CAS Information Engineering

Modul: Information Retrieval

Tag 6 – Fallstudien

Dr. Mirco Kocher

Frage

- Welches System ist besser?
  - A: konstante Präzision von 90%
  - B: für 9 von 10 Fragen 100% Präzision, aber versagt für jede zehnte Anfrage

Fallstricke bei der Interpretation der Masse

- Der Durchschnitt verwischt Leistungsunterschiede bei einzelnen Anfragen
- Gewisse Masse haben mathematische Unstabilitäten
- Das häufig verwendete Mass “Präzision nach 10 Dokumenten” kann keine Differenzierung liefern, wenn die Anzahl der relevanten Dokumente sehr hoch ist
- Desweitern kann im Fall sehr weniger relevanter Dokumente die maximale Präzision auf unterschiedliche Art erreicht werden, und ist schwer zu interpretieren, wenn sie gemittelt wird

Fallstricke bei der Interpretation der Masse Beispiele

- Total 1 relevantes Dokument
  - Die Präzision nach 10 Dokumenten ist 0.1, egal ob das relevante Dokument auf Rang 1 oder Rang 10 gefunden wird
  - Wird der Durchschnitt mit anderen Anfragen gebildet, so wird die Anfrage unnötig “benachteiligt”, da die maximal erreichbare Leistung nur 0.1 beträgt

- Total 250 relevante Dokumente
  - Die Präzision nach 10 Dokumenten ist 1.0, da vermutlich mindestens 10 Dokumente relative einfach zu finden sind
  - Systeme sind nicht differenzierbar, weil alle Systeme finden wohl mindestens 10 Dokumente

Beispiel: Average Precision

- Betrachten wir als Beispiel eine einzelne Average Precision einer gegebenen Anfrage:
  - Anfrage “Vegetables, Fruit and Cancer” (drei relevante Dokumente)

Rang	Okapi (A)		Okapi & PRF (B)	
1	**R**	1/1	nR	
2	**R**	2/2	**R**	1/2
3	nR		**R**	2/3
…	nR		nR	
35	nR		**R**	3/35
…	nR		nR	
108	**R**	3/108	nR	
	AP =	0.6759	AP =	0.4175
				-38.2%

Quelle: Jacques Savoy

Wahrnehmung der Suchqualität

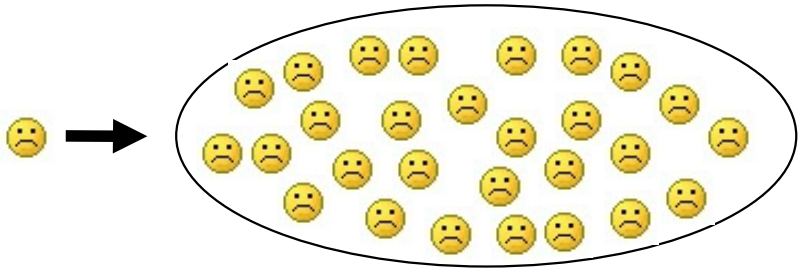
- EndbenutzerIn nimmt durchschnittliche Performance nicht unmittelbar wahr
- BenutzerIn sieht bloss die Performance der **eigenen** momentanen Anfrage
- Unzufriedene Kunden erzählen im Durchschnitt zehn Personen von ihrer schlechten Erfahrung. Zwölf Prozent erzählen sie bis zu 20 Personen
  - → Schlechte News verbreiten sich schnell. Mit der Nutzung des Internets (E-Mail, Blogs, etc.) kann es noch verheerender ausfallen
- Im Gegensatz dazu erzählen zufriedene Kunden im Durchschnitt nur fünf Personen von ihrer positiven Erfahrung
  - → Gute News verbreiten sich langsamer

- Quelle: Michael A. Aun, http://www.nsacentralflorida.com/Articles/Thirteencsfacts.pdf

“Verbreitung der Stimmung”

- Bei **1** Feedback eines unzufriedenen Kunden haben **20** andere Kunden bereits dasselbe Problem gehabt, geben aber kein Feedback
- Wie viele (potentielle) Kunden wissen nun über die negative Erfahrung, bis das erste Feedback kommt?
- → (**1**+**20**)*10 = 210 (potentielle) Kunden

Im Durchschnitt erzählt man 10 Personen von seiner negativen Erfahrung



Robustheit

- Folge:
  - Das System darf sich keine groben Ausrutscher erlauben
  - Durchschnittliche Performance kompensiert nicht für Ausrutscher
  - Man sollte sich nicht allzu viele Gedanken über die positiven Meldungen machen
- → Mean Average Precision kann irreführend sein
- Eine Betrachtung der Ausreisser ist nötig
  - z.B. Average Precision einzelner Anfragen
- Ein Mass für “Robustheit” wäre interessant
  - In diesem Bereich wird momentan aktiv geforscht

Evaluation eines IR Projektes (Fallstudie)

- Für Projektverantwortliche stellt sich die Frage, ob die richtige Suchtechnologie verwendet wird
- Konzentrieren wir uns auf die Retrievaleffektivität (“Suchqualität”), so stellen sich dieselben Fragen, welche die Cranfield-Methode zu beantworten hilft
- Wir lernen aus TREC/CLEF wie gut Systeme in Standardsituationen arbeiten
  - → hilft bei der grundsätzlichen Wahl der richtigen Suchtechnologie, der richtigen Suchparadigmen etc.
- Wenn unsere konkrete Situation bedeutend abweicht, müssen wir uns weitere Fragen stellen

Evaluation eines IR Projektes (Fallstudie)

- Ein solches IR-Projekt sollte nie unabhängig von Benutzern, Bedürfnissen, und den konkreten Daten durchgeführt werden
- Grundsätzlich werden vorhandene Testkollektionen diesen Rahmenbedingungen aber nicht gerecht
- Es muss eine pragmatische Alternative zu einer vollen Evaluation nach Cranfield-Methode gewählt werden (aus Gründen der Durchführbarkeit), welche die Benutzer und ihre Bedürfnisse reflektiert

Fallstudie 1: “Known Item Retrieval”

- Gegeben zwei Systeme, welches eine Sammlung von homogenen Dokumenten erschliessen (z. B. 100‘000 Patentbeschreibungen)
- Es werden 10 Dokumente als Anfragen ausgewählt
- Es werden 5 Personen für den Test rekrutiert
- Die 5 Personen konstruieren pro Ziel-Dokument eine Anfrage, mit dem Ziel, das Dokument “wiederzufinden”
- Die Resultate werden wie folgt bewertet:
  - Gefunden (Top-20): 1 Pkt.
  - Bonus:
    - Position 1: 4 Pkt.
    - Position 2 bis 5: 3 Pkt.
    - Position 6-10: 2 Pkt.
    - Position 11-20: 1 Pkt.

Aufwand der Evaluation

- Es werden 50 Anfrageinstanzen (5 Personen mal 10 Ziel-Dokumente) in beiden Systemen ausgewertet
- Dabei werden je 20 Dokumente “beurteilt”
  - Insgesamt müssen 50*20*2 = 2000 Dokumente auf Übereinstimmung mit den 10 gesuchten Ziel-Dokumenten getestet werden
- Es können dabei maximal 50*5 = 250 Punkte erreicht werden
- Systeme werden aufgrund ihrer erreichten Punktzahl beurteilt

Kritische Besprechung der Evaluation

- Besprechen wir das Vorgehen
- Ihr(e) Kommentar(e)?

Kritische Besprechung der Evaluation

- Einige Ansätze zur Besprechung:
  - 10 Ziel-Dokumente:
    - Genug?
    - Repräsentativ?
  - 5 Personen:
    - Genug?
    - Repräsentativ?
    - Ist es sinnvoll, die beiden Faktoren “Personen” und “Ziel-Dokumente” zu vermischen?
  - “Selbstgestricktes” Mass:
    - Begründung?
    - Interpretation?
    - Vergleichbarkeit?
    - Praxisrelevant?

“Known Item Retrieval”

- Idee: es wird mit der Suchmaschine nach “bekannten” Dokumenten gesucht
- Simuliert “Da war doch was”-Informationsbedürfnis
- Der Erfolg der Suchmaschine wird bewertet nach der Erfolgsquote, die gesuchten Dokumente (wieder-)zufinden
- Aber Achtung: widerspricht der Annahme, dass unbekannte Information gesucht wird
- Es muss pro Anfrage nur sehr wenig Auswertungsarbeit gemacht werden: das “relevante” Dokument ist bekannt, und muss in der Rangliste lokalisiert werden → Evaluation ohne Relevance Assessments
- Mass für Effektivität “Mean Reciprocal Rank” (MRR): MRR = 1/Rang
- Es wird der Durchschnitt über eine Anzahl von Anfragen ermittelt

“Known Item Retrieval”

- Das Ziel bei Known Item Retrieval ist es, ein “bekanntes” Dokument wieder zu finden
- Inwiefern ist dieses Ziel erfüllt, wenn weitere Dokumente mit der (fast) gleichen Information vom System gefunden werden?
- Ein IR-System kann nur die explizite Formulierung eines Informationsbedürfnisses zur Suche einsetzen
- Passt diese Formulierung auf mehr als ein Dokument “gleich” gut, wie verhält sich dann das Suchresultat?

“Known Item Retrieval”

- Wenn die Anfragen aufgrund von Begriffen aus den gesuchten Dokumenten konstruiert wird, welches der folgenden System ist dann bevorzugt?
  - jenes mit Wortnormalisierung
  - jenes ohne Wortnormalisierung

- → die gesuchten Dokumente müssen “einzigartig” sein
- → die Anfrage soll nicht “reverse engineered” werden

Fallstudie 2 (Kürzere Zusammenfassung)

- Vertikale Suche oder Google?
- Es soll verglichen werden, ob die verwendete vertikale Suche “besser” als Google funktioniert
- Vertikale Suche = Suche in einer spezifischen Domäne, für eine bestimmte Klasse von Benutzern, mit Informationsbedürfnissen aus einem eingeschränkten Bereich

Fallstudie 2 (Kürzere Zusammenfassung)

- Es werden die häufigsten Suchanfragen ermittelt (aus Logfiles)
- Hinter jeder Anfrage steht ein Informationsbedürfnis
- Innerhalb der Domäne können diese Bedürfnisse eingegrenzt werden
- Für jede Anfrage werden diejenigen Aspekte ermittelt, welche in den bestrangierten Dokumenten behandelt werden (sollen)

Auswertung Fallstudie 2

- Für jede Anfrage wird dann intellektuell bestimmt, welches System mehr Aspekte abdeckt, die den zu erwartenden Informationsbedürfnissen gerecht werden
- Im Zweifelsfall werden die Systeme als gleichwertig beurteilt
- Die Anzahl der Anfragen mit einem klaren Vorteil für das eine oder andere System wird ermittelt → “Aspectual Recall”
- Aufwand:
  - Beispiel: 50 häufigste Anfragen (decken typischerweise eine hohe Anzahl der tatsächlichen Anfrageinstanzen ab), 10 bestrangierte Dokumente = 500 Dokumente, die auf Aspekte untersucht werden müssen
- Diese Evaluation kann keine Aussage über Ausbeute machen, sondern nur über die Vollständigkeit des Resultates

Kritische Betrachtung

- Beide Alternativen
  - Known Item Retrieval
  - Aspectual Recall
- zielen darauf ab, den Relevance Assesment Aufwand zu limitieren
- Sie liefern nur beschränkt die gleiche Aussage wie die Ausbeute

Schlussfolgerungen

- Das ausführliche Evaluieren einer Suchmaschine ist sehr aufwändig
- Systementwickler evaluieren im Rahmen von Evaluationskampagnen
  - CLEF, TREC
- Ergebnisse solcher Evaluationskampagnen können zur grundsätzlichen Wahl der richtigen Suchmethodiken herangezogen werden
- Machmal muss ein Kompromiss gewählt werden
  - Wichtig:
    - Wahl einer korrekten Methodik
    - Vergleichbarkeit (Wahl der Masse)
    - Korrekte Interpretation des Resultats (speziell der absoluten Werte)

Schlussfolgerungen

- Die Evaluation sollte sich auf die wichtigen Aspekte beschränken
- Aspekte nicht vermischen
- Vergleiche liefern stabilere Ergebnisse als absolute Werte
- Absolute Werte sind im Allgemeinen nicht isoliert interpretierbar