

Information Engineering 1: Information Retrieval

Evaluation

Kapitel 4

Martin Braschler

Ziele

- Warum sollte man IR Systeme evaluieren?
- Wie setzt man eine gute Evaluation auf? (Theorie)
- Was ist die gebräuchlichste Vorgehensweise bei einer IR Evaluation?
- Was sind pragmatische Kompromisse bei einer IR Evaluation?
- Fallstudien

Warum sollte man Information Retrieval Systeme evaluieren?

- IR Systeme unterscheiden sich grundsätzlich in ihren „Grunddimensionen“:
 - Was sind die Dokumente/Informationen, welche erschlossen werden (heterogen, homogen, horizontal, vertikal, ...)?
 - Was sind die Paradigmen des Zugriffs (Pull, Push, Katalog, ...)?
 - Was sind die Benutzerkategorien, welche angesprochen werden?
 - Was sind die Bedürfniskategorien, welche befriedigt werden sollen?
 - Was ist das Geschäftsmodell?
- Klassiker #1: „ich finde nichts!“
- Klassiker #2: „Google (Web Search) funktioniert viel besser!“
- ➔ Es muss evaluiert werden, ob die richtige Suchtechnologie am richtigen Ort eingesetzt wird!

Einleitung

- Evaluation ist nötig, um die „Leistung“ des Systems zu bewerten.
- Da Information kein „greifbares“ Gut ist, ist eine Kosten/Nutzen-Analyse nur schwer zu erstellen.
- Es ist schwierig, den Anteil der einzelnen Komponenten (Faktoren) an einem positiven oder negativen Suchresultat zu ermitteln (Datenabdeckung, Dateneinteilung, Anfrageformulierung, ...)
 - Wie gut ist das System?
 - Wie gut könnte ein generisches System gleicher Bauart sein?
 - Wie gut wäre ein optimales System?
 - → wird das richtige System eingesetzt?
- Achtung: es ist schwierig, aus vergangener Leistung auf zukünftige Leistung zu schliessen!

„Zutaten einer Evaluation“:

1. Aufgabenstellung

- Eine seriöse Evaluation setzt eine klare Aufgabenstellung voraus:
 - Was sind die Motivation, Ziele, und Vorgehensweise der Evaluation?
 - Welches sind die an den Resultaten interessierten Parteien?
 - Wird eine „interne“ oder „externe“ Evaluation angestrebt?
 - Ist das Vorgehen „investigativ“, oder wird „experimentell“ gearbeitet?
 - Wird das System als „black box“ oder als „glass box“ behandelt?
 - Was ist der Richtwert (Benchmark)? Sonstige Massstäbe?
 - Wird erschöpfend oder indikativ analysiert?
 - Werden qualitative oder quantitative Kriterien untersucht?

„Zutaten einer Evaluation“:

2. Ausgestaltung der Evaluation

- Es folgt das „Design“ der Evaluation, welche der Aufgabenstellung gerecht werden muss:
- Was sind die Leistungsfaktoren?
 - Umgebungsvariablen
 - Systemparameter
- Was sind die Leistungsmassstäbe?
 - Durchschnitte (→ Problem?), Minimum, Maximum, etc.
- Was sind die Leistungsmasse?
 - Effektivität
 - Effizienz
 - Akzeptanz
- Was sind die Daten?
- Wie ist das Prozedere zu gestalten?

„Zutaten einer Evaluation“:

3. Testdaten

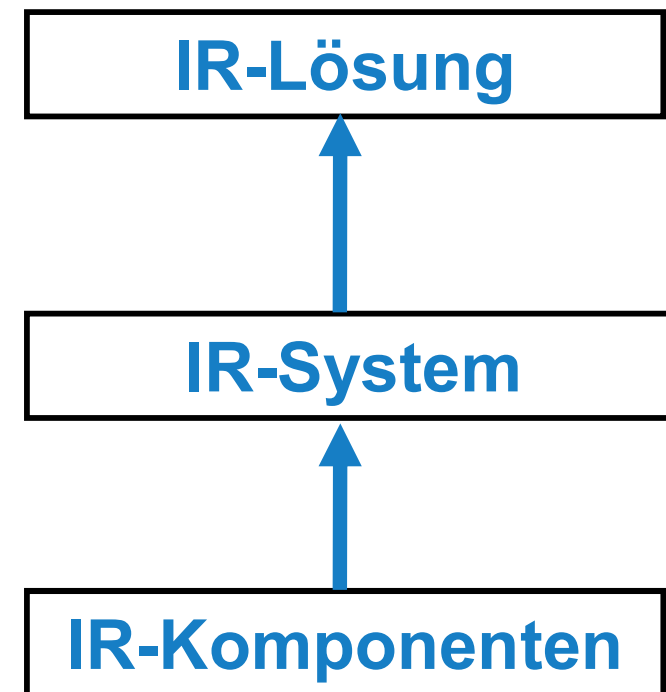
- Für die Evaluation eines IR-Systems werden in der Regel Testdaten gebraucht.
- Daten über die Benutzer
- Daten über Informationsbedürfnisse
- Dokumentdaten
- Daten über die Relevanz der Dokumente in Hinsicht auf die Informationsbedürfnisse

IR-Evaluation nach Cranfield/SMART

- Im akademischen Umfeld der Disziplin Information Retrieval hat sich das „Cranfield-Paradigma“ durchgesetzt (Implementation des Paradigma auch als SMART-Auswertung bekannt).
- Wurde in den 1960ern (1963 & 1967) eingeführt (Cleverdon).
- Wurde seit den späten 1960ern in den SMART Experimenten eingesetzt.
- Ist heute das meistverbreitete Paradigma in akademischen Kreisen.
- Aktive Forschung zu diesem Paradigma, insbesondere zur Frage, wie es auf sehr grosse Datenbestände („Big Data“) ausgeweitet werden kann

3 Stufen der Evaluation der Retrievaleffektivität

- Ganzer Wissensbeschaffungsprozess (UI-Fragen, etc)
- Ganzes Retrievalsystem (Anfrage -> Dokument)
- Komponenten des Retrievalsystems (Stemmer, etc.)



Ebene

IR-Evaluation nach Cranfield/SMART: Aufgabenstellung

- Zu evaluieren ist die Retrievaleffektivität des IR-Systems, mittels geeigneter Masse
- Es wird eine „interne“, „direkte“ Evaluation angestrebt, d. h., die Effektivität soll direkt gemessen werden, und nicht über ihren Anteil an einem grösseren Resultat, d.h. der IR-Lösung bewertet werden.
- Die Evaluation soll experimentell erfolgen.
- Das System wird während der Evaluation als „black box“ behandelt.
- Die Evaluation erfolgt im Vergleich zu einem vorgegebenen Richtresultat (das optimale Retrievalresultat).
- Es soll eine quantitative Evaluation durchgeführt werden (Effektivität).
- Ihr Semesterbeitrag ist Teil einer "Cranfield-Evaluation"!

IR-Evaluation nach Cranfield/SMART: Ausgestaltung (Design)

- Es wird ein Labortest verwendet.
- Vorgehen:
 - Das Setup ist von operationellen Umgebungen abgekoppelt.
 - Es wird von spezifischen Benutzern und Interpretationen ihrer Informationsbedürfnisse abstrahiert (→ Die Anzahl der Umgebungsvariablen wird minimiert).
 - Leistung wird als durchschnittliche Leistung über eine Anzahl von Retrievalvorgängen gemessen.
 - Als Leistungsmasse kommen Präzision und Ausbeute zum Einsatz („Retrievaleffektivität“).
 - Die Dokumentdaten werden geeignet bestimmt und „eingefroren“.
 - Es wird ein „Batch-Setup“ verwendet, und die Experimente sind beliebig nachvollziehbar und **wiederholbar**.

IR-Evaluation nach Cranfield/SMART: Testdaten

- Verwendet wird eine so genannte „Testkollektion“.
- Diese umfasst:
 - Eine Reihe von Informationsbedürfnissen („Topics“) von fiktiven Benutzern.
 - Eine „eingefrorene“ Menge an Dokumenten als Suchdaten.
 - Relevanzbeurteilungen von Dokumenten in Hinsicht auf die Informationsbedürfnisse (Interpretation der Informationsbedürfnisse).

IR-Evaluation nach Cranfield/SMART: Ausbeute/Präzision

- Zwei allgemein anerkannte Masse für Retrievalqualität sind Ausbeute und Präzision.
- Diese Eigenschaften dieser Masse sind gut analysiert und verstanden.
- Beide Masse modellieren die Annahme, dass möglichst viel relevante, und möglichst wenig irrelevante Information gefunden werden soll.

$$\text{Präzision} := \frac{\text{\#relevante Dokumente im Resultat}}{\text{\#Dokumente im Resultat}}$$

$$\text{Ausbeute} := \frac{\text{\# relevante Dokumente im Resultat}}{\text{\#relevante Dokumente in der Kollektion}}$$

- Ziel: Optimierung eines oder beider Kriterien!
- Die beiden Masse sind mengenbasiert.
- Die beiden Masse widersprechen sich:
 - hohe Ausbeute → niedrige Präzision
 - hohe Präzision → niedrige Ausbeute

Mathematische Definition

- Ausbeute : $\rho_r(q) := \frac{|D_r^{rel}(q)|}{|D^{rel}(q)|}$

- Präzision : $\pi_r(q) := \frac{|D_r^{rel}(q)|}{|D_r(q)|}$

- Interpolierte Ausbeute/Präzisionsfunktion : $\Pi_q(\rho) := \max \{ \pi_r(q) \mid \rho_r(q) \geq \rho \}$

- Wobei:

$D_r^{rel}(q)$ Relevante Dokumente in der Resultatmenge

$D^{rel}(q)$ Menge aller relevanter Dokumente

$D_r(q)$ Alle Dokumente der Resultatmenge

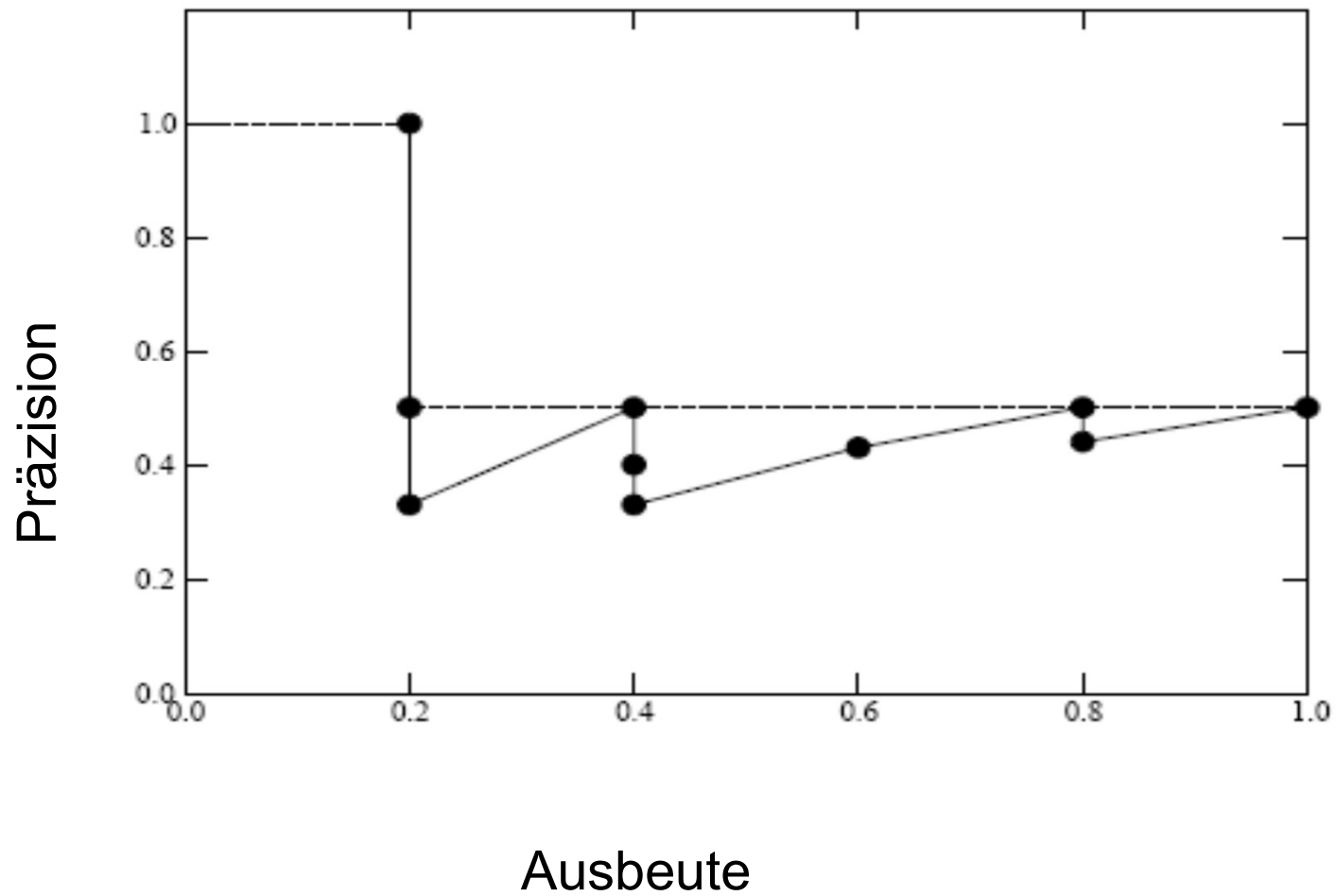
Ausbeute/Präzision

■ Berechnung von Ausbeute und Präzision auf Ranglisten

Rang	Relevant?	Ausbeute	Präzision	Prz. Interpoliert
1	+	0.20	1.00	1.00
2	-	0.20	0.50	0.50
3	-	0.20	0.33	0.50
4	+	0.40	0.50	0.50
5	-	0.40	0.40	0.50
6	-	0.40	0.33	0.50
7	+	0.60	0.43	0.50
8	+	0.80	0.50	0.50
9	-	0.80	0.44	0.50
10	+	1.00	0.50	0.50

Verständnisfrage: welche Annahme zur Ausbeute wurde hier getroffen?

Ausbeute/Präzisions-Graph



„Average Precision“

- „non-interpolated“: Durchschnitt der Präzisionswerte an den Rängen aller relevanten Dokumente.
- „interpolated“: Durchschnitt der Präzisionswerte für spezielle Ausbeutegrade.
- Für das verwendete Beispiel gilt:
- $AVGPREC = (Prec(0.25) + Prec(0.50) + Prec(0.75)) / 3 = (0.50 + 0.50 + 0.50) / 3 = 0.50$
- Ist das beliebteste „Ein-Zahl-Mass“. Konvergiert gegen die Fläche unter dem Ausbeute/Präzisions-Graph.
- Aber: die Widersprüchlichkeit von Ausbeute und Präzision bedingt, dass „Average Precision“ nicht immer ein praxisrelevantes Mass darstellt.

„Mean Average Precision“

- Durchschnittliche Average Precision über alle Anfragen

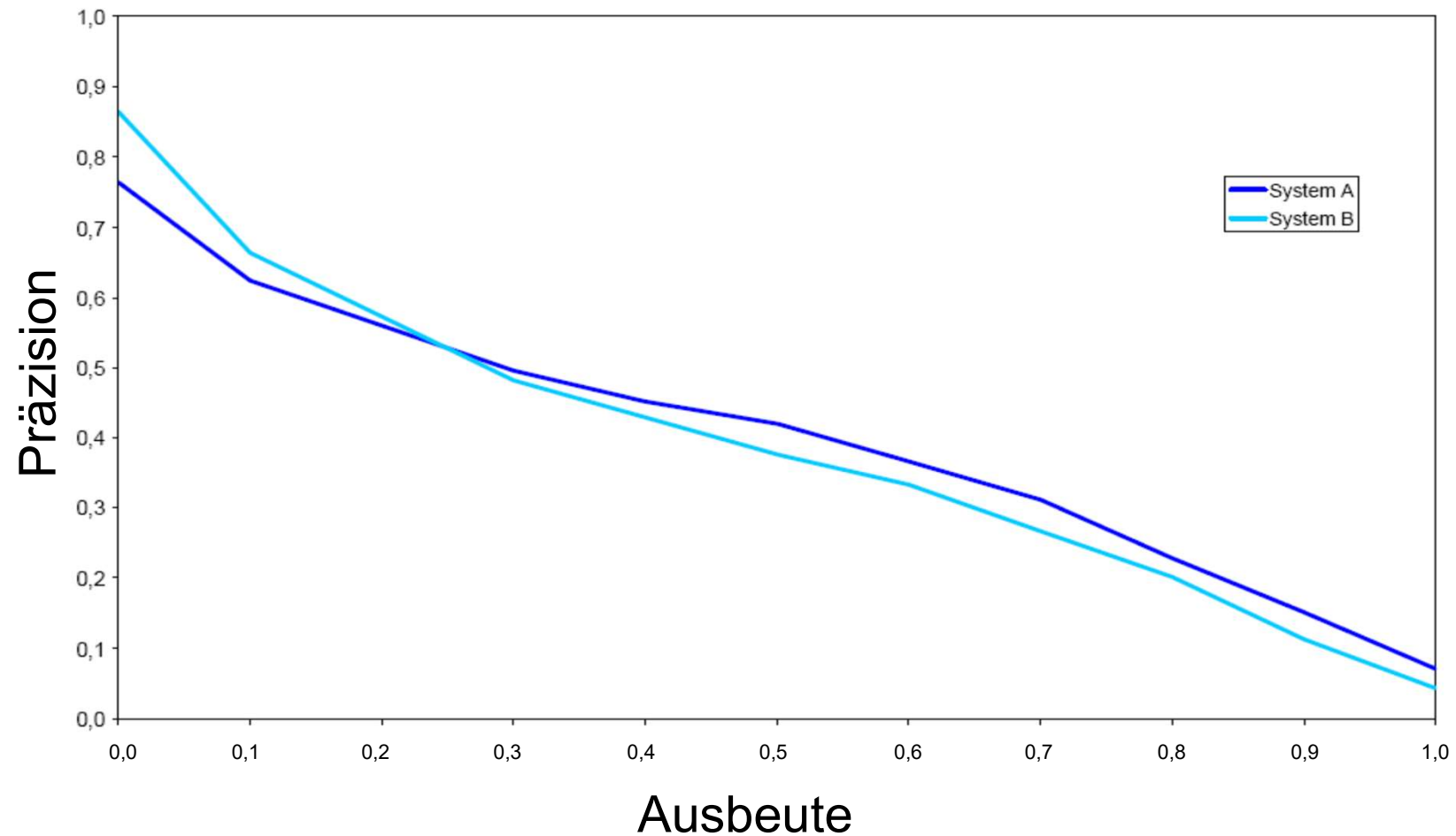
- MAP : $\Pi(\rho) := \frac{1}{|Q_0|} \sum_{q \in Q_0} \Pi_q(\rho)$

- Wobei:

q Einzelne Anfrage

Q_0 Menge aller Anfragen

Ausbeute/Präzisions-Graph



MAP = Fläche unter der Kurve («Area under the Curve», AUC)

Weitere Masse

F-Score

- Mengenbasiertes Mass. Balanciert Präzision ($P@k$, an Ranglistenstelle k) und Ausbeute ($R@k$):

$$F_{\beta} = \frac{(\beta^2 + 1) \times P@k \times R@k}{\beta^2 \times P@k + R@k}$$

- Ergibt im Sonderfall $\beta=1$ das harmonische Mittel von $P@k$ und $R@k$.

$$F_1 = \frac{2 \times P@k \times R@k}{P@k + R@k}$$

Weitere Masse

«Normalized Discounted Cumulative Gain» – nDCG

- Idee: Dokumente können auf einer Skala unterschiedlich relevant sein (z.B. «hochrelevant», «teilrelevant», ...)
- Wir summieren eine numerische Interpretation dieser Werte
- Spezialfall: Relevanz binär: $\{0,1\}$
- Relevante Dokumente auf tiefen Ranglistenplätzen tragen weniger zum Resultat bei → Discounted
- Je nach Verteilung der relevanten Dokumente schwankt der Maximalwert von Query zu Query → Normalisierung durch das Optimum

Normalized Discounted Cumulative Gain

Berechnung wie folgt:

$$\text{DCG}_v = \sum_{i=1}^v \frac{rel_i}{\log(i + 1)}$$

$$\text{IDCG}_v = \sum_{k \in \text{REL}_v} \frac{rel_k}{\log(k + 1)}$$

$$\text{nDCG}_v = \frac{\text{DCG}_v}{\text{IDCG}_v}$$

v = Anzahl Dokumente in der Rangliste, rel_i = Relevanzscore des Dokuments an Rang i , REL_v = «ideale Rangliste», nach rel_i sortiert.

Durchführung

- Der Labortest wird im Allgemeinen vollautomatisch durchgeführt:
 - Das System wird mit den gewählten Testdokumenten aufgesetzt und konfiguriert.
 - Die Informationsbedürfnisse werden zu Anfragen umgewandelt und an das System gesendet.
 - Die Resultate werden gesammelt.
 - Die Resultate werden auf ihre Relevanz bewertet.
 - Die Präzisions- und Ausbeutewerte werden ermittelt.

Fallbeispiel Google vs. Fireball

- Wir simulieren dieses Vorgehensweise mit einem Minimalbeispiel.

Aufgabe Google

- Ermitteln Sie die Ausbeute/Präzisionswerte und die Average Precision für eine Testanfrage in Google:
- „verhältnis china mongolei“

Resultate Google

Google Web Bilder Groups Verzeichnis News Froogle Desktop Mehr »

verhältnis china mongolei [Erweiterte Suche](#)
[Einstellungen](#)

Suche: ☐ Das Web ☒ Seiten auf Deutsch ☐ Seiten aus Deutschland

Web Ergebnisse **1 - 10** von ungefähr **67.400** Seiten auf **Deutsch** für **verhältnis china mongolei** . (0,22 Sekunden)

[Republik China - Wikipedia](#)
Das **Verhältnis** Taiwans zur Volksrepublik **China** (VRC) ist recht kompliziert. ... Faktisch werden diese Forderungen aber (zB im Fall der **Mongolei**) nicht mehr ...
de.wikipedia.org/wiki/Republik_China - 72k - [Im Cache](#) - [Ähnliche Seiten](#)

[Republik China - Wikipedia](#)
Das politische und staatsrechtliche **Verhältnis** der Republik **China** zur ... Faktisch werden diese Forderungen aber (zB im Fall der **Mongolei**) nicht mehr ...
de.wikipedia.org/wiki/Nationalchina - 71k - [Im Cache](#) - [Ähnliche Seiten](#)

[Eurasischer Verlag](#)
MONGOLEI. Feiern zur Reichsgründung durch Dschingis Khan ... sich vor trojanischem Mikrochip aus **China** · Finnland hat die EU-Ratspräsidentschaft übernommen ...
www.eurasischesmagazin.de/ - 25k - [Im Cache](#) - [Ähnliche Seiten](#)

[Auswärtiges Amt - Mongolei: Außenpolitik](#)
Ein gutes **Verhältnis** zu **China** ist für die **Mongolei** gerade im Hinblick auf den Außenhandel von entscheidender Bedeutung: Der einzig nutzbare Weg zum Meer ...
www.auswaertiges-amt.de/diplo/de/Laenderinformationen/Mongolei/Aussenpolitik.html - 19k - [Im Cache](#) - [Ähnliche Seiten](#)

[Das Parlament, Nr. 30-31 2006, 24.07.2006 - In der Mongolei ist ...](#)
In der **Mongolei** ist ein wachsendes Misstrauen gegen **China** zu beobachten. ... Beide Seiten normalisierten ihr **Verhältnis** zueinander, schufen einen ...
www.bundestag.de/dasparlament/2006/30-31/Thema/D17.html - 16k - [Im Cache](#) - [Ähnliche Seiten](#)

[Weltpolitik - Links zum Verhältnis VR China - Taiwan](#)
Angeboten werden Informationen zum Souveränitätskonflikt zwischen der VR **China** und Taiwan. Außerdem finden sich Links zu aktuellen Umfragen zum **Verhältnis** ...
www.weltpolitik.net/Regionen/AsienPazifik/ChinaTaiwan/Links.html - 16k - [Im Cache](#) - [Ähnliche Seiten](#)


Anzeigen

Mongolei-Forum
Alles Rund um Reisen und Leben in der **Mongolei**. Schnelle antworten!
www.mongolei-forum.de

Aufgabe Fireball: Vergleich mit Google

- Ermitteln Sie nochmals die Präzisions- und Ausbeutewerte für dieselbe Anfrage, mit der Suchmaschine Fireball (fireball.de)
- „verhältnis china mongolei“

Resultate Fireball


FIREBALL

Profisuche | Livesuche | Hilfe

Deutsch | Weltweit | Bilder | Nachrichten | Produkte

1 - 10 von 2.683 deutschsprachigen Treffern

Republik China - Wikipedia
... Republik **China** bis heute Anspruch auf ganz **China**, das Gebiet der **Mongolei**, fast alle Inseln im ... und staatsrechtliche **Verhältnis** der ...
<http://de.wikipedia.org/wiki/Republ...>

Mongolei
... Nationalflagge – national flag. Seiten**verhältnis** – ratio = 1:2 ... 1963 · Grenzvertrag zwischen **China** und **Mongolei**. 1989 · ...
<http://www.flaggenlexikon.de/fmongo...>

Erdkunde-Wissen
... Ein gutes **Verhältnis** zu **China** ist für die **Mongolei** gerade im Hinblick auf den Außenhandel von ... Tang Jiaxuan besuchte im Juli 2001 die ...
<http://www.erdkunde-wissen.de/erdku...>

Die Südliche **Mongolei** - Autonomes Gebiet Innere **Mongolei**
... der Volksrepublik **China**. Es heißt auf Mongolisch "Südliche **Mongolei**", die Bezeichnung "Innere **Mongolei**" (im ... keine Angaben über das ...
<http://userpage.fu-berlin.de/~corff...>

Weltpolitik - Links zum **Verhältnis** VR **China** - Taiwan
Government Information Office www.gio.gov.tw Webpage der Regierung der Republik **China**, Taiwan. Auf der web-page der taiwanischen Regierung finden sich ...
<http://www.weltpolitik.net/Regionen...>

Lexikon Republik **China**
... Republik **China** bis heute Anspruch auf ganz **China**, das Gebiet der **Mongolei**, fast alle Inseln im ... und staatsrechtliche **Verhältnis** der ...
<http://lexikon.freenet.de/Republik...>

Anzeigen:

Auswertung Google vs. Fireball

Google				Fireball			
#	Relevant	Ausbeute	Präzision	#	Relevant	Ausbeute	Präzision
1	Ja (1)	0.1	1.00	1	Ja (1)	0.1	1
2	Ja (2)	0.2	1.00	2	Nein	0.1	0.5
3	Ja (3)	0.3	1.00	3	Ja (2)	0.2	0.67
4	Ja (4)	0.4	1.00	4	Ja (3)	0.3	0.75
5	Nein	0.4	0.80	5	Nein	0.3	0.60
6	Nein	0.4	0.67	6	Nein	0.3	0.5
7	Ja (5)	0.5	0.71	7	Ja (4)	0.4	0.57
8	Nein	0.5	0.63	8	Nein	0.4	0.5
9	Nein	0.5	0.56	9	Nein	0.4	0.44
10	Nein	0.5	0.50	10	Nein	0.4	0.4

Auswertung Google vs. Fireball

Google				Fireball			
#	Relevant	Ausbeute	Präzision	#	Relevant	Ausbeute	Präzision
11	Nein	0.5	0.45	11	Nein	0.4	0.36
..	Nein	0.5	Nein	0.4	..
15	Ja (6)	0.6	0.40	15	Ja (5)	0.5	0.33
16	Ja (7)	0.7	0.44	16	Ja (6)	0.6	0.40
..	Nein	0.7	Nein	0.6	..
19	Ja(8)	0.8	0.42	19	Nein	0.6	0.32
				20	Ja (7)	0.7	0.35
				..	Nein	0.7	..
				32	Ja (8)	0.8	0.25

Resultate der „Evaluation“ (Google)

- Die Resultate für die Auswertung der Anfrage:
 - Präzision nach 10 Dokumenten: 0.5
 - Ausbeute nach 10 Dokumenten: 0.5
 - Average Precision: 0.71
 - (Annahme: 10 relevante Dokumente)
- Aber: was bedeutet das? Ist das ein gutes Resultat?

Resultate der „Evaluation“ (Fireball)

- Die Resultate für die Auswertung der Anfrage:
 - Präzision nach 10 Dokumenten: 0.4
 - Ausbeute nach 10 Dokumenten: 0.4
 - Average Precision: 0.47
 - (Annahme: 10 relevante Dokumente)

Frage



- Was kann man von solchen Resultaten lernen?
- Ist Google doppelt so gut wie Fireball?

Einschränkungen unseres Vorgehens

- Nur eine Anfrage
- Präzisionsorientiertes Szenario
- Können Ausbeute nicht bestimmen
- Datenbasis nicht eingefroren → nicht wiederholbar
- Datenbasis nicht identisch (→ will man diesen Aspekt in den Test einbeziehen?)
- Wie sind die Relevanzbeurteilungen zu betrachten?

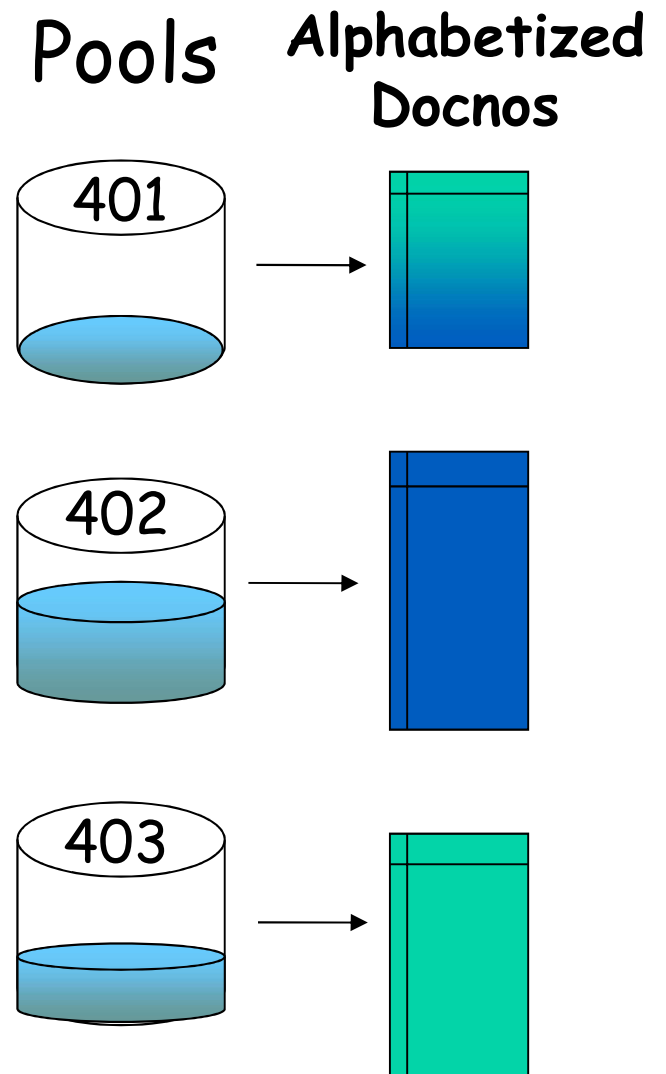
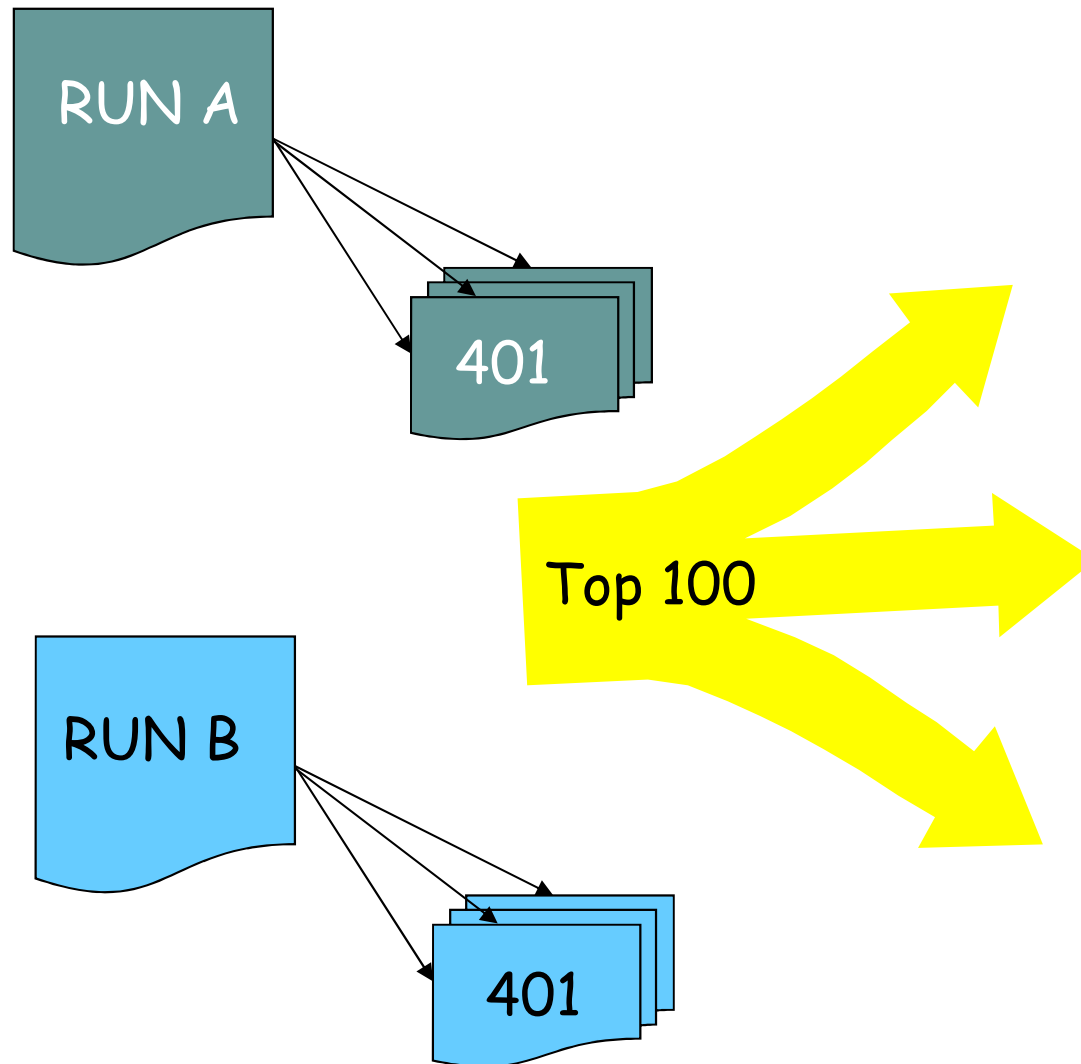
Bestimmung der Ausbeute

- Um die Ausbeute zu bestimmen, müssen sämtliche Dokumente nach Relevanz bewertet werden.
- Dies war für frühe Testkollektionen praktikabel (z. B. Cranfield 1400 Dokumente, 225 Anfragen in den 1960ern), aber heute nicht mehr realistisch (daher auch unsere Annahme in dem vorangegangenen Beispiel)
- Der Aufwand, Relevanzbewertungen durchzuführen, explodiert mit der Grösse der Testkollektionen. Sogar im Cranfield-Fall: 315000 Bewertungen!
(1400 Dokumente * 225 Anfragen = 315000 Bewertungen)
- Es folgen zwei Konsequenzen:
 - Es sollten falls möglich vorhandene Testkollektionen (wieder-)verwendet werden
 - Die Ausbeute muss auf eine alternative Art bestimmt werden

Evaluations-Kampagnen

- Das Problem des enormen Aufwands, Testkollektionen herzustellen, wird durch Evaluationskampagnen begegnet: interessierte Parteien nehmen an einer Kampagne teil, in deren Rahmen mehrere Systeme evaluiert werden.
- Idee: wenn genügend viele verschiedene Teilnehmer mit unterschiedlichen Systemen die Evaluation durchführen, werden (fast) alle relevanten Dokumente von mindestens einem System gefunden – es müssen nicht alle Dokumente gelesen werden, um die Ausbeute zu bestimmen.
- Zudem: richtige Interpretation der Ergebnisse → siehe später
- Die grössten Foren:
 - TREC – <http://trec.nist.gov>
 - CLEF – <http://www.clef-campaign.org>

Pooling



Klassische Ad-Hoc Testkollektionen (TREC/CLEF) – Pooling Methodik

	# part.	# lg.	# docs.	Size in MB	# Assess.	# topics	# ass. per topic
CLEF 2003	33	9	1,611,178	4124	188,475	60 (37)	~3100
CLEF 2002	34	8	1,138,650	3011	140,043	50 (30)	~2900
CLEF 2001	31	6	940,487	2522	97,398	50	1948
CLEF 2000	20	4	368,763	1158	43,566	40	1089
TREC8 CLIR	12	4	698,773	1620	23,156	28	827
TREC8 AdHoc	41	1	528,155	1904	86,830	50	1736
TREC7 AdHoc	42+4	1	528,155	1904	~80,000	50	~1600

Legende: #part. = Anzahl Teilnehmer
 #lg. = Anzahl Sprachen
 #docs. = Anzahl Dokumente
 #assess. = Anzahl Bewertungen
 #ass. per topic = Anzahl Bewertungen pro Thema/Anfrage

ClueWeb – 1 Milliarde Dokumente

- Experimente in Hinblick auf Big Data-Phänomene werden erst mit sehr grossen Testdatensets möglich.
- ClueWeb mit ~1 Milliarde Dokumenten/Webseiten (multilingual)
- Topics aus Logs von Websuchdiensten hergeleitet
- Pooling ersetzt durch alternative Vorgehensweisen, wie z.B. die «Minimal Test Collection Method» [Carterette et al., 2006]



The ClueWeb09 Dataset

The ClueWeb09 dataset was created to support research on information retrieval and related human language technologies. It consists of about 1 billion web pages in ten languages that were collected in January and February 2009. The dataset is used by several tracks of the TREC conference.

Dataset Specifications

Web Pages:

- 1,040,809,705 web pages, in 10 languages
- 5 TB, compressed. (25 TB, uncompressed.)

See the [Record Counts Section](#) on the [Dataset Information and Sample Files](#) page for detailed information on the distribution of records and languages.

Frage



- Was ist das Problem mit Relevanzbewertungen?

Relevanzbewertungen

- Das Informationsbedürfnis hinter der Testanfrage:
- *Titel:* Verhältnis China-Mongolei (Codierte Anfrage)
- *Beschreibung:* Finde Informationen über das Verhältnis und Kooperationsvereinbarungen zwischen China und der Mongolei in der jüngeren Geschichte.
- *Erläuterung:* Dokumente, die Informationen über politische und/oder wirtschaftliche Beziehungen zwischen China und der Mongolei im 20. Jahrhundert liefern, sind relevant.

Verbalisierung des
Informationsbedürfnis

Relevanz

- Der Begriff der Relevanz
- Das Verständnis einer Anfrage und eines Dokuments hängt immer auch vom konkreten Benutzer ab:
 - Vor-/Hintergrundwissen
 - Reihenfolge des Auffindens
 - Wandelnde Informationsbedürfnisse
 - Persönliche Präferenzen
 - Vollständigkeit der Antwort

Stabilität von Relevanzbeurteilungen

- Beurteilt jeder Benutzer die Relevanz gleich? → Nein!
- Was bedeutet das für die Resultate, welche wir in einem Labortest erhalten?
 - absolute Werte der verwendeten Leistungsmasse können variieren
 - relative Vergleiche zwischen Systemen sind sehr stabil
 - → Konsequenzen?
- Erkenntnisse aus den TREC Kampagnen: Übereinstimmung zwischen Relevanzbeurteilern („Relevance Assessors“) können sehr gering sein, ohne zu einer Änderung der grundsätzlichen Schlüsse aus der Evaluation zu führen.
- Es gibt also per Definition kein 100% Retrievalresultat

Assessors

Cranfield legacy

TREC assessors



Photo from NIST

Interpretation der Resultate

- Die Variabilität der Resultate über verschiedene Anfragen ist im Allgemeinen sehr hoch.
- Es ist schwierig, statistisch signifikante Resultate zu erhalten (TREC/CLEF: oft sind dutzende Systeme in der gleichen „statistischen Leistungsklasse“)
- → es ist eine gewisse Grundanzahl an Anfragen notwendig, um überhaupt Aussagen machen zu können. Als anerkanntes Minimum gilt 25, wobei eigentlich mindestens 50 Anfragen verwendet werden sollten.
- Werte für Ausbeute sind unter dem Vorbehalt zu interpretieren, dass nur ein Teil der Dokumente bewertet wurde. Hier sind die Anzeichen sehr ermutigend, dass dies auf die grundsätzlichen Schlussfolgerungen wenig Einfluss haben sollte.

«Million Query Track»

- Untersucht die Frage: «mehr Queries mit wenigen Relevanzbewertungen oder wenige Queries mit (vollständigen?) Relevanzbewertungen»
- Nicht wirklich «eine Million», sondern 40,000 Queries, aus denen ein paar Hundert für die Bewertung selektiert wurden.
- Nur durchschnittlich 34 Bewertungen pro Query
- Masse wie Average Precision können nur geschätzt werden resp. es kann nur ein Ranking hergeleitet werden.

Frage



- Welches System ist besser? Jenes mit konstanter Präzision von 90%, oder jenes, welches für 9 von 10 Fragen 100% Präzision liefert, aber für jede zehnte Anfrage versagt?

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. Desweiteren kann im Fall sehr weniger relevanter Dokumente die maximale Präzision auf unterschiedliche Art erreicht werden, und ist schwer zu interpretieren, wenn sie gemittelt wird.
- Beispiel: 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.

Beispiel: Average Precision

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

rank	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

Frage



- Was für Schlüsse kann der Benutzer aus diesen Average Precision-Werten schliessen?
- Was ist der Unterschied zwischen diesen beiden Average Precision-Werten?

Wahrnehmung der Suchqualität

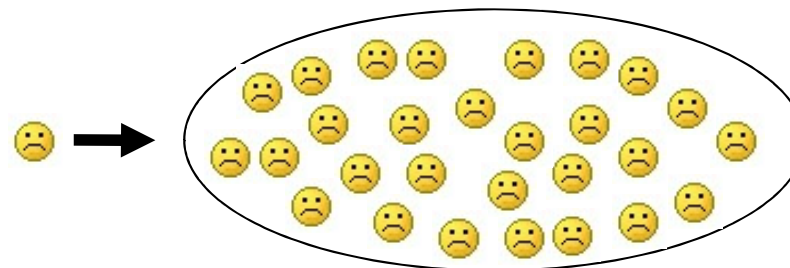
- Der Endbenutzer nimmt eine durchschnittliche Performance nicht unmittelbar wahr. Er sieht bloss die Performance **seiner** momentanen Anfrage
- Unzufriedene Kunden erzählen im Durchschnitt zehn Personen von ihrer schlechten Erfahrung. Zwölf Prozent erzählen sie bis zu 20 Personen.
 - → Schlechte News verbreiten sich schnell. Mit der Nutzung des Internets (Email, 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 das selbe Problem gehabt, geben aber kein Feedback (→ 20 mal Faust-im-Sack)
- 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



Quelle: http://www.ara.com.au/240.html?&tx_ttnews%5btt_news%5d=70&tx_ttnews%5bbackPid%5d=179&cHash=ca4543de3

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 positive Meldungen machen.
- → Mean Average Precision kann irreführend sein.
- Eine Betrachtung der Ausreisser ist nötig (z.B. Average Precision einzelner Anfragen, aber: siehe Beispiel oben)
- Ein Mass für "Robustheit" wäre interessant (in diesem Bereich wird momentan aktiv geforscht!)

Evaluation eines IR Projektes (Fallstudie)

- Für einen Projektverantwortlichen 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 Dokument eine Anfrage, mit dem Ziel, das Dokument „wiederzufinden“
- Die Resultate werden wie folgt bewertet:
 - Gefunden (innerhalb 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 Anfragen) in beiden Systemen ausgewertet.
- Dabei werden je 20 Dokumente „beurteilt“: insgesamt müssen $50 \cdot 20 \cdot 2 = 2000$ Dokumente auf Übereinstimmung mit den 10 gesuchten Dokumenten getestet werden
- Es können dabei maximal $50 \cdot 5 = 250$ Punkte erreicht werden. Die 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 Anfragen: Genug? Repräsentativ?
- 5 Personen: Genug? Repräsentativ? Ist es sinnvoll, die beiden Faktoren „Personen“ und „Anfragen“ zu vermischen?
- Das “selbstgestrickte” 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): $MRR = 1/\text{Rang}$
- 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?
- Wenn die Anfragen aufgrund von Begriffen aus den gesuchten Dokumenten konstruiert wird, welches System ist dann bevorzugt: jenes mit Wortnormalisierung, oder 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
- 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.

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 Assessment 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.
- Für die Projektarbeit muss ein Kompromiss gewählt werden. Wichtig sind die Wahl einer korrekten Methodik, die Vergleichbarkeit (Wahl der Masse), sowie die korrekte Interpretation des Resultats (speziell der absoluten Werte!)
- Die Evaluation sollte sich auf die wichtigen Aspekte beschränken, und diese Aspekte nicht vermischen.
- Vergleiche liefern stabilere Ergebnisse als absolute Werte; absolute Werte sind im Allgemeinen nicht isoliert interpretierbar.