

## **Feld-Spezifische Indexierung von Internet-Dokumenten im Rahmen von WebCLEF 2006**

*Ben Heuwing, Robert Strötgen*

Universität Hildesheim  
Informationswissenschaft  
Marienburger Platz 22  
31141 Hildesheim  
ben.heuwing@uni-hildesheim.de

### **Zusammenfassung:**

Im Rahmen von WebCLEF 2006 wurde an der Universität Hildesheim mit dem sehr umfangreichen, multilingualen EuroGOV-Korpus experimentiert. Im Vordergrund stand die feldspezifische Indexierung anhand von HTML Strukturelementen. Zusätzlich wurde der Einsatz von Blind Relevance Feedback evaluiert. Wie 2005 wurde ein sprachunabhängiger Indexierungsansatz verwendet. Experimentiert wurde mit dem HTML-Title Element, dem H1 Element und anderen Auszeichnungen, die Text hervorheben. Blind Relevance Feedback wurde für alle Felder außer für das Volltextfeld ‚content‘ implementiert. Die besten Resultate wurden mit einer starken Gewichtung der HTML-Title und H1 Elemente erreicht und stellten eine geringfügige Verbesserung gegenüber den Ergebnissen aus den letztjährigen Postexperimenten dar. Der Einsatz von Blind Relevance Feedback führte nicht zu Verbesserungen. Für WebCLEF 2006 wurden verbesserte Ergebnisse mit den manuell erstellten Anfragen erreicht, während von den Veranstaltern automatisch erstellte Anfragen zu Ergebnissen führten, die wesentlich unter denen der manuell erstellten lagen. Dies war bei allen teilnehmenden Gruppen der Fall.

### **Abstract:**

For WebCLEF 2006 we experimented with the large, multilingual EuroGOV-Collection. Fieldspecific Indexing using the HTML structure of the web documents was evaluated. In addition, blind relevance feedback was applied in the search process. As in 2005, the experiments were carried out with a language independent indexing strategy. We experimented with HTML title, H1 element and other elements emphasizing text. Blind relevance feedback was implemented for all index fields except for the full content. The best results with the WebCLEF 2005 topics were achieved with a strong weight on the title-element accomplishing a marginal improvement over the best post submission runs for the mixed-monolingual task at WebCLEF 2005. Blind relevance feedback could not yet improve results. For the WebCLEF 2006 topics, improved results were achieved with the manually generated topics, while those automatically generated led to results far below average for all groups participating.

## 1. Einleitung

Vor dem Hintergrund eines sich erweiternden Europas mit gesteigener Sprachenvielfalt und der Bedeutung von Suchmaschinen als für viele Nutzer wichtigstem Zugangspunkt zum Internet ist die Entwicklung und Evaluierung von Systemen für multilinguales Web-Retrieval von steigender Bedeutung. Der 2005 als Teil der CLEF-Initiative entstandene WebCLEF-Track<sup>1</sup> bietet eine Plattform, um sich mit den spezifischen Herausforderungen zu befassen, die Mehrsprachigkeit und ein sehr großer Korpus an das Information Retrieval stellen. Die Universität Hildesheim hatte nach der erfolgreichen Teilnahme bei WebCLEF 2005 den Anspruch, das entstandene System zu verbessern und zur Evaluierung weiterer IR-Methoden zu nutzen.

Die für das Web-Retrieval typischen Herausforderungen sind die vergleichsweise hohe Heterogenität und der große Umfang der zu durchsuchenden Kollektion. Bei näherer Betrachtung des für WebCLEF erstellten EuroGOV-Korpus werden auch die Herausforderungen der Multilingualität deutlich. Der Korpus umfasst 3,6 Millionen Dokumente in mehr als 25 Sprachen. Zunächst werden daher der EuroGOV-Korpus und die für WebCLEF entwickelten Aufgabenstellungen vorgestellt. Abschnitt 2 gibt dann einen Überblick über die Systeme und Verfahren, die im Kontext von Web Retrieval gute Ergebnisse in verschiedenen Evaluierungsinitiativen gezeigt haben. Auf der Grundlage der an der Universität Hildesheim für den CLEF Ad-hoc Track entwickelten Systeme konnte für WebCLEF 2005 mit einem sprachunabhängigen Indexierungsansatz das beste System für mehrsprachiges Retrieval entwickelt werden (JENSEN ET AL. 2006 und JENSEN 2005a/b). Das System wurde für WebCLEF 2006 in Hinblick auf die Vorverarbeitung der Daten, die Nutzung weiterer HTML-Strukturelemente und die Implementierung einer Blind Relevance Feedback-Funktion zur Anfrageerweiterung weiterentwickelt. In Abschnitt 4 werden die zentralen Elemente des Systems vorgestellt. Danach werden die erzielten Ergebnisse ausgewertet und in Bezug zu den Ergebnissen der anderen Teilnehmer gesetzt. Im letzten Punkt sollen die Ergebnisse bewertet und ein Ausblick auf die Maßnahmen gegeben werden, die für die nächste Teilnahme geplant sind.

## 2 WebCLEF Organisation

### 2.1 EuroGOV-Korpus

Für den ersten Durchgang von WebCLEF 2005 wurde aus den Internetseiten der Europäischen Union sowie den Seiten der Regierungseinrichtungen der europäischen Mitgliedsstaaten und Russlands eine Dokumentensammlung erstellt. Eine detaillierte Übersicht geben SIGURBJÖRNSSON ET AL. (2005a). Dieser Korpus wurde 2006 ohne Änderungen wieder verwendet. Der 80GB große EuroGOV-Korpus umfasst ca. 3.6 Mio. Internetseiten. Beim Zusammenstellen des Korpus wurde versucht, ein möglichst vollständiges Abbild des gewählten Ausschnittes des Internets zu erstellen. Dabei stellte es sich aufgrund von unterschiedlichen Namenskonventionen und komplexer Strukturen als

---

<sup>1</sup> <http://ilps.science.uva.nl/WebCLEF/>

problematisch heraus, die Seiten der Regierungsinstitutionen korrekt zu identifizieren. Bei der Erstellung wurde daher kein Anspruch auf Vollständigkeit erhoben. Stattdessen wurde versucht jeweils die Seiten der Regierung und die der wichtigsten Ministerien mit aufzunehmen. Gespeichert wurden reine Text/HTML Formate und Rich Document Types wie .doc, .pdf und .ps, jedoch keine Grafiken. Während des Crawling-Vorgangs, bei dem die Seiten eingesammelt wurden, entstanden aus unbekanntem Grund außerdem 70.000 leere Dokumente. Der Korpus enthält mehr als 20 verschiedene Sprachen, wobei die Verteilung der Sprachen eng an die Domains gebunden ist. Besonders interessant ist daher auch die europäische Domain eu.int, da sie Dokumente in allen Sprachen der Mitgliedsstaaten enthält.

Der Korpus besteht aus 157 Dateien in einem XML-ähnlichen Format mit jeweils maximal 25.000 Dokumenten. Zu jedem Dokument gibt es einen Eintrag, der Metadaten und den Inhalt des Dokuments enthält. Vorhandene Metadaten sind neben einer eindeutigen Identifikationsnummer beispielsweise die Internetadresse (URL) des Dokuments und Angaben über den Dokumenttyp, wenn diese durch den ausgebenden Server übertragen wurden. Die eigentlichen Dokumente befinden sich in Form von Text innerhalb eines speziellen XML-Elements, dem CDATA-Element. Innerhalb dieses Elements können beliebige andere XML-Elemente auftauchen, ohne als solche verarbeitet zu werden. Dies würde unweigerlich zu Fehlermeldungen führen, da die Internetsprache HTML in hohem Maße XML-Charakteristiken aufweist und dabei verschiedenste Fehlerquellen enthält. Nachteilig war, dass so in den CDATA-Elementen andere CDATA-Elemente aus den Internetdokumenten auftraten. Diese Verschachtelung von CDATA-Elementen ist in XML jedoch nicht zulässig. Dies ist einer der Gründe, warum der Korpus nur XML-ähnlich ist, das Format des Korpus also nicht wohlgeformtes XML ist. Ein weiterer Grund hierfür ist, dass die URLs, die als Metadaten angegeben wurden, Sonderzeichen enthalten, die nicht XML konform sind. Problematisch ist auch die Verarbeitung der Rich Document Types, deren binäre Bestandteile ebenfalls in Text umgewandelt wurden, wodurch im Korpus sinnlose Zeichenfolgen entstanden sind, welche auch Zeichen enthalten, die in einem XML-Dokument nicht auftauchen dürfen. Den Teilnehmern wurden Listen zur Verfügung gestellt, welche die leeren Dokumente und alle Dokumente der Typen .doc und .pdf identifizieren und mit denen diese dann herausgefiltert werden können, da diese Dokumente für die gestellten Aufgaben nicht relevant waren. Eine weitere Herausforderung für den Prozess der Vorverarbeitung sind die zahlreichen unterschiedlichen Zeichenkodierungen der Textdokumente.

Der Korpus ist insofern in Bezug auf die enthaltenen Sprachen, die Dokumenttypen und die Zeichenkodierungen, stark heterogen. Dies entspricht den realen Gegebenheiten im Internet und der Aufgabenstellung einer Internetsuche.

## **2.2 Topics und Topic-Erstellung**

Die als Topics bereitgestellten Suchanfragen für WebCLEF 2006, auf deren Grundlage die teilnehmenden Systeme verglichen wurden, bestanden aus 319 manuell erstellten Anfragen (124 neu erstellte und 195 der für WebCLEF 2005 erstellten Topics) und 1620 automatisch erstellten. Die manuellen Topics umfassen insgesamt 11 Sprachen, die automatischen repräsentieren dagegen eher die Sprachvielfalt des Korpus, da sie mit Hilfe von Dokumenten aus allen Domains erstellt wurden (BALOG ET AL. 2006a). Das für die Erstellung der automatischen Topics verwendete Verfahren wurde im Nachhinein bekannt gegeben.

Dabei wurden aus zufällig ausgewählten Zieldokumenten über ein probabilistisches Modell Anfragen aus einem oder zwei Termen erstellt. Beim automatischen Ablauf der Erstellung sollte der Suchvorgang eines echten Nutzers simuliert werden. Daher wurde zusätzlich auch ein gewisser Anteil an Fehlern (Noise) eingebracht (BALOG ET AL. 2006a), was teilweise zu gemischtsprachigen Topics oder auf andere Art weniger realistischen Anfragen führte, etwa Topic WC0600346 „*bundesministerium der marginalspalte*“.

Die Topics wurden in einem XML-Format zur Verfügung gestellt (vgl. Abb. 1) und enthielten Metadaten zum Topic, wie die Sprache, eine englische Übersetzung, die Zieldomain und ein Nutzerprofil desjenigen, der das Topic erstellt hatte (in Hinblick auf bevorzugte Sprachen). Die Übersetzung und Angaben zum Nutzerprofil enthielten nur die manuell erstellten Topics. Die Metadaten konnten für die Suche eingesetzt werden, wobei dies jeweils angegeben werden musste.

```
- <topic>
  <num>WC0600001</num>
  <title>Arbeiten Deutschen Bundestag</title>
  - <metadata>
    - <topicprofile>
      <language language="DE" />
      <translation language="EN">Job opportunities German Bundestag</translation>
    </topicprofile>
    - <targetprofile>
      <language language="DE" />
      <domain domain="de" />
    </targetprofile>
    - <userprofile>
      <native language="NL" />
      <active language="EN" />
      <active language="DE" />
      <countryofbirth country="NL" />
      <countryofresidence country="NL" />
    </userprofile>
  </metadata>
</topic>
```

Abb. 1: WebCLEF 2006-Topic

### 2.3 WebCLEF-Tasks und Bewertung

Im Mittelpunkt der Evaluierung von WebCLEF stehen zwei Bereiche, die typische Aufgaben einer mehrsprachigen Suchmaschine repräsentieren, der *Mixed Monolingual Task* und der *Multilingual Task*. Für beide werden dieselben Topics eingesetzt. Beim Mixed Monolingual Task sollen Ergebnisse jeweils nur in der Sprache geliefert werden, in der die Suchanfrage formuliert wurde. Im multilingualen Task sollen zu jeder Anfrage relevante Ergebnisse in allen Sprachen gefunden werden. Zu jeder Anfrage wird dabei bereits eine englische Übersetzung angegeben. Als relevant gewertet werden dann neben dem ursprünglichen Zieldokument auch ähnliche Dokumente in den verschiedenen Sprachen. Vorstellbare Anwendungsszenarien für eine solche sprachübergreifende Suche sind etwa der Vergleich von Gesetzgebungen verschiedener Länder, Migranten, die Informationen über ein Land suchen, oder Informationssuche im Vorfeld der Eröffnung einer Firma im Ausland (DE RIJKE & SANTOS 2005). Dabei wird vorausgesetzt, dass der Nutzer über aktive oder passive Sprachkenntnisse in mehreren Zielsprachen verfügt. Da die Ergebnisse des multilingualen Tasks von WebCLEF 2005 insgesamt jedoch wenig erfolgsversprechend

waren, sprachen sich viele der Teilnehmer gegen eine Wiederaufnahme aus. Bei WebCLEF 2006 wurde daher nur der Mixed Monolingual Task durchgeführt. Es stand den Teilnehmern jedoch offen, zur Evaluierungszwecken auch Ergebnisse für den multilingualen Task einzureichen.

Aufgrund eingeschränkter Ressourcen für Topic-Erstellung und Relevanzbewertung wurde für die WebCLEF-Tasks das Prinzip der *known-item*-Suche (SIGURBJÖRNSSON ET AL. 2005b) gewählt. Dabei wird angenommen, dass ein Nutzer mit seiner Suchanfrage das Ziel verfolgt, eine bestimmte, bereits bekannte Seite wieder zu finden. Dabei sollte entweder die Homepage, also die Einstiegsseite einer Website, oder eine bestimmte Seite innerhalb einer Website gefunden werden (Homepage Topics bzw. Named Page Topics). Zu welcher Gruppe ein Topic gehört war zum Zeitpunkt der Suche nicht bekannt. Diese Aufgabenstellung ist typisch für die Nutzung von Internetsuchmaschinen. Das *known-item* Prinzip vereinfacht aber auch die Bewertung, da zu jedem Topic nur ein gültiges Dokument existiert, welches schon bei der Erstellung des Topics mit angegeben werden kann, wobei jedoch noch Duplikate und Übersetzungen berücksichtigt werden müssen. So ist es für die Bewertung nur von Bedeutung, an welcher Stelle der Ergebnisliste dieses Dokument auftaucht. Dabei bleibt allerdings unberücksichtigt, ob noch andere Dokumente in der Ergebnisliste für die Anfrage relevant sind. Es wurde vorgeschlagen, einen *Ad-hoc Task*, bei dem auch diese Dokumente in die Bewertung eingehen, einzuführen. Ein solcher kann aber nur durchgeführt werden, wenn Mittel für die Relevanzbewertung zur Verfügung gestellt werden (SIGURBJÖRNSSON ET AL. 2005b).

Aufgrund des *known-item* Ansatzes wird als primäres Evaluierungsmaß der Mean Reciprocal Rank (MRR) verwendet. Dieser gibt den Rang des relevanten Dokuments innerhalb der Ergebnisliste (Reciprocal Rank =  $1 / \text{Rang des ersten relevanten Dokuments in der Ergebnisliste}$ ) als Durchschnitt über die Anfragen an. Ein MRR von 1 würde demnach bedeuten, dass das relevante Dokument immer an erster Stelle zurückgegeben wurde, ein MRR von 0.25, dass es durchschnittlich an vierter Stelle liegt. Der Umfang der Ergebnisliste zu jeder Anfrage war bei WebCLEF auf 50 Treffer beschränkt. Ein weiteres Evaluierungsmaß ist die durchschnittliche Rate, mit dem das relevante Dokument in den ersten  $x$  Ergebnissen (Average Success @ 1, 5, 10, 20, 50) enthalten ist. Diese Maße erlauben Bewertungen in Hinblick auf eine hohe Präzision der Suche und sind im Kontext von Web Retrieval üblich, da angenommen wird, dass Nutzer im Internet häufig nur die ersten Ergebnisse einer Suche beachten (MISHNE & DE RIJKE 2006,504). Um die Bewertung und Postexperimente zu erleichtern wurden von den Veranstaltern Dateien mit den Ergebnislisten und ein Perl-Skript zur automatischen Überprüfung der Ergebnisse zur Verfügung gestellt.

### 3 Aktuelle Entwicklungen im Web Information Retrieval

Als Überblick zum aktuellen Forschungsstand im Bereich des Web Information Retrieval und des mehrsprachigen Information Retrieval sollen einige Systeme und Methoden aus den entsprechenden Tracks innerhalb der Evaluierungsinitiativen TREC<sup>2</sup> und CLEF<sup>3</sup> vorgestellt werden.

---

<sup>2</sup> <http://trec.nist.gov/>

### 3.1 TREC: Terabyte Track

Innerhalb der TREC-Initiative beschäftigt sich der Terabyte Track<sup>4</sup> mit der Suche in sehr großen Datensammlungen. Neben einem klassischen Ad-Hoc Task und einem Task in Hinblick auf die Recheneffizienz bei der Verarbeitung von Suchanfragen, gibt es in dieser Initiative auch einen den WebCLEF Tasks ähnlichen Task (Named Page Finding Task, vgl. Abschnitt 2: WebCLEF-Tasks und Bewertung). Die hierbei erfolgreichen Systeme setzen größtenteils auch webspezifische Retrieval Methoden ein, wie die Analyse der Linkstruktur (z.B. PageRank-Algorithmus<sup>5</sup> von Google), Berücksichtigung der HTML-Struktur der Dokumente oder der Texte von Links, die auf eine Seite zeigen (In-links).

So setzt beispielsweise das im Terabyte Track 2005 mit einem MRR von 0,441 zweitbeste System auf alle drei Varianten (CLARKE ET AL. 2005). Das von der der Universität Massachusetts erstellte System benutzt in seinem erfolgreichsten Run für den Named Page Finding Task eine Mischung aus verschiedenen Language Modelling-Techniken. Die Struktur der HTML-Dokumente wird analysiert und in Feldern für HTML-Title, Headings und Body indiziert. Ein weiteres Indexfeld enthält die Texte der Links, die auf das Dokument verweisen. Außerdem wird der PageRank-Algorithmus eingesetzt und die Zahl der In-Links berücksichtigt (METZLER ET AL. 2005).

Dass der Einsatz der genannten webspezifischen Methoden keine notwendige Voraussetzung für den Erfolg ist, zeigt das erstplazierte System der Tsinghua Universität, das hier von nur die Indexierung der In-link Texte einsetzt und zusätzlich eine Analyse von Wortpaaren in den Anfragen vornimmt (ZHAO ET AL. 2005).

HAWKING & CRASWELL (2004) fassen für den Home Page Finding Task in TREC 2001 zusammen, dass der Einsatz von In-link Texten sehr effektiv ist, während die Berücksichtigung der Linkstruktur, wie die Zählung von In-links oder der PageRank-Algorithmus, weniger erfolgreich waren. Als vorteilhaft stellte sich die Analyse der URL in Hinblick auf die Position der Seite in der Hierarchie der Website heraus.

### 3.2 WebCLEF 2005: Mixed Monolingual Task

Die Systeme, die beim Mixed Monolingual Task von WebCLEF 2005 die besten Ergebnisse zeigten, waren das der Universität Glasgow und das der Firma Hummingbird. Die Universität Glasgow setzte auf Feld-Spezifische Indexierung und sprachspezifisches Stemming (das Zurückführen der Terme auf ihre Stammformen). Allerdings konnte auch ohne Stemming oder mit Stemmingansätzen, die eigentlich für das Englische entwickelt wurden, vergleichbare, nur wenig schlechtere Ergebnisse erzielt werden. Das sprachspezifische Stemming konnte die Qualität der Ergebnisse sogar einschränken, wenn Fehler bei der Sprachidentifizierung auftraten. Während der Vorverarbeitung des Korpus wird bei diesem System die wahrscheinlichste Zeichenkodierung der Dokumente heuristisch auf der Basis von HTTP-Header und enthaltener Metadaten ermittelt und in UTF-8 kodiert (MACDONALD ET AL. 2005).

---

<sup>3</sup> <http://www.clef-campaign.org/>

<sup>4</sup> <http://www-nlpir.nist.gov/projects/terabyte/>

<sup>5</sup> vgl. Sergey Brin, Lawrence Page: The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: Computer Networks and ISDN Systems 1998. <http://www-db.stanford.edu/~backrub/google.html>

Das System von Hummingbird benutzt nur das Title-Element aus der HTML-Struktur der Dokumente. Außerdem werden die URLs der Dokumente analysiert. Es wurde sowohl mit sprachunabhängigen Indexierungsmethoden als auch mit der Erstellung einzelner Indizes für elf Sprachen experimentiert. Eine höhere Gewichtung von Dokumenten, die höher in der Verzeichnishierarchie standen, hatte einen positiven Effekt für einen Teil der Topics (Homepage Finding, siehe Punkt 2: WebCLEF-Tasks und Bewertung). Auch sprachspezifische Stopwortlisten führten im Durchschnitt zu Verbesserungen. Wie auch für das System der Universität Glasgow war eine starke Gewichtung des HTML Title Elementes sehr effektiv (TOMLINSON 2005).

### **3.3 Multilinguale Verfahren im Kontext von WebCLEF**

Für das sprachübergreifende Retrieval ist neben der Beherrschung verschiedener Sprachen auch die Übertragung zwischen den Sprachen von elementarer Bedeutung. Die bei WebCLEF 2005 eingesetzten Verfahren zur Übersetzung führten in diesem Kontext nicht zu Verbesserungen. Stattdessen wurden die besten Ergebnisse mit dem sprachunabhängigen Ansatz (Suche anhand der Anfragen und ihrer englischen Übersetzungen) der Universität Hildesheim erreicht. Auch die zweitplazierte Gruppe Miracle verwendete keine Übersetzungsverfahren. Hierbei zeigte sich auch, dass Anfragen, die Eigennamen enthielten, am erfolgreichsten behandelt werden konnten (SIGURBJÖRNSSON ET AL. 2005b). In welcher Form die in den anderen CLEF-Tracks entwickelten sprachspezifischen Verfahren für die komplexe Situation eines vielsprachigen Web-Korpus skalierbar sind, ist somit noch offen. Im Vordergrund der CLEF-Tracks stehen Verfahren aus den Bereichen der Übersetzung von Anfragen (hierbei beste Ergebnisse mit mehreren verschiedenen Ressourcen), der Anpassung von Indexierung und Anfrageverarbeitung an einzelne Sprachen und des Zusammenfassens der Ergebnisse aus den verschiedenen Sprachen (GONZALO & PETERS 2005).

## **4 Retrievalsystem der Universität Hildesheim**

Die Universität Hildesheim hatte im Rahmen der CLEF-Initiative bereits seit mehreren Jahren Erfahrungen im multilingualen Retrieval gesammelt, so beispielsweise im Ad-hoc Track (HACKL et. al 2005). Vor diesem Hintergrund wurde 2005 ein System für den ersten Durchgang des WebCLEF-Tracks entwickelt, das mit seinem sprachunabhängigen Ansatz bei der multilingualen Suche das erfolgreichste im Feld der Teilnehmer war.

Für WebCLEF 2006 stand die Überarbeitung von grundlegenden Funktionen des Systems im Vordergrund, um Verbesserungen in beiden Bereichen, also auch für den Mixed Monolingual Track zu erreichen. Dabei wurde darauf Wert gelegt die Vorverarbeitung zu verbessern, um einen umfassenderen Index erstellen und mit mehreren Indexfeldern experimentieren zu können. Im Bereich des Retrieval wurde eine Blind Relevance Feedback Funktion zur Anfrageerweiterung implementiert. Die direkte Herangehensweise mit einem multilingualen Index wurde beibehalten, ebenso die multilinguale Suche ohne Übersetzung der Anfragen, die im letzten Durchgang zum Erfolg beim multilingualen Task geführt hatte. Obwohl dieser Task bei WebCLEF 2006 nicht stattfand, wurde zu Testzwecken auch mit dem neuen System ein multilingualer Run erstellt und eingereicht. Das entwickelte

System setzt als Basis-Suchmaschine die sehr effiziente Programmiersprache Apache Lucene<sup>6</sup> ein und ist vollständig in der Programmiersprache Java implementiert.

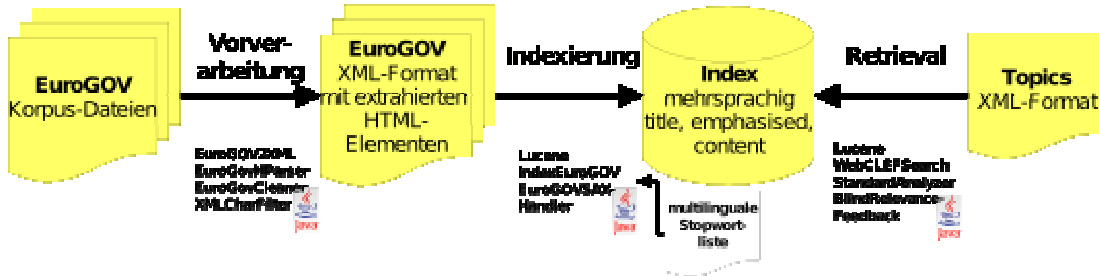


Abb. 2: System der Universität Hildesheim für WebCLEF 2006

#### 4.1 Vorverarbeitung des EuroGOV-Korpus

Die Dateien, aus denen der Korpus besteht, werden zunächst in ein XML-konformes Format gebracht, damit bei der Indexierung mit einem XML-Parser darauf zugegriffen werden kann. Bei dem System zu WebCLEF 2005 waren nach der Vorverarbeitung nicht alle Dateien vollständig XML-konform. Dies führte dazu, dass einzelne Dokumente bei der Indexierung nicht erfasst werden konnten. Bei der Vorverarbeitung werden daher nun zunächst alle Zeichen gefiltert, die nicht XML-konform sind<sup>7</sup>. Solche Zeichen treten im Korpus häufig auf. Der Filter arbeitet direkt auf der Ebene des Zeichenstroms und damit sehr effizient. Eine vermutete Fehlerquelle sind die vielen unterschiedlichen Zeichenkodierungen, die im Korpus auftreten. Vor allem bei problematischen Kodierungen dürfte dies Auswirkungen auf die Retrievalqualität haben. Das Problem wird wohl nur durch die spezielle Behandlung jedes einzelnen Dokuments in Hinblick auf seine Kodierung behoben werden können. Nach der Filterung dieser Zeichen werden die Sonderzeichen in den URLs maskiert, und die in wohlgeformten XML-Dokumenten nicht zugelassenen verschachtelten CDATA-Elemente entfernt.

Um mittels des XML-Parsers auf Inhalte von einzelnen in den Dokumenten enthaltenen HTML-Elementen zugreifen zu können, werden diese aus dem CDATA-Bereich extrahiert und als neue Elemente in die XML-Struktur eingefügt. Berücksichtigt werden folgende Elemente: <title>, <h[1-6]>, <strong>, <em>, <b> und <i>, da sie auf verschiedene Weise dazu dienen, Inhalte im HTML-Code inhaltlich oder optisch besonders hervorzuheben und diese daher potentiell bedeutungstragender als andere Inhalte sind. Im Laufe der Entwicklung wurde festgestellt, dass bei diesem Schritt wiederum Fehler im Dokument entstehen können, die daher in einem zweiten Arbeitsschritt beseitigt werden. Vor allem das mehrfache Ersetzen von Sonderzeichen durch XML-Entities (von & zu &amp; zu &amp;amp; etc.), dessen Folgen häufig im Internet zu finden sind, führte zu unerwarteten Problemen. Aufgrund dieser Maßnahmen konnten alle Text- bzw. HTML-Dokumente indiziert werden.

<sup>6</sup> <http://lucene.apache.org>

<sup>7</sup> <http://www.w3.org/TR/2004/REC-xml111-20040204/#charsets>

## 4.2 Feldspezifische Indexierung

Für die Teilnahme 2006 sollte die Indexierungsstrategie durch die Berücksichtigung der HTML-Struktur der Dokumente erweitert werden. Die Inhalte wichtiger HTML-Elemente sollten in unterschiedlichen Feldern indiziert werden, um ihnen dann bei der Suche unterschiedliche Gewichtungen zuteilen zu können. Die Nutzung des HTML Titel-Elements hatte sich bereits für den Durchgang im Jahr 2005 als sehr effizient herausgestellt. Für die Implementierung war es eine wichtige Voraussetzung, dass die eingesetzte Suchmaschine Lucene die Erstellung von Indizes mit mehreren Feldern erlaubt. Aufgrund einer vorherigen Analyse der Häufigkeit, mit der die verschiedenen HTML-Elemente auftreten, wurde entschieden, die Inhalte von <title> und <h1>-Elementen zu einem Indexfeld ‚title‘, die Inhalte der anderen extrahierten Elemente in einem Feld ‚emphasised‘ zusammenzufassen, und dann beim Retrieval mit unterschiedlichen Gewichtungen der Indexfelder zu experimentieren. Der Inhalt der Dokumente wird einmal in einem Feld ‚content‘ komplett indiziert. Die Termvektoren, die für die anderen Felder berechnet und für das Blind Relevance Feedback benötigt werden, werden für dieses Volltext-Feld nicht berechnet, da der Schritt für den gesamten Dokumentinhalt zu rechenintensiv erschien. Stattdessen werden hierfür zusätzlich 50 Tokens aus dem Inhalt in einem eigenen Feld indiziert (content\_cutoff). Der Text für dieses Feld wird aus der Mitte des Dokumentes genommen, da sich der bedeutungstragende Text einer Internetseite häufig nicht an ihrem Anfang befindet. Den Anfang einer Seite bilden häufig Menus und andere Navigationselemente, die in einem Webauftreten immer wieder in der gleichen Form auftreten und nicht spezifisch für ein einzelnes Dokument sind (CHEN ET AL. 2006). Aufwendigere Verfahren zum Aussortieren dieser Templates, wie beispielsweise durch den Vergleich verschiedener Seiten in Hinblick auf übereinstimmende Bestandteile, wurden nicht eingesetzt, um die Recheneffizienz des Systems nicht zu beeinträchtigen. Der Indexierungsvorgang auf einem Server mit 2 AMD Opteron 2,4 GHz Prozessoren und 8Gb Arbeitsspeicher dauerte 83 Stunden und der komplette Index mit Termvektoren beansprucht ungefähr 6Gb auf der Festplatte.

Vor der Indexierung werden an den Termen nur geringfügige Änderungen durchgeführt. Die Verarbeitung wird einem effizienten Standardmodul von Lucene überlassen, dem StandardAnalyzer. So wird außer der Entfernung von ‚s‘ -Endungen kein Stemming (das Zurückführen auf Wortstammformen) vorgenommen. Entsprechend dem Ansatz, mit einem multilingualen Index zu arbeiten, wurden keine sprachspezifischen Stemming-Algorithmen eingesetzt. Für WebCLEF 2005 hatte sich gezeigt, dass diese einfache Methode auch sprachunabhängig gute Ergebnisse liefern kann. Bestimmte Konstruktionen wie Akronyme oder E-Mail Adressen werden jedoch von der Analyzer-Klasse gesondert behandelt. Zusätzlich wurde eine Stopwortliste eingesetzt, um besonders häufig auftretende und damit wenig bedeutungstragende Wörter zu entfernen. Die bereits vorhandene multilinguale Stopwortliste, die 13 Sprachen umfasste, wurde um die im Korpus am häufigsten auftretenden Wörter erweitert. Diese erweiterte Liste beinhaltete 4722 Wörter. Die Idee, eine weitere titel-spezifische Stopwortliste einzusetzen, die z.B. auch automatisch erstellte und daher nicht bedeutungstragende Konstruktionen wie ‚no title‘ oder ähnliche Ausdrücke in anderen Sprachen entfernt hätte, wurde nicht verwirklicht. Eine Analyse der häufigsten Titel im EuroGOV-Korpus zeigte überraschenderweise, dass zwar einzelne Titel relativ häufig auftreten, diese aber trotzdem bedeutungstragend sind.

### 4.3 Gewichtung und Blind Relevance Feedback im Retrievalprozess

Beim Retrievalvorgang werden mit Hilfe des zuvor erstellten Indexes die besten Ergebnisse zu den Suchanfragen gesucht. Die Indexierung in mehrere Indexfelder schafft die Voraussetzung, um während des Retrievalvorganges mit unterschiedlichen Gewichtungen experimentieren zu können. Durch das Indexieren aller Dokumente mit Termvektoren konnte zusätzlich der Einsatz von Methoden des Blind Relevance Feedback (BRF) evaluiert werden. Termvektoren geben zu jedem Dokument die enthaltenen Terme und die Häufigkeit ihres Auftretens an. Für das BRF wird zunächst ein erster Suchdurchgang durchgeführt, um dann mit ausgewählten Termen aus den besten gefundenen Dokumenten (hierzu werden die Termvektoren benötigt) die ursprüngliche Suchanfrage zu erweitern und damit eine erneute Suche durchzuführen. Dahinter steht die Annahme, dass die im ersten Durchgang gefundenen besten Ergebnisse bereits der Anfrage entsprechen und weitere relevante Dokumente Ähnlichkeiten mit ihnen aufweisen sollten (GROSSMAN & FRIEDER 1998,84).

Die Experimente wurden durch die Tatsache vereinfacht, dass die Topics und die Ergebnisse aus dem Vorjahr als Testumgebung zur Verfügung standen und das von den Veranstaltern zur Verfügung gestellte Skript zur Überprüfung der Ergebnisse automatisiert zum Abschluss eines Suchvorgangs ausgeführt werden konnte, wodurch sofort Vergleichswerte zur Verfügung standen.

Um die Anfragen in das interne Format von Lucene zu übertragen, wird der Lucene QueryParser eingesetzt. Das Ranking der Ergebnisse basiert auf einer längennormalisierten tf-idf-Formel. Für eine Anfrage (Query) ‚q‘ und ein Dokument ‚d‘ wird der Rankingwert ‚score‘ wie folgt berechnet<sup>8</sup>:

$$\text{score}(q,d) = \frac{\sum_{t \text{ in } q} (\text{tf} \cdot \text{idf}^2 \cdot \{\text{Gewicht } t \text{ in } q\} \cdot \{\text{Längennormalisierung}\})}{\{\text{Überschneidung Anfrage/Dokument}\} \cdot \{\text{Query-Normalisierungsfaktor}\}}$$

wobei:

- t = Term
- tf = term frequency (Anzahl des Auftretens des Terms im Dokument)
- idf = inversed document frequency:  $\log(\text{Anzahl Dokumente}/(\text{Anzahl Dokumente mit Term} + 1)) + 1$
- {Gewicht t in q} = Gewichtung des Terms in der Anfrage
- {Längennormalisierung} =  $1 / \text{Wurzel}(\text{Anzahl Terme im Feld})$
- {Überschneidung Anfrage/Dokument} = Faktor, der einbezieht, wie viele Terme der Anfrage auch im Dokument auftauchen: Terme, die in Anfrage und im Dokument enthalten sind / Anzahl der Worte in der Anfrage
- {Query-Normalisierungsfaktor}: hat keine Auswirkungen auf das Ranking, macht Rankingwerte verschiedener Anfragen vergleichbar

Gesucht wurde jeweils auf den Feldern ‚content‘, ‚emphasised‘ und ‚title‘. Die Verwendung von ‚content\_cutoff‘ statt des ‚content‘-Feldes zeigte keine Vorteile, dieses Feld wurde daher nur für das BRF genutzt.

Damit die einzelnen Felder beliebig gewichtet werden konnten, wurde die Gewichtungsoption auch über die Kommandozeile verfügbar gemacht. Eine hohe Gewichtung des title-

<sup>8</sup> <http://lucene.apache.org/java/docs/api/org/apache/lucene/search/Similarity.html>

Feldes (20:0.1:1 im Verhältnis zu den beiden anderen Inhaltsfeldern, aber auch 10:5:1) brachte bei ersten Experimenten die größten Vorteile. Eine Suche nur auf dem Titelfeld führte im Vergleich zu schlechteren Resultaten. Die höhere Gewichtung der gemeinsam indexierten Inhalte von HTML-Überschriften (H2-H6) und anderer Elemente, die entweder der semantischen Hervorhebung oder der Hervorhebung im Schriftbild dienen (strong, em, bold und i), brachte hingegen nicht die gewünschten Ergebnisse. Eine differenzierte Indexierung der einzelnen Elemente könnte hier zu weiteren Erkenntnissen führen. Es könnten aber auch Verzerrungen in den Ergebnissen aufgetreten sein. Einmal, weil die Terme in den ‚title‘ und ‚emphasised‘ Feldern mehrmals (nämlich zusätzlich noch im Volltextfeld) indiziert wurden, was Auswirkungen auf die Gewichtung hat. Weiterhin sieht die verwendete Rankingformel von Lucene eine Längennormalisierung vor: Je kleiner der Inhalt eines Feldes ist, desto mehr Gewicht wird den einzelnen Worten gegeben. Dadurch, dass die ‚title‘ und ‚emphasised‘-Felder kleiner sind als das ‚content‘ Feld, wird ihr relatives Gewicht bereits verstärkt. Entsprechende Versuche deuten im Nachhinein darauf hin, dass es effektiv ist, bei diesen Feldern mit sehr niedrigen Gewichtungen zu arbeiten, um die oben beschriebenen Effekte auszugleichen.

Vor dem Hintergrund des Mixed Monolingual Tasks werden zusätzlich die in den Metadaten der Topics angegebenen Internetdomains der gesuchten Seiten ausgewertet, um die Ergebnisse auf Dokumente in dieser Domain einzuschränken. Dies ist durch den Query-Filter von Lucene ohne großen Rechenaufwand möglich. Dieser hat zusätzlich den Vorteil, dass die Rankingergebnisse nicht weiter beeinflusst werden. Die Domains sind Teil der Dokument-IDs und werden während des Indexierens extrahiert und in ein eigenes Feld geschrieben. Ohne diesen Schritt musste eine Suche mit Platzhaltern auf den IDs ausgeführt werden, was zu Performance-Problemen führte. Dieser Domain-Filter wurde in allen Runs genutzt.

Ebenfalls zu Evaluierungszwecken können folgende Parameter von der Kommandozeile aus verändert werden: Die Anzahl der für das BRF eingesetzten Dokumente, die Anzahl der für die Anfrageerweiterung verwendeten Terme sowie die Gewichtung der Erweiterung relativ zur ursprünglichen Anfrage. Für das BRF wird eine an der Universität Hildesheim für den CLEF Ad-Hoc Track erstellte Implementierung eingesetzt, die auf den von Lucene bereitgestellten Termvektoren arbeitet und die Termgewichte über einen Robertson Selection Value berechnet (HACKL ET AL. 2005).

#### **4.4 Vergleich zum System für WebCLEF 2005 und Ergebnisse aus den Experimenten**

Zu den für WebCLEF 2006 implementierten Veränderungen am System gehören:

- die erschöpfendere Indizierung des Korpus (durch Volltextindizierung und Bereinigung von XML-Fehlerquellen für den Parser)
- die Feld-spezifische Indexierung weiterer HTML-Elemente (‚emphasised‘)
- eine Blind Relevance Feedback Funktion
- Einsatz von Metadaten (der Domain-Filter)

Insgesamt ist ein großer Teil der in den Experimenten festgestellten Verbesserungen (vgl. Abb. 3) auf den Einsatz des Domain-Filters zurückzuführen. Die Verbesserungen basieren

also auf dem Gebrauch von Metadaten und weniger auf den Verbesserungen am System. Um den Einfluss des Domain-Filters bereinigt, ergeben sich im Vergleich zu den Ergebnissen der letzten Teilnahme 2005 leicht verbesserte Werte für den MRR und deutliche Verbesserungen für den Average Success @ 50, also den durchschnittlichen Erfolg innerhalb der ersten 50 Ergebnisse. Zum Vergleich wurde einer der Runs auch ohne den Domain-Filter angegeben (Abb. 3: UhiTwoDF). Der Vergleich der beiden Runs ergibt, dass der Domain-Filter eindeutig zur Verbesserung des MRR beiträgt. Es kann angenommen werden, dass der Domain-Filter vor allem dazu beiträgt, die Genauigkeit der Suche zu erhöhen, indem die Position der gesuchten Ergebnisse in der Ergebnisliste gesteigert wird, da viele Dokumente, die nicht gültig sein können, ausgeschlossen werden. Die Tatsache, dass mit dem neuen System auch ohne Filter mehr gültige Dokumente gefunden werden (abzulesen an dem Erfolg nach 50 Ergebnissen), ist wahrscheinlich auf die erschöpfendere Indexierung zurückzuführen, da weder der Einsatz weiterer HTML-Elemente noch das Blind Relevance Feedback zu signifikanten Steigerungen führten.

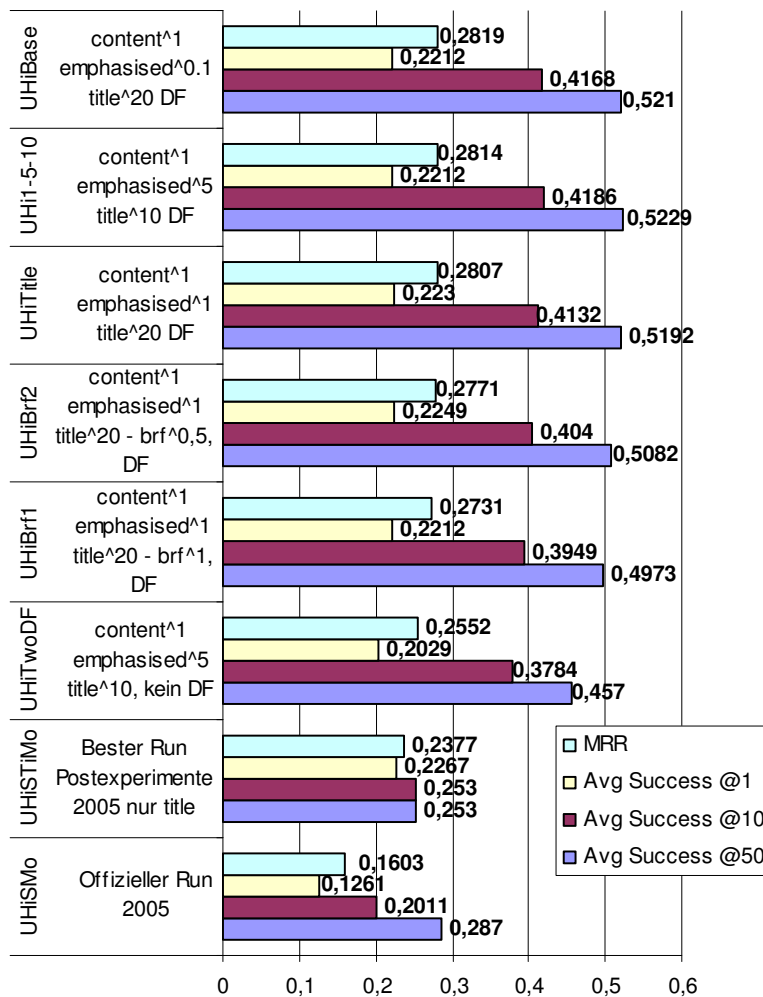


Abb. 3: Mixed Monolingual Task – WebCLEF 2005 und Postexperimente im Vergleich zu den diesjährigen Experimenten (DF=Domain-Filter)

Die Veränderungen am System führten auch im multilingualen Task zu Verbesserungen (vgl. Abb. 4). Hier zeigen sich die Verbesserungen bei der Indexierung, da keine weiteren Optimierungen für diesen Task durchgeführt wurden, und die Filterung nach Domains der Aufgabe entsprechend nicht in Frage kam. Besonders deutlich ist die Steigerung der insgesamt gefundenen Dokumente (Average Success @ 50).

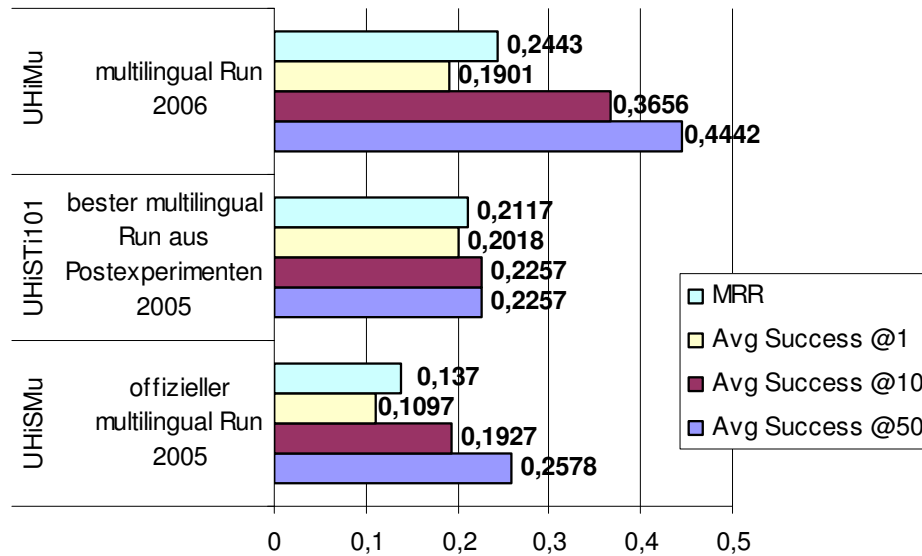


Abb. 4: Multilingual Task – WebCLEF 2005 und Postexperimente im Vergleich zu den diesjährigen Experimenten

## 5 Ergebnisse WebCLEF 2006

### 5.1 Ergebnisse der Universität Hildesheim

Für die Teilnahme an WebCLEF 2006 wurden nach Experimenten mit den Topics des Vorjahres fünf Runs für den offiziellen Mixed Monolingual Task eingereicht sowie einer außer Konkurrenz für den multilingualen. Dabei wurde mit unterschiedlichen Gewichtungen der ‚emphasised‘ und ‚title‘ Felder sowie mit verschiedenen Parametern für BRF experimentiert (vgl. Tabelle 1). Bei allen Runs kam der Domainfilter zum Einsatz. Als Basis-Run wurde der beste Run aus den Experimenten gewählt.

Run	Details
UHiBase	Gewichte: content^1 emphasised^0.1 title^20, Domainfilter
UHiTitle	Gewichte: content^1 emphasised^1 title^20, Domainfilter
UHi1-5-10	Gewichte: content^1 emphasised^5 title^10, Domainfilter
UHiBrf1	Gewichte: content^1 emphasised^1 title^20 blind relevance feedback (Gewicht der Anfrageerweiterung: 1), Domainfilter
UHiBrf2	Gewicht: content^1 emphasised^1 title^20 blind relevance feedback (Gewicht der Anfrageerweiterung: 0.5), Domainfilter
UHiMu	(multilingual) content^1 emphasised^1 title^20 – Engl. Übersetzung^10

Tabelle 1: Parameter der eingereichten Runs der Universität Hildesheim für WebCLEF 2006

Beim Vergleich der Ergebnisse der einzelnen Runs fällt auf, dass diese relativ nah beieinander liegen (vgl. Tabelle 2). Die Variationen bei den Gewichtungen führten zu geringen Abweichungen. Dabei bestätigten sich die Erkenntnisse aus den Experimenten (vgl. Abschnitt 4: Vergleich zum System für WebCLEF 2005): Die starke Gewichtung des ‚title‘-Feldes brachte Vorteile, die des ‚emphasised‘-Feldes jedoch nicht. Dies reduziert die Wahrscheinlichkeit, dass die zusätzlich gewählten HTML-Elemente eine stärker diskriminierende Wirkung haben als sonstige Inhalte. Die Anfrageerweiterung mittels BRF brachte keine Vorteile, verschlechterte jedoch das Ergebnis nicht maßgeblich. Ob der Einsatz von BRF für die known-item Aufgabenstellung nützlich sein kann, ist somit noch offen. Die Verbesserungen gegenüber dem System aus WebCLEF 2005 lassen sich vor allem auf die Nutzung von Metadaten und die erschöpfendere Indexierung zurückführen.

Starke Unterschiede ergaben sich jedoch zwischen den unterschiedlichen Arten von Topics. So führten die automatisch erstellten Topics zu Ergebnissen, die weit unter denen der manuell erstellten Topics lagen (vgl. Tabelle 2). Dies kann daran liegen, dass die Topics teilweise problematisch waren (vgl. Abschnitt 2: Topics). Ein weiterer Grund dürfte aber auch die größere Sprachenvielfalt sein. Statt der elf Sprachen der manuellen Topics sind bei den automatischen erstellten potentiell alle Sprachen des Korpus enthalten, da bei der Erstellung alle 27 Domains genutzt wurden (BALOG ET AL. 2006a). Darauf weist auch der Unterschied innerhalb der manuell erstellten Topics. Die für 2006 neu hinzugekommen manuellen Topics, die weniger Sprachen (nur Niederländisch, Englisch, Deutsch, Ungarisch, Spanisch) umfassen als die Topics aus dem Vorjahr, führten zu besseren Ergebnissen. Die von der Art der Topics abhängigen Unterschiede in den Ergebnissen zeigten sich bei allen teilnehmenden Gruppen (vgl. Abb. 5).

	<i>alle Topics</i>		<i>automatisch erstellte Topics</i>		<i>manuell erstellte Topics</i>	
	MRR	Average success at 10	MRR	Average success at 10	MRR	Average success at 10
UHiBase	0,0795	0,1377	0,0346	0,0772	0,3076	0,4451
UHiTitle	0,0724	0,1253	0,0264	0,0630	0,3061	0,4420
UHi1-5-10	0,0718	0,1233	0,0242	0,0574	0,3134	0,4577
UHiBrf1	0,0677	0,1104	0,0220	0,0475	0,3000	0,4295
UHiBrf2	0,0676	0,1124	0,0221	0,0500	0,2989	0,4295
UHiMu	0,0489	0,0758	0,0083	0,0154	0,2553	0,3824

Tabelle 2: Ergebnisse Universität Hildesheim WebCLEF 2006 (ursprüngliche Topicauswahl), beste Ergebnisse grau hinterlegt

## 5.2 Überblick über die Teilnehmer

Aufgrund der Probleme mit den automatisch erstellten Runs wurde von den Veranstaltern im Nachhinein eine neue Topicauswahl erstellt, welche um alle Topics beschnitten wurde, in denen keine der Gruppen das korrekte Ergebnis liefern konnte. Die Ergebnisse wurden damit neu berechnet (Abb. 5). Aufgrund der zahlenmäßigen Dominanz der automatischen Topics und der mit diesen Topics verbundenen Probleme wurde zum Vergleich zusätzlich der Durchschnitt aus den beiden Ergebnissen für die automatischen und die manuellen Topics angegeben.

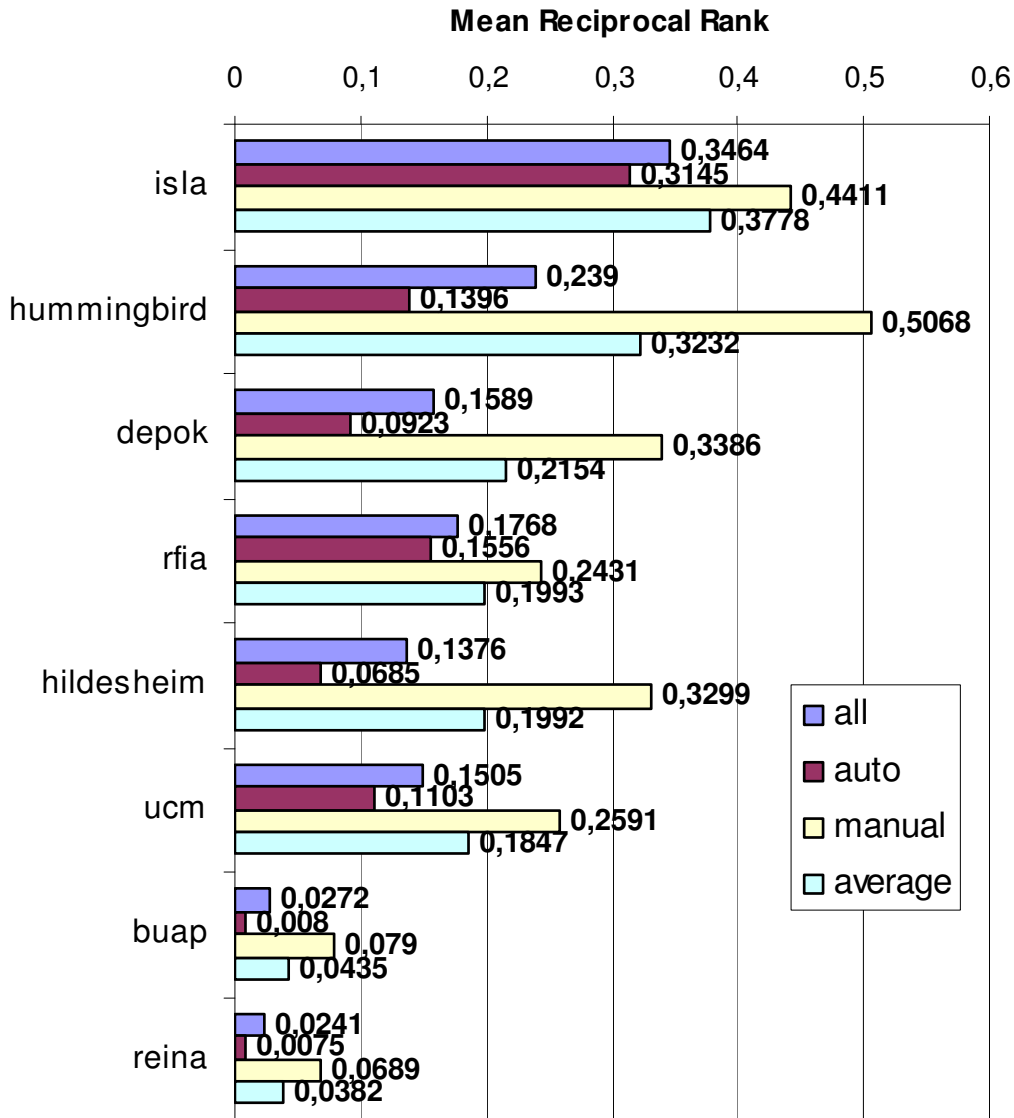


Abb. 5: Beste Runs der Teilnehmer in WebCLEF 2006 (MRR) mit der eingeschränkten Topicauswahl, *average* ist der Durchschnitt aus den Ergebnissen für die manuellen und automatischen Topics (BALOG ET AL. 2006)

Die Ergebnisse von zwei der acht teilnehmenden Gruppen setzen sich sowohl nach Durchschnittswert als auch nach absoluten Ergebnissen von den anderen ab (vgl. Abb. 5). Dies sind die Universität Amsterdam und die Firma Hummingbird. Die Ergebnisse von Hummingbird gehörten bereits bei WebCLEF 2005 zu den besten. Die vier folgenden Gruppen, u.a. die Universität Hildesheim, zeigen untereinander vergleichbare Ergebnisse, welche aber deutlich hinter denen der beiden ersten Gruppen zurückfallen. Der Teilnehmer mit den besten Ergebnissen im Mixed Monolingual Track des Vorjahres, die Universität Glasgow, nahm in diesem Jahr nicht teil.

Das System der Universität Amsterdam ist auf Basis von Lucene implementiert. Es wird kein Stemming vorgenommen. Die einzige sprachspezifische Maßnahme ist der Umgang mit griechischen und russischen Zeichenkodierungen. Es wurde ein Index für den kompletten Dokumentinhalt und einer für die ‚title‘-Elemente erstellt und die Ergebnisse kombiniert, wobei mit verschiedenen Fusionsverfahren experimentiert wurde. (BALOG & DE RIJKE 2006)

Das System der Firma Hummingbird experimentierte mit der Gewichtung des HTML-Titel Feldes und der Analyse der URLs. Phrasen aus der Anfrage, die im Titel auftauchten, und Überschriften wurden höher gewichtet. Dies brachte positive Resultate, allerdings nur bei den manuell erstellten Topics. Stemming wurde nicht eingesetzt. (TOMLINSON 2006)

## **6 Zusammenfassung und Ausblick**

Für die zweite Teilnahme am CLEF Web Track sollte das bestehende System durch die Indexierung mit verschiedenen Feldern unter Verwendung von HTML-Elementen verbessert werden. Die Verwendung des HTML-Titel Elementes erwies sich auch diesmal als vorteilhaft. Weitere Experimente mit der Feld-spezifischen Indexierung wären in Bezug auf die Gewichtung der Felder und dem Zusammenspiel mit der Rankingformel von Lucene sinnvoll.

Weiterhin konnte eine Blind Relevance Feedback Funktion implementiert werden. Die Verwendung von BRF hat allerdings nicht zu Verbesserungen der Retrievalergebnisse geführt. Es ist fraglich, ob die Anwendung im Kontext der known-item Suche sinnvoll ist. Allerdings bleibt auch hier Raum für Experimente mit wichtigen Parametern, wie der Verwendung verschiedener Felder und der Gewichtung der Anfrageerweiterung unabhängig von der Gewichtung der ursprünglichen Anfrage.

Die Wirksamkeit des gewählten sprachunabhängige Indexierungsansatzes bestätigte sich sowohl durch die von uns durchgeführten Experimente, als auch dadurch, dass die bei WebCLEF 2006 besten Ergebnisse mit ganz oder weitgehend sprachunabhängigen Methoden erzielt wurden. Weiterhin konnten Erfahrungen mit der Implementierung eines Domain-Filters in Lucene gesammelt werden. Insgesamt zeigte das System der Universität Hildesheim geringfügige Verbesserungen im Vergleich zu 2005.

Zur weiteren Verbesserung des Systems sollte der Umgang mit den verschiedenen Zeichenkodierungen während der Vorverarbeitung verbessert werden. In zukünftigen Experimenten sollen außerdem qualitätsbasierte Bewertungsmethoden eingeführt werden, bei denen Informationen zum Dokumentlayout mit in die Bewertung der Dokumente eingehen (MANDL 2006). Auch hierfür ist eine verbesserte Vorverarbeitung des EuroGOV-Korpus notwendig, um während der Indexierung einen direkten Zugriff auf die HTML-Struktur der Dokumente zu ermöglichen.

## Literaturverzeichnis

- Balog, Krisztian; Azzopardi, Leif; Kamps, Jaap; de Rijke, Maarten (2006): Overview of WebCLEF 2006. [http://www.clef-campaign.org/2006/working\\_notes/workingnotes2006/balogOCLEF2006.pdf](http://www.clef-campaign.org/2006/working_notes/workingnotes2006/balogOCLEF2006.pdf)  
Verifiziert 15.9.2006
- Balog, Krisztian; de Rijke, Maarten (2006): The University of Amsterdam at WebCLEF 2006 [http://www.clef-campaign.org/2006/working\\_notes/workingnotes2006/balogCLEF2006.pdf](http://www.clef-campaign.org/2006/working_notes/workingnotes2006/balogCLEF2006.pdf)  
Verifiziert 15.9.2006
- Chen, L.; Ye, S.; Li, X. (2006): Template Detection for Large Scale Search Engines. In: Proceedings ACM Symposium on Applied Computing ACM Press. S. 1094-1098.
- Clarke, Charles L. A.; Scholer, Falk, Soboro, Ian (2005): The TREC 2005 Terabyte Track. <http://trec.nist.gov/pubs/trec14/papers/TERABYTE.OVERVIEW.pdf> Verifiziert 15.9.2006
- Gonzalo, Julio; Peters, Carol (2005): The Impact of Evaluation on Multilingual Text Retrieval. In: SIGIR '05: Proceedings of the 28<sup>th</sup> annual international ACM SIGIR conference on Research and development in information retrieval. ACM Press, New York 2005 S. 603-604
- Grossman, David A.; Frieder, Ophir (1998): Information retrieval : algorithms and heuristics. Kluwer, Boston
- Hackl, René; Mandl, Thomas; Womser-Hacker, Christa (2005): Ad-hoc Mono- and Multilingual Retrieval Experiments at the University of Hildesheim. In: Working Notes of the 6<sup>th</sup> Workshop of the Cross-Language Evaluation Forum, CLEF 2005. Sept. 2005, Wien.
- Hawking, David; Craswell, Nick (2004): Very Large Scale Retrieval and Web Search (Preprint version). [http://es.csiro.au/pubs/trecbook\\_for\\_website.pdf](http://es.csiro.au/pubs/trecbook_for_website.pdf) Verifiziert 15.9.2006
- Heuwing, Ben; Mandl, Thomas; Strötgen, Robert (2006): Multilingual Web Retrieval Experiments with Field Specific Indexing Strategies for CLEF 2006 at the University of Hildesheim. [http://www.clef-campaign.org/2006/working\\_notes/workingnotes2006/heuwingCLEF2006.pdf](http://www.clef-campaign.org/2006/working_notes/workingnotes2006/heuwingCLEF2006.pdf) Verifiziert 19.9.2006
- Jensen, Niels (2005a) Web Information Retrieval am Beispiel des WEB-GOV Korpus. Magisterarbeit Internationales Informationsmanagement, Universität Hildesheim.
- Jensen, Niels (2005b) Mehrsprachiges Information Retrieval mit einem WEB-Korpus. In: Mandl, Thomas; Womser-Hacker, Christa (Hrsg.) (2006): Effektive Information Retrieval Verfahren in Theorie und Praxis: Ausgewählte und erweiterte Beiträge des Vierten Hildesheimer Evaluierungs- und Retrievalworkshop (HIER 2005) Hildesheim, 20.7.2005. Konstanz: Universitätsverlag [Schriften zur Informationswissenschaft 45] S. 235-244.
- Jensen, Niels; Hackl, René; Mandl, Thomas; Strötgen, Robert (2006): Web Retrieval Experiments with the EuroGOV Corpus at the University of Hildesheim. In: Accessing Multilingual Information Repositories: 6<sup>th</sup> Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, Revised Selected Papers. Springer [LNCS 4022]
- Macdonald, Craig; Plachouras, Vassilis; He, Ben; Lioma, Christina; Ounis, Iadh (2005): University of Glasgow at WebCLEF 2005: Experiments in per-field normalisation and language specific stemming. [http://www.clef-campaign.org/2005/working\\_notes/workingnotes2005/macdonald05.pdf](http://www.clef-campaign.org/2005/working_notes/workingnotes2005/macdonald05.pdf)  
Verifiziert 15.9.2006
- Mandl, Thomas (2006): Implementation and Evaluation of a Quality Based Search Engine. In: Proceedings of the 17<sup>th</sup> ACM Conference on Hypertext and Hypermedia (HT '06) Odense, Denmark, 22.-25 August. ACM Press.
- Metzler, Donald; Strohman, Trevor; Zhou, Yun; Croft, W.B. (2005): Indri at TREC 2005: Terabyte Track. <http://trec.nist.gov/pubs/trec14/papers/umass-tera.pdf> Verifiziert am 15.9.2006
- Mishne, Gilda; de Rijke, Maarten (2005): Boosting Web Retrieval Through Query Operations. In: Advances in Information Retrieval: 27<sup>th</sup> European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 2005 Proceedings. Springer [LCNS 3408] S. 502-516
- Santos, Diana; de Rijke, Maarten (2005): WebCLEF 2005 workshop report. <http://ilps.science.uva.nl/WebCLEF/WebCLEF2005/Report/index.html> Verifiziert 8.9.2006
- Sigurbjörnsson, Börkur; Kamps, Jaap; de Rijke, Maarten (2005a): Blueprint of a Cross-Lingual Web Retrieval Collection. In: Journal of Digital Information Management, vol. 3 (1) S. 9-13.

- Sigurbjörnsson, Börkur; Kamps, Jaap; de Rijke, Maarten (2005b): Overview of WebCLEF 2005. [http://www.clef-campaign.org/2005/working\\_notes/workingnotes2005/sigurbjornsson05.pdf](http://www.clef-campaign.org/2005/working_notes/workingnotes2005/sigurbjornsson05.pdf)  
Verifiziert 15.9.2006
- Tomlinson, Stephen (2005): European Web Retrieval Experiments with Hummingbird SearchServerTM at CLEF 2005. [http://www.clef-campaign.org/2005/working\\_notes/workingnotes2005/tomlinson05.pdf](http://www.clef-campaign.org/2005/working_notes/workingnotes2005/tomlinson05.pdf)  
Verifiziert 15.9.2006
- Tomlinson, Stephen (2006): European Web Retrieval Experiments at WebCLEF 2006 [http://www.clef-campaign.org/2006/working\\_notes/workingnotes2006/tomlinsonWebCLEF2006.pdf](http://www.clef-campaign.org/2006/working_notes/workingnotes2006/tomlinsonWebCLEF2006.pdf)  
Verifiziert 15.9.2006
- Zhao, Le; Ceng, Rongwei; Ma, Shaoping; Jin, Yijiang; Zhang, Min (2005): THUIR at TREC 2005 Terabyte Track. <http://trec.nist.gov/pubs/trec14/papers/tsinghuau-ma.tera.pdf> Verifiziert 15.9.2006