Joachim Griesbaum, Thomas Mandl, Christa Womser-Hacker (Hrsg.)

Information und Wissen: global, sozial und frei?

Proceedings des 12. Internationalen Symposiums für Informationswissenschaft (ISI 2011)

Hildesheim, 9.—11. März 2011

vwh
Verlag Werner Hülsbusch
Fachverlag für Medientechnik und -wirtschaft

Vergleich von IR-Systemkonfigurationen auf Komponentenebene

Jens Kürsten, Thomas Wilhelm und Maximilian Eibl

Technische Universität Chemnitz
Straße der Nationen 62, 09111 Chemnitz
{jens.kuersten, eibl}@informatik.tu-chemnitz.de

Kurzfassung

In der klassischen Evaluationsforschung werden Verfahren anhand der Performance der Gesamtsysteme bewertet. Dies macht es schwer, die Wirkung einzelner Systemkomponenten und ihr Zusammenspiel zu analysieren. Auf Basis einer umfangreichen Evaluation mit mehreren englischsprachigen Testkorpora wird ein Werkzeug zum interaktiven Vergleich von Systemkonfigurationen vorgestellt. Neben der Möglichkeit, den Einfluss einzelner Komponenten auf das Rechercheergebnis zu untersuchen, wird ein Ansatz zur Normierung der Ergebnisse beschrieben. Dieser wird genutzt, um den Einsatz des Visualisierungswerkzeugs für den korpusübergreifenden Vergleich von Systemkonfigurationen zu skizzieren.

Abstract

In traditional information retrieval (IR) evaluation new models are assessed based on system effectiveness in terms of a metric suitable to the problem of interest. Drawing conclusions about the impact of system components and possible interaction effects is almost impossible in this setting. We address this problem and present a tool for interactive comparison of system configurations on component-level based on a large-scale evaluation across several English test collections. An approach to normalize effectiveness measures is applied in order to allow comparison across test collections. The provided visual presentation enables researchers and practitioners to study the impact of system components on retrieval effectiveness in an efficient way.

1 Motivation

Evaluationsforschung ist eines der wesentlichen Instrumente der Informationswissenschaft. Im Information Retrieval ermöglicht sie eine Bewertung von Systemen anhand festgelegter Kriterien, das heißt einer Menge von Anfragen und deren Relevanzbewertungen bezüglich einer Sammlung von Dokumenten. Diese Vorgehensweise für die Evaluation von IR Systemen, die als Cranfield Paradigma bezeichnet wird, hat sich im Verlauf der letzten Jahrzehnte im Bereich des textuellen Retrievals nicht wesentlich weiterentwickelt. Im gleichen Zeitraum haben sich die eingesetzten Systeme in ihrem Aufbau sehr wohl verändert. Nicht nur die Komponenten, aus denen ein IR System besteht, sind für sich gesehen komplexer geworden. Auch die Vielfalt verfügbarer theoretischer Modelle ist wesentlich größer als zu Zeiten des Cranfield Paradigmas. Dies macht es zum heutigen Zeitpunkt nahezu unmöglich, für eine vorliegende Dokumentsammlung das Optimum im Sinne der Systemgenauigkeit aus dem Portfolio der wissenschaftlichen Modelle und Komponenten auszuwählen. Für Inhaltsanbieter, egal ob regionaler Zeitungsverlag oder Fernsehsender mit nationaler Reichweite, stellt sich unter anderem genau dieses Problem bei der Verwertung ihres Archivbestandes im Geschäftsfeld des Internets. Daher werfen diese Fakten im Kontext der textbasierten Suche die Frage nach neuen Methoden der vergleichenden Systembewertung auf.

An dieser Schnittstelle setzt der hier vorgestellte Ansatz zur Evaluation auf Komponentenebene an. Die Idee des Vergleichs von Systemen auf Basis ihrer wesentlichen Bestandteile wurde bereits im Rahmen verschiedener Evaluationskampagnen vorgeschlagen. Ein allgemeiner Überblick wird in (Hanburry 2010) gegeben. Als Ausgangspunkt dient der Grid@CLEF Task (Ferro 2009). Dort wurde ein vierstufiges Konzept für die Evaluation von bilingualen Systemen vorgeschlagen. Basierend auf dieser Idee werden hier unterschiedliche Methoden für drei wesentliche Systemkomponenten anhand verschiedener Textkorpora verglichen. Als Testdaten dienen zwölf englischsprachige Kollektionen aus den CLEF[1] und TREC[2] Kampagnen mit entsprechenden Anfragen und Relevanzbewertungen.

1 Cross-Language Evaluation Forum: http://www.clef-campaign.org/, abgerufen am 08.11.2010

2 Text Retrieval Conference: http://trec.nist.gov/, abgerufen am 08.11.2010

Die Bewertung von IR Systemen beruht im Normalfall auf einer Kennzahl wie Mean Average Precision (MAP) oder Geometric Mean Average Precision (GMAP). Beide Metriken repräsentieren die durchschnittliche Güte eines Systems auf einer Menge von Anfragen. Wir verwenden MAP und GMAP für den Vergleich der Konfigurationen. Diese Auswahl wird in Abschnitt vier näher begründet. Da wir zusätzlich über verschiedene Kollektionen hinweg vergleichen wollen, wenden wir eine Methode zur Normierung der Werte an, die ebenso in Abschnitt vier beschrieben wird. Die wesentlichen Beiträge dieser Forschungsarbeit sind die Vorstellung des experimentellen Aufbaus und der Ergebnisse einer umfassenden Evaluation auf Komponentenebene über mehrere englischsprachige Korpora hinweg. Darüber hinaus wird ein Werkzeug zur Visualisierung vorgestellt, mit dem sich die umfangreichen Evaluationsergebnisse vergleichen und interpretieren lassen.

2 Experimentaufbau

In den vergangenen fünf Jahren wurde ein hochgradig flexibles Retrieval Framework entwickelt, dessen Konzept in (Kürsten 2008) dargestellt ist. Das System integriert die beiden wissenschaftlichen Tools Terrier (Ounis 2007) und Lemur (Ogilvie 2002) sowie das Open-Source Projekt Apache Lucene[3]. Im Rahmen der CLEF Kampagne wurden Erfahrungen gesammelt, um Einblicke in die Funktionsweise von IR Systemen zu gewinnen und das Zusammenspiel der Komponenten besser zu verstehen. Die in den Vergleichen erzielten Resultate hatten dabei durchweg unterschiedliche Güte und machen die eingangs formulierte These der Schwierigkeit der Auswahl einer möglichst optimalen Systemkonfiguration anhand eines vorgegebenen Datenkorpus offensichtlich.

2.1 Parameter der Evaluation

Nachfolgend werden Evaluationsergebnisse vorgestellt, die über einen dreidimensionalen Parameterraum von Systemkonfigurationen aufgespannt wer-

3 Apache Lucene Suchmaschine: http://lucene.apache.org/, abgerufen am 08.11.2010

den. Die erste wesentliche Komponente ist die Wortstammreduktion. In das verwendete Framework wurden fünf Varianten integriert und getestet:

- Porter Stemmer (Porter 1997)
- Krovetz Stemmer (Krovetz 1993)
- UeaLite Stemmer (Jenkins 2005)
- N-Gram Stemmer, mit N=4 und N=5

Die Auswahl der angegebenen Algorithmen zur Wortstammreduktion deckt drei unterschiedliche Ansätze ab. Der wohl am weitesten verbreitete Algorithmus nach (Porter 1997) steht für die Klasse der regelbasierten Verfahren. Der von (Krovetz 1993) beschriebene Algorithmus steht für die Gruppe der Ansätze, die die Probleme von zu starker oder zu schwacher Reduktion durch den Einsatz eines Wortbuchs abschwächen. Der UeaLite Stemmer aus (Jenkins 2005) setzt auf eine generell weniger stark ausgeprägte Reduktion. Der N-Gram Stemmer ist sprachenunabhängig aber dadurch gleichzeitig auch ungenauer. Aufgrund dessen ist er für den Einsatz auf mehrsprachigen Kollektionen besonders geeignet. In einer umfassenden Studie (McNamee 2009) über eine Reihe von Testkollektionen haben sich die angegebenen Varianten mit N=4 und N=5 auf Korpora in englischer Sprache im Vergleich als besonders robust erwiesen.

Der Ranking Algorithmus, der in der Forschung von allen Komponenten am häufigsten untersucht wird, bildet die zweite Dimension des untersuchten Parameterraums. In der vorliegenden Studie werden 13 verschiedene Modelle, die zum Großteil im Terrier Framework (Ounis 2007) integriert sind, zur Evaluation ausgewählt. Die nachfolgende Liste enthält alle getesteten Varianten in drei wesentlichen Gruppen:

- Klassische Modelle: TF-IDF, BM25 und Lucene
- Modelle aus dem Divergence from Randomness (DFR) Framework: DFRee, DFR_BM25, DLH, DPH, BB2, IFB2, In_ExpB2 und PL2
- Linguistisch motivierte (LM) Modelle: HiemstraLM, DirichletLM

Die erste Gruppe enthält mit dem Ranking Algorithmus von Lucene und TF-IDF zwei ähnliche Modelle, die beide im Wesentlichen auf einer Kombination von Term- und inverser Dokumentfrequenz basieren. Der Hauptunterschied liegt in der flexiblen Gewichtungsmöglichkeit mithilfe der Lucene Programmierschnittstelle. Die aufgeführte Klasse der wahrscheinlichkeitsbasierten DFR Modelle wird bereits umfangreich im Rahmen der Dokumen-

tation[4] des Terrier Frameworks beschrieben. Erwähnenswert ist hier die Sonderstellung der Modelle DLH und DPH, die aufgrund einer abgewandelten mathematischen Grundlage streng genommen in eine eigene Kategorie gehören. Die theoretische Basis der Gruppe der LM Algorithmen ist eine für jeden Dokumentkorpus spezifische Analyse der Wortverteilungen, die dann wiederum mit wahrscheinlichkeitsbasierten Annahmen in eine Dokumentrangfolge überführt wird.

Die letzte Dimension des angedeuteten Parameterraums entsteht durch den Einsatz unterschiedlicher Modelle für automatisches Pseudo-Relevanzfeedback (PRF). In dieser Studie wurden die zwei Ansätze Kullback-Leiber und Bose-Einstein2, die ihre theoretische Grundlage ebenfalls im DFR Modell haben, mit Konfigurationen ohne PRF verglichen. Beim Einsatz von PRF haben die beiden zusätzlichen Parameter Dokumentanzahl und Termanzahl einen Einfluss auf die Güte der Ergebnisse. Daher wurden hierfür sieben verschiedene Ausprägungen der Dokumentanzahl und 13 Varianten für die Gesamtanzahl der letztlichen Erweiterungsterme ausgewählt.

Die Variation der Ausprägungen der angegebenen Parameter ergibt 11,895 Systemkonfigurationen, die entsprechend für jeden der im nachfolgenden Abschnitt aufgeführten Testkorpora getestet wurden. Insgesamt wurden daher gut 140,000 Retrievalexperimente durchgeführt.

2.2 Testkorpora

Für die Evaluation wurden unterschiedliche Korpora englischer Sprache ausgewählt, um Zusammenhänge zwischen der Art des Korpus in Bezug auf linguistische Merkmale und der Güte der Systemkonfigurationen zu untersuchen. In Tabelle 1 sind die verwendeten Korpora mit entsprechenden Metadaten gelistet. In der Menge der Testkorpora wurden vier wesentliche Typen identifiziert: (a) bibliothekarische Kataloge (LIB); (b) Beschreibungen multimedialer Daten (MM); (c) Nachrichtenartikel (NEWS); und (d) eine Sammlung manuell transkribierter Sprache (SPTR). Im Rest dieses Beitrags verwenden wir die Identifikation (ID) aus Tabelle 1, um ein entsprechendes Testkorpus zu benennen. Die Abkürzung KPN steht für die Evaluationskampagne, bei der die entsprechende Dokumentsammlung verwendet wurde.

4 Dokumentation zum Terrier Framework: http://terrier.org/docs/v3.0/dfr_description.html, abgerufen am 08.11.2010

Die Spalte #QU gibt die Größe der für den jeweiligen Testkorpus verfügbaren Anfragemenge an.

Tabelle 1. *Testkorpora und deren Eigenschaften*

ID	Name	KPN	Jahr	#Dok.	#QU
LIB1	CSA-EN	CLEF	2007	20,000	50
LIB2	TEL (British Library)	CLEF	2008	~1,000,000	100
LIB3	Federal Register	TREC	1997	~55,000	150
LIB4	GIRT4-EN (GESIS)	CLEF	2003	~150,000	150
SPTR	103rd Congress Rec.	TREC	1997	~30,000	50
MM1	Belga Image Captions	CLEF	2009	~500,000	50
MM2	IAPR-TC12 Annotat.	CLEF	2007	20,000	60
MM3	Wiki Images	INEX	2006	~150,000	106
NEWS1	Financial Times	TREC	1997	~210,000	150
NEWS2	Foreign Broadcast IS	TREC	1997	~130,000	150
NEWS3	LA Times 1994	CLEF	2009	~110,000	89
NEWS4	LA Times 1989/90	TREC	1997	~130,000	150

3 Ergebnisse

Eine detaillierte Auswertung der gut 140,000 durchgeführten Experimente stellt aufgrund der schieren Datenmenge eine Herausforderung dar. Als Kennzahlen für die Systembewertung werden MAP und GMAP eingesetzt. In Tabelle 1 wird deutlich, dass für die Testkorpora unterschiedlich große Mengen von Anfragen verwendet wurden. In (Robertson 2006) wurde argumentiert, dass die Stabilität der MAP insbesondere auf kleineren Anfragemengen am höchsten ist. Ausgehend von dieser These verwenden wir daher MAP als Referenzmaß. Darüber hinaus führen wir die Güte der Systeme zusätzlich anhand der GMAP auf. Die Gründe dafür sind die ebenfalls in (Robertson 2006) angeführte Fokussierung auf die Robustheit von Systemen und den Aspekt, dass keines der beiden Maße besser oder schlechter ist als das jeweils andere. Ferner soll später beim Vergleich der Systemkonfigurationen die Möglichkeit bestehen, die für eine jeweilige Forschungsfrage passendere Kennzahl auszuwählen. In Tabelle 2 werden die besten Ergebnisse je

Testkorpus dem durchschnittlichen Ergebnis aller Systemkonfigurationen (AMAP, AGMAP) gegenübergestellt.

Die Analyse der Resultate zeigt, dass sowohl die durchschnittliche Güte der Systemkonfigurationen als auch deren Verhältnis zur jeweils besten Konfiguration in Abhängigkeit von Korpus und Anfragemenge variiert. Betrachten wir jedoch das Verhältnis der beiden aufgelisteten Größen, so lässt sich feststellen, dass die beste Konfiguration für die MAP im Bereich von 22 bis 42 Prozent oberhalb der AMAP liegt. Wobei hier jedoch vier Testkorpora die obere Grenze um 40 Prozent markieren. Betrachtet man die Ergebnisse im Sinne der GMAP, so zeigt sich, dass genau diese vier Testkorpora die AGMAP mit 74 bis 130 Prozent besonders deutlich übertreffen.

Tabelle 2. *Durchschnittliche Güte der Systemkonfigurationen je Testkorpus im Verhältnis zur besten getesteten Systemkonfiguration nach MAP und GMAP*

ID	AMAP	MAP	AGMAP	GMAP
LIB1	0.2878	0.3776 (+31.20%)	0.2021	0.2971 (+47.00%)
LIB2	0.2958	0.4187 (+41.56%)	0.1584	0.2759 (+74.18%)
LIB3	0.2403	0.3361 (+39.84%)	0.0356	0.0817 (+129.63%)
LIB4	0.3248	0.4183 (+28.80%)	0.1911	0.2802 (+46.61%)
SPTR	0.2225	0.3203 (+43.92%)	0.0656	0.1390 (+112.08%)
MM1	0.4198	0.5309 (+26.48%)	0.2995	0.4685 (+56.46%)
MM2	0.2380	0.2916 (+22.49%)	0.0653	0.1028 (+57.46%)
MM3	0.2168	0.2781 (+28.30%)	0.0798	0.1292 (+61.83%)
NEWS1	0.2717	0.3306 (+21.68%)	0.0864	0.1430 (+65.55%)
NEWS2	0.2360	0.3112 (+31.83%)	0.0574	0.1191 (+107.48%)
NEWS3	0.4521	0.5864 (+29.70%)	0.2616	0.4239 (+62.02%)
NEWS4	0.2215	0.2876 (+29.85%)	0.0901	0.1450 (+60.99%)

Aus Tabelle 1 kann man wiederum ablesen, dass die Anzahl der Anfragen allein nicht für diese Schwankungen verantwortlich sein kann. Denn die Anfragemenge deckt das volle Spektrum zwischen 50 und 150 ab. Die restlichen Testkorpora schwanken beim Verhältnis GMAP zu AGMAP zwischen gut 47 und knapp 66 Prozent. Dies entspricht einem ähnlichen Bereich wie beim Verhältnis zwischen MAP und AMAP. Die absoluten Werte sind jedoch deutlich höher, was dafür spricht, dass die besten Systemkonfigurationen wesentlich robuster sind. Die hier dargelegten Daten ermöglichen noch keine

Bewertung des Einflusses einzelner Komponenten auf die Güte der Ergebnisse. Die Voraussetzungen für eine vergleichende Bewertung der Systemkonfigurationen auch über mehrere Testkorpora werden im nachfolgenden Abschnitt diskutiert.

4 Vergleich der Systemkonfigurationen

Aus Forschungssicht interessanter als die Güte des besten Systems für jedes Testkorpus zu bestimmen, ist die vergleichende Bewertung der Konfigurationen über eine Menge von Korpora. Darüber hinaus stellt sich dann vielmehr die Frage, welche Zusammenhänge zwischen Testkorpus und Systemkonfiguration in Bezug auf die Güte der Retrievalergebnisse existieren. Um sich dieser Fragestellung anzunehmen und eine Vergleichbarkeit über Korpora hinweg zu gewährleisten, müssen die Ergebnisse in geeigneter Form normiert werden. Zu diesem Thema existieren bereits Forschungsarbeiten, deren Kernideen nachfolgend kurz skizziert werden. Ist diese Normierung erfolgt, können die Ergebnisse in entsprechender Form aufbereitet werden, um einen Vergleich auch über verschiedene Korpora hinweg zu ermöglichen. Da die Darstellung von knapp 12,000 Experimenten in traditioneller Form mittels Tabellen oder statischen Grafiken nicht gelingen kann, stellen wir ein interaktives Werkzeug zur visuellen Interpretation vor.

4.1 Normierung der Ergebnisse

In (Mizarro 2007) werden die Evaluationsdaten für eine spätere Netzwerkanalyse normiert. Dabei erfolgt die Standardisierung der Ergebnisse anhand der Systemgüte im Sinne von MAP oder GMAP, einerseits durch Subtraktion des durchschnittlichen Ergebnisses für eine Anfrage und andererseits durch Subtraktion des Durchschnittswerts für ein System. In weiteren Publikationen wurde diskutiert, die Normierung anhand des besten Wertes je Anfrage durchzuführen. Ein weiterer Ansatz zur Standardisierung (Webber 2008) schlägt hingegen vor, für ein System vergleichbare Ergebnisse auch auf unterschiedlichen Korpora zu erzeugen. Dazu wird zusätzlich zur einfachen Normierung auch die Varianz von Retrievalergebnissen betrachtet, was eine

Auswahl von Anfragen ermöglicht, mit der dann vergleichbare Gesamtergebnisse erzielt werden können.

In unserem konkreten Fall ist die Varianz der Ergebnisse der Systemkonfigurationen von Interesse. Denn sie ist der womöglich einzige Anhaltspunkt für die Zusammenhänge zwischen Korpuseigenschaften und Systemkonfigurationen. Daher wurde die in (Mizarro 2007) vorgeschlagene Strategie angewendet und die Ergebnisse für MAP und GMAP jeweils anhand der durchschnittlichen Güte aller Systemkonfigurationen je Anfrage normiert. Damit wird für jede Konfiguration jeweils ein normierter Wert MAP_n und $GMAP_n$ ermittelt. Die Formeln (1) und (2) verdeutlichen den Prozess zur Normierung anhand der MAP_n. Dabei steht $AP(s_i, t_j)$ für die Average Precision von Systemkonfiguration *i* für Anfrage *j* und $AAP(t_j)$ für die durchschnittliche Average Precision aller Konfigurationen für Anfrage *j*.

$$AP_n(s_i, t_j) = AP(s_i, t_j) - AAP(t_j) \quad (1)$$

$$MAP_n = \frac{1}{n}\sum_{j=1}^{n} AP_n(s_i, t_j) \quad (2)$$

Der Nachweis der Äquivalenz zwischen MAP und MAP_n respektive GMAP und $GMAP_n$ wurde ebenfalls in (Mizarro 2007) geführt. Die Äquivalenzbeziehung zwischen MAP und MAP_n respektive GMAP und $GMAP_n$ macht eine Auflistung der normierten Ergebnisse analog zu Tabelle 2 überflüssig.

4.2 Interaktive Visualisierung

Aus Sicht der Autoren ist ein Vergleich von Systemkonfigurationen und deren Komponenten am besten visuell und interaktiv realisierbar. Die Datenvisualisierung ist ein eigenes Forschungsgebiet, deren Inhalte hier nicht näher erläutert werden sollen. Vielmehr wird es als Mittel zum Zweck eingesetzt. Zur Visualisierung von mehrdimensionalen Daten wird in der Literatur häufig das Prinzip paralleler Koordinaten (Wegman 1990) vorgeschlagen. Daher soll dieser Ansatz hier als Grundlage dienen. Zur Realisierung wird die freie Programmbibliothek Protovis[5] der Forschergruppe Visualisierung an der Stanford Universität verwendet.

5 JavaScript Bibliothek Protovis: http://vis.stanford.edu/protovis/, abgerufen am 08.11.2010

Abbildung 1 zeigt das Visualisierungswerkzeug. Die Säule ganz rechts steht für die Güte der Konfiguration in Bezug auf die Rechercheergebnisse. Alle weiteren Säulen spannen den Parameterraum auf. Die Farbgebung[6] verdeutlicht den Einfluss der Konfiguration auf die Recherchequalität, dabei steht grün für gute und rot für schlechte Ergebnisse. Die Nutzung des Werkzeugs erfolgt in zwei Schritten. Zuerst wird aus der Datenbasis eine gewünschte Untermenge selektiert. Diese Auswahl erfolgt durch Einschränkung der Parameter, der Güte der Ergebnisse oder einer Kombination aus beiden.

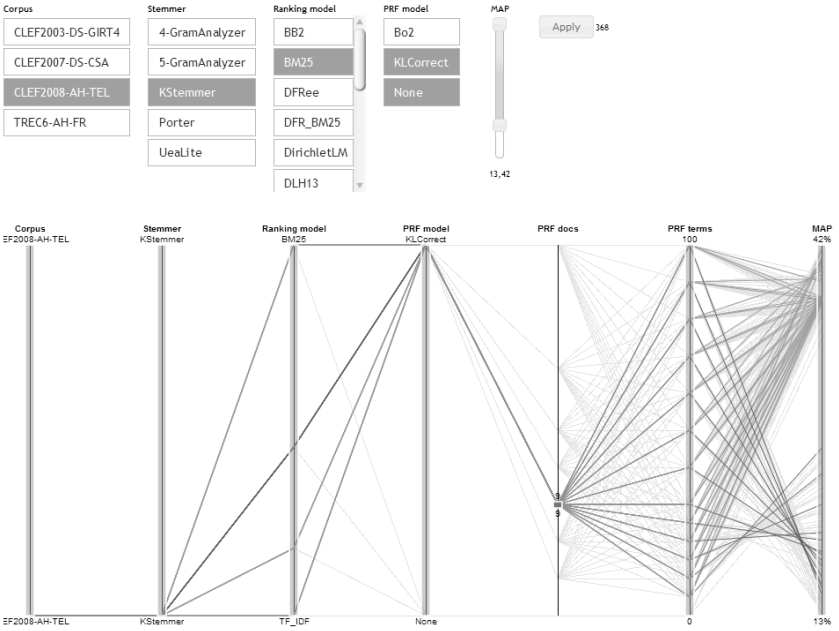

Abbildung 1: Visualisierung der Ergebnisse der mehrdimensionalen Systemkonfigurationen für ein ausgewähltes Testkorpus. Die Säulen 1–6 stellen den Parameterraum dar, Säule 7 die Güte (MAP) der Retrievalergebnisse.

Im folgenden Schritt können die ausgewählten Daten interaktiv verglichen werden. Dazu kann auf jeder Säule ein Bereich markiert werden, der sich verschieben lässt, um unterschiedliche Konfigurationen der entsprechen-

6 Der Prototyp zur Visualisierung ist online erreichbar unter: http://sachsmedia.tv/compeval/, abgerufen am 08.11.2010

den Komponente zu vergleichen. Im dargestellten Beispiel wurden der Korpus CLEF2008-AH-TEL, der Krovetz Stemmer, vier Ranking Modelle sowie kein PRF und das PRF Modell KLCorrect ausgewählt. Exemplarisch wurde in der Visualisierung auf der Säule PRF Dokumente eine Einschränkung auf 9 Dokumente vorgenommen. Die MAP für die selektierten Systemkonfigurationen liegt zwischen 13 und 42 Prozent. Die Ergebnisse der im Beispiel ausgewählten Konfigurationen sind farblich hervorgehoben. Zu erkennen ist eine Kumulation im Bereich von 13 bis 26 Prozent MAP und eine zweite im Bereich von 36 bis 42 Prozent. Über die MAP Säule kann die Darstellung auf ausgewählte Retrievalergebnisse reduziert werden, um beispielsweise eine besonders gute oder schlechte Konfiguration zu identifizieren.

5 Fazit und Ausblick

Zur vergleichenden Bewertung der am Retrievalprozess beteiligten Systemkomponenten wurden die umfangreichen Ergebnisse der vorgestellten experimentellen Studie in eine interaktive, grafische Darstellung überführt. Das Werkzeug erlaubt es, jede Dimension und die Zielmetrik des untersuchten Parameterraumes gezielt einzuschränken. Das in Abschnitt vier dargelegte Beispiel zeigt, dass sowohl die Auswirkungen einzelner Parameter als auch die Wechselwirkungen zwischen den Komponenten auf einen oder mehrere Testkorpora analysiert werden können. Damit lassen sich Rückschlüsse auf die Robustheit der jeweiligen Systemkonfigurationen ziehen. Für weitere Arbeiten existieren bereits wesentliche Ansatzpunkte. So ist beispielsweise eine Verbesserung der Auswahlmechanismen des Werkzeugs geplant. Die Möglichkeit Koordinaten aus der Darstellung zu entfernen, würde es erlauben, auf bestimmte Aspekte der Konfiguration konzentrierter eingehen zu können.

Danksagung

Diese Arbeit wurde teilweise von Mitarbeitern der Forschungsinitiative *sachsMedia* (www.sachsmedia.tv) realisiert, die im Rahmen des Förderprogramms *Unternehmen Region* vom BMBF finanziert wird.

Referenzen

Ferro, N. und Harman, D. (2009). CLEF 2009: Grid@CLEF pilot track overview. In: Multilingual Information Access Evaluation I. Text Retrieval Experiments, 10th CLEF Workshop, Corfu, Greece

Hanburry, A. und Müller, H. (2010). Automated Component-Level Evaluation: Present and Future. In: Multilingual and Multimodal Information Access Evaluation, Padua, Italy

Jenkins, M. C. und Smith, D. (2005). Conservative stemming for search and indexing. In: Proceedings of the 28th international ACM SIGIR conference, Salvador, Brazil

Kürsten, J., Wilhelm, T., und Eibl, M. (2008). Extensible retrieval and evaluation framework: Xtrieval. LWA 2008: Lernen – Wissen – Adaption, Workshop Proceedings, Germany

Krovetz, R. (1993). Viewing morphology as inference process. In: Proceedings of the 16th international ACM SIGIR conference, pp. 191–202, Pittsburgh, USA

McNamee, P., Nicholas, C., und Mayfield, J. (2009). Addressing morphological variation in alphabetic languages. In: Proceedings of the 32nd international ACM SIGIR conference, pp. 75–82, July 19–23, Boston, USA

Mizarro, S. und Robertson, S. (2007). Hits hits TREC: exploring IR evaluation results with network analysis. In: Proceedings of the 30th international ACM SIGIR conference, pp. 479–486, Amsterdam, Netherlands

Ogilvie, P. und Callan, J. (2002). Experiments using the Lemur toolkit. In: Proceedings of the 2001 Text Retrieval Conference, pp. 103–108. National Institute of Standards and Technology, special publication 500-250, USA

Ounis, I., Lioma, C., Macdonald, C., und Plachouras, V. (2007). Research directions in terrier: a search engine for advanced retrieval on the Web. Novatica/UPGRADE Special Issue on Next Generation Web Search, pp. 49–56

Porter, M. F. (1997). An algorithm for suffix stripping. In: Multimedia information and systems series – Readings in information retrieval, pp. 313–316, San Francisco, USA

Robertson, S. (2006) On GMAP: and other transformations. In: Proceedings of the 15th ACM CIKM conference, pp. 78–83, Arlington, USA

Webber, W., Moffat, A., und Zobel, J. (2008). Score standardization for inter-collection comparison of retrieval systems. In: Proceedings of the 31st international ACM SIGIR conference, pp. 51–58, Singapore

Wegman, E. J. (1990). Hyperdimensional data analysis using parallel coordinates. In: Journal oft he American Statistical Association, Vol. 85, No. 411, pp. 664–675, USA