำถามจะทำให้เกิดเอกสารที่เกี่ยวข้องและเอกสารที่ไม่เกี่ยวข้อง ดังนั้น, สำหรับแต่ละข้อคำถามนั้น เราจะได้ตัวเลขของ Precision และ Recall คู่ของตัวเลข Recall Precision นั้น สามารถที่จะถูกเปรียบเทียบสำหรับการค้นหาครั้งที่ i และ j เมื่อใดก็ตามที่

$$RECALL_i < RECALL_j$$

และ $$PRECISION_i < PRECISION_j$$

ผลลัพธ์ของการค้นหาครั้งที่ j จะถูกพิจารณาว่า ดีกว่าการค้นหาครั้งที่ i แต่จะมีปัญหาเกิดขึ้นเมื่อ

$$RECALLi < RECALLj$$

และ $$PRECISION_i > PRECISION_j$$

หรือในทางกลับกัน

$$RECALL_i > RECALL_j$$

และ $$PRECISION_i < PRECISION_j$$

ในกรณีเหล่านี้ การจะพิจารณาว่าการค้นหาครั้งไหนจะดีกว่านั้น จะขึ้นอยู่กับผู้ใช้, กล่าวคือ, ผู้ใช้จะต้องกำหนดว่า สิ่งที่ผู้ใช้สนใจนั้นเป็น Recall หรือ Precision และประเมินความสำคัญของความแตกต่างระหว่างค่าของ Recall และ Precision ในระบบค้นคืนสารสนเทศที่เป็นมาตรฐานนั้น Recall จะเพิ่มขึ้นในขณะที่จำนวนของเอกสารที่ถูกดึงออกมานั้น

6.7 ปัญหาของวิธีการวัด Recall และ Precision

วิธีการวัดด้วย Recall และ Precision นั้น เป็นประโยชน์ต่อผู้ใช้ เนื่องจากจะแสดงให้เห็นถึงความสำเร็จที่สัมพันธ์กันของระบบที่สนองตอบต่อความต้องการต่าง ๆ ของผู้ใช้ นอกไปจากนี้, วิธีการวัดสามารถที่จะถูกตีความหมายโดยตรงในการแสดงดึงประสบการณ์ของผู้ใช้อีกด้วย ดังนั้น Precision Performance 0.2 ที่ Recall เท่ากับ 0.5 แสดงว่า ผู้ใช้ได้รับเอกสารที่เกี่ยวข้องครึ่งหนึ่ง และจะต้องตรวจสอบ 4 เอกสารที่ไม่เกี่ยวข้องสำหรับทุก ๆ เอกสารที่เกี่ยวข้องที่ได้มา

มีข้อสังเกตบางอย่างเดี่ยวกับวิธีการวัด Recall และ Precision แบบมาตรฐาน ดังต่อไปนี้

1.จากนิยามพื้นฐาน, วิธีการวัด Recall และ Precision ได้ถูกผูกเข้าด้วยกันตามปกติ (Normally Tied) ระหว่างกลุ่มเอกสารที่กำหนดให้มากับ Query Set ที่กำหนดภายในสิ่งแวดล้อมที่ไม่เปลี่ยนแปลง เป็นไปได้ที่จะมีการเปลี่ยนแปลงวิธีการสร้างดัชนีหรือภาษาดัชนีหรือวิธการค้นหาและเพื่อที่จะระบุในภายหลังได้ว่า การเปลี่ยนแปลงเหล่านี้ จะมีผลต่อการกระทำของระบบในแง่ของ Recall และ Precision อย่างไร, อีกแง่หนึ่ง, Recall และ Precision จะต้องถูกใช้อย่างระมัดระวังในการเปรียบเทียบกรกระทำของสองระบบที่แตกต่างกันอย่างสิ้นเชิงซึ่งกระทำการต่อกลุ่มเอกสารต่าง ๆ ,ชุดของข้อคำถามและกลุ่มของผู้ใช้ที่แตกต่างกัน

	Doc is Relevant	Doc is NOT relevant
Doc is retrieved	a	b
Doc is NOT retrieved	c	d

- Accuracy: (a+d) / (a+b+c+d)
- Precision: a/(a+b)
- Recall: ?

ถ้าพิจารณาโดยละเอียดเกี่ยวกับความเปลี่ยนแปลงที่คาดหวังได้จากค่าของตัววัด Recall และ Precision แล้ว เมื่อขนาดของกลุ่มเอกสารได้เพิ่มขึ้น หรือ เมื่อจำนวนเฉลี่ยของเอกสารที่เกี่ยวข้องต่อหนึ่งข้อคำถามได้ลดหายไปจากแต่ละกลุ่มเอกสาร ทั้งใน 2 กรณีนี้ Recall และ Precision จะถูกทำให้เสื่อมคุณค่าลง เนื่องจากจำนวนของเอกสารที่เกี่ยวข้องและถูกดึงออกมานั้น มีท่าทีจะไม่เพิ่มขึ้นเป็นอัตราส่วนกันกับขนาดของกลุ่มเอกสาร (Collection) จึงต้องมีการระมัดระวังที่จะทำให้กลุ่มเอกสารและคุณสมบัติความเกี่ยวข้องระหว่างเอกสารกับข้อคำถามนั้น มีความเสมอภาคกันก่อนที่จะใช้วิธีการวัด Recall และ Precision ในการประเมินผลของกลุ่มเอกสารต่าง ๆ ที่แตกต่างกัน

2. ปัญหาอื่น ๆ จะเกิดขึ้นเกี่ยวเนื่องกับการประเมินความเกี่ยวข้องของเอกสารกับข้อคำถามของผู้ใช้ การประเมินค่าของความเกี่ยวข้องนั้นเป็นที่ต้องการ ถ้าเอกสารที่เกี่ยวข้องได้ถูกแบ่งแยกให้เห็นเด่นชัดต่างหากจากเอกสารที่ไม่เกี่ยวข้อง ผู้สังเกตการณ์บางคนได้ยืนยันว่าวิธีการวัด Recall และ Precision นั้นจะถูกนำไปใช้กับสิ่งแวดล้อมของผู้ใช้ ในกรณีที่ซึ่งสามารถพิจารณาหาความเกี่ยวข้องระหว่างเอกสารกับข้อคำถามได้ จากผลของการวิจัยหลาย ๆ ครั้งได้แสดงว่า แม้ว่าผู้ประเมินคนละคนนั้นจะให้การประเมินความเกี่ยวข้องที่แตกต่างกันสำหรับเอกสารที่ถูกเลือกมาอย่างสุ่ม ๆ, มีผลต่อตัววัด Recall และ Precision เพียงเล็กน้อย ดังนั้น เราอาจจะประเมินความเกี่ยวข้องระหว่างเอกสารกับข้อคำถามต่าง ๆ จากผู้ประเมินที่มีความเห็นโดยอิสระหลาย ๆ คนได้

3. ปัญหาที่น่าสนใจเกี่ยวกับการใช้วิธีการวัด Recall และ Precision นั้นก็คือ อิทธิพลของประเภทของข้อคำถามบนผลที่ได้จากการประเมินผล (Evaluation Outcome) ในจุดนี้เราอาจแบ่งประเภทของข้อคำถาม อย่างเช่น

1) ข้อคำถามประเภทที่มีหัวข้อเรื่องอย่างสั้น ๆ (Short Subject Heading Queries) ซึ่งหัวข้อเรื่องเป็นคำอธิบายสั้น ๆ เช่น คำว่า “Thermal Control”, “Turbulance Studies” เป็นต้น

2) ข้อคำถามประเภทที่ใช้ชื่อเรื่องทั้งหมด (Title-Length Queries) ซึ่งชื่อเรื่องได้อธิบายสาขาวิชาของเอกสารนั้น

3) ข้อคำถามประเภทที่ใช้ข้อความที่สมบูรณ์ (Full-Text Queries)

ซึ่งข้อคำถามประเภทต่าง ๆ เหล่านี้ ได้ถูกใช้ในการที่จะสร้างข้อเรียกร้องในการค้นหา

ข้อมูล และข้อคำถามหลาย ๆ ชนิดเหล่านี้ อาจจะถูกสร้างในระบบที่ Query Formulations สุดท้าย ได้ถูกกำหนดโดยผู้ใช้ซึ่งเป็นสื่อกลางของวิธีการค้นหา แม้ว่าข้อคำถามที่สั้นและเป็น Subject-Heading Queries ได้จัดการเกี่ยวกับหัวข้อเรื่องโดยทั่วไป, ในขณะที่ Full-Text Queries ส่วนใหญ่ในบางครั้ง จะหาความเจาะจงได้มากกว่า แต่ก็ไม่จำเป็นเสมอไปที่ความยาวของข้อคำถามจะเกี่ยวกับความเจาะจงโดยตรง ในกรณีใดก็ตาม, ระบบควรที่จะถูกทดสอบโดยการใช้ข้อคำถามที่สามารถเป็นจริงผสมกับการสะท้อนให้เห็นประเภทของข้อคำถามที่ถูกส่งเข้าไปจริงในสถานการณ์ปฏิบัติ

4.ข้อพิจารณาข้อสุดท้ายที่เกี่ยวกับการคำนวณ Recall และ Precision ก็คือ การกำหนดระดับของความเกี่ยวข้องที่มีต่อเอกสารต่าง ๆ ของกลุ่มเอกสารที่กำหนดให้มา และการเลือกระดับของเอกสาร (Document Rank) สำหรับการแสดงผลลัพธ์ ภายใต้สถานการณ์ปกติ, จะมีการประเมินค่าความเกี่ยวข้อง 2 ค่านี้เป็นปกติ ซึ่งจะประเมินดูว่าเอกสารหนึ่งนั้นเกี่ยวข้องหรือไม่เกี่ยวข้อง ในสถานการณ์เหล่านี้ การคำนวณค่า Recall และ Precision จะไม่ชัดเจน ถ้าระบบหนึ่งใช้ความเกี่ยวข้องหลาย ๆ ระดับ และเอกสารต่าง ๆ ที่ถูกดึงออกมานั้นถูกจัดลำดับ, หลาย ๆ เอกสารอาจจะดูคล้ายคลึงพอ ๆ กันกับข้อคำถามที่กำหนดให้มาและถูกแทนที่ในตำแหน่งที่ต่อเนื่องกันบนรายการของผลลัพธ์ อย่างไรก็ตาม, เนื่องจากลำดับนั้น ๆ จะมีผลต่อการประเมินค่า Precision-Recall , จึงจำเป็นที่จะต้องมีวิธีการที่จะทดแทนการจัดลำดับของเอกสารที่มีความคล้ายคลึงพอ ๆ กัน วิธีการหนึ่งอาจจะทำให้โดยการกำหนดระดับของความเดี่ยวข้องของเอกสารที่มีความเดี่ยวข้องพอ ๆ กันด้วยระดับเฉลี่ยของความเกี่ยวข้องของเอกสารชุดนี้

ในทางปฏิบัติ, ปรากฏว่า ผู้ประเมินค่าความเกี่ยวข้องนี้ค่อนข้างที่จะยุ่งยากในการใช้ความเกี่ยวข้องหลาย ๆ ระดับ แทนที่เพียงแต่ตัดสินว่าเอกสารใดเกี่ยวข้องหรือไม่เกี่ยวข้องเท่านั้น นอกไปจากนี้, อาจจะเกิดความผิดพลาดและความไม่แน่นอนในการประเมินค่าความเกี่ยวข้องหลาย ๆ ชนิดที่ผู้ใช้ถูกบีบบังคับให้แยกแยะระหว่างเอกสารชนิดที่เกี่ยวข้องอย่างสมบูรณ์, อาจจะเกี่ยวข้อง, เกี่ยวข้องบ้างเล็กน้อย หรือไม่เกี่ยวข้องกันเลย สิ่งเหล่านี้ อันที่จริงอาจจะทำให้ดูมีน้ำหนักมาขึ้นและมีความถูกต้องมากกว่าที่จะถูกค้นหาโดยการใช้หลาย ๆ ระดับของความเกี่ยวข้อง อย่างไรก็ตาม, มีหลาย ๆ วิธีการประเมินผลและพารามิเตอร์ต่าง ๆ ได้ถูกเสนอขึ้นมา เพื่อใช้กับน้ำหนักความเกี่ยวข้องที่เปลี่ยนแปลงได้ และสำหรับระบบที่อนุญาตให้มี Ties ในการจัดลำดับ (Rank) ของเอกสารที่ถูกดึงออกมา วิธีการวัดเหล่านี้ได้ถูกแนะนำพร้อมกับหลักเกณฑ์การประเมินผลที่เพิ่มเติมในตอนต่อไป

การแก้ไขได้มีผู้ทำการวิจัยและได้เสนอวิธีการไว้ดังนี้

(1) The E-Measure เป็นการรวม Recall และ Precision ให้เป็นจำนวนเดียว(van Rijsbergen(79) โดยมีสูตรในการคำนวณดังนี้

$$E = 1 - \frac{1}{\alpha\left(\frac{1}{P}\right) + (1-\alpha)\frac{1}{R}}$$

$$\alpha = 1/(\beta^2 + 1)$$

โดย P = precision ,R = recall ผลลัพธ์ของการคำนวณจะใช้สำหรับวัดความสัมพันธ์ที่สำคัญของ P และ R (measure of relative importance of P or R)
ตัวอย่าง ถ้าผลลัพธ์จากการคำนวณ

E = 1 หมายความว่าผู้ใช้มีความสนใจใน precision และ recall เท่ากัน

E= ∞ หมายความว่าผู้ใช้ไม่สนใจเกี่ยวกับ precision

E= 0 หมายความว่าผู้ใช้ไม่สนใจเกี่ยวกับ recall

(2) F Measure (Harmonic Mean) เป็นอีกวิธีการหนึ่ง มีสูตรในการคำนวณดังนี้

$$F(j) = \frac{2}{\frac{1}{r(j)} + \frac{1}{P(j)}}$$

โดย r(j) คือ recall สำหรับเอกสาร j

P(j) คือ precision สำหรับเอกสาร j

6.8 Fallout

ได้มีการแสดงให้เห็นมาก่อนแล้วว่า วิธีการวัด Recall และ Precision ได้ถูกตีความหมายจากผู้ใช้โดยตรงในแง่ของความพอใจในการค้นหา แต่บางครั้งก็เป็นการยากที่จะคำนวณ ยกตัวอย่าง เช่น Recall อาจจะไม่ได้ถูกกำหนดเนื่องจากไม่มีเอกสารที่เกี่ยวข้องกับข้อคำถามเลยในเอกสารกลุ่มนั้น ในทำนองเดียวกัน, Precision ไม่สามารถที่จะกำหนดได้ ถ้าไม่มีเอกสารใด ๆ ถูกดึงออกมาเลย

ข้อบกพร่องอื่น ๆ ในการใช้ Precision และ Recall นั้นก็คือ การขาดลักษณะที่เป็นไปในแนวขนานในคุณสมบัติของ 2 ตัววัดที่ว่านี้ สมมุติว่า ประสิทธิผลการดึงข้อมูลเพิ่มขึ้นพร้อมกับจำนวนเอกสารที่เกี่ยวข้องที่ได้รับเป็นคำตอบข้อเรียกร้อง และลดลงตามจำนวนเอกสารที่ไม่เกี่ยวข้องที่ถูกดึงออกมา ดังนั้นจึงมีการเสนอตัววัดที่เรียกว่า Fallout ซึ่งจะสะท้อนให้เห็นการกระทำสำหรับเอกสารที่ไม่เกี่ยวข้องอยู่ในรูปแบบที่สอดคล้องกับตัววัดการกระทำของเอกสารที่เกี่ยวข้องด้วยตัววัด Recall โดยปกติ Fallout จะถูกกำหนดดังต่อไปนี้

FALLOUT = RETNREL / (RETNREL + NRETNREL)

$$\text{Fallout} = \frac{\text{จำนวนของเอกสารที่ไม่เกี่ยวข้องที่ถูกดึงออกมา}}{\text{จำนวนทั้งหมดของเอกสารที่ไม่เกียวข้องในกลุ่มเอกสาร}}$$

ถ้า Recall ได้ถูกแสดงในรูปแบบของความน่าจะเป็น เช่นเป็นความน่าจะเป็นของเอกสารหนึ่งที่ถูกดึงออกมาและมีความเกี่ยวข้องกับข้อเรียกร้อง Fallout ก็จะเป็นความน่าจะเป็นของเอกสารหนึ่งที่ถูกดึงออกมาและไม่เกี่ยวข้อง ดังนั้น ระบบค้นคืนสารสนเทศที่มีประสิทธิผลดีจะมี Recall ที่สูงสุดและ Fallout ต่ำที่สุด

ในสถานการณ์การดึงข้อมูลโดยปกติ, Recall, Precision และ Fallout จะขึ้นอยู่กับ Generality Factor ซึ่งเป็นจำนวนเฉลี่ยของเอกสารที่เกี่ยวข้องที่มีอยู่ในกลุ่มเอกสารต่อหนึ่งข้อคำถาม จากตาราง 6.2 และตาราง 6.3 จะสังเกตเห็นว่า 3 ตัววัดใด ๆ จากตัววัด R, P, F และ G จะกำหนดตัววัดตัวที่ 4 ได้อย่างอัตโนมัติ ยกตัวอย่าง, Precision อาจจะถูกกำหนดในรูปศัพท์ของ Recall, Fallout และ Generality ได้ดังนี้

$$P = \frac{R.G}{(R.G) + F(1 - G)}$$

	Relevant	Nonrelevant	
Retrieved	RETREL	RETNREL	RETREL+RETNREL
Not retrieved	NRETREL	NRETREL	NRETREL+ RETNREL
	RETREL +NRETREL	RETREL +NRETNREL	RETREL+ RETNREL+ NRETREL RETNREL

ตารางที่ 6.1 : Contingency Table

อย่างไรก็ตามเนื่องจากจำนวนทั้งหมดของเอกสารที่ไม่เกี่ยวข้อง(RETNREL+ NRETNREL) ในทางปฏิบัติ จะใหญ่กว่าเอกสารที่เกี่ยวข้อง (RETREL + NRETERL) มาก การเปลี่ยนแปลงใด ๆ ใน Generality ของเอกสารกลุ่มหนึ่ง ดูจะมีผลต่อ Fallout น้อยกว่า Precision โดยเฉพาะ ในขณะที่ Generality ลดลง ไม่ว่าจะเนื่องจากจำนวนของเอกสารที่เกี่ยวข้องลดลง หรือเนื่องมาจากจำนวนเอกสารทั้งหมดได้เพิ่มขึ้น, จำนวนของเอกสารที่ถูกดึงหรือเกี่ยวข้อง (RETREL) มีแนวโน้มว่าจะลดลง, แต่จำนวนทั้งหมดของเอกสารที่ถูกดึงออกมา (RETERL+RETNREL) เช่นเดียวกับจำนวนเอกสารที่ไม่เกี่ยวข้อง (RETREL+RETNREL) เช่นเดียวกับจำนวนเอกสารที่ไม่เกี่ยวข้อง (RETNERL+NRETNREL) อาจจะยังคงมีค่าคงที่ ดังนั้น, Precision จะมีค่าที่มีความเปลี่ยนแปลงมากกว่า Fallout นอกไปจากนี้, Fallout จะถูกกำหนดให้เป็นศูนย์เมื่อไม่มีเอกสารใดถูกดึงออกมา (RETREL+RETNREL = 0) เนื่องจากอาจจะปลอดภัยในการที่สมมุติให้จำนวนเอกสารที่ไม่เกี่ยวข้องในกลุ่มเอกสารนั้นมีค่าเป็นศูนย์

Symbol	Evaluation measure	Formula	Explanation
R	Recall	$\frac{RETREL}{RETREL + NRETREL}$	Proportion of relevant actually retrieved
P	Precision	$\frac{RETREL}{RETREL + RETNREL}$	Proportion of retrieved actually relevant
F	Fallout	$\frac{RETNREL}{RETNREL + NRETNREL}$	Proportion of nonrelevant actually retrieved
G	Generality	$\frac{RETREL + NRETREL}{RETREL + NRETREL + NRETREL + NRETNREL}$	Proportion of relevant per query

ตารางที่ 6.2 Typical Retrieval Evaluation Measures

ได้มีการเสนอให้คำนวณค่า Recall-Fallout แทนที่วิธีการวัดด้วย Recall-Precision ตัวอย่างการแสดงถึง Recall-Fallout ได้ถูกแสดงไว้ในรูป 6.2 กราฟในรูป 6.2 ได้แสดงให้เห็นว่า Fallout ไม่ได้ถูกตีความหมายโดยผู้ใช้อย่างง่ายดายเหมือนกับ Precision อันที่จริงเราอาจจะใช้ Recall-Precision และ Recall-Fallout ในการแสดงถึงประสิทธิผลของระบบในสถานการณ์ดึงข้อมูลที่มีความต้องการแตกต่างกัน เนื่องจาก Recall จะแสดงถึงอัตราส่วนของเอกสารที่เกี่ยวข้องที่ถูกดึงออกมาจากการค้นหา ในขณะที่ Precision เป็นตัววัดประสิทธิภาพที่ความเกี่ยวข้องเหล่านี้ได้ถูกดึงออกมา Recall-Precision Output จะเป็น User-Oriented เนื่องจากผู้ใช้โดยปกติจะสนใจในการ Optimizing การดึงเอกสารที่เกี่ยวข้อง ส่วน Fallout เป็นการวัดของประสิทธิผลในการขจัดเอกสารที่ไม่เกี่ยวข้องกับข้อเรียกร้องในกลุ่มของเอกสาร (ซึ่งในหลาย ๆ กรณี, เราสมมุติให้มีขนาดเท่า ๆ กันโดยประมาณ) ด้วยเหตุผลนี้, การแสดงในรูปแบบของ Recall-Fallout อาจจะถูกพิจารณาให้เป็นรูป System Oriented

ตัวอย่างที่ 6.1 การประเมินการค้นคืนสารสนเทศ

สมมุติว่าจะทำการประเมินระบบสามระบบคือ “Bear”, “Cardinal” และ “Wolf” โดยใช้ความเกี่ยวข้องของสารสนเทศและระดับ (relevance information and rankings) สำหรับ สามข้อเรียกร้อง(three queries) กระทำบน Excel spreadsheet (HTTP://sims.berkeley.edu/courses/is202/f04/eval-data_blank.xls)

สิ่งสำคัญที่ต้องการในการประเมินสารสนเทศประกอบด้วย

1. การรวบรวมข้อมูลที่ใช้สำหรับทดสอบ
2. ระบบค้นคืนสารสนเทศ(IR systems) ที่สามารถให้ผลการจัดลำดับที่ตอบสนองต่อข้อสอบถาม(queries)
3. ข้อสอบถามที่ต้องการทดสอบ
4. การประเมินผลลัพธ์และความเกี่ยวพันธ์ที่ประเมินจากข้อสอบถาม(RELEVANCE JUDGEMENTS on the Queries) ซึ่งส่วนมากจะได้ผลลัพธ์เป็นเลขไบนารี (relevant or not relevant) หรือใช้ two scales

Relevance	Judgements		
Document	Query 0	Query 1	Query 2
1	1	0	0
2	1	0	0
3	1	0	0
4	1	0	0
5	1	0	0
6	0	0	0
7	0	0	0
8	0	1	0
9	0	1	0
10	0	1	0
11	0	1	0
12	0	1	0
13	0	0	0
14	0	0	1
15	0	0	1
16	0	0	1
17	0	0	1
18	0	0	1
19	0	0	0
20	0	0	0
21	0	0	0
22	0	0	0
23	0	0	0
24	0	0	0
25	0	0	0

รูปที่ 6.4 : Relevance Information

Query 0 for all Three Systems

Rank	Query 0		Rankings			
	Bear	Rel	**Cardinal**	Rel	**Wolf**	Rel
1	5		7		11	
2	2		16		2	
3	1		9		10	
4	15		18		21	
5	9		4		1	
6	19		17		5	
7	3		8		22	
8	6		5		9	
9	18		11		20	
10	4		15		3	
11	12		12		23	
12	8		10		4	
13	11		6		19	
14	17		2		6	
15	7		19		15	
16	20		20		25	
17	10		21		18	
18	21		25		16	
19	16		22		7	
20	23		1		8	
21	13		23		17	
22	22		13		12	
23	24		24		13	
24	25		3		14	
25	14		14		24	

Query 1 for all Three Systems

	Query 1		Rankings			
Rank	**Bear**	Rel	**Cardinal**	Rel	**Wolf**	Rel
1	12		16		7	
2	14		23		23	
3	8		3		3	
4	23		13		13	
5	9		4		4	
6	5		14		8	
7	20		17		17	
8	21		11		25	
9	4		7		16	
10	19		21		21	
11	3		18		18	
12	10		10		22	
13	7		19		19	
14	18		22		10	
15	11		20		20	
16	17		1		5	
17	24		2		2	
18	13		12		12	
19	2		15		15	
20	16		5		1	
21	6		8		14	
22	22		6		24	
23	15		9		9	
24	25		25		11	
25	1		24		6	

Query 2 for all Three Systems

	Query 2		Rankings			
Rank	Bear	Rel	Cardinal	Rel	Wolf	Rel
1	1		5		23	
2	18		24		9	
3	11		13		25	
4	16		19		18	
5	19		17		1	
6	10		12		2	
7	23		22		11	
8	21		4		10	
9	3		1		19	
10	6		8		16	
11	22		11		12	
12	12		3		22	
13	17		16		15	
14	2		23		8	
15	13		25		24	
16	24		18		14	
17	25		2		7	
18	14		20		13	
19	5		14		17	
20	4		10		20	
21	7		6		21	
22	9		21		4	
23	15		15		5	
24	20		7		6	
25	8		9		3	

We have the number of RELEVANT documents at each ranking level

Rank	Query 0		Rankings			
	Bear	Rel	**Cardinal**	Rel	**Wolf**	Rel
1	5	1	7	0	11	0
2	2	2	16	0	2	1
3	1	3	9	0	10	1
4	15	3	18	0	21	1
5	9	3	4	1	1	2
6	19	3	17	1	5	3
7	3	4	8	1	22	3
8	6	4	5	2	9	3
9	18	4	11	2	20	3
10	4	5	15	2	3	4
11	12	5	12	2	23	4
12	8	5	10	2	4	5
13	11	5	6	2	19	5
14	17	5	2	3	6	5
15	7	5	19	3	15	5
16	20	5	20	3	25	5
17	10	5	21	3	18	5
18	21	5	25	3	16	5
19	16	5	22	3	7	5
20	23	5	1	4	8	5
21	13	5	23	4	17	5
22	22	5	13	4	12	5
23	24	5	24	4	13	5
24	25	5	3	5	14	5
25	14	5	14	5	24	5

	Query 1		Rankings		
Rank	**Bear**	Rel	**Cardinal**	Rel	**Wolf**
1	12	1	16	0	7
2	14	1	23	0	23
3	8	2	3	0	3
4	23	2	13	0	13
5	9	3	4	0	4
6	5	3	14	0	8
7	20	3	17	0	17
8	21	3	11	1	25
9	4	3	7	1	16
10	19	3	21	1	21
11	3	3	18	1	18
12	10	4	10	2	22
13	7	4	19	2	19
14	18	4	22	2	10
15	11	5	20	2	20
16	17	5	1	2	5
17	24	5	2	2	2
18	13	5	12	3	12
19	2	5	15	3	15
20	16	5	5	3	1
21	6	5	8	4	14
22	22	5	6	4	24
23	15	5	9	5	9
24	25	5	25	5	11
25	1	5	24	5	6

	Query 2		Rankings			
Rank	Bear	Rel	Cardinal	Rel	Wolf	Rel
1	1	0	5	0	23	0
2	18	1	24	0	9	0
3	11	1	13	0	25	0
4	16	2	19	0	18	1
5	19	2	17	1	1	1
6	10	2	12	1	2	1
7	23	2	22	1	11	1
8	21	2	4	1	10	1
9	3	2	1	1	19	1
10	6	2	8	1	16	2
11	22	2	11	1	12	2
12	12	2	3	1	22	2
13	17	3	16	2	15	3
14	2	3	23	2	8	3
15	13	3	25	2	24	3
16	24	3	18	3	14	4
17	25	3	2	3	7	4
18	14	4	20	3	13	4
19	5	4	14	4	17	5
20	4	4	10	4	20	5
21	7	4	6	4	21	5
22	9	4	21	4	4	5
23	15	5	15	5	5	5
24	20	5	7	5	6	5
25	8	5	9	5	3	5

ทำการคำนวณหาค่า recall และ precision จากสูตร

$$\text{Precision} = \frac{|\,\text{RelRetrieved}\,|}{|\,\text{Retrieved}\,|}$$

$$\text{Recall} = \frac{|\,\text{RelRetrieved}\,|}{|\,\text{Rel in Collection}\,|}$$

By systems… for Bear…

	Query 0		Query 1		Query 2	
Bear	Recall	Prec	Recall	Prec	Recall	Prec
1	0.20	1.00	0.20	1.00	0.00	0.00
2	0.40	1.00	0.20	0.50	0.20	0.50
3	0.60	1.00	0.40	0.67	0.20	0.33
4	0.60	0.75	0.40	0.50	0.40	0.50
5	0.60	0.60	0.60	0.60	0.40	0.40
6	0.60	0.50	0.60	0.50	0.40	0.33
7	0.80	0.57	0.60	0.43	0.40	0.29
8	0.80	0.50	0.60	0.38	0.40	0.25
9	0.80	0.44	0.60	0.33	0.40	0.22
10	1.00	0.50	0.60	0.30	0.40	0.20
11	1.00	0.45	0.60	0.27	0.40	0.18
12	1.00	0.42	0.80	0.33	0.40	0.17
13	1.00	0.38	0.80	0.31	0.60	0.23
14	1.00	0.36	0.80	0.29	0.60	0.21
15	1.00	0.33	1.00	0.33	0.60	0.20
16	1.00	0.31	1.00	0.31	0.60	0.19
17	1.00	0.29	1.00	0.29	0.60	0.18
18	1.00	0.28	1.00	0.28	0.80	0.22
19	1.00	0.26	1.00	0.26	0.80	0.21
20	1.00	0.25	1.00	0.25	0.80	0.20
21	1.00	0.24	1.00	0.24	0.80	0.19
22	1.00	0.23	1.00	0.23	0.80	0.18
23	1.00	0.22	1.00	0.22	1.00	0.22
24	1.00	0.21	1.00	0.21	1.00	0.21
25	1.00	0.20	1.00	0.20	1.00	0.20

For Cardinal

	Query 0		Query 1		Query 2	
Cardinal	Recall	Prec	Recall	Prec	Recall	Prec
1	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00
5	0.20	0.20	0.00	0.00	0.20	0.20
6	0.20	0.17	0.00	0.00	0.20	0.17
7	0.20	0.14	0.00	0.00	0.20	0.14
8	0.40	0.25	0.20	0.13	0.20	0.13
9	0.40	0.22	0.20	0.11	0.20	0.11
10	0.40	0.20	0.20	0.10	0.20	0.10
11	0.40	0.18	0.20	0.09	0.20	0.09
12	0.40	0.17	0.40	0.17	0.20	0.08
13	0.40	0.15	0.40	0.15	0.40	0.15
14	0.60	0.21	0.40	0.14	0.40	0.14
15	0.60	0.20	0.40	0.13	0.40	0.13
16	0.60	0.19	0.40	0.13	0.60	0.19
17	0.60	0.18	0.40	0.12	0.60	0.18
18	0.60	0.17	0.60	0.17	0.60	0.17
19	0.60	0.16	0.60	0.16	0.80	0.21
20	0.80	0.20	0.60	0.15	0.80	0.20
21	0.80	0.19	0.80	0.19	0.80	0.19
22	0.80	0.18	0.80	0.18	0.80	0.18
23	0.80	0.17	1.00	0.22	1.00	0.22
24	1.00	0.21	1.00	0.21	1.00	0.21
25	1.00	0.20	1.00	0.20	1.00	0.20

For Wolf

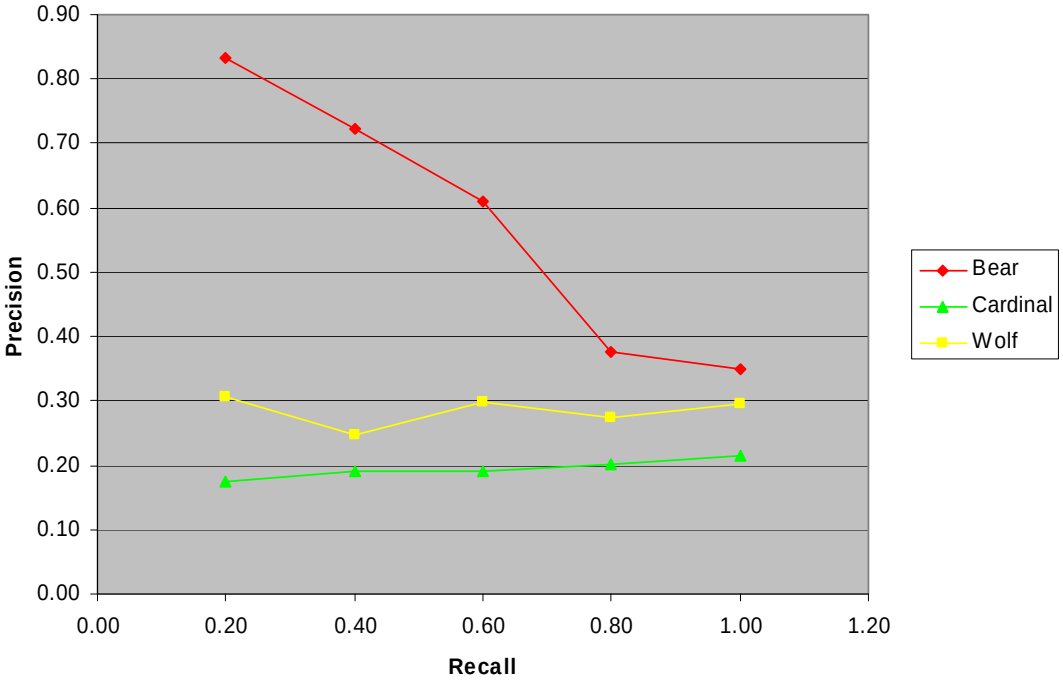
	Query 0		Query 1		Query 2	
Wolf	Recall	Prec	Recall	Prec	Recall	Prec
1	0.00	0.00	0.00	0.00	0.00	0.00
2	0.20	0.50	0.00	0.00	0.00	0.00
3	0.20	0.33	0.00	0.00	0.00	0.00
4	0.20	0.25	0.00	0.00	0.20	0.25
5	0.40	0.40	0.00	0.00	0.20	0.20
6	0.60	0.50	0.20	0.17	0.20	0.17
7	0.60	0.43	0.20	0.14	0.20	0.14
8	0.60	0.38	0.20	0.13	0.20	0.13
9	0.60	0.33	0.20	0.11	0.20	0.11
10	0.80	0.40	0.20	0.10	0.40	0.20
11	0.80	0.36	0.20	0.09	0.40	0.18
12	1.00	0.42	0.20	0.08	0.40	0.17
13	1.00	0.38	0.20	0.08	0.60	0.23
14	1.00	0.36	0.40	0.14	0.60	0.21
15	1.00	0.33	0.40	0.13	0.60	0.20
16	1.00	0.31	0.40	0.13	0.80	0.25
17	1.00	0.29	0.40	0.12	0.80	0.24
18	1.00	0.28	0.60	0.17	0.80	0.22
19	1.00	0.26	0.60	0.16	1.00	0.26
20	1.00	0.25	0.60	0.15	1.00	0.25
21	1.00	0.24	0.60	0.14	1.00	0.24
22	1.00	0.23	0.60	0.14	1.00	0.23
23	1.00	0.22	0.80	0.17	1.00	0.22
24	1.00	0.21	1.00	0.21	1.00	0.21
25	1.00	0.20	1.00	0.20	1.00	0.20

Averages for all systems

ต่อจากนั้นทำการหาค่าเฉลี่ยของprecision ที่ระดับของ recall ที่กำหนด ของระบบทั้งหมดกำหนดไว้ 5 ระดับได้แก่ 20% หรือ 0.20, 40% หรือ 0.40, 60% หรือ 0.60, 80% หรือ 0.80,100% หรือ 1.00 การคำนวณแต่ละข้อสอบถามนั้น หาจากค่าของ precision ในลำดับที่สูงที่สุด(HIGHEST RANKED precision value) ที่จับคู่กับค่าของ Recall ในระดับที่ยอมรับ และนำค่าเฉลี่ยของค่า precision ข้ามข้อสอบถามทั้งหมด

	Bear	Cardinal	Wolf
0.20	0.83	0.18	0.31
0.40	0.72	0.19	0.25
0.60	0.61	0.19	0.30
0.80	0.38	0.20	0.27
1.00	0.35	0.21	0.30

Graphing the results



Cutoff Levelsfor all systems

Cutoff	Bear	Cardinal	Wolf
5	0.53	0.13	0.2
10	0.33	0.13	0.23
20	0.23	0.18	0.22

6.9 บทสรุป

การประเมินผลระบบค้นคืนสารสนเทศเป็นการพิจารณาดูว่า ระบบค้นคืนสารสนเทศจะมีประสิทธิผลมากน้อยเพียงใด ซึ่งประสิทธิภาพของระบบจะวัดได้จากการใช้ทรัพยากรของระบบ เช่น การใช้เนื้อที่หน่วยความจำ, การใช้ CPU Time เป็นต้น ส่วนประสิทธิผลการวัดด้วยค่า Recall และ Precision ทำได้โดยใช้สูตรต่อไปนี้คือ

$$\text{Recall} = \frac{\text{จำนวนเอกสารที่เกี่ยวข้องที่ถูกดึงออกมา}}{\text{จำนวนเอกสารที่เกี่ยวข้องทั้งหมด}}$$

$$\text{Precision} = \frac{\text{จำนวนเอกสารที่เกี่ยวข้องที่ถูกดึงออกมา}}{\text{จำนวนเอกสารที่ถูกดึงออกมาทั้งหมด}}$$

เราจะทำการประเมินผลระบบเมื่อต้องการเปรียบเทียบความสามารถของระบบที่มีอยู่กับระบบอื่น ๆ, เมื่อส่วนประกอบบางส่วนของระบบได้เปลี่ยนไป และเมื่อมีการนำเอาส่วนประกอบใหม่ของระบบเข้ามารวมกับระบบที่มีอยู่เดิม

การที่จะประเมินผลระบบนั้น จำเป็นต้องมีการทดสอบระบบ การวัดความสามารถการทำงานของระบบจะต้องใช้หลักเกณฑ์การวัดที่สามารถแสดงเป็นปริมาณตัวเลขได้ เราเรียกว่า Objective Measurement ซึ่งจะหาได้โดยการบันทึก, การสังเกตโดยตรงซึ่งจะใช้แบบสอบถาม, เทคนิคในการสัมภาษณ์ หรือใช้เครื่องกลในการรวบรวมข้อมูล ในกรณีที่ไม่อาจจะหาค่าของพารามิเตอร์ได้ เราก็อาจจะคาดคะเนด้วยการใช้ Sampling Techniques

ส่วนประกอบของระบบค้นคืนสารสนเทศที่มีอิทธิพลต่อการทำงานของระบบได้แก่ ข้อมูลนำเข้า, โครงสร้างของแฟ้มข้อมูล, ภาษาดัชนี, วิธีการสร้างดัชนี, การแสดงรูปแบบของเอกสาร, การวิเคราะห์ข้อคำถาม,วิธีการค้นหาข้อมูล และรูปแบบของผลลัพธ์ที่ได้จากระบบ

หลักเกณฑ์การประเมินผลที่น่าสนใจเกี่ยวกับผู้ใช้ ได้แก่ Recall, Precision, ความพยายามของผู้ใช้, Response Time, การแสดงรูปแบบของผลลัพธ์ที่ได้จากการค้นหา และ Collection Coverage

การคำนวณค่า Recall และ Precision นั้น จะมีปัญหาในการวัดปริมาณของเอกสารที่เกี่ยวข้อง และขาดลักษณะที่เป็นไปในแนวขนานในคุณสมบัติของ 2 ตัวที่ว่านี้ Fallout เป็นตัววัดที่จะสะท้อนให้เห็นหารกระทำของเอกสารที่ไม่เกี่ยวข้องอยู่ในรูปแบบที่สอดคล้องกับRecall ซึ่ง

$$\text{Fallout} = \frac{\text{จำนวนของเอกสารที่ไม่เกี่ยวข้องที่ถูกดึงออกมา}}{\text{จำนวนทั้งหมดของเอกสารที่ไม่เกียวข้องในกลุ่มเอกสาร}}$$

ระบบค้นคืนสารสนเทศที่มีประสิทธิผลดี จะมี Recall ที่สูงสุดและ Fallout ต่ำที่สุด

แบบฝึกหัด

1. การประเมินผลระบบ IR จะประเมินจากอะไรบ้าง เหตุใดจึงต้องมีการประเมินผล
2. ส่วนประกอบต่าง ๆ ของระบบ IR ที่เกี่ยวข้องกับการประเมินผล ได้แก่อะไรบ้าง
3. Collection Coverage หมายถึงอะไร
4. หลักเกณฑ์ที่ใช้ในการประเมินผลที่เกี่ยวกับผู้ใช้ ได้แก่อะไรบ้าง จงอธิบาย
5. เหตุใด การวัด Recall และ Precision จึงเป็นเรื่องค่อนข้างยากในทางปฏิบัติ
6. ในกรณีที่การค้นหาครั้งที่ i และ j ปรากฏผลว่า

$$RECALL_i > RECALL_j$$

$$PRECISION_i < PRECISION_j$$

   การหาครั้งใดจะดีกว่ากัน จงให้เหตุผลประกอบ
7. การวัด Recall และ Precision ในทางปฏิบัตินั้น จะเกิดปัญหาเกี่ยวกับวิธีการวัดอย่างไร
8. ข้อคำถามที่ผู้ใช้ระบุเพื่อเรียกร้องหาเอกสารที่ต้องการนั้น พอจะแบ่งออกได้เป็นกี่ประเภทอะไรบ้าง
9. Fallout หมายถึงอะไร เกี่ยวข้องกับ Recall และ Precision อย่างไร

บรรณานุกรม

สุมาลี เมืองไพศาล ,"**การจัดการข้อมูล และการเรียกใช้ข้อมูล**" ,สำนักพิมพ์มหาวิทยาลัยรามคำแหง ,CS337 ,2535

Sirak Kaewjamnong, "**Text Properties and Languages**" , ภาควิชาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยศิลปากร ,http://www.cs.su.ac.th/~sirak/517632/IR%20Model.ppt,

Prof. Ray Larson & Prof. Marc Davis ,**" Information OrganizationAnd Retrieval : IR Evaluation "** , http://www.sims.berkeley.edu/academics/courses/is202/f04/

HTTP://sims.berkeley.edu/courses/is202/f04/eval-data_blank.xls