

Methoden für Robustes Information Retrieval und dessen Evaluierung

Thomas Mandl, Daniela Wilczek

Institut für Informationswissenschaft und Sprachtechnologie
Universität Hildesheim
Marienburger Platz 22
31141 Hildesheim, Deutschland
mandl@uni-hildesheim.de

Abstract

Information Retrieval Systeme sollen möglichst robust arbeiten und ihren Benutzern unter einer Vielzahl von Bedingungen zufriedenstellende Ergebnisse liefern. Der Beitrag referiert Ansätze zur Entwicklung von robusten Systemen und zeigt, wie die Robustheit evaluiert werden kann. Die Ergebnisse des Robust Task des Cross Language Evaluation Forum (CLEF) 2008 werden vorgestellt und unter verschiedenen Gesichtspunkten analysiert.

1 Robustheit im Information Retrieval

Wissen muss in vielerlei Kontexten und unter sehr heterogenen Bedingungen aktiviert werden und als Information den Benutzer erreichen. Robustheit ist daher ein Desiderat für Information Retrieval Systeme.

Eine in der Praxis häufig angewandte Form der Evaluierung von Information Retrieval Systemen ist die Ego-Search. Dabei gibt jemand seinen eigenen Namen in das Eingabefeld für die Suche und wirft einen Blick auf die Ergebnisse. Diese Form der Evaluierung hält wissenschaftlichen Kriterien nicht stand. Prüfen wir einige dieser Kriterien ab und vergleichen wir die Ego-Search mit einer wissenschaftlichen Evaluierung nach dem Cranfield-Paradigma. Die Nutzung nur einer Anfrage führt zu einem völlig unzuverlässigem Ergebnis. Typischerweise nutzen Evaluierungen 50 Anfragen und mitteln die Einzelergebnisse. Wird nur ein System betrachtet, so lässt sich überhaupt keine Aussage über Qualität machen. Betrachtet der Juror in der Ego-Search nur die obersten Treffer, so stellt dies ebenfalls keinen gerechten Maßstab dar und schränkt das Ergebnis auf Anwendungsfälle ein, die eine hohe Präzision erfordern. Viele Szenarien erfordern aber einen hohen Recall.

Die Evaluierungsforschung kann seit einigen Jahren auf umfangreiche Daten aus den Initiativen Text Retrieval Conference (TREC), Cross Language Evaluation Forum (CLEF) und NTCIR zurückgreifen. So konnten genaue Analysen der Zuverlässigkeit von Retrieval-Tests erfolgen und viele der früher eher heuristischen Annahmen einer wissenschaftlichen Überprüfung unterziehen.

2 Aktuelle Forschung zur Evaluierung

Die Evaluierungen von Information Retrieval Systemen folgen meist dem Cranfield-Paradigma, welches die

Bestandteile einer zuverlässigen vergleichenden Bewertung benennt [Robertson 2008]. Mehrere Systeme bearbeiten identische Aufgaben (Topics) für eine Menge von Dokumenten. Im Anschluss werden die Treffer zusammengefasst und von menschlichen Juroren bewertet. Je nach der Position der relevanten und nicht-relevanten Dokumente in den Trefferlisten der Systeme können Qualitäts-Maße berechnet werden, anhand derer die Systeme verglichen werden.

Die Validität des Cranfield-Paradigmas wurde in den letzten Jahren aus mehreren Blickwinkeln analysiert. Wie viele Experimente, Systeme, Dokumente, Anfragen, Juroren und Relevanzurteile sind eigentlich nötig, um zwei Systeme verlässlich vergleichen zu können und Rückschlüsse auf ihre Performanz außerhalb der Evaluierung ziehen zu können?

Häufig wurde etwa die Subjektivität der Juroren bemängelt. Manche Forscher fanden weitere, aus ihrer Sicht relevant Dokumente unter den bisher nicht bewerteten oder den als negativ bewerteten Dokumenten. Die scheint den Relevanz-Urteilen als Fundament der Bewertung die Glaubwürdigkeit zu entziehen. Anhand von mehrfach bewerteten Dokumenten für TREC 4 nachgewiesen werden, dass tatsächlich bei den relevanten Dokumenten tatsächlich nur eine Übereinstimmung von zwischen 30% und 40% erreicht wurde [Voorhees 1998]. Gleichwohl wirkt sich dies aber nicht auf die Reihenfolge der Systeme aus. Der Vergleich zwischen den Systemen fällt unabhängig von Juroren sehr ähnlich aus [Voorhees 1998].

Die Variabilität zwischen den Topics ist bei allen Evaluierungen meist größer als die zwischen den Systemen. Dies wirft erneut berechtigten Zweifel an der Zuverlässigkeit der Evaluierung auf. Wenn die Topics sich sehr stark voneinander unterscheiden, dann könnte die zufällige Auswahl von einigen anderen Topics doch zu einem stark abweichendem Evaluierungsergebnis führen. Auch dies wurde mehrfach wissenschaftlich untersucht. Dazu geht man vom Original-Ranking der Systeme aus, lässt einzelne Topics weg und betrachtet die Ergebnisse der Evaluierung ohne diese Topics. Das Ergebnis ist eine Rangfolge von Systemen. Unter der zwar fragwürdigen, aber plausiblen Annahme, dass das Ranking mit allen Topics das Optimum darstellt, kann nun das optimale mit dem neuen, mit weniger Topics erstellten Ranking verglichen werden. Korrelationen oder Fehler-Maße, die zählen, wie oft ein „schlechteres“ System vor einem

besseren platziert ist, liefern Maßzahlen für den Vergleich.

Hier zeigte sich immer, dass die Rangfolgen der Systeme weitgehend stabil blieben und sich bis zu einer Menge von etwa 25 Topics wenig an dem Ergebnis der Evaluierung ändert [Sanderson & Zobel 2005, Mandl 2008]. Auch hinsichtlich der Anzahl der Topics sind die typischen Evaluierung also sehr zuverlässig. Ab 25 Anfragen kann man mit einem guten Level an Zuverlässigkeit rechnen.

Robustheit bedeutet für Produkte gemeinhin, dass sie auch unter wechselnden, schwierigen und unvorhergesehenen Bedingungen noch einigermaßen zufriedenstellend funktionieren, wobei in Kauf genommen wird, dass sie nicht unbedingt besonders glänzen. Für Information Retrieval Systeme kommen hierfür unterschiedliche Kollektionen oder Benutzer in Frage, jedoch liefert vor allem die bereits besprochene Variabilität zwischen einzelnen Aufgaben die größte Herausforderung.

Die Qualität von Antworten im Information Retrieval schwankt zwischen einzelnen Anfragen sehr stark. Die Evaluierung im Information Retrieval zielt in der Regel auf eine Optimierung der durchschnittlichen Retrieval-Qualität über mehrere Testanfragen (Topics). Sehr schlecht beantwortete Anfragen wirken sich besonders negativ auf die Zufriedenheit des Benutzers aus. Für die Steigerung der Robustheit ist es erforderlich, besonders die Qualität der Systeme für die schwierigen Topics zu erhöhen [Voorhees 2005]. Dazu ist sowohl die automatische Vorhersage der Schwierigkeit als auch die Analyse und besondere Behandlung der vermutlich schwierigen Topics nötig. Dementsprechend verspricht die Analyse der einzelnen Topics großes Potential für die Verbesserung der Retrieval-Ergebnisse [Mandl 2008].

3 Entwicklung Robuster Systeme

Für die Steigerung der Robustheit von Information Retrieval Systemen wurden bisher explizit einige Verfahren getestet. Bereits die Grundformreduktion kann als robuste Technik angesehen werden. Die robusten Techniken dienen meist zur der Steigerung der Performanz bei schwierigen Topics, jedoch können auch andere generelle Strategien zur Klassifizierung von Topics und ihrer spezifischen Behandlung durch bestimmte Systemkomponenten als robuste Techniken gelten.

Schwierige Anfragen entstehen oft dadurch, dass der Suchbegriff in der Kollektion gar nicht enthalten ist (*out of vocabulary* Problem). Dadurch entfalten Strategien zur Erhöhung des Recall wie automatisch generierten Ähnlichkeitsthesauri und Blind Relevance Feedback gar nicht die notwendige Wirkung. Als Strategie hat sich hier die Termerweiterung in einer anderen Kollektion bewährt. Dabei wird der nicht gefundene Begriff z.B. im Web gesucht, aus den Treffern werden häufig mit ihm in Beziehung stehende Terme extrahiert und mit diesen wird dann in der gewünschten Kollektion recherchiert.

Aber auch andere Parameter der Termerweiterung werden analysiert [Tomlinson 2007]. In einem Experiment im Rahmen von CLEF (siehe Abschnitt 5) berichten [Pérez-Agüera & Zaragoza 2008], dass Blind Relevance Feedback (BRF) zwar die durchschnittliche Qualität des Systems steigert (*Mean Average Precision*, MAP), jedoch die Robustheit verringert (*Geometric Mean Average Precision*, GMAP).

Dieses Phänomen hat [Kwok 2005] ebenfalls für das Englische untersucht und dabei die Auswirkungen von Blind Relevance Feedback auf unterschiedlich schwierige Anfragen analysiert. Es zeigte sich, dass BRF für mittelschwere Aufgaben zu positiven Effekte führte, während es bei schwierigen und leichten Anfragen eher Verschlechterungen bewirkte. Dies lässt sich möglicherweise dadurch erklären, dass leichte Anfragen bereits im ersten Durchlauf relative gut beantwortet werden und BRF weniger relevante Terme hinzufügt. Dagegen funktioniert BRF bei sehr schwierigen Anfragen nicht, weil das erste Ergebnis so schlecht ist, dass keine guten Erweiterungsterme gefunden werden [Kwok 2005].

Darauf aufbauende Arbeiten versuchen, die Kohärenz der besten Treffer zu untersuchen. Die Kohärenz in der gesamten Kollektion und den inhaltlichen Zusammenhang der besten Treffer vergleichen [He et al. 2008]. Dabei bemerken sie, dass BRF besser wirkt, wenn die Treffermenge thematisch sehr homogen ist. Treten Dokumente zu mehreren Themen in der Treffern auf, so kann das BRF eine thematische Verschiebung bewirken. Eine Messung der Kohärenz in der Treffermenge kann über die Anwendung des BRF entscheiden.

Auch ambige Terme können bei der Termerweiterung eher schaden als nutzen. In einem System bestimmen [Cronen-Townsend et al. 2002] ein Klarheitsmaß, welches die Ambiguität eines Terms im Vergleich zum *language model* der Kollektion misst. Nur nicht ambige Terme werden zur Termerweiterung eingesetzt.

Ein weiterer Ansatz zur Verbesserung der Robustheit von Systemen liegt in der automatischen Disambiguierung. Ambiguität tritt in der natürlichen Sprache häufig auf und führt zu Schwierigkeiten beim Retrieval. Eine Analyse des Kontexts kann helfen, aus mehreren Bedeutungen eines Wortes die gemeinte zu erschließen. Dieser Ansatz wurde u.a. in CLEF 2008 untersucht.

4 Robuste Evaluierung

Die CLEF-Initiative (www.clef-campaign.org) etablierte sich im Jahr 2000. Seitdem steigt die Zahl der Teilnehmer bei CLEF stetig an und die Forschung zu mehrsprachigen Information Retrieval Systemen gewinnt an Dynamik. CLEF folgt dem Modell von TREC und schuf eine mehrsprachige Kollektion mit Zeitungstexten. Inzwischen umfasst die Dokument-Kollektion für das ad-hoc Retrieval die Sprachen Englisch, Französisch, Spanisch, Italienisch, Deutsch, Holländisch, Schwedisch, Finnisch, Portugiesisch, Bulgarisch, Ungarisch und Russisch. Mehrere weitere Tracks wie *Question Answering*, *Web-Retrieval*, *Spoken Dokument Retrieval* oder *Geographic CLEF* untersuchen bestimmte Aspekte des mehrsprachigen Retrieval.

Im Rahmen von CLEF wird seit drei Jahren ein Robust Task durchgeführt, der benutzerorientierte Maße wie GMAP in den Mittelpunkt rückt [Mandl 2006]. Für CLEF 2008 standen erstmals Disambiguierungs-Daten für die Dokumente zur Verfügung, so dass überprüft werden konnte, ob zusätzliches semantisches Wissen das Retrieval robuster gestalten kann [Agirre et al. 2009].

5 Ergebnisse des Robust Task 2008

Die Ergebnisse der Robust Task 2008 haben gezeigt, dass sich die Hoffnung nicht unmittelbar bestätigt, dass Disambiguierungs-Daten (Word Sense Disambiguation,

WSD) das Retrieval verbessern. Im mono-lingualen Retrieval für das Englische berichteten zwar einige Forschungsgruppen, dass sie durch WSD bessere Ergebnisse erzielen konnten. Das beste Ergebnis war aber ohne WSD erzielt worden. Allerdings galt dies nur für MAP. Berücksichtigt man den GMAP, so ergibt sich eine Veränderung auf der ersten Position und das WSD-Experiment der gleichen Gruppe rückt auf den ersten Platz [Agirre et al. 2009].

Im Folgenden werden die Ergebnisse für den bi-lingualen Task vom Spanischen ins Englisch betrachtet. Hier sieht die Bilanz für WSD noch negativer aus. Abbildung 1 zeigt die Verteilung der *Average Precision* über den Wertebereich. Der Box-Plot zeigt an den Antennen den minimalen und den maximalen Wert auf, während die Box mit dem Strich die mittleren 50% der Werte mit dem Median markiert. Zum einen ist die Spannbreite der Topics und die der Systeme aufgezeichnet. Jeweils ein Box-Plot zeigt die Systeme, die WSD benutzen und die Systeme, die ohne WSD arbeiten. Bei den Topics sind jeweils alle Topics einmal für die Systeme mit und einmal die ohne WSD gezeigt.

Die Abbildung zeigt deutlich, dass wie bei jeder Evaluierung die Varianz bei den Topics sehr viel höher ist als die bei den Systemen. Bei den Topics (Aufgaben) sind immer sowohl sehr schwere als auch sehr leichte Topics vertreten. Das Minimum der Topics liegt immer bei 0 und das Maximum immer über 0,8. Die Spannbreite der Systeme ist sehr viel geringer. An beiden Verteilungen ist jeweils aber zu erkennen, dass WSD anscheinend eher schadet. Der Median und das Maximum liegen bei Benutzung von WSD niedriger. Disambiguierungs-Daten scheinen die Robustheit nicht zu erhöhen.

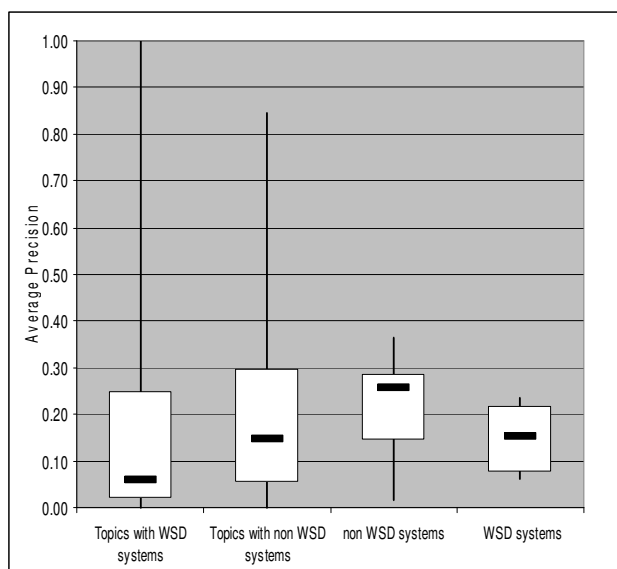


Abb. 1: Box-Plots für die Verteilung der AP für Topics und Systeme

Bei der vergleichenden Betrachtung der MAP und des GMAP ergeben sich besonders unter Berücksichtigung kleinerer Topic-Mengen interessante Effekte. Besonders sollte untersucht werden, ob bestimmte Topics besonders von der WSD profitiert haben und welche Topics unter der WSD gelitten haben. Die folgenden Tabellen listen diese Topics. Diese können für die Analyse der Gründe

für das Scheitern oder den Nutzen von WSD gute Hinweise bieten. Diese liefern möglicherweise Hinweise für weitere Verbesserungsansätze.

Insgesamt standen 160 Testfragen zur Verfügung. Daraus wurden mehrere kleinere Mengen erzeugt. Zum einen zwei 50er und eine 60er Gruppe, wobei die ersten 50 eine Gruppe bildeten und die zweiten 50 eine zweite und der Rest die 60er Gruppe. Für die Erzeugung zweier 100er und zweier 120er Gruppen wurde der gleiche Ansatz gewählt, wobei beide Gruppen sich in der Mitte jeweils überlappen. Variiert wurde die Reihenfolge der Topics, so dass mehrere verschiedene Versionen v.a. der 50er Aufteilung erzeugt wurden.

Tabelle 1: Verschlechterung durch WSD (-0,48 bis -0,26 MAP absolut)

170	Find documents about French plans for reducing the number of official languages in the European Union to five languages.
333	Find information on the trial of the French war criminal, Paul Touvier.
185	What happened to the photographs and films that Dutch soldiers made in Srebrenica which provided evidence of violations of human rights?
289	Find documents giving information on the Falkland Islands
162	Find documents about the problems posed by Greece concerning the abolishment of customs restrictions between the European Union and Turkey

Tabelle 2: Verbesserung durch WSD (+0,16 bis +0,26 MAP absolut)

183	In what parts of Asia have dinosaur remains been found?
173	Find reports on the experimental proof of top quarks by US researchers
294	What is the speed of winds in a hurricane?
253	In which countries or states is the death penalty still practiced or at least permitted by the constitution?
196	Find reports on the merger of the Japanese banks Mitsubishi and Bank of Tokyo into the largest bank in the world.

Jede dieser Gruppen wurde als individueller Retrieval-Test betrachtet und daraus eine Rangfolge der Systeme sowohl mit MAP als auch mit GMAP ermittelt. Insgesamt waren die Korrelationen zwischen Rangfolgen sehr hoch. Für die 160 Topics ergab sich eine Korrelation von 0.90. Das bedeutet, dass es minimale Änderungen bei den Positionen gab. Interessanterweise liegen die meisten Korrelationen für kleinere Mengen höher. Geringere Korrelationswerte ergeben sich lediglich, wenn die 50er Untermengen aus einer Sortierung nach der Verbesserung der AP durch die WSD erzeugt werden. Für die beiden

letzten Mengen, also die Gruppen der Topics welche durch WSD eher profitieren, liegt die Korrelation nur bei 0,87. Bei diesen 50 bzw. 60 Topics macht es also durchaus Sinn, den GMAP zu betrachten und damit ein robustes Maß anzuwenden.

6 Fazit

Dieser Aufsatz erläutert die robuste Systementwicklung ebenso wie die robuste Evaluierung. Einige Ergebnisse aus der Analyse von Evaluierungsergebnissen werden vorgestellt. Auch im Rahmen von CLEF 2009 findet wieder ein Robust Task statt.

Literatur

- Agirre, E.; Di Nunzio, G.; Ferro, N.; Mandl, T.; Peters, C. (2009). CLEF 2008: Ad Hoc Track Overview. In: Evaluating Systems for Multilingual and Multimodal Access: 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, Revised Selected Papers. Berlin et al.: Springer [LNCS]. erscheint. Preprint: <http://www.clef-campaign.org>
- [Cronen-Townsend et al. 2002] Cronen-Townsend, S.; Zhou, Y.; Croft, B. (2002). Predicting query performance. In: 25th Annual Intl. ACM conference on Research and development in information retrieval (SIGIR). Tampere. S. 299-306.
- [He et al. 2008] He, J.; Larson, M.; de Rijke, M. (2008). On the Topical Structure of the Relevance Feedback Set. In: Proc. LWA: Workshop Information Retrieval. <http://ki.informatik.uni-wuerzburg.de/papers/baumeister/2008/LWA2008-Proc.pdf>
- [Kwok 2005] Kwok, K. (2005). An Attempt to Identify Weakest and Strongest Queries. In: SIGIR Workshop Predicting Query Difficulty. Salvador, Brazil. <http://www.haifa.il.ibm.com/sigir05-qp>
- [Mandl 2006] Mandl, T. (2006). Benutzerorientierte Bewertungsmaßstäbe für Information Retrieval Systeme: Der Robust Task bei CLEF 2006. In: Mandl, T.; Womser-Hacker, C. (Hrsg.): Effektive Information Retrieval Verfahren in Theorie und Praxis: Proc. des Fünften Hildesheimer Evaluierungs- und Retrievalworkshop (HIER 2006). Hildesheim: Universitätsbibliothek. S. 79-91. <http://web1.bib.uni-hildesheim.de/edocs/2006/519937899/meta/>
- [Mandl 2008] Mandl, T. (2008). Die Reliabilität der Evaluierung von Information Retrieval Systemen am Beispiel von GeoCLEF. In: Datenbank-Spektrum: Zeitschrift für Datenbanktechnologie und Information Retrieval. Heft 24. S. 40-47
- [Pérez-Agüera & Zaragoza 2008] Pérez-Agüera, J.; Zaragoza, H. (2008). UCM-Y!R at CLEF 2008 Robust and WSD tasks. In: CLEF Working Notes. <http://www.clef-campaign.org>
- [Robertson 2008] Robertson, S. (2008): On the history of evaluation in IR. In: Journal of Information Science. <http://jis.sagepub.com/cgi/reprint/34/4/439>
- [Sanderson & Zobel 2005] Sanderson, M.; Zobel, J. (2005). Information retrieval system evaluation: effort, sensitivity, and reliability. In: 28th Annual Intl. ACM Conference on Research and Development in Information Retrieval (SIGIR) Salvador, Brazil. S. 162-169
- [Tomlinson 2007] Tomlinson, S. (2007). Comparing the Robustness of Expansion Techniques and Retrieval Measures. In: Evaluation of Multilingual and Multimodal Information Retrieval. 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, Revised Selected Papers [LNCS 4730] S. 129-136.
- [Voorhees 1998] Voorhees, E. (1998). Variations in relevance judgments and the measurement of retrieval effectiveness. In: 21st Annual Intl. ACM Conference on Research and Development in Information Retrieval (SIGIR) Melbourne. S. 315-323
- [Voorhees 2005] Voorhees, E. (2005). The TREC robust retrieval track. In: ACM SIGIR Forum 39 (1) 11-20.