Fachbereich III Informations- und Kommunikationswissenschaften
Institut für Angewandte Sprachwissenschaft

Magisterarbeit

INTERNATIONALES INFORMATIONSMANAGEMENT

Multilinguales Webretrieval am Beispiel des EuroGOV Korpus

eingereicht von Niels Jensen (186148)
nielsolejensen@gmail.com

Erstgutachten Dr. Thomas Mandl
Zweitgutachten Prof. Dr. Christa Womser-Hacker

Hildesheim, im September 2005

Inhaltsverzeichnis

Zusammenfassung

Zusammenfassung Die vorliegende Arbeit befasst sich mit Multilingualem Webretrieval. Am Anfang werden verschiedene Retrieval Evaluation Initiativen beschrieben. Das Hauptaugenmerk liegt auf dem Cross Language Evaluation Forum (CLEF), mit dem in diesem Jahr gestarteten Web Track WebCLEF. Der Web Track WebCLEF ist in Anlehnung an die Web Tracks der TREC und NTCIR Initiativen entwickelt worden. Der entscheidene Unterschied zu diesen Tracks ist der multilinguale Ansatz, der im WebCLEF Track verfolgt wird. Allen Teilnehmern wurde eine Testkollektion bestehend aus dem EuroGOV Korpus, 547 Topics und der dazugehörigen Relevanzbewertung zur Verfügung gestellt. Neben dem Vergleich des WebCLEF Tracks zu den anderen Initiativen steht die aktive Teilnahme im Mittelpunkt dieser Arbeit. Aufgezeigt werden die Rahmenbedingungen für die Teilnahme, Eigenschaften der Testkollektion, die Vorgehensweise und Ziele des Institutes für Angewandte Sprachwissenschaften (IFAS) der Universität Hildesheim, Erfahrungen und Schwierigkeiten beim Generieren der verschiedenen Indizes, der eigentliche Retrievalprozess und die dazugehörigen Ergebnisse aller Experimente. Als Abschluss dieser Arbeit wird die Teilnahme ausgewertet, Verbesserungen zum eigentlichen Web Track und ein Ausblick für die erneute Teilnahme am WebCLEF Track 2006 dargestellt.

Schlüsselbegriffe: CLEF, TREC, NTCIR, SEWM-2004, XML, Datapreprocessing, Topicentwicklung, Lucene Index, multilinguale Stoppwortliste, Lucene Query Engine und multilinguales Web Retrieval.

Abstract This thesis deals with multilingual web retrieval. The main focus is on the active participation of the University of Hildesheim at the first web track of the Cross Language Evaluation Forum (CLEF). The web track WebCLEF originated from the two evaluation initiatives TREC and NTCIR. The main distinction between WebCLEF and the other initiatives is the multilingual approach. Every participating group received a

generated test collection for this web track. This collection consists of a multilingual web corpus (EuroGOV), 547 topics and the relevance assessments for those topics. This thesis starts with a description of the WebCLEF track and its belonging to the CLEF initiative. In addition three different web tracks will be specified and compared to the WebCLEF track. The main part of this thesis covers up the procedure of the experiments with a multilingual index at the University of Hildesheim together with all the experiences that have been gained from the participation. Finally, the experiences and the results of the submitted runs as well as the post experiments will be evaluated, so that suggestions for improving the track and the next participation in 2006 can be achieved.

Keywords: CLEF, TREC, NTCIR, SEWM-2004, XML, data preprocessing, topic development, Lucene index, multilingual stopwordlist, Lucene Query Engine and multilingual web retrieval.

Einleitung

Information Retrieval ist in den letzten Jahren zu einem wichtigen Bestandteil unseres Lebens geworden. Internetsuchmaschinen, digitale Bibliothekskataloge, Desktopsearch Programme und Online Auktionshäuser z.B. sind in den Tagesablauf eines PC Users fest integriert. Der Themenbereich Information Retrieval mit seinen benachbarten Disziplinen beschäftigt sich mit der Problemstellung der Informationsflut und wie diese für den User geordnet, gelenkt und gefiltert werden kann. Der Ursprung dieser Wissenschaft kann bis Juli 1945 zurückgeführt werden. Damals veröffentlichte Dr. Vannevar Bush im "Atlantic Monthly" den Artikel *As We May Think* [Bus45]. In diesem Artikel beschreibt Bush die Notwendigkeit bzw. Vorteile, alle Bücher einer Bibliothek auf kleinstem Raum abzuspeichern, um sie transportabel und schnell verfügbar zu machen. Nachdem Bush in diesem Artikel über verschiedene Möglichkeiten des effizienten Speicherns von Büchern und Artikeln geschrieben hatte, formulierte er die bekannte Vision des *Memex* Systems.

> *Consider a future device for individual use, which is a sort of mechanized private file and library. It needs a name, and, to coin one at random, "memex" will do. A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory. [Bus45]*

Memex sollte laut Bush Wissenschaftler in die Lage versetzen, auf Knopfdruck bestimmte Textpassagen, Grafiken oder Zeitungsartikel aufzurufen, um den Inhalt lesen und verarbeiten zu können. Die Retrievalkomponente wurde am Beispiel von Memex allerdings sehr naiv behandelt.

> *There is, of course, provision for consultation of the record by the usual scheme of indexing. If the user wishes to consult a certain book, he taps its code on the keyboard, and the title page of the book promptly appears before him, projected onto one of his viewing positions. Frequently-used codes are mnemonic, so*

> *that he seldom consults his code book; but when he does, a single tap of a key projects it for his use. [Bus45]*

Mit dieser Beschreibung wird die Grenze von Bush's Vision deutlich. Bush's Retrievalprozess setzte ein komplettes Wissen über die abgespeicherten Dokumente voraus. Benutzt werden konnte Memex nur mit präzisen Suchanfragen, die keine Fehler beinhalten durften. Trotz dieser einfachen Sichtweise des Wiederauffindens von Dokumenten und den damit verbundenen Informationen, wird dieses von Bush beschriebene Szenario als wegweisend für die Informationswissenschaft beschrieben. In dem Artikel *The Seven Ages of Information Retrieval* [Les96] stellt Michael Lesk die These auf, dass Bush mit dem Formulieren der Memex-Idee einer der Gründungsväter des Information Retrieval sei. Die Idee, das gesamte auf der Welt befindliche Wissen in einen Schreibtisch zu speichern und jederzeit für jedermann zugänglich zu machen, scheint mit einigen Abwandlungen in der heutigen Zeit erfüllt zu sein. Lesk sagt:

> *Rereading the original Bush paper, and looking at it from today's standpoint, the hardware seems mostly out of date, but the software goals have not been achieved. [Les96]*

Lesk beschreibt in seinem Artikel die Entwicklung des Information Retrievals in sieben Phasen, die an Shakespeare's *Seven Ages of Man* [Sha99] angelehnt sind. Aktuell befindet sich die Entwicklung in der Phase der Erfüllung (eng. Fulfillment), der vorletzten Phase. Information Retrieval ist nun ca. 65 Jahre alt, die Hardware Probleme sind zum größten Teil gelöst und die neuen Herausforderungen des Information Retrievals sind die Heterogenität von Daten, extrem große Datenmengen und unwissenden Usern. In diesem Umfeld versucht die Wissenschaft des Web Retrievals Antworten auf diese Probleme zu liefern. Das Internet, die eigentliche Umsetzung des Memex Szenario, speichert Wissen in unüberschaubarem Umfang ab, die Menge wächst täglich und ein Ende bzw. kontrollierbares Ansteigen dieser Entwicklung ist nicht abzusehen. Sogenannte Suchmaschinen bilden eine Art Leuchtturm im großen Meer der Informationen und versuchen User so effektiv wie möglich zu den benötigten Informationen zu bringen. Um diese Effektivität zu steigern und die relevanten Dokumente aus dem fast unüberschaubaren Sumpf heraus zu filtern, beschäftigen sich viele Forschungsprojekte mit dem Thema Information- bzw. Web Retrieval. Neben den Problemen der Datenflut stellt die Vielsprachlichkeit der User im Internet eine weitere große Herausforderung dar.
Die Effektivität eines Retrievalsystems auf Basis einer mehrsprachigen Datenmenge zu verbessern, ist eines der Hauptziele des europäischen *Cross Language Evaluation Forums*

(CLEF). Die Universität Hildesheim nimmt schon seit mehreren Jahren an den verschiedenen Tracks der CLEF Initiative teil.

Zielsetzung dieser Arbeit Ziel dieser Arbeit ist die aktive Teilnahme an dem ersten Web Track der CLEF Initiative. Dabei sollen Erfahrungen aus vergangen CLEF Teilnahmen der Universität Hildesheim in den Bereichen des Data Preprocessing (Datenkonvertierung), Indexieren von verschiedenen Korpora und dem Durchführen von Retrievalprozessen mit einfließen. Da der Zeitrahmen bis zur offiziellen Abgabe der Ergebnislisten sehr knapp berechnet war (2 1/2 Monate von Beginn der Arbeit bis zur Abgabe), war das vorgegebene Ziel einen technisch robusten Prototypen aufzubauen. Dies bedeutet, dass in erster Linie das Aufstellen eines laufenden Systems im Vordergrund stand, sodass die gewonnenen Erkenntnisse in der nächsten Teilnahme im Jahr 2006 verwendet werden können. Diese Arbeit gliedert sich in fünf Kapitel.

Kapitel 1 beschreibt das Cross Language Evaluation Forum und seine unterschiedlichen Tracks. Der WebCLEF Track und die dazugehörende Testkollektion werden besonders beleuchtet.

Kapitel 2 versucht drei weitere Retrieval Evaluationsinitiativen mit ihren dazugehörigen Web Tracks vorzustellen, um eine Vergleichsbasis für den WebCLEF Track zu schaffen. Am Ende von diesem Kapitel werden alle Tracks mit einander verglichen und vorgestellt. Eine ausgewählte Anzahl von teilnehmenden Systemen werden anhand von Ergebnissen und Eigenschaften analysiert und gegenübergestellt. Die Einzigartigkeit der WebCLEF Testkollektion wird in diesem Teil herausgearbeitet.

Kapitel 3 gibt einen kurzen Überblick über die Lucene 1.4 Klassenbibliothek, da diese Bibliothek die Softwaregrundlage aller WebCLEF Experimente der Universität Hildesheim bildet.

Kapitel 4 umfasst die Vorgehensweise und Erfahrungen, die während der WebCLEF Experimente gesammelt wurden. Des Weiteren werden die Ergebnisse der offiziell eingereichten Runs ausgewertet und eine ausführliche Reihe an Postexperimenten durchgeführt und ebenfalls ausgewertet.

Kapitel 5 wertet die Erfahrungen, die während der Experimente gesammelt worden sind aus und bereitet diese für die nächste Teilnahme im Jahr 2006 auf. Außerdem beinhaltet dieses Kapitel Verbesserungsvorschläge zum WebCLEF Track in den Bereichen der Topicentwicklung und der Relevanzbewertung.

Kapitel 1

WebCLEF: Der Cross-Linguale Web Track der CLEF Initiative

> *In the light of the linguistic diversity of the European web and European searchers, a cross-lingual web retrieval task, called WebCLEF, was launched at CLEF2005.[Sig05b]*

Mit dem kontinuierlichen Wachsen der Europäischen Union entstehen immer neue Anforderungen an den ständigen Austausch von Informationen für jeden Sprachraum. Mit dieser Herausforderung der Cross-Lingualität des Internets möchte sich der Web Track der CLEF Initiative befassen. Ziel ist es, Informationen sprachunabhängig für jeden zugänglich zu machen. Dies bedeutet, dass ein zu entwickelndes cross-linguales Retrievalsystem alle Sprachen weitestgehend abdecken muss und ebenfalls in der Lage sein sollte, sich auf die sprachlichen Fähigkeiten des Users einzustellen. Die sprachliche Vielfalt Europas und die weitverbreitete Multilingualität der Europäer bilden einen guten Testraum, um sich der Aufgabenstellung dieses Web Tracks zu stellen.

> *The purpose of the web track is to form a quality test collection for World Wide Web searching. To build a test collection, we need to use a static collection of documents since web sites are continually changing. We will use a collection of web pages from various European government sites (covering a range of domains and languages) collected in October 2004 to January 2005.[Sig05d]*

1.1 Cross Language Evaluation Forum (CLEF)

Der WebCLEF Track ist einer von acht unterschiedlichen Retrievaltasks des europäischen Cross Language Evaluation Forum (CLEF).

> *The Cross-Language Evaluation Forum (CLEF) supports global digital library applications by (i) developing an infrastructure for the testing, tuning and evaluation of information retrieval systems operating on European languages in both monolingual and cross-language contexts, and (ii) creating test-suites of reusable data which can be employed by system developers for benchmarking purposes. [CLE05]*

Der erste CLEF Durchgang wurde im Jahr 2000 ins Leben gerufen. Seitdem wachsen die Teilnehmerzahl und der Organisationsaufwand stetig. In diesem Jahr beinhaltet CLEF mehrere unterschiedliche Retrievaltasks, die alle gemeinsam den cross-lingualen Ansatz auf unterschiedliche Art und Weise verfolgen.

Mono-, Bi- and Multilingual Document Retrieval on News Collections (Ad-Hoc) Der ad hoc Track bietet die Möglichkeit, multilinguale Retrievalsysteme anhand einer multilingualen Testkollektion zu testen. Dieser Track unterteilt sich in den Monolingual-, Bilingual- und Multilingual Retrieval Task. Der Multilingual Retrieval Task basiert auf der CLEF 2003 multilingual-8 Testkollektion. Hierbei kann jede im Korpus enthaltene Sprache als Ausgangssprache verwendet werden. Das Ziel dieses Tasks ist es, alle relevanten Dokumente unabhängig von der Sprache in der Kollektion zu finden. Für den ad hoc Track wurden 50 Topics zur Verfügung gestellt. [CLE05]

Mono- and Cross-Language IR on Structured Scientific Data (Domain-Specific) Der Domain-Specific Track stellt einen themenspezifischen Korpus zur Verfügung, um mittels dieses Korpus und der dazugehörigen Topics themenbegrenzt Retrievalsysteme zu prüfen. Als Korpus dient die GIRT-4 sozialwissenschaftliche Datenbank. In dieser Datenbank sind die Sprachen Deutsch und Englisch vertreten, wobei ein Parallelkorpus in russisch ebenfalls vorliegt. Mit diesen Vorgaben sind die zu verwendenden Sprachen ebenfalls festgelegt. Auch in diesem Task gibt es einen Monolingual-, einen Bilingual- und einen Multilingual Task. [CLE05]

Interactive Cross-Language IR (iCLEF) Der iCLEF Track beinhaltet die Herausforderung ein System zu bauen, dass es Usern ermöglicht, Informationen in Sprachen zu finden, die sie nicht beherrschen. Der Track teilt sich in die beiden Subtasks interactive Question Answering (iCLEF QA) Task und den interactive Image Retrieval Task. Der iCLEF QA Task soll Usern helfen, anhand einer Frage in der eigenen Sprache in fremdsprachlichen Kollektionen nach Antworten zu suchen. Der iCLEF Image Retrieval Task hingegen soll Usern helfen Bilder zu finden, die durch

fremdsprachliche Bildbeschreibungen (Captions) gekennzeichnet sind. Für den iCLEF QA Task wurde das Korpus des QA@CLEF Tasks verwendet, ebenso wie der iCLEF Image Retrieval Task das Korpus des ImageCLEF Tasks verwendet. [CLE05]

Multiple Language Question Answering (QA@CLEF) Der QA@CLEF Task testet Systeme anhand von ausformulierten Fragen, die mittels der QA Systeme formulierte Antworten zurückliefern müssen. Hierbei wird zwischen *factoid* (als Antwort werden reine Fakten erwartet) und *definition* (als Antwort wird eine Definition erwartet) Fragen unterschieden. Das QA@CLEF Korpus besteht aus Dokumenten in neun verschiedenen Sprachen. Für den Durchgang 2005 wurden 81 Fragen formuliert, 24 Gruppen haben teilgenommen und 67 Runs (beantwortete Fragen) wurden eingereicht. [CLE05]

Cross-Language Retrieval in Image Collections (ImageCLEF) Der ImageCLEF Task zielt darauf ab Systeme zu testen, die in der Lage sind, mittels Bildbeschreibungen (Captions) und image matching Techniken als Antwort auf Suchanfragen relevante Bilder zu liefern. Bilingual ad hoc retrieval, interactive search, medical image retrieval und automatic annotation Task for medical images sind die vier Subtasks des imageCLEF Tasks. Hierfür wurden die drei Testkollektionen St. Andrews University historical photographic collection, ImageCLEFmed collection und die IRMA Datenbank (10.000 medizinische Bilder) entwickelt bzw. zur Verfügung gestellt. [CLE05]

Cross-Language Spoken Document Retrieval (CL-SR) Ziel des CL-SR Tracks ist es, Retrieval Systeme für Audiodateien zu evaluieren. Diese Audiodateien bestehen aus Aufzeichnungen von Dialogen in englischer Sprache. Für diesen Track wurde ein Korpus von ca. 750h zusammengesetzt. 25 Topics in sechs verschiedenen Sprachen wurden entwickelt und den Teilnehmern zur Verfügung gestellt. [CLE05]

Cross-Language Geographical Retrieval (GeoCLEF) Der GeoCLEF Track konzentriert sich auf Systeme, die speziell geographische Begriffe verarbeiten können. Ziel eines GeoCLEF Systems ist es, Fragen zu Adressen, Karten, Orten und Ländern anhand von Retrievalmethoden aus dem Testkorpus zu beantworten. Für den aktuellen Durchgang wurden nur Texte in Englisch und Deutsch zusammengestellt. Alle Texte sind Zeitungsartikel aus den Jahren 1994 und 1995 und beinhalten geographische Spezifikationen. [CLE05]

Multilingual Web Track (WebCLEF) Der WebCLEF Track wurde 2005 zum ersten Mal durchgeführt und hält sich an die Beispiele des Web Tracks der TREC und

NTCIR Initiativen. Für diesen Track musste eine neue Testkollektion bestehend aus Korpus, Topics und ihren Bewertungen sowie angepassten Tasks entwickelt werden. Alle weiteren Einzelheiten zum WebCLEF Track werden in diesem Kapitel näher beschrieben. [CLE05]

1.2 Rahmenbedingungen und Ablauf des Web Tracks

Der Web Track WebCLEF wurde federführend von der Universität Amsterdam organisiert und betreut. Die Universität Amsterdam war in dieser Aufgabe verantwortlich für die Erstellung der Richtlinien, Teilnehmerkontrolle, Aufbau der Testkollektion, Koordination der Topicentwicklung, Zusammenführen der Relevanzbewertung der Topics sowie die Pflege der Mailingliste. Von den Teilnehmern wurde erwartet, eine bestimmte Anzahl von Topics in der zugewiesenen Sprache zu entwickeln, sowie die mehr oder weniger aktive Teilnahme am eigentlichen Web Track. Insgesamt haben 15 Gruppen teilgenommen. Der größte Teil stammte aus Europa, ca. ein Drittel der Teilnehmer aus anderen Teilen der Welt. Unter den Teilnehmern waren auch Unternehmen wie z.B. Hummingbird und Metacarta vertreten. Für den Web Track gab es folgende Meilensteine:

- Oktober 2004 bis Dezember 2004: erzeugen des EuroGOV Korpus
- Januar 2005: aufschließendes Festlegen der verschiedenen Disziplinen (Tasks)
- 20. Januar 2005: Freigabe des EuroGOV Korpus
- 01. bis 31. Mai 2005: Topic Entwicklung
- 15. Mai 2005: Freigabe der 547 Topics
- 15. Juni 2005: ursprünglicher Termin zur Abgabe der Run-Ergebnislisten
- 20. Juni 2005: verlängerter Termin zur Abgabe der Run-Ergebnislisten
- 15. Juli 2005: Veröffentlichung der Relevanzbewertung in Form von QREL Dateien
- 21. August 2005: Abgabe der zu WebCLEF gehörenden Papers
- 21. bis 23. September 2005: CLEF Workshop in Wien

Die Organisatoren stellten eine sehr umfangreiche Testkollektion für die Teilnehmer zur Verfügung. Die Kollektion besteht aus dem EuroGOV Korpus, einer Datei die alle doppelten Seiten dieses Korpus enthält, eine Liste der identifizierten Sprache pro Dokument,

die WebCLEF Topics im XML Format und - nach Abgabe der Runs - eine vollständige Sammlung von QREL Dateien mit Auswertungsskript. Um den gesamten Umfang der Testkollektion zu erhalten, musste jede teilnehmende Gruppe und individuell beteiligte Personen dieser Gruppen eine Vertraulichkeitsvereinbarung zur Wahrung des Copyrights und dem angemessenen Gebrauch der Daten unterschreiben.

1.3 Das EuroGOV Korpus

Um die Multilinguale Vielfalt des WebCLEF Retrieval Tasks zu ermöglichen, musste eine neue Testkollektion erstellt werden. Diese Testkollektion, bestehend aus Korpus, Topics und deren Relevanzbewertung, wurde unter dem Namen EuroGOV zusammengeführt und den WebCLEF Teilnehmern zur Verfügung gestellt. Wie der Name der Testkollektion schon aussagt, basiert das Korpus auf den Webseiten des Portals der Europäischen Union, den Regierungsseiten von EU Mitgliedstaaten und Russland. Man beschränkte sich auf die besagten Regierungsseiten, um die Urheberrechte anderer Webseiten bzw. Domänen während des Generierens bzw. Sammelns des Korpus nicht zu verletzen. Das Korpus beinhaltet 3.6 Mio Webseiten in mehr als 20 verschiedenen Sprachen.

Main Domains			Additional Domains		
Domain	Pages	Size (gz)	Domain	Pages	Size(gz)
.cz	324.496	519M	.at	10.065	24M
.de	444.496	1,1G	.be	69.011	115m
.es	35.168	298M	.cy	1.972	7,9M
.eu.int	374.484	1,9G	.dk	2.144	5,4M
.fi	661.559	1,3G	.ee	16.768	44M
.fr	156.450	545M	.gr	303	416K
.hu	330.822	1,5G	.ie	12.754	32M
.it	89.836	324M	.lt	10.756	8,8M
.nl	149.949	434M	.lu	8.521	33M
.pt	147.445	753M	.lv	317.404	675M
.ru	104.659	479M	.mt	13.991	57M
.se	102.457	155M	.pl	66.885	330M
.uk	66.345	331M	.si	12.434	27M
			.sk	58.020	128M

Tab. 1.1: Anteilige Verteilung der versch. Domänen im EuroGOV Korpus [Sig05b]

EuroGOV wurde in den Monaten Oktober 2004 bis Januar 2005 von der Universität Amsterdam aufgebaut. Der Prozess des Crawlens zielte darauf hinaus, die Hauptregie-

rungsportale und die wichtigsten Ministerien der einzelnen Mitgliedsstaaten und Russlands zu erfassen. In Folge des Umfangs der Topleveldomänen war eine Vollständigkeit des Korpus nur bedingt zu gewährleisten.

> *These differences in domain naming traditions will make it difficult to guarantee completeness of some governments. As a result, what we should realistically aim for is that EuroGOV contains the fairly complete content of the main government portals, and the main ministries. [Sig05b]*

In der Zusammensetzung des EuroGOV Korpus sind Besonderheiten entstanden, die sich von anderen Web Datenkorpora unterscheiden. Diese Besonderheiten sollen im folgenden Abschnitt näher beleuchtet werden.

1.3.1 Domänen und Sprachen

Die EuroGOV Kollektion beherbergt 27 Topleveldomänen, von denen 13 als Hauptdomänen behandelt werden (Tab. 1.1). Diese 13 Domänen wurden aus den Bedürfnissen und Anforderungen der CLEF 2005 Initiative entwickelt. Die Hauptdomänen beanspruchen für sich einen sehr hohen Grad der Vollständigkeit. Die 14 zusätzlichen Domänen hingegen decken nur einen repräsentativen Teil ab. Die Sprachen dieser 14 Domänen fließen jedoch auch in die 13 Hauptdomänen ein, sodass der Ansatz, möglichst viele Sprachen mittels des Korpus abzudecken, gewährleistet ist. Als Begründung kann hierfür die .eu.int Domäne verwendet werden, da diese Site die 20 offiziellen Sprachen der Europäischen Union beherbergt. Des Weiteren beinhaltet das EuroGOV Korpus mehr Sprachen bzw. Länderdomänen als der WebCLEF 2005 Evaluations Rahmen abdeckt.

> *The EuroGOV collection features more languages and countries than are being used in the WebCLEF 2005 evaluation tasks. We made a deliberate choice to go for this extended list, of countries and domains. This will facilitate future task extensions for cross-lingual web retrieval, or re-use of the collection for other purposes. We feel that this reflects the natural situation when building a 'European' search engine. [Sig05b]*

Das gesamte EuroGOV Korpus beinhaltet 3.589.501 Seiten, die in 157 Dateien zu maximal 25.000 Seiten / Dokumenten zusammengefasst sind. Dieses Korpus benötigt einen Speichervolumen von 82 GB (komprimiert ca. 11 GB).

Die kleinste Testmenge lieferte die griechische Domäne (.gr) mit 303 Seiten / Dokumenten, wobei diese Menge ausreichte, um ein repräsentatives Korpus zu erstellen. Die kleinste Teilmenge für die 13 Hauptdomänen liefert Spanien (.es) mit 35.168 Seiten.

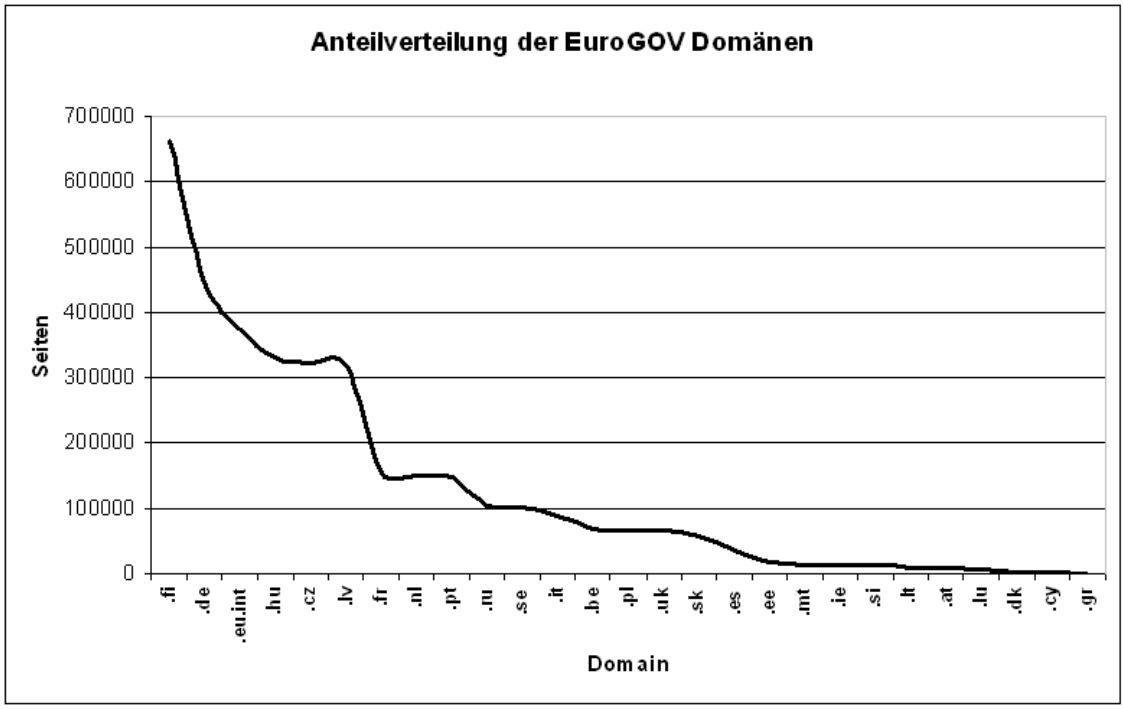


Abb. 1.1: Anzahl der Dokumente pro Topleveldomäne [Sig05b]

> *It is unclear, at this point, to what extent the varying numbers of pages per domain are a result of the available web content, different link structure of different governmental sites, or of our particular choices in crawler software or seed points. [Sig05b]*

Wichtig für die gesamte Kollektion ist, dass die vorhandenen Domänen bzw. die Ursprungsdomänen eines Dokuments nicht automatisch auf die Sprache schließen lassen. Domänen wie Frankreich (.fr), Deutschland (.de) oder Großbritannien (.uk) haben im Durchschnitt einen 97%igen Anteil der eigenen Landessprachen auf ihren Internetseiten vertreten. Länderdomänen wie z.B. Belgien (.be) oder Finnland (.fi) hingegen teilen sich im Sprachanteil ausgeglichener auf. (Tab. 1.2)

Als Folge dieser Erkenntnis wird die Abweichung von der Anzahl der einzelnen Länderdomänen zur gegebenen Landessprache deutlich. Der Gesamttrend ist hierbei abweichend von der .eu.int Domäne zu sehen. Überraschend ist die deutliche Dominanz der finnisch-ungarischen Sprachfamilie. (Tab. 1.3)

Domäne .de		Domäne .fr		Domäne .uk	
Sprache	Prozent	Sprache	Prozent	Sprache	Prozent
deutsch	97,7%	französisch	94,25%	englisch	99,05%
englisch	1,37%	deutsch	2,49%		
französisch	0,74%	englisch	2,24%		
		spanisch	0,81%		
Domäne .be		Domäne .fi			
Sprache	Prozent	Sprache	Prozent		
französisch	36,78%	finnisch	81.15%		
holländisch	24.32%	schwedisch	11,52%		
deutsch	21,61%	englisch	7,26%		
englisch	16,74%				

Tab. 1.2: Sprachverteilung der Domänen .de, .fr, .uk, .be und .fi [Sig05b]

1.3.2 Aufbau und Beispiel eines EuroGOV Dokuments

Der Aufbau eines EuroGOV Dokuments ist sehr strukturiert und wohlgeformtem XML ähnlich. Die Struktur besteht aus folgenden Elementen und den dazu gehörigen Attributen (Abb. 1.2):

```
<EuroGOV:bin domain="se" id="001">
<EuroGOV:doc
 url="http://www.regeringen.se/"
 id="Ese-001-35"
 md5="659b462005b40f04bde5946b2beaad71"
 fetchDate="Wed Sep 22 10:57:39 MEST 2004"
 contentType="text/html">
<EuroGOV:content>
<![CDATA[
<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">
<html lang="sv">
<head>
    <title>Regeringen och Regeringskansliet</title>
    <meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
    <meta http-equiv="Content-Script-Type" content="text/javascript">
    <meta http-equiv="Content-Style-Type" content="text/css">
    <script language="javascript" type="text/javascript" src="/js/popup.js"></script>
    <script language="javascript" type="text/javascript" src="/js/validationTexts_sv.js"></script>
    <script language="javascript" type="text/javascript" src="/js/formFunctions.js"></script>
    <link rel="stylesheet" type="text/css" href="/css/deprecatedstyle.css">
```

Abb. 1.2: Beispiel der Dokumentstruktur im EuroGOV Korpus [Web05]

EuroGOV:bin ist das Wurzelelement einer jeden Datei. Jede Datei umfasst höchstens 25.000 Dokumente.

Das Attribut *Domain* beinhaltet die Topleveldomäne der heruntergeladenen Dokumente. Jedes Dokument bekommt eine dreistellige *ID*-Nummer zugewiesen, die sich pro Topleveldomäne jeweils hochzählt.

Sprachverteilung			
Domäne .eu.int		EuroGOV	
Sprache	Prozent	Sprache	Prozent
englisch	33,26%	finnisch	20,28%
französisch	18,08%	deutsch	18,20%
deutsch	9,08%	ungarisch	12,58%
finnisch	6,24%	englisch	10,16%
spanisch	5,75%	lettisch	8,80%
holländisch	5,29%	französisch	6,98%
dänisch	5,13%	schwedisch	5,32%
portugiesisch	4,47%	portugiesisch	3,93%
schwedisch	3,26%	holländisch	3,91%
griechisch	2,92%	polnisch	2,14%
italienisch	2,64%	italienisch	1,70%
lettisch	1,13%	spanisch	1,39%
polnisch	1,05%	tschechisch-iso8859_2	1,13%
estnisch	0,60%	slowakisch-windows1251	0,89%
litauisch	0,51%	russisch-windows1251	0,60%
ungarisch	0,41%	dänisch	0,49%
tschechisch-iso8859_2	0,05%	estnisch	0,39%
slowakisch-windows1251	0,04%	russisch_koi8_r	0,30%
rumänisch	0,03%	slowakisch-ascii	0,27%
slowakisch-ascii	0,02%	griechisch	0,27%
russisch_koi8_r	0,02%	litauisch	0,19%
isländisch	0,01%	irisch	0,03%
russisch-windows1251	0,01%	walisisch	0,01%

Tab. 1.3: Sprachverteilung der Domäne .eu.int und des gesamten EuroGOV Korpus [Sig05b]

EuroGOV:doc beinhaltet alle Metadaten für jedes Dokument, die während des Crawlens generiert bzw. erkannt wurden.

Das Attribut *URL* beinhaltet die Internetadresse des Dokuments. Jedes Dokument bekommt eine eigene *ID*-Nummer nach folgendem Muster zugewiesen: Exx-yyy-z. E steht für EuroGOV, xx für die Topleveldomäne, yyy ist der Dateiname bzw. die *ID*-Nummer aus dem EuroGOV:bin Element und z ist die laufende Nummer des Dokuments. *md5* ist die errechnete Quersumme des Dokumentinhalts. Das *fetchDate* ist der Zeitpunkt des eigentlichen Dokumentendownloads. Das *contentType* Attribut beinhaltet das Datenformat des Dokuments.

EuroGOV:content beherbergt den eigentlichen Inhalt des Dokuments. Der gesamte In-

halt wurde in einem CDATA Bereich abgespeichert.

1.3.3 Herausforderungen des EuroGOV Korpus

Bei der Zusammenstellung des EuroGOV Korpus sind verschiedene Herausforderungen und Schwierigkeiten aufgetreten.

Vollständigkeit Trotz großer Anstrengungen bezüglich der vollständigen Abdeckung der einzelnen Domänen durch mehrere Crawldurchläufe, konnte kein 100%ges Ergebnis erzielt werden.

Unvollständige Beschreibung der Linkanalyse Eine umfassende Linkanalyse der EuroGOV Kollektion wurde noch nicht durchgeführt. Dies hat zwar keine Auswirkungen auf die Kollektion an sich, würde jedoch beim Evaluationsprozess sehr hilfreiche Informationen liefern, um herausfinden zu können, wie repräsentativ das Korpus im Vergleich zum realen Internet wirklich ist.

Leere Seiten Das EuroGOV Korpus beinhaltet ca. 70.000 leere Seiten. Es ist nicht bekannt, wodurch dieser Crawlfehler hervorgerufen wurde.

Rich Documents Types EuroGOV beinhaltet neben HTML Seiten auch Dateien in den Formaten .pdf, .ps und .doc. Mit diesen Seiten wurden jedoch keine Textextraktion durchgeführt, so dass die Seiten in ihren Originalformaten vorliegen. Dies erschwert die Weiterverarbeitung der Kollektion und erfordert einen erhöhten Preprocessing Aufwand der Daten. Im Grundsatz spiegelt dieser Punkt jedoch ein reales Webretrieval Szenario bzw. die Heterogenität der Daten wieder.

1.3.4 Ausblick für das EuroGOV Korpus

Aktuell ist keine weitere Aktualisierung bzw. Ausweitung des Korpus geplant, da es viele interessante Fragen zu dem Thema Cross-Linguales Information Retrieval beinhaltet. Laut Universität Amsterdam ist mit einem neuen Korpus für WebCLEF 2007 bzw. 2008 zu rechnen, in dem verbesserte Vollständigkeit und Rich Document Types (Abschnitt 1.3.3) besser umgesetzt werden sollen. Die weitere Entwicklung des EuroGOV Korpus ist jedoch eng mit dem Fortbestehen der WebCLEF Initiative verbunden.

1.4 Topicentwicklung

Zur Durchführung des Web Tracks mussten multilinguale Queries erstellt werden. Queries werden im CLEF Kontext als Topics bezeichnet. Aufgabe jeden Teilnehmers war es, mindestens 30 Topics in der zugewiesenen Sprache zu generieren. Hierbei teilte sich die Anzahl in 15 *homepages* und 15 *named page* Topics auf. Homepage Topics beziehen sich auf Einstiegs- bzw. Startseiten einer Domäne oder einer Website. Es können mehrere Homepage Seiten innerhalb einer Website auftreten. Z.B. haben Universitätsdomänen innerhalb ihrer Website Startseiten von einzelnen Fachbereichen, Bibliotheken oder Dozenten.
Named page Topics beinhalten spezielle Internetseiten, die sich auf ein spezifisches Thema beziehen. Named page Topics müssen keine Start- bzw. Homepage Seiten sein. Als Beispiel für eine named page Topic Seite wäre der Mensaspeiseplan einer Universität zu nennen. [Sig05d]
Parallel zur Entwicklung von mindestens 30 Topics pro Teilnehmer mussten folgende Zielstellungen erfüllt werden:

- Inhaltlich zum Topic übereinstimmende und mehrfach vorkommende Seiten mussten identifiziert und als *Duplicate* (Abschnitt 1.5) registriert werden. Dies galt auch für Seiten, die mehrsprachig angeboten wurden (Translations).

- Jeder Teilnehmer musste eine englische Übersetzung seiner generierten Topics mitliefern.

Die Topic Autoren wurden aufgefordert, persönliche Angaben zu ihren aktiven und passiven Sprachkenntnissen, Geburtsland und ihrem aktuellen Wohnland zu machen. Diese Metadaten wurden dem Topic neben der eigentlichen Query und der dazugehörigen englischen Übersetzung hinzugefügt. Des Weiteren wurde die Sprache und die Topleveldomäne des Topics identifiziert und zugewiesen.

Zur Generierung der Topics wurden jedem Teilnehmer mehrere hilfreichen Tools zur Verfügung gestellt. Die gesamte Abwicklung wurde über ein Webinterface durchgeführt und überwacht. Auf diesem Webinterface, welches durch die Universität Amsterdam gehostet wurde, konnte das EuroGOV Korpus anhand einer Lucene Search Box und der Terrier Searchengine (Universität Glasgow) durchsucht werden. Zusätzlich gab es Zugang zum Korpus durch das direkte Aufrufen der Dokument-IDs. Anhand dieser Suchmöglichkeiten und des vorgegebenen Phasenablaufes konnte jeder Teilnehmer durch das gezielte Anfragen bestimmter Themengebiete Topics entwickeln. Die Topicgenerierung staffelt sich in fünf Phasen.

Collection Exploration Phase Durchsuchen des Korpus nach bestimmten Seiten durch das Eingeben von zufällig gewählten Queries in die bereitgestellten Suchmaschinen. Die Ergebnislisten der beiden zur Verfügung gestellten Suchmaschinen umfassten jeweils 25 Dokumente. Ist eines dieser Dokumente von Interesse oder spiegelt es die eingetippte Query wieder, kann diese Query als mögliches Topic behandelt werden. Daher muss die Dokumenten-ID zur Identifikation der Zielseite notiert werden. Für Topics stehen die Varianten der named paged Topics und der homepage Topics zur Verfügung. Stimmen die Anfragen mit den Ergebnissen nicht überein, muss eine neue Query formuliert und getestet werden.

Query Formulation Phase Wenn in der Exploration Phase ein möglicher Topic Kandidat gefunden wird, muss die Query umformuliert werden. Diese Umformulierung sollte nach dem Muster eines Bookmarks durchgeführt werden. Ein Bookmark ist der Lesezeichentitel eines Favoriten in einem Internetbrowser. Das umformulierte Bookmark muss erneut als Query an die Kollektion geschickt werden. Falls nun die gleiche Seite, identifiziert durch die Dokumenten-ID, wie im ersten Queryversuch in der Ergebnisliste erscheint, kann diese Bookmarkformulierung als Topictitle verwendet werden.

Duplicate Detection Phase In dieser Phase muss die Ergebnisliste, in der die gewünschte Zielseite vorliegt, nach Duplikaten und Übersetzungen durchsucht werden. Die Seiten, die inhaltlich zum Topic gehören, müssen ebenfalls in Topicsubmission Form angegeben werden. Dies ist für die spätere Relevanzbewertung von entscheidender Bedeutung.

Query Translation Phase Nach erfolgreichem Generieren eines Topictitles muss das Topic in die englische Sprache als Basis aller multilingualen Retrievaldurchläufe übersetzt werden.

Topic Submission Alle gesammelten Informationen eines Topics, Topic Title als Bookmark formuliert, die Dokumenten-ID, Duplikate und Übersetzungen sowie die englische Übersetzung wurden über ein Webformular zur Topicgenerierung eingereicht.

Wie in Abb. 1.3 sichtbar, wurden den Teilnehmern die Topics in wohlgeformtem XML Format übergeben. Insgesamt wurden 547 Topics generiert, von denen 242 homepage Topics und 305 named page Topics sind (Abschnitt 1.4). Insgesamt beinhalten die Zielseiten und die dazugehörigen Topics 11 verschiedenen Sprachen (Tab. 1.4).

```
<topics task="WebCLEF" year="2005">
  <topic>
    <num>WC0001</num>
  <title>road safety in europe</title>
  <metadata>
    <topicprofile>
      <language language="EN" />
      <translation language="EN">road safety in europe</translation>
    </topicprofile>
    <targetprofile>
      <language language="EN" />
      <domain domain="eu.int"/>
    </targetprofile>
    <userprofile>
      <native language="EN" />
      <native_other>Tok Pisin</native_other>
      <active language="FR" />
      <active_other>Bahasa Indonesia, Thai, Lao, Korean</active_other>
      <passive language="FR" />
      <passive language="DE" />
      <passive_other>Bahasa Indonesia, Thai, Lao, Korean</passive_other>
      <countryofbirth country="AU" />
      <countryofresidence country="AU" />
    </userprofile>
  </metadata>
</topic>
```

Abb. 1.3: Aufbau und Struktur eines WebCLEF 2005 Topics

	Total	ES	EN	NL	PT	DE	HU	DA	RU	GR	IS	FR
Total	547	134	121	59	59	57	35	30	30	16	5	1
HP	242	67	50	25	29	23	16	11	15	5	1	
NP	305	67	71	34	30	34	19	19	15	11	4	1
Duplicates (topics)	191	37	47	21	15	38	11	12	8	1	1	
Duplicates (total)	473	82	109	40	95	90	18	26	11	1	1	
Translations (topics)	114	25	24	9	4	13	6	15	6	7	5	
Translations (total)	387	100	47	18	7	39	17	101	11	19	28	

Tab. 1.4: Verhältnis der WebCLEF Topics 2005 zu NP: named page Topics; HP: homepage Topics und den Sprachen ES: Spanisch; EN: Englisch; NL: Holländisch; PT: Portugiesisch; DE: Deutsch; HU: Ungarisch; DA: Dänisch; RU: Russisch; GR: Griechisch; IS: Isländisch; FR: Französisch [Sig05c]

1.5 Relevanzbewertung der WebCLEF Topics

Während des Erstellens der Topics wurden neben dem Identifizieren von Duplikaten (inhaltlicher und sprachlicher Art) auch die dazugehörigen URLs und IDs der bestimmten Zielseiten gesammelt. Diese drei Felder wurden als Basis für die Relevanzbewertung be-

nutzt. Da beide WebCLEF Tasks, named page finding und homepage finding, auf das Finden bestimmter Dokumente hinzielen, ist die Relevanzbewertung nur mittels *binary judging* durchzuführen. Binary judging bedeutet, dass ein gefundenes Dokument nur relevant oder irrelevant sein kann. Es gibt beim WebCLEF Track keine unterschiedlichen Relevanzstufen. Als relevant werden alle Dokumente eingestuft, die während der Topicentwicklung im Zusammenhang mit den Topics eingereicht wurden. Somit zählen die eigentliche Zielseite, die dazugehörigen Duplikate und die Seitenübersetzungen als relevante Seiten für ein Topic. Auf Basis dieser Topicauswertung können im Durchschnitt pro Topic 1,56 relevante Dokumente erwartet werden (Tab. 1.4). Diese Zahl errechnet sich aus dem Durchschnitt aller relevanten Dokumente zu allen Topics.

1.6 Tasks und Submission

Der WebCLEF 2005 Durchgang konzentrierte sich auf *known item* Suchanfragen, auch *navigational Search* genannt. Hierbei versucht der Suchende bzw. User eine bestimmte Seite zu finden, die er entweder schon einmal besucht hat oder von der er in Kenntnis gesetzt wurde. Known item Search ist auch der Grund für die Query Formulation Phase. (Abschnitt 1.4) Die Formulierung der Query in eine Art Bookmark unterstreicht die Kenntnis des Users bezüglich des gesuchten Dokuments. Da das beschriebene Ziel das Finden einer speziellen Seite ist, kann damit auch das Vorgehen des binary judging (Abschnitt 1.5) der Relevanzbewertung erklärt werden. Diese bestimmte Seite wiederzufinden ist die Aufgabe des WebCLEF Durchgangs im Jahre 2005. Diese Aufgabe wurde in drei Teilbereiche unterteilt.

- Mixed Monolingual Task
- Multilingual Task
- Bilingual English to Spanish Task

Der Mixed Monolingual Task zielt darauf hin, nur in der gegebenen Topicsprache den EuroGOV Korpus zu durchsuchen. Die anzunehmende Simulation wäre, dass ein User eine Internetseite in der gleichen Sprache wie die dazugehörige Query sucht. Übersetzte Seiten sind bei dieser Simulation also nicht von Belang und somit irrelevant.
Dies gilt nicht für den Multilingual Task. Hier sucht der User unabhängig von der Sprache anhand seiner Query alle relevanten Internetseiten des EuroGOV Korpus. In diesem Task kommen also die übersetzten Seiten einer Internetseite zum Tragen. Der Multilingual Task ist der Mittelpunkt der WebCLEF Initiative.

Als dritter Task wurde der Bilingual English to Spanish Task angeboten. Für diesen Task wurde ein gesondertes Topicset entwickelt. Das dazugehörige Szenario beinhaltet einen Englisch sprechenden User, der seine Suchanfragen auf Englisch eingibt, um dann eine Ergebnisliste zurückzubekommen, die alle relevanten spanischen Dokumente liefert.

Jeder Teilnehmer durfte fünf Runs pro Disziplin abgeben. Die Ergebnislisten jedes Runs durften maximal 50 Treffer beinhalten. Von den fünf pro Disziplin erlaubten Runs musste mindestens ein Baseline Run abgegeben werden. Baseline Runs dürfen nicht mit Hilfe der - in den WebCLEF Topics enthaltenen - Metadaten durchgeführt werden, sondern lediglich auf Basis des title (monolingual) und translation (multilingual) Felds. Die Ergebnislisten mussten in einem von den Organisatoren festgelegtem Format eingereicht werden, damit eine schnelle und einheitliche Evaluation gewährleistet werden konnte. Zur Überprüfung der Ergebnislisten wurde eine Perlskript von Seiten der Organisatoren zur Verfügung gestellt. Der Aufbau der Ergebnislisten stellte sich wie folgt dar:

WC	0001-0547	Exx-yyy-z	1-50	Mean Reciprocal Rank	Run
WebCLEF	Topic #	Doc ID	Rang	MRR	Name max. 12 Stellen

Tab. 1.5: Aufbau einer WebCLEF Ergebnisliste [Web05]

Die generierten Ergebnisslisten mussten als Textdatei zusammen mit einer näheren Beschreibung der Vorgehensweise über das Webinterface eingereicht werden. Folgende Angaben mussten mit den Ergebnislisten abgegeben werden:

- Index Einheit
- Indexiertechnik
- Indextyp
- Retrievalmethode

Die eingereichten Runs konnten bis zum Abgabetermin jederzeit über das Webinterface eingesehen und korrigiert werden. [Web05]

1.7 Evaluierung der Ergebnisse

Die Ergebnisse wurden nach Abgabe aller Runs und Fertigstellung der Relevanzbewertung von Seiten des Organisators geprüft. Da alle WebCLEF Tasks auf der Grundlage

von known item Search (Abschnitt 1.6) durchgeführt wurden, wurde als Hauptevaluierungsmaß der Mean Reciprocal Rank (MRR) verwendet. Ziel der Evaluierung ist es, die Effektivität eines Systems zu messen. Aufgrund des known item Search Ansatzes ist nur das erste relevante Dokument in der Ergebnisliste von Entscheidung für die Evaluierung. Die relevanten Dokumente sind aufgrund der schon abgeschlossenen Relevanzbewertung bekannt und müssen nun in der Ergebnisliste lokalisiert werden. Mit der Position des ersten relevanten Dokuments in der Ergebnisliste müsste daraufhin die Effektivität ermittelt werden. Mittels des Mean Reciprocal Rank kann dies durchgeführt werden. Der MRR berechnet sich wie folgt:

MRR = 1 / Rang des ersten relevanten Dokuments in der Ergebnisliste [Web05]

1.8 Fazit

Die für den WebCLEF neu geschaffene Testkollektion liefert viele neue Herausforderungen in Größe und Multilingualität. In dem ersten Durchgang müssen die einzelnen Schwierigkeiten lokalisiert werden, um im nächsten Durchgang auf sie reagieren zu können. Allgemein kann zum WebCLEF 2005 Track gesagt werden, dass aufgrund von einigen unterschiedlichen Aufgabenstellungen die Erfahrungen aus anderen Web Tracks nur bedingt berücksichtigt werden können. Somit ist der erste Durchgang entscheidend für die erfolgreiche Fortführung dieses ersten Multilingualen Web Tracks. Erfahrungen mit WebCLEF und mögliche Verbesserungs- bzw. Änderungsvorschläge folgen als Ausblick und Fazit im letzten Kapitel (Kap. 5).

Kapitel 2

Verschiedene Web Track Initiativen

In den vergangenen Jahren wurden verschiedene Initiativen in den Bereichen der Evaluierung von Retrieval Systemen durchgeführt. Der Vergleich dieser Tracks mit dem WebCLEF Track gibt Aufschluss über die Gemeinsamkeiten und Unterschiede. Im Vergleich wird deutlich, dass zum einen eine regionale Komponente bei jedem Web Track mit einspielt und sich zum anderen die Aufgabenstellungen meistens ähneln. Um den für den WebCLEF Track dieses Jahr neu entwickelten EuroGOV Korpus richtig zu bewerten und seine Besonderheiten herausarbeiten zu können, ist ein Blick auf die anderen Korpora von ausschlaggebender Bedeutung. In diesem Kapitel werden die Retrievalevaluationsforen TREC (Nordamerika), NTCIR (Japan) und SEWM-2004 (China) näher beschrieben. Es folgt erst ein kurzer Gesamtüberblick der einzelnen Initiativen, um dann die dazugehörigen Web Tracks genauer zu beleuchten und einige teilnehmende Systeme mit ihren Ergebnissen vorzustellen. Manche Tracks der japanischen und chinesischen Initiative konnten jedoch aufgrund von mangelnden Übersetzungen der einzelnen Papers bzw. Internetseiten in die englische Sprache nicht in aller Ausführlichkeit beschrieben werden. Dies gilt insbesondere für den chinesischen Track. Im Anschluss an die Beschreibung der separaten Initiativen werden der WebCLEF Track und andere vergleichbare Tracks einander gegenüber gestellt und verglichen.

2.1 Text Retrieval Conference (TREC)

Die Text Retrieval Conference ist eine Forschungsinitiative zur Förderung der Erkenntnisse im Bereich des Information Retrieval und der dazugehörigen Technologien. Hauptorganisator dieses Forums ist das nordamerikanische National Institute of Standards and Technology (NIST). Als weitere Partner treten die U.S. Behörden US Department of Defence Advanced Research and Development Activity(ARDA) und die Defence Advanced

Research Projects Agency (DRAPA) auf. Die TREC Initiative verfolgt seit 1992 vier Hauptziele:

- Forschung im Bereich Information Retrieval mit Hilfe von großen Testkollektionen zu fördern,
- den Informations- und Erfahrungsaustausch zwischen Industrie, Wissenschaft und Regierung mittels eines offenen Forums zu verbessern,
- den Austausch von neugewonnenen Erkenntnissen zwischen Wissenschaft und Industrie zu beschleunigen, um daraus kommerzielle Produkte zu entwickeln und
- das Bereitstellen von adäquaten Evaluierungsmethoden für Wissenschaft und Industrie, sowie des schnellen Reagierens auf aktuelle Problemstellungen.

An dem 13. Durchgang von TREC im Jahr 2004 haben insgesamt 103 Organisationen teilgenommen. Die teilnehmenden Gruppen beschränkten sich nicht nur auf Nordamerika, sondern waren weltweit zu finden; 67 % aller Teilnehmer hatten einen universitären Hintergrund. Im TREC 2004 Durchgang wurden sieben verschiedene Aufgabenstellungen (Tasks) zur Verfügung gestellt. Den einzelnen Tasks dienten angepasste Testkollektionen als Basis. Zu einer Testkollektion gehören die drei Teilelemente Datenbestand, bestehend aus unterschiedlichen Arten von Dokumenten, formulierte Suchanfrage in strukturierter Form (TREC Kontext Topics genannt) und eine Relevanzbewertung des Datenbestands, bezogen auf die Suchanfragen.

> *A test collection is an abstraction of an operational retrieval environment that provides a means for researchers to explore the relative benefits of different retrieval strategies in a laboratory setting. [Voo05]*

TREC Topics erlauben ein breites Spektrum an unterschiedlichen Querykonstruktionen und eine detaillierte Beschreibung der Kriterien, die ein Dokument erfüllen muss, um für ein Topic relevant zu sein. Bei der Entwicklung von Suchanfragen unterscheidet TREC zwischen dem Ausdruck eines Informationsbedürfnisses - im TREC Kontext Topic genannt - und der eigentlichen Datenstruktur eines Topics, die an die Retrievalsysteme geschickt werden. Die Datenstruktur wird Query genannt. Die allgemeine Struktur eines Topics bzw. einer Query sieht wie folgt aus:

num Topic ID Nummer

title Suchanfrage formuliert in maximal drei Worten

```
<num> Number:  656
<title> lead poisoning children
<desc>
How are young children being protected against lead poisoning from paint and
water pipes?
<narr>
Documents describing the extent of the problem, including suits against
manufacturers and product recalls, are relevant.  Descriptions of future plans
for lead poisoning abatement projects are also relevant.  Worker problems with
lead are not relevant.  Other poison hazards for children are not relevant.
```

Abb. 2.1: Ein Beispieltopic aus dem TREC 2004 Robust Track [Voo05]

desc Nähere Beschreibung der Suchanfrage (title) in einem Satz.

narr Präzise Beschreibung relevanter Dokumente in Verbindung mit dem aktuellen Topic

Die Topicentwicklung liegt in der Verantwortung und im Ermessensspielraum der einzelnen Teilnehmer und ist nicht durch die Organisatoren vorgegeben. Die NIST ist die Sammelstelle für alle Topics. Topicentwickler sind in der Lage, mittels des NIST PRISE Systems, ein Vector Space Retrieval System [Dim05, Oar99] und potentielle Suchanfragen auf dem Korpus zu testen. Wenn die Suchanfragen relevante Dokumente zurückliefern, können diese Suchanfragen für offizielle TREC Topics verwendet werden. Während dieses Prozesses der Topicsuche werden auch die relevanten Dokumente pro Topic gesammelt, um die Relevanzbewertung pro Topic gleichzeitig durchzuführen. Damit sind Topicentwickler auch gleichzeitig verantwortlich für die Relevanzbewertung ihrer eigenen Topics. Die Bewertung eines Topics bzw. der zurückgegebenen Dokumente läuft nach folgendem Schema ab: Die Bewerter, bei der NIST Assessors genannt, sollen sich die Frage stellen, ob das zu bewertende Dokument dazu verwendet werden könnte, einen Artikel über das Topic zu schreiben. Wenn dies der Fall ist, gilt das Dokument als relevant, ganz gleich ob die Information in diesem Dokument schon von vorherigen Dokumenten abgedeckt wurde oder nicht. Da die Anforderungen der sieben einzelnen Tasks an die Testkollektionen, den Korpus und die Topics unterschiedlicher Natur sind, wurden die Kollektionen den Bedürfnissen der einzelnen Tasks angepasst. Diese notwendigen Veränderungen wurden durch die Parameter Größe und Anzahl der Dokumente, sowie unterschiedliche Struktur der Topics umgesetzt. Die Grundstruktur der Topics wurde in den meisten Fällen beibehalten (Abb. 2.1). TREC 2004 beherbergte folgende Aufgabenstellungen [Voo05]:

Genomics Track Der Genomics Track wurde 2003 zum ersten Mal offiziell durchgeführt. Das Ziel dieses Tracks ist es, Retrievalsysteme nur in themenspezifischen Domänen agieren zu lassen, um herauszufinden, ob themenspezifische Informationen die Effektivität eines Retrievalsystems in einer den Themen zugeordneten Domäne steigern

können. Basis des Genomics Track 2004 waren drei verschiedene Varianten der Kategorisierung und ein ad hoc Retrieval Task. Als Korpus wurden Daten aus der MEDLINE Datenbank [Med05], einer bibliographischen Datenbank für biomedizinische Literatur, entnommen. Die Kollektion beinhaltet 4,5 Mio. Dokumente mit einem Speicheraufkommen von neun GB an Daten. 50 Topics wurden von Fachleuten der Biomedizin entwickelt und bewertet. [Voo05]

Hard Track Der HARD Track steht für "High Accuracy Retrieval from Documents" und wurde das erste Mal 2003 eröffnet. Ziel dieses Tracks war es, die Retrievalperformance in den ersten Positionen der Ergebnislisten zu verbessern. Umgesetzt wurde diese Thematik mittels personalisierter und spezifizierter Suchanfragen. Neben den üblichen TREC Topics wurden weitere Metadaten hinzugefügt. Die Metadaten lieferten Informationen zum gesuchten Themenbereich des Topics, wie z.B. Geographie und den persönlichen Daten des Assessors (Topicentwickler und Bewerter). Die Daten des Assessors wurden aufgenommen, weil der Topicentwickler gleichzeitig auch die Rolle des Informationssuchenden / Users übernahm. Der HARD Track wurde auf Basis eines eigenen Korpus anhand von ad hoc Retrieval durchgeführt. Das Korpus setzte sich aus englischen Zeitungsartikeln des Jahres 2003 zusammen. Die Größe dieser Kollektion betrug ca. 1500 MB und umfasste ca. 650.000 Dokumente. 50 Topics wurden für den HARD Track entwickelt. Die für den HARD Track spezifischen Metadaten durften erst nach Abgabe eines Baseline Runs in die Systeme integriert werden. Die Baseline Runs bestanden somit aus den vier Grundelementen eines TREC Topics. 16 Gruppen haben 2004 135 Runs eingereicht. [Voo05]

Novelty Track Der Novelty (Neuheit) Track hat die Aufgabe, Systeme auf die Fähigkeit relevante und neue (nicht mehrfachvorkommende) Informationen innerhalb einer Dokumentenkollektion hin, zu untersuchen. In diesem Task muss das System nicht die relevanten Dokumente herausfiltern, sondern lediglich die wichtigen Informationen hervorheben. Der User findet die wichtigen Dokumente durch das Skimming (gleiten) der Dokumentenkollektion. Lediglich die hervorgehobenen Informationen innerhalb eines Dokuments geben ihm Aufschluss über die Relevanz eines Dokuments. Als Dokumentenkollektion wurden Teile des AQUAINT Corpus of English News Text [Aqu05] gewählt. Dieser auf den Novelty Track zugeschnittene Korpus beinhaltet ca. 1.033.000 Dokumente und nimmt drei GB Speicher in Anspruch. Die Assessors bewerteten die Kollektion intellektuell. Es wurden 25 relevante Dokumente pro Topic gewählt und manuell so bearbeitet, dass die relevanten Informationen, bezogen auf die jeweiligen Topics, hervorgehoben wurden. Der Novelty Task stellte

allen Teilnehmern vier Aufgaben:

- Identifiziere auf Basis des gesamten Korpus alle relevanten und neuen Sätze, die Informationen zum jeweiligen Topic beinhalten,
- Identifiziere auf Basis des gesamten Korpus nur alle neuen Sätze, die Informationen zum jeweiligen Topic beinhalten,
- Identifiziere auf Basis der ersten fünf Dokumente des gesamten Korpus alle relevanten und neuen Sätze, die Informationen zum jeweiligen Topic beinhalten und
- Identifiziere auf Basis der ersten fünf Dokumente des gesamten Korpus alle neuen Sätze, die Informationen zum jeweiligen Topic beinhalten.

14 Gruppen nahmen 2004 an diesem Track teil und reichten 183 Runs ein. [Voo05]

Question Answering (QA) Track Der QA Track adressiert das Problem der zunehmende Informationsflut als Produkt von Retrievalsystemen. Die QA Systeme sollen im Gegensatz zu gewichteten und geordneten Ergebnislisten reale Antworten zurückliefern. Die Topickollektion dieses Tracks besteht aus Fakten- (FACTOID), Listen- (LIST) und Defintionsfragen (OTHER). Als Korpus wurde erneut der AQUAINT Corpus of English News Texts [Aqu05] gewählt. 65 Fragenserien mit insgesamt 230 FACTOID, 56 LIST und 65 OTHER wurden entwickelt. Die Antworten pro Serie wurden mit 0,5 pro FACTOID, 0,25 pro LIST und 0,25 pro OTHER gewichtet. 28 verschiedene Gruppen nahmen mit 63 Runs an dem Track 2004 teil. [Voo05]

Robust Track Der Robust Track konzentriert sich auf die Verbesserung von Systemen, die mit schwierigen oder alten Topics arbeiten und die in den vergangenen Durchgängen mit schlechter Performance auffielen. Der Robust Track wurde 2003 ins Leben gerufen.

> *The initial track provided strong evidence that optimizing average effectiveness using the standard methodology and current evaluation measures further improves the effectiveness of the already-effective topics, sometimes at the expense of the poor performers. [Voo05]*

Um das Ziel diesen Tracks zu erreichen, wurden insgesamt 250 Topics zusammengestellt. 200 dieser Topics waren sogenannte schlechte Performer aus vergangenen Jahren. Um diese Topics dennoch verwenden zu können, musste die alte TREC Datenkollektion (TREC Disk 4 und 5) eingesetzt werden. Zu den 200 alten Topics

wurden 50 Neue hinzugefügt. 50 der alten 200 Topics wurden von den Assessors als *schwer* eingestuft, da die durchschnittlichen Precision Werte unterhalb des Gesamtdurchschnitts des 2003-Durchgangs lagen und mindestens einen sehr schlechten Wert hervorriefen. Der Robust Track war ein klassischer ad hoc Retrievaltask. 14 Gruppen nahmen an diesem Track teil und reichten 110 Runs ein. [Voo05]

Terabyte Track Der Terabyte Track wurde zum ersten mal 2004 durchgeführt. Ziel diesen Tracks ist es, eine Evaluierungsmethode für sehr große Kollektionen mit Speicheraufkommen im Terabyte Bereich zu entwickeln und teilnehmenden Gruppen die Möglichkeit zu bieten, die eigenen Systeme an großen Datenmengen zu testen. Als Datenkollektion wurde das .GOV2 Korpus verwendet. Das .GOV2 Korpus wurde im Frühjahr 2004 nach umfangreichen Crawlprozessen der US amerikanischen Regierungsdomäne (.gov) entnommen und umfasst 25 Mio. Dokumente, die ein Speicheraufkommen von 426GB benötigen. Auf Basis dieses Korpus kann ein realgetreues Retrievalszenario im Internet simuliert werden. Die Aufgabe für die Teilnehmer war ein klassischer ad hoc Retrievaltask. Auf Basis des .GOV2 Korpus wurden 50 Topics entwickelt. Die Topics bestanden aus einfachen Suchanfragen, die ein Informationsdefizit decken sollten. Die Retrievalsysteme mussten ihre Prozesse bei 10.000 Dokumenten pro Topic abbrechen. 17 Gruppen nahmen an diesem Track mit insgesamt 70 Runs teil. [Voo05]

Web Track Das Ziel des Web Tracks ist das Untersuchen des Retrievalverhaltens einer großen Datenkollektion, die eine große Anzahl an Hyperlinkstrukturen aufweist. Hier steht, wie im Terabyte Track, die Größe der Datenkollektion im Vordergrund, wobei die Struktur der Dokumente als weitere Komponente mit eingebracht werden soll. Als Korpus diente die .GOV Kollektion. Das .GOV Korpus entstand wie der.GOV2 Korpus durch umfangreiche Crawlprozesse der US amerikanischen Regierungsdomäne (.gov) im Frühjahr 2002 und beinhaltet 1.247.753 englischsprachige Dokumente, wobei sich die größte Anzahl aus HTML Seiten zusammensetzt. Neben den HTML Dokumenten nehmen Dateiformate wie .txt, .doc, .pdf, .ps etc. 15,6 % der Gesamtmenge ein. Jedes Dokument im Korpus wurde bei einer maximalen Größe von 100 K abgeschnitten, sodass die Gesamtgröße des .GOV Korpus 18,1 GB beträgt. Das Ziel einer realen Nachbildung des WWW mittels des .GOV Korpus zu entwickeln gelang. Mit der Weiterentwicklung des im Terabyte Track benutzten .GOV2 Korpus konnte dieses Ziel im Jahr 2004 sogar ausgebaut und verbessert umgesetzt werden. Die einzelnen Dokumente des Korpus beinhalten den eigentlichen Inhalt der Internetseite sowie einige Metadaten zur näheren Beschreibung der Dokumente, wie

z.B. URLs, Dokument IDs etc.. In vergangenen TREC Web Tracks wurden *topic destillation, named page finding* und *homepage finding* Tasks betrieben. Da Suchmaschinen des WWW diese drei Suchanforderungen im realen Gebrauch für den User vereinen müssen, wurden im TREC Web Track 2004 alle Tasks mittels einer gemischten Topicsammlung miteinander vereint. In diesem Zusammenhang wird im TREC Kontext von dem Mixed Query Task gesprochen.

Topic destillation Die Suchanfrage beschreibt ein allgemeines Thema. Das System sollte passend zu dem Thema eine relevante Homepage / Startseite zurückliefern.

Homepage finding Ein User möchte eine ihm bekannte Seite wiederfinden oder hat von einer interessenten Seite gehört (known item Search). Die Suchanfrage stellt sich aus dem Namen einer Seite zusammen, die der User sucht. Das System sollte die URL der relevanten Homepage / Startseite angeben.

Named page finding Ein User möchte eine ihm bekannte Seite wiederfinden oder hat von einer interessenten Seite gehört (known item Search). Die Suchanfrage besteht aus dem Namen einer speziellen Seite (keiner Startseite), die der User sucht. Das System sollte die URL dieser speziellen Seite bzw. einer Seite mit relevanten Informationen finden.

Aufgrund der unterschiedlichen Zielergebnisse zwischen den Querytypen topic destillation und den namedpage / homepage finding müssen unterschiedliche Bewertungsmaße verwendet werden. Dies liegt daran, dass bei der Bewertung von named page und homepage Topics nur die URL der relevanten Seiten und ihre Duplikate identifiziert werden müssen. Bei den destillation Topics müssen aufwändigere Prozesse zur Findung der Startseite einer relevanten Seite durchgeführt werden. Die gesamte Relevanzbewertung wurde manuell mit Hilfe der Assessors erledigt. Bei den named pages und den homepage Topics mussten lediglich die URLs der als relevant bewerteten Seiten registriert werden. Zur Findung der relevanten Seiten eines topic destillation Tasks musste nach erfolgreicher Findung einer Seite, die relevante Informationen beinhaltet, die Startseite durch Browsingprozesse gefunden werden. Insgesamt fanden die Assessors 1763 relevante Seiten für die TREC Web Track 2004 Topics. Davon waren 80 auf Basis der named page Topics und 83 auf Basis der homepage Topics. Für die destillation Topics wurden 1600 relevante Seiten gefunden. Diese Werte versprechen, pro named page Topic im Schnitt 1,06 Seiten, pro homepage Topic 1,106 Seiten und 21,33 Seiten pro destillation Topic

zu finden. Named page und homepage Topics wurden mit dem Mean Reciprocal Rank (Abschnitt 1.7) bewertet, da nur die erste relevante Seite pro Ergebnisliste entscheidend war. Bei der Bewertung von destillation Topics hingegen wurden mehrere relevante Dokumente als Ergebnis erwartet und mussten daher auch in die Bewertung mit einfließen. Deswegen bewerteten die TREC Organisatoren die destillation Topics mittels des Mean Average Precision (MAP) Wertes. Der MAP berechnet den durchschnittlichen Precision Wert pro Query. Der Precision Wert spiegelt die relevanten Dokumente unter den gefundenen Dokumenten wieder.

Precison = gefundene relevante Dokumente / gesamte Anzahl der gefundenen Dokumente

Für den Web Track wurden 225 Topics entwickelt, die nur aus einem Title bestanden und nicht, wie andere Standard TREC Topics, ein Description bzw. Narrative Feld beinhalteten. Der Suchanfragentyp (z.B. named page finding) wurde in die Topics mit aufgenommen. Der Retrievalprozess musste eine aus maximal 1000 Stellen bestehende Ergebnisliste produzieren. An diesem Web Track nahmen 18 Gruppen mit 83 Runs teil. Der TREC Web Track wird im Jahr 2005 eingestellt und durch den *Enterprise Track* ersetzt. Der Enterprise Track konzentriert sich auf die Suche von Informationen innerhalb eines Unternehmens zur Lösung von Aufgaben. Damit ist die eigentliche Aufgabenstellung des Web Tracks nicht wieder aufgenommen worden [Ent05]. Der Terabyte Track beinhaltet die Aufgabenstellungen des Web Track auf einer abstrakteren Ebene und ist somit der eigentliche Nachfolger bzw. die Weiterentwicklung des TREC Web Tracks. [Voo05, Cra05]

Tracks	TREC 2003	TREC 2004
QA	33	28
WEB	27	18
Novelty	14	14
Genomics	29	33
HARD	14	16
Robust	16	14
Terabyte		17
Summe aller Teilnehmer	93	103

Tab. 2.1: Verteilung von TREC Runs auf die unterschiedlichen Tasks in den Jahren 2003 und 2004 [Voo05]

2.1.1 Systeme und Ergebnisse des TREC Web Track 2004

Die 225 Topics stellten sich zu gleichen Anteilen zu jeweils 75 named page, home page und distillation Topics zusammen. Zur Umsetzung des Web Tracks bedienten sich die teilnehmenden Gruppen unterschiedlicher Ansätze. Da die zu indexierende Sprache des .GOV Korpus Englisch ist, wurden allgemein bekannte Stemming Methoden, wie z.B. der Porter Stemmer [Por05], zum Indexieren verwendet. Im Hauptaugenmerk des Web Tracks lagen die Retrievalsysteme mit den zu prüfenden Parametern. In der Auswertung der einzelnen Runs konnten die sechs wichtigsten Parameter der Systeme herausgefiltert werden. Während des Retrievalprozesses wurden unter den besten Runs die indexierten Anchor Texte verwendet, die Dokumenten- und Link- Strukturen berücksichtigt, URL-Längen und Eigenschaften mit einbezogen und die unterschiedliche Verwendung von Querytypen implementiert.

Run	Avg	TD	NP	HP	Anc	Lnk	Doc	ULen	Uoth	QCls
MSRC04C12	0.97	0.92	0.99	1.00	yes	yes	yes	yes	no	no
MSRAx2	0.96	0.99	0.92	0.97	yes	yes	yes	yes	yes	no
uogWebSelAn	0.86	0.92	0.84	0.82	yes	no	yes	yes	no	yes
UAmsT04MWScb	0.84	0.82	0.85	0.86	yes	yes	yes	yes	no	no
THUIRmix045	0.79	0.70	0.85	0.84	yes	no	yes	no	no	no
ICT04CIIS1AT	0.78	0.79	0.83	0.73	yes	no	yes	no	no	no
humW04rdpl	0.74	0.91	0.66	0.64	no	no	yes	yes	yes	no
SJTUINCMIX3	0.70	0.70	0.74	0.65	yes	no	yes	no	no	yes
MeijiHILw1	0.69	0.61	0.84	0.63	yes	yes	yes	yes	no	no
csiroatnist	0.67	0.62	0.62	0.76	yes	yes	yes	yes	yes	no
MU04web1	0.63	0.64	0.50	0.74	yes	yes	yes	yes	yes	no
wdf3oks0arr1	0.59	0.47	0.74	0.54	yes	no	yes	yes	yes	no
VTOK5	0.54	0.56	0.70	0.36	yes	no	yes	no	yes	no
mpi04web08	0.52	0.46	0.58	0.51	yes	yes	yes	yes	yes	no
fdwiedf0	0.46	0.50	0.38	0.51	no	no	no	yes	yes	no
LamMcm1	0.38	0.27	0.44	0.44	yes	yes	yes	yes	yes	no
irtbow	0.13	0.07	0.22	0.11	no	no	no	no	no	no
XLDBTumba01	0.04	0.01	0.09	0.01						
Total	0.3165	0.61	0.65	0.61	83%	50%	88%	72%	55%	16%

Tab. 2.2: Die besten Runs der 18 TREC Web Track 2004 Teilnehmer und die unterschiedlich verwendeten Retrievalparameter. TD: Topic destillation in MAP; NP: Named page Topic in MRR; HP: Homepage Topic in MRR Anc: Anchor text; Lnk: Link Struktur; Doc: Dokument Struktur; ULen: URL Länge; Uoth: andere URL Eigenschaften; QCls: Querytyp processing [Cra05]

Die Tabelle 2.2 zeigt deutlich, dass auf die ersten vier Parameter nur in den seltensten Fällen verzichtet worden ist und die besten Ergebnisse auch auf der Verwendung dieser Strategien basieren. Andere URL Eigenschaften und das Anpassen der Querys aufgrund des Querytyps schienen nicht von ausschlaggebendem Wert zu sein, obwohl sie hilfreich zu sein schienen. Die besten fünf teilnehmenden Gruppen beschrieben ihre Systeme in Stichpunkten wie folgt:

Microsoft Cambridge (MSRC04C12) Verschränkung von gestemmten und nicht gestemmten Runs unter jeweiliger Verwendung von Dokumentenstruktur, URL Länge und PageRank. [Cra05]

Microsoft Research Asia (MSRAx2) Man fügte relevance scores auf die title, body, anchor und URL Felder ein und fusionierte diese vier Felder. Die Score Funktion beinhaltete BM25, Distanzwerte und eine neu eingeführte URL Gewichtungsfunktion. Die finale Bewertung basierte auf der Zusammenführung des relevance score und des HostRank, welches ein dem PageRank ähnlicher Wert ist. [Cra05]

Universität Glasgow (uogWebSelAn) Content und anchor-text retrieval, Porter Stemmer, Divergence From Randomness PL2 weighting Modell, URL Längen reranking, Wahl zwischen content und anchor-text retrieval, oder content mit anchor-text und URL Längen retrieval. [Cra05]

Universität Amsterdam (UAmsT04MWScb) CombMNZ (non-normalized, non-weighted) von gestemmten und nicht gestemmten Runs, unter jeweiliger Verwendung eine Kreuzung von Sprachmodellen beim Volltextstemming der title und anchor Textfelder. [Cra05]

Universität Tsinghua (THUIRmix045) Wortpaargewichtung basierend auf anderen Runs, die Contentretrieval auf Volltext und Inlink Anchor mit höheren Gewichtungen auf dem title, head und fett gedruckten bzw. den ersten Zeilen eines Dokuments nutzen. [Cra05]

2.2 NTCIR

Der japanische NTCIR Workshop ist eine Serie von Evaluierungsworkshops, um die Forschung in den Bereichen Information Access, Information Retrieval, Cross-Linguales Information Retrieval, Text Summarization, Question Answering und Text Mining voran zu bringen [NTC05]. Organisiert und gefördert wird diese Initiative von dem National

Institute of Informatics (NII) und Japans MEXT Grant-in-Aid for Scientific Research Institute. Die Initiative zielt darauf hin, eine Plattform von umfangreichen Testkollektionen für die genannten Forschungsbereiche zur Verfügung zustellen. Ziele des NTCIR Workshops sind seit 1997 das

- Fördern von Forschungstätigkeiten in dem Bereich der Information Access Technologie durch das Bereitstellen von wiederverwendbaren sowie umfangreichen Testkollektionen,
- Organisieren einer Plattform für Forschungsgruppen, die daran interessiert sind, Erkenntnisse und Methoden in den Bereichen von Cross Systemen in formloser Atmosphäre auszutauschen und das
- Untersuchen von Methoden und Metriken zur Evaluation von Information Access Technologien und das Aufbauen und Pflegen von Testkollektionen.

Der NTCIR Workshop zielt darauf hin, ein umfangreiches Evaluationsforum in den beschriebenen Bereichen zur Verfügung zu stellen. Die ersten Workshops konzentrierten sich auf die ostasiatischen Sprachen mit dem Hauptaugenmerk auf die japanische Sprache. Der Workshop findet in 1 1/2 jährlichem Rhythmus statt. Der Begriff Information Access, der in den Beschreibungen zum Workshop öfter auftritt, wird von den Organisatoren als Mantel um alle weiteren Forschungsbereiche, die für den Workshop von Interesse sind, gesetzt. Information Access bezieht sich auf den gesamten Informationsbeschaffungsprozess. Dies beinhaltet die Wahrnehmung eines Users, dass er ein Informationsdefizit hat, bis hin zur Interpretation der gefundenen Daten. [Kan04]
Der vierte NTCIR Workshop, der von April 2003 bis Juni 2004 stattfand, formulierte fünf unterschiedliche Retrievaldisziplinen. Im NTCIR Kontext spricht man von Tasks oder Challenges. Allen fünf Tasks wurde eine auf die Bedürfnisse der Aufgaben angepasste Testkollektion, bestehend aus Korpus und Topics, bereit gestellt. Mancher Korpus wurde aus vergangenen Workshops wiederverwendet bzw. erweitert.

CLIR Der Cross-Linguale Information Retrieval Task beschäftigt sich, wie der Name schon sagt, mit dem multilingualen Ansatz des Information Retrievals. Der CLIR Task verwendete eine Testkollektion namens NTCIR-4CLIR. Das Korpus dieser Kollektion setzte sich aus Zeitungsartikeln der Jahre 1998 bis 1999 zusammen. In diesem Korpus sind traditionelles Chinesisch, Koreanisch, Japanisch und Englisch vertreten. Der Umfang dieses Korpus beträgt 1.576.825 Dokumente die insgesamt ca. 3GB Speicher beanspruchen.

Sprachen	Anteil in %	Dokumente
trad. Chinesisch	24%	381.376
Koreanisch	16%	254.438
Japanisch	37,6%	593.636
Englisch	22,4%	337.172
Summe	100%	1.576.826

Tab. 2.3: Sprach- und Dokumentenverteilung im NTCIR-4CLIR Korpus [Kan04]

Für diesen Korpus wurden für den vierten NTCIR Durchgang 60 Topics generiert. Die Topics sind denen ähnlich, die in den TREC (Abschnitt 2.1) und CLEF (Abschnitt 1.1) Workshops aufgebaut sind. Alle CLIR Topics bestehen aus Title, Description, Narrative, Target Language und Source Language Feldern. Die Relevanzbewertung der Topics wurde in highly relevant, relevant, partially relevant und irrelevant eingestuft (multi grade judging). Die Bewertung wurde manuell durch die Topicentwickler durchgeführt. Für den CLIR Task wurden vier Subtasks entwickelt:

- Single Language (Sprache X zu Sprache X)
- Bilingual CLIR (Sprache X zu Sprache Y)
- Pivoted Bilingual CLIR
- Multilingual CLIR (Sprache X zu Chinesisch, Koreanisch, Japanisch und Englisch)

Bevor alle Topicfelder von den einzelnen Gruppen verwendet werden durften, musste erst ein Mandatory Run pro Teilnehmergruppe eingereicht werden. Mandatory Run für den CLIR Task bedeutete, die Query nur auf Basis des Title und Description Felds zu generieren. 26 Gruppen nahmen am NTCIR-4CLIR Task teil.

Patent Der Patent Retrieval Task beschäftigt sich mit dem Durchsuchen von einzelnen Patentdokumenten. Organisiert wurde dieser Task von der Japan Intellectual Property Association (JIPA) und dem National Institute of Informatics (NII). Als Grundlage dieses Tasks dient die NTCIR-4PATENT Kollektion. Diese Kollektion besteht aus einem Korpus und 34 Haupttopics, die um 69 Topics ergänzt wurden. Das Korpus besteht aus 3,5 Mio japanischen Patent Dokumenten aus den Jahren 1993 bis 2000. Zu allen Dokumenten gibt es einen japanisch-englischen Parallelkorpus, der alle Dokumente in Form eines Abstracts umfasst. Beide Korpora nehmen 55 GB an Speichervolumen ein. Die Patenttopics unterscheiden sich im Aufbau nur wenig von den CLIR Topics. Sie beinhalten ein Title, ein Description und ein

Narrative Feld gleich den Standardtopics verschiedener Retrieval Initiativen. Der Unterschied liegt in den patentspezifischen Feldern COMP, CNUM und DOC. Das Feld COMP steht für Component of a CLAIM und umfasst die Hauptaussage bzw. Haupteigenschaft eines Patents. Das Feld CNUM (CLAIM component ID) wird mit Identifikationsnummern des Patents gefüllt. Das DOC Feld ist mit der gesamten Patentbeschreibung belegt. Der Mandatory Run des Patent Tasks darf nur mit Hilfe des COMP Feldes durchgeführt werden. Die Relevanzbewertung dieses Tasks wurde von professionellen Patentfachleuten semi-automatisch durchgeführt. Der Task unterteilte sich in die beiden Subtasks Invalidity Search (Suche nach Patenten anhand von Patenten, die die Topics darstellen) und einer automatischen Patent Map Creation. An dem Patent Task nahmen im vierten Workshop 10 Gruppen teil.

Question Answering Der Question Answering Challenge umfasst, wie der TREC QA Track, die Aufgabe, anhand von Fragen, die als Topics formuliert sind, Antworten aus dem Datenkorpus zu generieren. Hierfür wurde das NTCIR-4QA Korpus verwendet, dass aus japanischen Zeitungsartikeln der Jahre 1998 bis 1999 zusammengestellt wurde. Das Korpus setzt sich aus 593.636 Dokumenten zusammen, die gemeinsam 776 MB Speichervolumen umfassen. Mit diesem Korpus als Grundlage wurden drei Subtasks entworfen, die jeweils eine bestimmte Anzahl von Topics zugeteilt bekamen:

- Subtask 1: Fünf mögliche Antworten (197 Topics)
- Subtask 2: Eine Liste mit allen Antworten (199 Topics)
- Subtask 3: Eine Serie von Fragen (251 Topics)

Die Topics wurden von den Topicentwicklern bewertet, indem für jede Frage eine Antwort formuliert wurde. Die Antworten mussten eine zweistufige Prüfungsphase durchlaufen, um dann als relevante Antwort für ein Topic verwendet zu werden. An dem QA Task der NTCIR-4 nahmen 18 Gruppen teil.

Text Summarization Der Text Summarization Task erwartet von einem System, dass es mittels einer Query (Topic) eine Themenvorgabe bekommt. Mit dieser Vorgabe wird die Datenkollektion durchsucht, worauf als Ergebnis eine Dokumentensammlung bzw. ein Text zu dem Querythema produziert wird. Dieser Task wird im NTCIR als Challenge bezeichnet, um die Komplexität von Anfang an zu verdeutlichen. Die Teilnehmer bekamen als Kollektion eine Sammlung von Zeitungsartikeln (NTCIR-4SUMM) sowie Themenanfragen in Form von Topics. Die Relevanzbewertung wurde durch das National Institute of Informatics (NII) überwacht. Im Vorfeld wurden die

nötigen Zusammenfassungen zusammengestellt und bewertet. Der Task hieß Multi Document Summarization und konnte in zwei verschiedenen Formen (kurz oder ausführlich) eingereicht werden. An dieser Challenge beteiligten sich neun Gruppen.

Web Retrieval Von dem TREC Web Track angeregt, wurde 2002 der NTCIR Web Task ins Leben gerufen. Ziel war es, neben dem Evaluieren von Web Retrieval Systemen eine größere Testkollektion als die .GOV Kollektion aufzubauen. Des Weiteren sollten neben der englischen auch die japanische Sprache mit in den Korpus einfließen. Das Korpus (NW100G-01) wurde für den dritten NTCIR WEB Task (2001) entwickelt und beinhaltet einen Webcrawl der japanischen .jp Topleveldomäne. Viele Dokumente, die sich nicht in der .jp Domäne befinden aber durch Inlinks der .jp Domäne in einem Zusammenhang mit ihr stehen, wurden ebenfalls mit in den Korpus aufgenommen. Dadurch liegen nicht nur Dokumente aus der japanischen Topleveldomäne im Korpus vor. Das Korpus besitzt ein Speichervolumen von ca. 100 GB mit ca.11 Mio. Dokumente. 11 Gruppen beteiligten sich an den Web Tasks des vierten Workshops.

	Statistics of NW100G-01	
1	# of crawled sites	97,561
2	max. # of pages within a site	1,300
3	# of crawled pages	11,038,720
4	# of pages for searching	15,364,404
5	# of links connected from 1 - 3	78,175,556
6	# of links connected from 1 - 3 to 1 - 4	64,365,554

Tab. 2.4: NW100G-01 Korpus der NTCIR-4WEB Kollektion [Egu04d]

Die Sprache des Korpus beschränkt sich hauptsächlich auf japanisch und englisch, wobei in der Beschreibung der Kollektion eingeräumt wurde, dass auch andere Sprachen vorkommen können. Die möglichen anderen Sprachen wurden laut Organisator für den Web Task nicht berücksichtigt [Egu04d]. Um die NTCIR-4WEB Testkollektion zu vervollständigen, wurden Topics generiert. Die allgemeine Struktur von NTCIR-4WEB Topics setzt sich aus den Feldern Title, Description, Narrative und RDOC (zum Topic gehörende relevante Dokumente) zusammen. Die Topics wurden den einzelnen Subtasks im Strukturaufbau angepasst. Die Relevanzbewertung wurde parallel zur Topicentwicklung durchgeführt und war von den den Subtasks gestellten Anforderungen abhängig. Alle relevanten Dokumente pro Topic wurden jedoch anhand ihrer Dokumenten-ID registriert und im RDOC Feld gespeichert. Beim

Einreichen eines Runs musste jede einzelne Gruppe 13 bestimmte Informationen bezüglich der Runs mit angeben. Diese Informationen waren für die Systemauswertung von ausschlaggebender Bedeutung und mussten deshalb vollständig ausgefüllt sein.

Topic Part verwendete Topicfelder

Query Method automatische oder interactive Queryverarbeitung

Query Unit Queryeinheit (Wort, einzelne Phrasen, N Gram etc.)

Query Expansions Techniken um Queries auszuweiten

Link Information Nutzung von Linkinformationen

Anchor Nutzung des indexierten Anchor Textes

IR Model verwendetes Retrievalmodell

Ranking Ranking Faktor um Gewichtungen zu berechnen

Index Unit Indexeinheit und Nutzung von Link und Tag Informationen im Index

Indexing Technique Indexierungsmethoden

Index Structure Indexstruktur

Filtering Methoden der Extraktion von relevanten Dokumenten

External Resources Nutzung externer Resourcen zur Indexierung, Gewichtung und Durchsuchung der Datenkorpora

Der NTCIR-4 WEB Task besteht aus den folgenden vier Subtasks:

Information Retrieval Task Der Information Retrieval Task wurde entwickelt, um die Effektivität und Performance von Suchmaschinen im Hinblick auf Dokumentrelevanzbewertung mit Hilfe von Inhalt und Linkstruktur zu testen. Basis dieses Tasks ist ein klassischer ad hoc Retrievaltask, wobei der Fokus auf der Bewertung und dem Nutzen von Hyperlinkstrukturen, sowie dem Eliminieren von doppelten Webdokumenten liegt. Das NW100g-01 Korpus wurde hier verwendet. Der Information Retrieval Task deckt zwei unterschiedliche Usermodelle ab: (i) Das erste Modell gibt einen User vor, der mittels einer Suchanfrage versucht, sein Informationsbedürfnis zu decken. (ii) Beim zweiten Modell hingegen erwartet der User von seiner Suchanfrage nur ein oder wenige relevante Dokumente innerhalb der ersten Positionen einer Ergebnisliste zurück. Der Information Retrieval Task kommt dem Topic Destillation Task des TREC Web Tracks (Abschnitt 2.1) sehr nah. Die Struktur der Information Retrieval Task

⟨TOPIC⟩
⟨NUM⟩0001⟨/NUM⟩
⟨TITLE CASE="c" RELAT="2-3"⟩offside, soccer, rule⟨/TITLE⟩
⟨DESC⟩ I want to find documents that explain the offside rule in soccer. ⟨/DESC⟩
⟨NARR⟩⟨BACK⟩ I want to know about the offside rule in soccer. ⟨/BACK⟩⟨TERM⟩ Offside is a foul committed by a member of the offense side. There are several patterns for situations in which the offside rule can be applied, and it is the most difficult soccer rule to understand. ⟨/TERM⟩⟨RELE⟩ Relevant documents must explain situations where the offside rule applies. ⟨/RELE⟩⟨/NARR⟩
⟨ALT0 CASE="b"⟩offside⟨/ALT0⟩
⟨ALT1 CASE="b"⟩offside, player, position⟨/ALT1⟩
⟨ALT2 CASE="b"⟩offside, soccer⟨/ALT2⟩
⟨ALT3 CASE="b"⟩soccer, offside, rule⟨/ALT3⟩
⟨USER⟩2nd year undergraduate student, male, 4 years of search experience, skill level 3, familiarity level 5⟨/USER⟩
⟨/TOPIC⟩

Abb. 2.2: NTCIR-4WEB IR Topic (englische Übersetzung) [Egu04b]

Topics setzte sich aus NUM (Topic Nummer), TITLE (3 Suchterme geordnet nach Wichtigkeit), DESC (kurze Erklärung der Suchanfrage), NARR (ausführliche Beschreibung des Suchanfragenhintergrunds, der verwendeten Terme und eines relevanten Dokuments) , ALT0, ALT1, ALT2, ALT3 und User - Informationen zu den Topicentwickler - Feldern zusammen. Das ALT0 Feld erfasst den ersten Term des TITLE Felds. Die ALT1 bis ALT3 Felder sind Information Retrieval Task spezifische Felder, in denen die Suchanfrage, unabhängig vom TITLE, von drei weiteren Usern neu formuliert wird.

Alle Topics wurden in japanischer Sprache entwickelt. Als Summe wurden 128 Topcis für den Task zugelassen. Die Relevanzbewertung dieser 128 Topics wurde von den Entwicklern intellektuell durchgeführt und basiert auf der *multi-grade* Struktur. Multi-grade Struktur bedeutet, dass es verschiedene Abstufungen von Relevanzen gibt. Diese Bewertung teilt sich in vier Bereiche auf.

- Höchst relevante Dokumente
- relevante Dokumente
- Weniger relevante Dokumente
- Irrelevante Dokumente

Die beiden Usermodelle Ui und Uii erforderten eine angepasste Evaluierung. 53 Topics wurden dem Ui Modell zugeordnet und mittels Precision und Recall Werten evaluiert. Zur Berechnung der Werte wurden die ersten 1000 Dokumente pro Topic berücksichtigt. Die restlichen 75 Topics gehörten zum Uii Modell und sind anhand des Weighted Reciprocal Rank (WRR) auf Basis der ersten 20 Dokumente pro Topic bewertet worden. Der Weighted

Reciprocal Rank ist eine Abänderung des MRR (1.7) und wird zur Bewertung von Topics mit multi-grade Relevanz verwendet. Dabei werden die MRR Werte pro einzelner Relevanzstufe ermittelt und einem vorher zugewiesenen Gewicht zugeordnet. Nach ermitteln dieser Werte wird der maximale Wert für den zu bewertenden Run ausgewählt. [Egu04b]

WRR = max. MRR der einzelnen Relevanzstufen

Navigational Retrieval Task Dieser Subtask basiert auf dem Szenario des *known item Search*. Der User sucht eine oder wenige relevanten Seiten zu einem Thema das er schon kennt bzw. von dem er schon gehört hat. Dieser Task ist mit den named page finding und homepage finding Tasks der TREC und CLEF Web Tracks vergleichbar. Als Ergebnis oder relevante Dokumente kommen nur eines bzw. eine sehr kleine Anzahl in Frage.

```
<TOPIC>
<NUM>Topic number</NUM>
<TYPE>Type code</TYPE>
<CATEGORY>Category code
</CATEGORY>
<TITLE>Search terms</TITLE>
<DESC>Search description sentence</DESC>
<NARR>
  <TERM>Explanation of terms (optional)
  </TERM>
  <BACK>Explanation of back ground
  (optional) </BACK>
  <RELE>Relevance criteria (optional)
  </RELE>
</NARR>
<USER SPECIALTY="Knowledge level
code">Attributes of searcher</USER>
</TOPIC>
```

Abb. 2.3: Navigational Retrieval Topicstruktur (englische Übersetzung)

Für den Navigational Retrieval Task kam das NW100g-01 Korpus erneut zum Einsatz. 300 Topics wurden für den vierten NTCIR Workshop entwickelt. Aufbau bzw. Struktur eines Navigational Retrieval Topic gleicht dem NTCIR Standard Topic bis auf die zwei zusätzlichen Felder: TYPE und CATEGORY. Im TYPE Feld wird die Anzahl und Kombination der Suchterme im TITLE näher spezifiziert. Im CATEGORY Feld wird durch den Topicentwickler der Themenbereich des Topics eingetragen. Anhand des CATEGORY Feldes können in den teilnehmenden Runs domänenabhängige bzw. themenorientierte Queries verfasst werden. Teilnehmenden Gruppen war erlaubt, vier verschiedene

Kombinationen von Topicfeldern für die Querygenerierung zu nutzen:

1. TITLE Feld (Mandatory Run)
2. jede mögliche Kombination aus TITLE, DESC und NARR/BACK
3. jede mögliche Kombination aus TYPE, CATEGORY und TITLE
4. jede mögliche Kombination aus TYPE, CATEGORY und 2.

Die Relevanzbewertung der Topics verlief während der Entwicklungsphase zur Hälfte parallel. Die zweite Hälfte musste später nachgereicht werden. Im Bewertungsprozess sollten neben den aktuell gefundenen Dokumenten auch die verlinkten Seiten auf Relevanz geprüft werden. Mit diesem Ansatz einigte man sich auf eine mehrstufige Bewertung von Dokumenten. Dokumente im Navigational Retrieval Task konnten entweder als *relevant, partially relevant* oder *non relevant* bewertet werden. Partially relevante Seiten sind Dokumente, die entweder einen Link zu einer relevanten Seite beherbergen oder als Ergänzungsseite einer relevanten Seite gelten. Als Bewertungsmaß wurde im Navigational Retrieval Task der Weighted Reciprocal Rank verwendet. [Egu04c]

Geographic Information Task Der Geographic Information Task ist für die unterschiedlichen Web Tracks einzigartig und zielt auf das konkrete Suchen von geographischen Informationen innerhalb relevanter Webdokumente mittels allgemein gehaltenen Suchanfragen. Das ideale GeoTask System würde auf Grundlage der Query *Universität Tokio*: Adressen, Anfahrtsskizzen, Karten und weitere relevante Informationen zum Standort liefern. Um diesen Task durchzuführen, wurde ein neuer Korpus auf Basis des NW100g-01 entwickelt. Das Korpus musste sich erheblich verkleinern, um die Probleme großer Korpora zu vermeiden und den wahren Fokus dieses Tasks zu fördern. Zusätzlich sollten die Topics des Tasks sich nur im Großraum Tokio bewegen und die Dokumente mussten mindestens ein Schlüsselwort des Topics beinhalten. Mit dieser Herangehensweise konnte die Entwicklung von Extraktionstechniken zur Sammlung von geografischen Informationen beschleunigt werden. Das erstellte GeoTASK Korpus umfasste 240.000 Dokumente. Die Topics für diesen Task waren als Fragen formuliert. Beim aktiven Durchführen des vierten NTCIR Workshops stellte sich heraus, dass auf Seiten der teilnehmenden Gruppen kein Interesse am Geographic Information Task vorlag, sodass Erfahrungen bzw. Evaluation Ergebnisse nicht näher beschrieben werden können. [Ari04]

Topical Classification Task Der Topical Classification Task evaluiert die unterschiedlichen Methoden Usern ein automatisch klassifiziertes Ergebnis zu liefern.

Dies soll dem User die kognitive Überlastung beim Browsingprozess ersparen. Hierbei sind Techniken des Clusterings und der Klassifikation anzuwenden. Als Korpus wurde ein weiteres Mal der NW100g-01 Korpus benutzt. Von Seiten der Organisatoren wurde keine Klassifikation vorgegeben, und es wurden auch keine Grenzen für Ergebnislisten der einzelnen Runs gesetzt. Die Ergebnislisten sollten allerdings pro Klasse bzw. Kategorie eine nach Relevanz geordnete Liste von Dokumenten beinhalten. Um die Testkollektion für den Topical Classification Task zu vervollständigen, wurden die 47 ersten Topics des Information Retrieval Tasks verwendet. Vorteile dieses Ansatzes sind die im Nachhinein möglichen Auswertungen der einzelnen Retrievalstrategien. Die Relevanzbewertung der Topics konnte im Allgemeinen von den erneut verwendeten Topics übernommen werden. Für relevant wurden nur japanische oder englische Dokumente bewertet. Alle weiteren Sprachen wurden vernachlässigt. Die vierstufige Bewertung des Information Retrieval Task wurde übernommen. Die Hauptevaluierung der einzelnen Runs wurde anhand von Precison und Recall Werten durchgeführt. Die Besonderheit, dass die Ergebnisliste neu entstandene Klassen bzw. Kategorien enthält, musste ebenfalls evaluiert werden. Dies geschah mit sehr hohem Zeitaufwand, wobei den Assessors die klassifizierten Ergebnislisten vorgelegt wurden. Als erster Schritt musste untersucht werden ob die einzelnen Klassifikationen dem Topicthema entsprachen. Zusätzlich wurden die Klassen mit der höchsten Anzahl an relevanten Dokumenten am höchsten gewichtet, um dann auf einen Durchschnittswert zu kommen. Pro Klasse wurden jeweils die ersten 20 Dokumente bewertet bzw. evaluiert. Im Voraus wurden keine möglichen Klassifikationen entwickelt, da man die Ergebnisse neutral bewerten wollte. Daher wurden die Klassifikationen nicht in Frage gestellt. Entscheidender Faktor für eine Klasse war lediglich die Anzahl relevanter Dokumente. [Egu04a]

Es nahmen 74 Gruppen aktiv an dem vierten NTCIR Workshop teil. Insgesamt kamen alle Gruppen aus zehn Ländern. Die sprachliche Komponente dieses Workshops wird in den meisten Papers als Bilingual beschrieben. Wichtig zu bemerken ist, dass der englische Anteil des Korpus unter 10 % lag. Die Topics wurden allerdings zu 100% in Englisch übersetzt. Hauptakteure kamen aus Japan, China, Korea und Singapur. Die Internationale Beteiligung gab es nur im CLIR Task. Die weiteren Tasks schienen für den nicht-asiatischen Raum aktuell nicht relevant zu sein. Wie die internationale Beteiligung im fünften NTCIR Workshop ausfällt, bleibt abzuwarten.

2.2.1 Systeme und Ergebnisse des NTCIR-4 Web Task

Der NTCIR-4WEB Task ist im Vergleich zu den anderen Web Tracks der umfangreichste Track. 145 Runs wurden eingereicht, wobei der größte Anteil von Seiten der Organisation vorgelegt wurde. Über die Vorgehensweise und Parameter können keine großen Gemeinsamkeiten unter den einzelnen Systemen entdeckt werden. Die Indexierungstechniken gehen von morphologischen Analysen über Normalization Prozesse bis hin zu verschiedenen N-Gram Ansätzen. Das Gleiche gilt für die unterschiedlich genutzten Retrieval- und Ranking-Modelle. Dies lässt darauf schließen, dass für japanisches Web Retrieval noch keine optimalen Systemkonfigurationen gefunden wurden. In der Tabelle 2.5 werden die unterschiedlichen Strategien zusammengefasst. Es werden jeweils die besten drei teilnehmenden Gruppen der einzelnen Subtasks mit ihren besten Runs dargestellt.

2.3 SEWM 2004

Im Jahr 2002 begann die Initiative zur Evaluierung chinesischer Information Retrieval Systeme. Organisatoren waren die Chinese Language Computing Society, Asian Federation of Natural Langauge Processing (AFNLP) und die Universität Peking. Grundlage für die Entwicklung eines chinesischen Web Tracks sind die TREC und NTCIR Workshops. SEWM nimmt für sich die TREC Ziele in Anspruch (Abschnitt 2.1). SEWM ist ein monolingualer Web Track. Als Voraussetzung für diesen Track wurde ein chinesischer Web-Korpus (CWT100g) erstellt. Die chinesische Tianwang Internet Suchmaschine stellte das Korpus zusammen. Das Korpus besteht aus 1.000.614 Websites die insgesamt 5.712.710 Dokumente umfassten. Das Speichervolumen dieses Crawls beträgt ca. 90 GB und beinhaltet ausschließlich HTML und plain Text Dateien.

Die Aufteilung eines CWT100g Dokuments ist ähnlich aufgebaut wie die TREC und WebCLEF Dokumente. Die Informationen URL, Datum, Content Type und Dokument ID sind im Dokument enthalten. Als Besonderheit gegenüber anderen Korpora gilt die Angabe von Dokumentlänge und Größe. Die SEWM Initiative [SEW05a] betreibt auch eine Homepage, die Informationen, Richtlinien zur Teilnahme und Evaluationsergebnisse anbietet. Da alle Seiten in chinesisch sind, ist eine detaillierte Beschreibung des chinesischen Web Tracks nicht möglich. Es wird eine sehr komprimierte Internetseite [SEW05b] für den SEWM 2004 Workshop auf Englisch angeboten.

IR Runs	DBLAB-tt-01	GRACE-tt-02	sstut-tt-02
a-prec RL1	0,2189	0,1985	0,1439
a-prec RL2	0,2438	0.2164	0,1672
STask	IR	IR	IR
TFeld	Ti & Content	Ti	Ti
IRModel	prob. model	prob. model	OKAPI, BM25
Ranking	OKAPI	OKAPI	OKAPI tf-idf, doc length
IUnit	Word & phrase	Character	segmented words
ITechnique	morph., POS, norm.	Word form norm.	morph. Analy. Chasen
Anc	no	no	no
Lnk	yes	yes	yes
NR Runs	ORGREF-AT40-P1	K3100-tt-02	TKB-01
a-WRR RL1	0,4750	0,4418	0,1742
a-WRR RL2	0,5614	0.5556	0,3496
STask	NR	NR	NR
TFeld	Ti	Ti	Ti
IRModel	boolean	unknown	OKAPI
Ranking	tf-idf	tf-idf	BM25
IUnit	Character	n-gram	word, bi-word
ITechnique	Character norm.	n-gram	morphology
Anc	yes	yes	no
Lnk	yes	yes	yes
TC Runs	ELRG-01	METAL-01	SRLAB-01
a-prec RL1	0,2454	0,3604	0,2167
a-prec RL2	0,1279	0.3011	0,2090
STask	TC	TC	TC
KL Model	clust. on term co-occurrence	cont. based clust.	bag of words model
IRModel	boolean	unknown	OKAPI
Ranking	meta search engine	meta search engine	probab. of the class given doc
IUnit	word, bi-word, tri-word	word	word
ITechnique	morphology	morphology	max. extension index
Lnk	no	no	no

Tab. 2.5: Die drei besten Runs des Information Retrieval Tasks NTCIR-4WEB
a-prec RL1: durchschnittl. Precision auf Basis der ersten 100 Dokumente (Rigid Level);
a-prec RL2: durchschnittl. Precision auf Basis der ersten 1000 Dokumente (Relaxed Level);
STask: Subtask des NTCIR-4WEB (IR: Information Retrieval Task, NR: Navigational Retrieval Task, TC: Topic Classification Task);
TFeld: Topicfeld (Ti: Title);
IRModel: Information Retrieval Modell;
IUnit: Index Einheit;
ITechnique: Indexierungsmethoden (morphologische Analyse, Part of Speech Tagging, Normalization);
Anc: Indexieren des Anchortextes;
Lnk: Indexieren der Linkinformationen
[Egu04d, Egu04c, Egu04a]

```
version: 1.0
url: http://ffff.363.net/
date: Fri, 04 Jun 2004 14:47:03 GMT
ip: 202.102.16.24
length: 1666

HTTP/1.1 200 OK
Date: Fri, 04 Jun 2004 02:46:13 GMT
Server: Apache/1.3.27 (Unix) PHP/4.2.3
Last-Modified: Sat, 12 Jul 2003 23:28:41 GMT
ETag: "27d19b-55b-3f1099a9"
Accept-Ranges: bytes
Content-Length: 1371
Keep-Alive: timeout=4, max=100
Connection: Keep-Alive
Content-Type: text/html

<html>

<head>
<meta http-equiv="Content-Type" content="text/html; charset=gb2312">
<title>¾²Ìì¼ø¼¼</title>
</head>

<body bgcolor="#77B7F7">
Document Content
```

Abb. 2.4: Beispiel eines CWT100g Dokuments

2.4 Fazit und Vergleich zu WebCLEF

Wenn man die in diesem Kapitel beschriebenen Kollektionen mit der WebCLEF Kollektion vergleicht, wird schnell die Einzigartigkeit der WebCLEF Kollektion deutlich. Die WebCLEF Kollektion ist die erste multilinguale Testkollektion. Sie besitzt neben der .GOV Kollektion aus dem Jahr 2002 den zweitkleinsten Korpus. Der Fokus der EuroGOV Kollektion ist, wie im Anfangszitat des ersten Kapitels deutlich wird, das Schaffen einer Evaluationsplattform für eine europäische Suchmaschine. Dieses Vorhaben konzentriert sich mehr auf die sprachliche Komponente im Unterschied zu den anderen hier beschriebenen Web Tracks. Mit dem NW100g Korpus gibt es zwar die Gelegenheit, bilinguale Systeme zu testen, der Fokus ist hier jedoch nicht auf eine sprachlich ausgeglichene Kollektion, sondern auf das Arbeiten mit japanischen Webdokumenten ausgerichtet. Neben den unterschiedlichen Korpora fallen insbesondere die große Vielfalt von Retrievaltasks auf. Hier zeigt der NTCIR Workshop die größte Auswahl. Mit den einzelnen Tasks treten unterschiedliche Evaluationsmethoden der Relevanzbewertung von Dokumenten und die

Tracks	Kollektion	Subtasks	Seitenanzahl	Größe	Sprachen
TREC Web Track	GOV	3	1,3 Mio	18,1GB	englisch
TREC Terabyte Track	GOV2	1	25 Mio	426GB	englisch
NTCIR-4WEB	NW100g-01	4	11 Mio	100GB	japanisch & englisch
SEWM 2004	CWT100g		5,7 Mio	90GB	chinesisch
WebCLEF	EuroGOV	3	3,6 Mio	82GB	$<$ 20 Sprachen aus EU

Tab. 2.6: Unterschiede der einzelnen Testkollektionen

einhergehende Auswertung der eingereichten Runs auf. Beim Lesen der Workshopproceedings entsteht der Eindruck, dass die Methoden sich aufgrund vergangener Erfahrungen entwickelt haben. So sind die Bewertungsansätze des WebCLEF Tasks nicht annähernd so detailliert wie die der NTCIR bzw. TREC Initiativen. Dies ist auf die im Vergleich nicht so umfangreichen Erfahrungen mit Web Tracks der WebCLEF Organisatoren zurückzuführen. Des Weiteren wird bei der Auswertung der unterschiedlichen Systeme deutlich, dass die Gruppen des TREC Web Tracks in den meisten Fällen sehr ähnliche Strategien im Bereich des Indexierens und der Wahl der Information Retrieval Systeme wählen. Beim näheren Betrachten des Tracks ist eine gewisse Ausgereiftheit der Systeme zu erkennen. Dies äußert sich im Vergleich der einzelnen Ergebnisse, die anhand der Teilnehmergruppen erzielt wurden. TREC Web Track Systeme schneiden mit durchschnittlich abschneidenden Systemen besser ab als gute NTCIR-4WEB Systeme. Die Ausgereiftheit des TREC Web Tracks und seiner Systeme scheint auch der Grund für das Einstellen dieses Tasks zu sein. Als alternativer Track wurde der Terabyte Track ins Leben gerufen. Diese Entwicklung ist im NTCIR-4WEB Track noch nicht zu beobachten. Ob sich die verwendeten Systeme im WebCLEF Track ähneln oder große Unterschiede auftreten werden, ist noch nicht bekannt. Anhand von Erfahrungen, die während der Teilnahme des WebCLEF Tracks und dem Untersuchen der Einzelheiten der weiteren Web Tracks gemacht wurden, können Verbesserungsvorschläge für den zweiten Durchgang des WebCLEF Tracks gesammelt werden. Die Erkenntnisse sind im letzten Kapitel (5) zusammengefasst.

Kapitel 3

Apache Lucene 1.4 Software

Die Softwarebasis aller WebCLEF Experimente des Instituts für Angewandte Sprachwissenschaft (IFAS) der Universität Hildesheim ist das Open-Source Projekt Lucene 1.4 [Luc05]. Lucene ist eine Open-Source-Java-Bibliothek zum Erzeugen und Durchsuchen von Indizes. Anhand dieser Klassenbibliothek wurden eigene Java Klassen entwickelt, um den Bedürfnissen des WebCLEF Tracks zu entsprechen.

> *Apache Lucene is a high-performance, full-featured text search engine library written entirely in Java. It is a technology suitable for nearly any application that requires full-text search, especially cross-platform. [Luc05]*

Lucene zeichnet sich durch hohe Performance und Skalierbarkeit bei beliebiger Projektgröße und beliebigen Anforderungen aus. Da die WebCLEF Experimente aufgrund der Kollektionsgrösse einen hohen Anspruch an Performance haben, bietet sich die Verwendung dieser Java Bibliothek an. Des Weiteren basieren die meisten Erfahrungen von Retrieval Experimenten an der Universität Hildesheim auf der Verwendung von Lucene Klassen [Hac03, Hac04, Hac05c]. Die Bibliothek setzt sich aus zwei Hauptbestandteilen zusammen, die gleichzeitig auch die Lucene eigenen Abläufe beschreiben:

1. Eine Komponente erzeugt den Index, wobei diesem beliebige, aber definierte Dokumente hinzugefügt werden.

2. Eine Query Engine durchsucht diesen Index.

Die Verwendung und Konzepte dieser beiden Komponenten der Lucene Klassenbibliothek werden in diesem Kapitel näher beschrieben.

3.1 Indexieren mit Lucene 1.4

Die erste Hauptkomponente der Lucene 1.4 Java Bibliothek ist das Indexieren von beliebig großen Dokumentkollektionen. Lucene 1.4 beinhaltet eine Sammlung von Klassen und Paketen, die das Indexieren steuert und die einzelnen Abläufe koordiniert. Lucene liefert allerdings nur eine begrenzte Anzahl an Analyse Klassen. Der Schwerpunkt von Lucene liegt im Steuern der Indexierungsprozesse. [Luc05] Das Grundkonzept des Lucene Index besteht aus Index, Dokument, Feld und Ausdruck. Ein Index enthält eine Folge von Dokumenten, ein Dokument ist eine Folge von Feldern, ein Feld ist eine benannte Folge von Ausdrücken und ein Ausdruck ist eine Zeichenkette in Form eines Strings. Dieses Konzept würde für einen EuroGOV Index wie folgt aussehen:

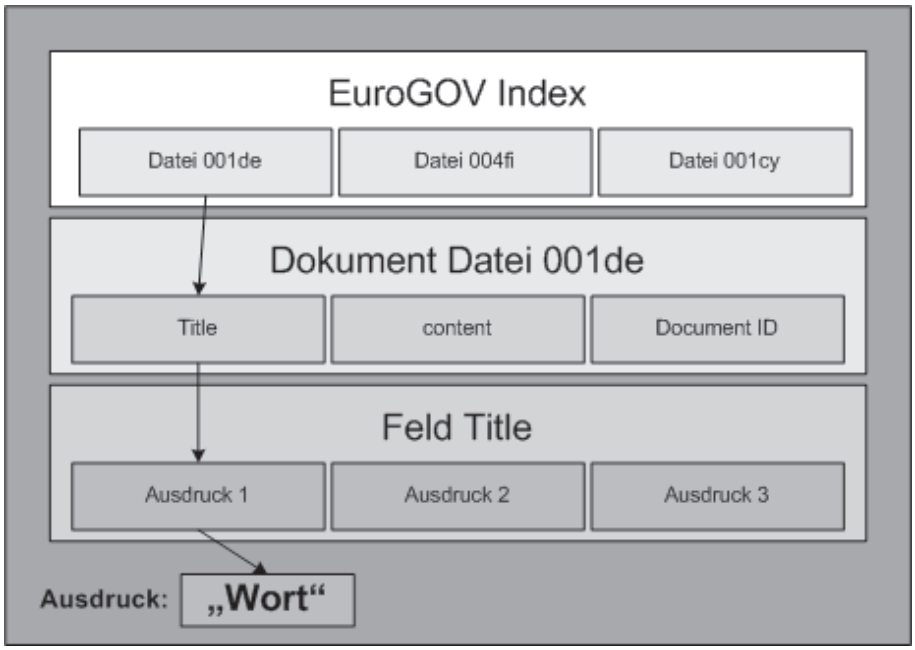


Abb. 3.1: Aufbau eines Lucene Index am Beispiel des EuroGOV Korpus

Der Indexierungsprozess nimmt einen durchschnittlichen Zeitaufwand von 20MB[1] pro Minute in Anspruch, wobei das Erweitern eines Indexes bzw. das Erstellen eines neuen Indexes den gleichen Zeitaufwand beinhaltet. Die Geschwindigkeit bei der Erstellung eines Indexes ist immer abhängig von der Rechenleistung (CPU), der Größe des Arbeitsspeichers (RAM) und dem zur Verfügung stehenden Festplattenplatz. Insgesamt kann gesagt werden, dass ein geringer Arbeitsspeicherbedarf beim Erstellen bzw. Erweitern eines Indexes besteht. Die Indexgröße nimmt ungefähr 30% des ursprünglichen Korpus ein. Anhand des oben dargestellten Konzepts (Abb. 3.1) wird deutlich, warum wohlgeformtes XML

[1] Diese Werte gelten für einen Pentium M 1,5GHz Prozessor [Luc05]

hervorragend zum Indexieren mit Lucene geeignet ist. Mit der Fähigkeit mehrere Indexfelder pro Dokument zu erstellen, erleichtert XML das Ansprechen bzw. Begrenzen von Feldern während des Indexierungsvorgangs und das Steuern von verschiedenen Indexierungsmethoden. Lucene bietet die Möglichkeit, invertierte Indizes zu erstellen. Bei einem invertierten Index werden für jeden Ausdruck die Dokumente, die es enthält, aufgelistet. Damit werden einzelne Statistiken pro Ausdruck gespeichert. Mit diesen Statistiken kann die darauf folgende Suche effizienter gestaltet werden, da anhand von diesen gespeicherten Statistiken bzw. Verweisen nicht der ganze Index durchsucht werden muss. Im Gegensatz dazu speichert ein nicht invertierter Index keine Statistiken oder Verweise von Ausdrücken. Felder in einem Lucene Index können auf zwei Arten gespeichert bzw. erstellt werden. Entweder wird ein Feld buchstäblich gespeichert (stored) oder es wird in Ausdrücke aufgeteilt (tokenized). Ein Feld, das mit einem Text buchstäblich gefüllt wird, wird nicht invertiert abgespeichert. Die meisten Felder eines Lucene Indexes sind ”tokenized”, also invertiert und in einzelnen Ausdrücken abgespeichert. Bei bestimmten Bezeichnungsfeldern, wie z.B. einem ID Feld, ist es besser buchstäblich zu indizieren, da diese Felder Schlüsselfunktionen übernehmen und komplett indexiert werden müssen, ohne bestimmten Vereinfachungen oder Reduktionen unterzogen zu werden. Sowohl den Indexierungsprozess, als auch für das Endprodukt Index hat Lucene eigene plattformunabhängige Dateiformate entwickelt. Hierbei handelt es sich - nach Beendigung der Indexierung - um die Formate segments, deletable und .cfs.

- Die *segments* Datei beinhaltet den eigentlichen Index in Form von mehreren Subindizes. Diese Subindizes basieren auf den einzelnen im Voraus bestimmten Indexfeldern. Jeder Subindex eines Segments beinhaltet den Feldnamen, gespeicherte Feldwerte (Statistik des invertierten Index), die Liste aller gespeicherten Ausdrücke, die Termfrequenzen aller vorkommenden Ausdrücke, Termpositionen in den vorhandenen Dokumenten, Normalisierungsfaktoren pro Feld zur Gewichtung der Retrievalergebnisse und die Termvektoren pro Ausdruck bzw. Term in jedem vorkommenden Feld. Termvektoren werden beim Prozess der Indexierung nicht standardmäßig erstellt, sondern müssen im Quellcode aktiviert werden. Im Quellcode des Lucene IndexWriters würde die Funktion zur Berechnung von Termvektoren mit den Parametern stored bzw. unstored aktiviert werden.

  public static final Field UnStored(String name, String value)
  bzw.
  public static final Field Stored(String name, String value) [Luc05]

- Die *deletable* Datei gibt an welche Dateien während des Indexierens eingelesen, indexiert, der Segments Datei hinzugefügt und dann anschließend gelöscht wurden.
- Die *.cfs* Datei speichert die Metadaten des Indexierungsprozesses. Sie beinhaltet die Werte Anzahl der Dokumente, Dokumentnamen und Rohdaten der Dokumente.

Die drei erwähnten Dateiformate sind die Ergebnisdateien eines Lucene Indexierungsprozesses. Durch die Fähigkeit, mehrere Informationsbereiche innerhalb einer Datei zu speichern, nennt man die oben beschriebenen Dateien auch Container Dateien [Luc05]. Alle drei Container Dateien entstehen während des Indexierens aus vielen verschiedenen Dateien, die zum Speichern der Informationen und Aktionen benötigt werden. Am Ende eines Indexiervorgangs liegen die drei oben erwähnten Dateien vor. Die Dateien, die während des Indexierens entstehen, werden "Pre-Index" Dateien genannt und haben folgende Formate:

Segmentdateien Diese Dateien speichern die aktiven Segmente des Index:

- Feldnamen (.fnm)
- Feldwerte (.fdx)
- Liste aller gespeicherten Ausdrücke (.tis, .tii)
- Termfrequenzen (.frq)
- Termpositionen (.prx)
- Normalisierungsfaktoren (.nrm)
- Termvektoren (.tvx, .tvd, .tvf)

Blockierdateien Bestimmte Dateien werden benutzt um anzuzeigen, dass ein anderer Prozess auf den Index zugreift.

- commit.lock: Ein anderer Prozess greift auf die "segments" Datei zu.
- index.lock: Ein anderer Prozess fügt Dateien zum Index hinzu oder entfernt welche.

Löschdateien Die Datei "deletable" gibt die Dateien an, die der Index nicht mehr benutzt, die aber nicht gelöscht werden können, da die Datei z.B. gerade geöffnet ist.

Um mit den Lucene eigenen Dateien arbeiten zu können und in der Lage zu sein, die generierten Indizes zu kontrollieren, wurde von Andrzey Bialeki [Bia05] ein Tool entwickelt, um schnell und einfach auf den Lucene Index zugreifen zu können. Dieses Tool heißt Luke

- Lucene Index Toolbox [Bia05]. Zusammengefasst ist Luke ein praktisches Entwicklungs- und Diagnose Tool, das dabei behilflich ist schon existierende Lucene Indizes darzustellen und den Inhalt auf verschiedenen Wegen zu bearbeiten. Luke befähigt Programmierer

- Dokumente anhand von Termen oder vergebenen Dokument IDs zu durchsuchen,
- Dokumente anzuschauen und in die Zwischenablage zu kopieren,
- eine gewichtete Liste aller vorkommenden Terme zu erstellen,
- eine Suchanfrage zu starten und die Ergebnisse zu analysieren und zu evaluieren,
- ausgewählte Dokumente aus dem Index zu entfernen und
- die Indexfelder zu bearbeiten und den Index zu optimieren.

Das Erstellen von Subindizes anhand der von Lucene zur Verfügung gestellten Klassen bietet eine Reihe von Vorteilen. Lucene ermöglicht ebenfalls das Zugreifen von mehreren unterschiedlich erstellten Indizes gleichzeitig. Mit diesen Eigenschaften ist es möglich, anhand einer Query auf mehrere Indizes zuzugreifen. Dieses Szenario ist bei der Verarbeitung von multilingualen Testkollektionen sehr verbreitet. Mit Lucene können sprachspezifische Indexiermethoden benutzt werden, wie dies in den CLEF 2005 ad-hoc multi-lingual Retrieval Track des IFAS [Hac05b] umgesetzt wurde. Die Lucene 1.4 Bibliothek beherbergt Klassen zur Indexierung der Sprachen Deutsch und Englisch. Weitere Indexiermethoden für andere Sprachen sind ohne großen Umfang in die Lucene Bibliothek implementierbar bzw. befinden sich in der Lucene Sandbox [Luc05].

Die Generierung eines Indexes wird anhand der Lucene IndexWriter Klasse koordiniert bzw. initialisiert. Der IndexWriter steuert die Tätigkeiten während des Indexierungsprozesses zwischen dem entstehenden Index und dem zu indexierenden Korpus. Er koordiniert den Ablauf durch das Aufrufen und Löschen der oben beschriebenen "Pre-Index" Dateien. Um einen neuen Index zu erstellen oder einen bestehenden Index zu erweitern müssen dem IndexWriter die Parameter Indexverzeichnis, Analyzer, Stoppwortliste und ein Boolscher Operator (true bzw. false) zugewiesen werden. In der IndexEuroGOV[2] Klasse wurden diese Parameter wie folgt initialisiert:

IndexWriter writer = (new Indexwriter(directory, analyzer, true));

[2] siehe DVD/IndexEuroGOV

indexDocs(writer, corpus directory); [Luc05]

Um den Lucene Indexierungsprozess zu starten, muss im Voraus ein neuer IndexWriter initialisiert werden. Dem IndexWriter *writer* wird das Verzeichnis genannt, in dem der zu erstellende Index gespeichert, ein Analyzer zum indexieren übergeben und mittels des Boolschen Operators true die Aufgabe einen neuen Index zu erstellen übertragen. Der Prozess wird dann mit Hilfe der indexDocs Methode gestartet. Dieser Methode werden der neu initialisierte IndexWriter und der zu indexierende Korpus übergeben.

3.2 Lucene 1.4 Query Engine

Die Lucene Query Engine ist die zweite Hauptkomponente der Lucene 1.4 Java Bibliothek. Sie beinhaltet die Verarbeitung von Queries, die Steuerelemente des Retrievalprozesses und die Fähigkeit, Ergebnislisten geordnet und gewichtet auszugeben. Folgende Eigenschaften sind in der Lucene Query Engine enthalten [Luc05]:

geordnete Suche Die besten Ergebnisse werden als erstes ausgegeben.

viele verschiedene Suchstrategien Die Lucene Query Engine beinhaltet standardmäßig sechs verschiedene Suchansätze, die durch eigene Syntax gebraucht werden.

Feldsuche Queries sind in der Lage, einzelne Felder des Indexes anzusprechen. Diese Felder können ebenfalls gewichtet werden.

Suche auf mehreren Indizes Es können mit einer Query mehrere Indizes gleichzeitig angesprochen werden. Die Ergebnisse jedes Indexes müssen jedoch separat fusioniert werden.

Lucene unterstützt standardmäßig reichhaltige Suchoptionen. Allerdings besteht auch die Möglichkeit, eine eigene Suchsyntax zu entwickeln. Eine Suchanfrage wird in einzelne Ausdrücke und Operatoren unterteilt. Ausdrücke bestehen aus einzelnen Wörtern. Eine Phrase ist eine Gruppe von Wörtern, die für die Query Engine in Anführungsstriche gesetzt werden muss. Mehrere Ausdrücke bzw. Phrasen können durch Operatoren mit einander verbunden werden. Lucene unterstützt das Suchen auf einzelnen Feldern wie z.B. Dokumenttitel und Inhalt. Dies erleichtert die Verarbeitung von Queries im XML Format und hilft beim Gewichten bzw. einzelnen Verwenden von Queryfeldern. Lucene beherbergt standardmäßig folgende Suchstrategien:

Wildcards Wildcards sind Platzhalter für ein oder mehrere Zeichen in einer Query. Dargestellt werden diese Platzhalter als Dollarzeichen $ und bei mehrfacher Verwendung von Zeichen mittels eines Sterns *. Wildcards dürfen nur innerhalb oder am Ende eines Ausdrucks verwendet werden.

Fuzzy Suche Beinhaltet das Suchen mittels einer undeutlichen Query. Hierbei wird am Ende eines einzelnen Ausdrucks eine Tilde gesetzt, um die Undeutlichkeit zu deklarieren. Aufgelöst bzw. verarbeitet wird die Undeutlichkeit mittels des Levenshteinschen Distanz Algorithmus [Lev65]. Der Algorithmus berechnet den Unterschied zwischen zwei Zeichenketten, um im Falle einer Fuzzy Suche den nächstmöglichen Term zur Vervollständigung oder Verbesserung des Queryterms zu finden. Ausdrücke, die über die Fuzzy Suche in der Lucene Query Engine gesucht werden, werden automatisch mit einem Verstärkungsfaktor von 0,2 gewichtet.

Distanzsuche Bei der Distanzsuche dürfen bestimmte Ausdrücke eine festgelegte maximale Distanz zueinander aufweisen. Anhand einer Tilde - gefolgt von der maximalen Anzahl an Ausdrücken die folgen dürfen - kann diese Distanzsuche an Lucene übergeben werden.

Verstärkungsfaktoren Lucene bietet die Möglichkeit, einzelnen Ausdrücken eine Gewichtung innerhalb der Query zu geben. Zugewiesen werden die Gewichtungen am Ende eines Ausdrucks nach einem ^ Zeichen. Der Standardfaktor ist 1, der Verstärkungsfaktor muss immer positiv sein, kann aber kleiner als 1 ausfallen.

Boolesche Operatoren Des Weiteren können die gängigen Boolschen Operatoren Oder (OR), Und (AND), Und Oder (AND OR) sowie negierende Operatoren (NOT) verwendet werden.

Gruppierung von Elementen Elemente bzw. Ausdrücke können mittels Klammern gruppiert und mit Boolschen Operatoren verknüpft werden.

Syntaxzeichen ausschließen Syntaxzeichen, die in den vorangegangenen Lucene Syntax Handhabungen erwähnt wurden, können durch einen vorgesetzten Backslash auch direkt für die Query verarbeitet werden.

Die Query Engine arbeitet nach einem strukturierten Ablauf. Alle Java Klassen, die für die Query Engine benötigt werden, befinden sich in den Paketen Search und QueryParser (org.apache.lucene.search und org.apache.lucene.queryparser) der Lucene 1.4 Bibliothek. Lucene Queries bestehen aus einzelnen Ausdrücken bzw. Phrasen und

Operatoren. Zuerst wird mit Hilfe der *IndexReader* Klasse der vorhandene Lucene Index eingelesen und an die *IndexSearcher* Klasse übergeben. Parallel dazu wird die Query generiert. Dies geschieht mit Hilfe des *Query* Objekts. Diesem Objekt wird das Verzeichnis zu den Queries bzw. Topics zugewiesen und mit Hilfe eines Analyzers indexiert. Hierbei ist wichtig, dass sich Index und Queryindex in ihrer Indexierungseinheit (n-Gram, Wort, etc.) gleichen, deswegen müssen die identischen Indexiermethoden gewählt werden. Die generierten Queries und der eingelesene Index werden durch die *search* Methode im *Hits* Objekt zusammen geführt. Mit diesem Ausdruck beginnt die Lucene Query Engine zu arbeiten und durchsucht den Index mit Hilfe des Queryindex.

IndexReader reader = IndexReader.open(index directory);
IndexSearcher searcher = new IndexSearcher(reader);
Query query = new TermQuery(new Term(topic directory, analyzer));
Hits hits = searcher.search(query); [Luc05]

Welche von den oben beschriebenen Suchstrategien angewendet werden, muss bereits beim Generieren der Query berücksichtigt worden sein. Die Ergebnisse eines Lucene Retrievalprozesses werden, wie am Anfang dieses Abschnitts beschrieben, in einer geordneten Liste ausgegeben. Die Lucene Ordnung einer Liste basiert auf einer internen Punktevergabe, die beim Ordnen der Ergebnisse die Bedingung für die Reihenfolge aller Ergebnisse liefert. Wie diese Ergebnisliste verwendet wird, ist bedarfsabhängig und damit auch offen gestaltet. Die Lucene Query Engine wird in der WebCLEFSearch[3] Klasse der IFAS WebCLEF Experimente eingesetzt.

[3] siehe DVD/WebCLEFSearch

Kapitel 4

Webretrieval Experimente der Universität Hildesheim im WebCLEF 2005 Kontext

4.1 Vorgehensweise und Ziel

Im Jahr 2005 startete der erste Web Track der CLEF Initiative. Die Besonderheit dieses Web Tracks ist im Vergleich zu anderen Tracks der cross-linguale Ansatz. Die Testkollektion EuroGOV ist somit einzigartig und birgt viele Herausforderungen im Rahmen der Cross-Lingualität und des Webretrievals. Mit den Erfahrungen aus den vergangenen CLEF Jahrgängen und der Motivation, mehr Erkenntnisse über das Verarbeiten von großen Testkollektionen zu sammeln, entschloss sich das IFAS der Universität Hildesheim an dem diesjährigen WebCLEF Durchgang teilzunehmen. Die praktische Teilnahme verlief im Rahmen dieser Magisterarbeit. Dieses Kapitel soll die Vorgehensweise dokumentieren, die für die Teilnahme an den Experimente im Zeitraum vom 1. April 2005 bis zum 20. Juni 2005 vorgesehen wurde, und Gründe für das Vorgehen näher erklären. Für das IFAS war die Teilnahme an einem Web Track Premiere und somit war die Zielvorgabe für den WebCLEF Durchgang 2005 eine erfolgreiche Teilnahme mit mindestens einem *Multilingual Run* und einem *Mixed Monolingual Run* (Abschnitt 1.6). Als Software Basis für die folgenden Experimente wurde das - in Kapitel 3 beschriebene - Open Source Projekt Lucene 1.4 [Luc05] verwendet. Für die Durchführung standen drei Server im Rechenzentrum der Universität Hildesheim zur Verfügung. Auf allen drei Servern konnte mittels des Netzwerkprotokolls SSH Remote zugegriffen werden.

Herangehensweise für die WebCLEF Experimente war es, die schon vorhandenen und

Server	Marke	OS	Java	CPU	RAM	Speicher
ir01	IBM	Linux 9,3	1.5	2 x AMD Opteron 250 2,4GHz	8 GB	215 GB
file01	IBM	Linux 9,3	1.5	2 x Intel Xeon 3,2GHz	2 GB	196 GB
app01	IBM	Linux 9,3	1.5	2 x Intel Xeon 3,2GHz	2 GB	196 GB

Tab. 4.1: Hardware zur Durchführung der WebCLEF Experimente an der Universität Hildesheim

in den letzten CLEF Durchgängen getesteten Systeme des IFAS so zu adaptieren, dass effektives Web Retrieval implementiert werden konnte. Aufgrund dieser Herangehensweise staffelten sich die Experimente in fünf Phasen, die jeweils ihren zeitlichen Rahmen durch den WebCLEF Zeitplan vorgegeben bekamen. Da der zeitliche Rahmen sehr eng gesteckt war, liefen einige Phasen parallel zu einander.

Abb. 4.1: Geplanter Zeitverlauf der WebCLEF Experimente am IFAS

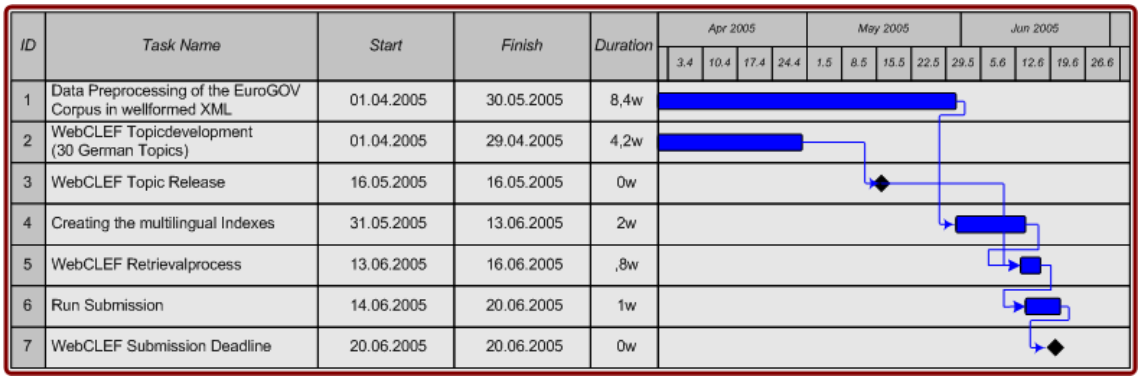


Abb. 4.2: Verschiebungen des Zeitplans der WebCLEF Experimente am IFAS

Die erste Phase bestand darin, den zur Verfügung gestellten EuroGOV Korpus so

zu bearbeiten, dass zum Indexieren wohlgeformtes XML vorliegen würde. Dieser Umformungsprozess sollte anhand eines Perlscripts durchgeführt werden. Aufgrund von nicht zufriedenstellenden Ergebnissen der Umformung durch das Perlscript wurde dieser Phasenabschnitt anhand einer Javaapplikation umgesetzt. Die zweite Phase bestand darin, 30 deutsche Queries - im CLEF Kontext Topics genannt - auf Basis des EuroGOV Korpus zu generieren. Nach Fertigstellung der zu entwickelnden Topics und dem erfolgreichen Umformen der Testkollektion in wohlgeformtes XML, konnte in der dritten Phase die Indexierung der EuroGOV Kollektion beginnen. Ziel war es, wie in den vergangenen CLEF Durchgängen des IFAS, für jede vorkommende Sprache einen eigenen Index zu entwickeln, um sprachspezifische Methoden der Indexierung anzuwenden und somit bessere Ergebnisse zu erzielen. Nach der offiziellen Veröffentlichung der WebCLEF Topics konnte dieses Ziel aufgrund von zeitlichen Engpässen nicht mehr wie erwartet verfolgt werden, da jedem Topic nicht die verwendete Sprache hinzugefügt wurde, wie dies in den einzelnen Dokumenten der Testkollektion der Fall war. Zur Verwirklichung mehrsprachiger Indizes hätte eine Sprachidentifizierung der Topics in den Retrievalprozess eingebaut werden müssen, um dann die richtigen Indizes ansprechen zu können. Aus diesem Grund entschied man sich einen multilingualen Index als Basis aller Retrievalprozesse zu verwenden. Infolgedessen mussten Methoden zur Indexierung gefunden werden, die sprachunabhängig und effektiv arbeiten würden. Mit diesen Vorgaben entschied man sich für den Lucene StandardAnalyzer [Luc05] und einen NGram Analyzer, der während des CLEF 2005 ad-hoc multi-lingual Retrieval Track der Universität Hildesheim [Hac05b] ebenfalls zum Einsatz kam. Anhand dieser Indexiermethoden sollten in der dritten Phase mehrere Indizes generiert werden. Nach Umsetzung der dritten Phase musste die Lucene 1.4 Searchengine [Luc05] (vierte Phase) den WebCLEF Anforderungen angepasst werden. Anhand der WebCLEF Topics sollte auf den Indexfeldern Title und Content gesucht werden. In der fünften und letzten Phase mussten alle erfolgreich durchgeführten Runs in das offizielle WebCLEF Format gebracht werden, um sie bis zum 20. Juni 2005 (offizielle Verschiebung des Abgabetermins vom 15. Juni 2005 auf den 20. Juni 2005 fand am 15. Juni 2005 statt) über das WebCLEF Webinterface einzureichen. Beim Betrachten des Zeitplans wird deutlich, dass der meiste Zeitaufwand für die Bereinigung der Kollektion gebraucht wurde. In den nächsten Abschnitten werden die einzelnen Phasen mit ihren Herausforderungen und Erfolgen der Webretrieval Experimentenreihe des IFAS genauer beschrieben. Die Abbildung 4.3 zeigt den nötigen zeitunabhängigen Ablauf von Erhalt des EuroGOV Korpus bis zum eigentlichen Retrievalprozesses und Erstellen der Ergebnislisten.

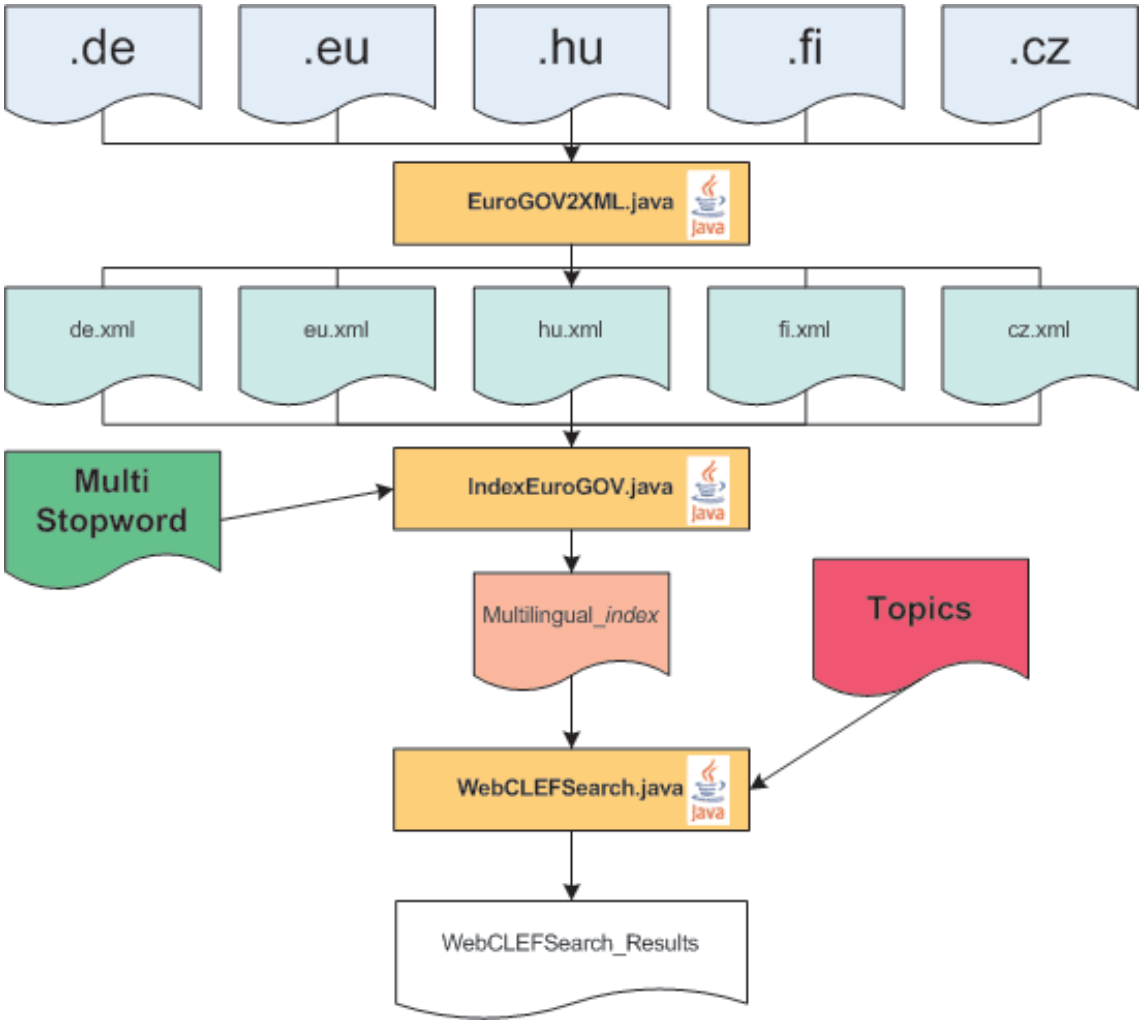


Abb. 4.3: Gesamter WebCLEF 2005 Prozess des IFAS

4.2 Topicgenerierung

Im Zeitraum vom 1. April 2005 bis 30. April 2005 hatten alle Teilnehmer die Möglichkeit, das von der Universität Amsterdam zur Verfügung gestellte Webinterface zur Topicgenerierung zu benutzen. Das IFAS war für die Generierung von 30 deutschen Topics (15 named paged Topics und 15 homepage Topics; Abschnitt 1.4) verantwortlich. In der Veröffentlichung *Inside the Evaluation Process of the Cross-Language Evaluation Forum (CLEF): Issues of Multilingual Topic Creation and Multilingual Relevance Assessment* [Klu02] wird eindrücklich die Bedeutung gute Topics zu entwickeln mit folgendem Einleitungssatz verdeutlicht:

> *Since the beginning of information retrieval evaluation topic generation is considered as one of the most important tasks. [Klu02]*

Um für den Web Track angemessene Topics zu generieren, stützte man sich auf vier Kriterien.

- Topics müssen ein Informationsdefizit auf Seiten des Users aufweisen. [Man03]
- Topics müssen immer in natürlicher Sprache formuliert sein. [Man03]
- Topics sollen einen realen Bezug zur Testkollektion haben. [Klu02]
- Topics müssen in einer strukturierten Form zur Verfügung gestellt werden. [Klu02]

4.2.1 Erfahrungen beim Generieren von deutschen WebCLEF Topics

Das Entwickeln von 30 Topics erwies sich als sehr viel zeitaufwendiger, als es von den Organisatoren angekündigt wurde. Schwierigkeiten stellten in erster Linie die zur Verfügung gestellten Tools sowie die Testkollektion an sich dar. Da die Universität Hildesheim sich ausschließlich mit deutschen Topics beschäftigt hat, konnten die Schwierigkeiten auch sprachspezifisch oder domänenspezifisch sein. Bisher konnte jedoch noch kein Erfahrungsaustausch bezüglich der Topicentwicklung durchgeführt werden. Für die Generierung bzw. Suche von möglichen Topics wurde in den Topleveldomänen .de, .at und .eu.int gesucht. Beim Suchen nach möglichen deutschen Topics traten folgende Herausforderungen auf:

erhöhtes Frameaufkommen In der .de Domäne konnte erhöhtes Auftreten von Frames registriert werden. Dies hat zur Folge, dass in der Testkollektion EuroGOV viele Dokumente nicht vollständig unter einer ID auftreten, sondern nur die Inhalte einzelner

Frames als eigenständige Dokumente mit eigener ID zur Verfügung stehen. Damit kann zum Teil erklärt werden, warum die deutsche Topleveldomäne im Vergleich zu anderen Domänen ein so hohes Dokumentaufkommen im EuroGOV Korpus hat (Abb. 1.1).

ILPS Lucene: EuroGOV Searchengine Die Suchmaschine, die von der Universität Amsterdam zur Topicentwicklung angefertigt wurde, produzierte schwer interpretierbare Ergebnislisten. Mit überwiegender Mehrheit beinhaltete die 25 stellige Liste Bilder, Grafiken und Forumeinträge, die sich für mögliche Topics nicht eigneten. Die Vorgabe, dass die Zielseite eines Topics in dieser Ergebnisliste auftreten musste, wurde erschwert durch das häufige Auftreten von nicht verwendbaren Seiten - aus oben genannten Gründen. Diese Tatsache wird wahrscheinlich auch Einfluss auf die Vollständigkeit der Relevanzbewertung deutscher WebCLEF Topics haben.

CDATA Problematik In der Beschreibung des EuroGOV Korpus wird zur Struktur gesagt, dass sie XML ähnlich ist, aber häufiger mehrere schließende CDATA Tags in einem CDATA Bereich gefunden wurden [Web05]. Dies würde bedeuten, dass der Inhalt nach dem ersten schließenden Tag für die Indexierung der ILPS Lucene: Searchengine nicht zur Verfügung stand und damit auch nicht zur Findung von Topics behilflich war. Es wurde jedoch keine konkrete Aussage zur Häufigkeit dieses Problems gemacht, sodass nicht abgeschätzt werden kann, wie vollständig der Index der ILPS Lucene: Searchengine wirklich war.

Terrier Searchengine Durch die Universität Glasgow wurde neben der oben genannten ILPS Lucene: EuroGOV Searchengine die Terrier Searchengine zur Verfügung gestellt. Terrier steht für TERabyte RetrIEver und konzentriert sich auf das Indexieren und Durchsuchen von großen Datenkorpora. Der Vorteil von Terrier ist, dass es durch wenige Schritte ein voll funktionfähiges Retrievalsystem für grosse Testkollektionen bereit stellt. Terrier wurde schon in den TREC Durchläufen (.GOV, .GOV2; Abschnitt 2.1) und in einigen CLEF Disziplinen getestet [Ter05]. Leider funktionierte die Bereitstellung der Universität Glasgow von der Terrier Searchengine während des Topicentwicklungszeitraums nicht, sodass keine praktischen Erfahrungen mit Terrier von Seiten des IFAS gemacht werden konnten. Von Interesse wäre die Fähigkeit der Terrier Suchmaschine Queries nach Sprache und Domänen zu filtern.

Bookmark Umformulierung Nachdem ein mögliches Topic gefunden wurde, musste dieses in einen Bookmark ähnlichen Titel umbenannt werden (Abschnitt 1.4). Nach erneutem Suchen der möglichen Zielseite durch den Bookmarktitel konnte diese Seite

in den meisten Fällen in der 25 stellige Ergebnisliste nicht mehr wiedergefunden werden (Hauptkriterium für ein Topic). Daher musste die Umformulierung diesen Gegebenheiten angepasst werden. Dies ist der Grund für die manchmal doch eher fragwürdige Formulierung einiger deutscher Topics.
Beispiele:

- Das Topic WC0014 *Bundeskanzleramt am Spreebogen* musste den Anhang *am Spreebogen* bekommen um die Startseite des Bundeskanzleramtes zu finden.
- Das Topic WC0133 *Bundespresseamt* musste umbenannt werden in *Bundespresseamt Bela Anda*, da der alleinige Begriff Bundespresseamt die gewünschte Startseite des Bundespresseamts nicht anzeigen würde.
- Das Topic WC0077 *Joschka Fischer auf der Afghanistan Konferenz 2004* musste mit *Joschka Fischer* in Verbindung gebracht werden, da unter Afghanistan Konferenz 2004 keine relevanten Seiten gefunden werden konnten.

Diese Problematik der Umformulierung erschwerte das Ziel, Queries mit realem Bezug zur Zielseite zu definieren [Klu02].

Rich Document Types Datenformate wie z.B. .pdf, .doc oder .ps konnten nicht gefunden werden und fielen damit aus der gesamten Topicgenerierung heraus.

Die oben beschriebenen Herausforderungen sind beim Entwickeln von deutschen Topics aufgetreten; die Frage, die sich stellt ist, ob diese Probleme sprach- bzw. domänenabhängig oder allgemeine Herausforderungen beim Generieren von Topics sind. Beim aktiven Entwickeln und Suchen von Topics wurde deutlich, dass das Finden von möglichen named page Topics sich wesentlich einfacher gestaltete, als das Finden von homepage Topics. Die Schwierigkeiten beim Finden von deutschen homepage Topics lag darin, dass in der deutschen Topleveldomäne der Inhalt von Startseiten oft aus Grafiken, Flashanimationen oder sehr wenig Text bestand. Daher sind die Indizes von Startseiten der deutschen Topleveldomäne eher unvollständig bzw. klein, was den Retrievalprozess nicht gerade erleichterte. Aufgrund dieser Schwierigkeiten hat man sich entschieden, 50 % der homepage Topics in der eu.int Domäne zu generieren, da diese inhaltsreichere Startseiten beherbergt. Der Zeitaufwand für die Entwicklung von 30 deutschen Topics lag für eine Person bei ca. 1 1/2 Wochen. Trotz aller Schwierigkeiten beim Suchen und Finden von deutschen Topics konnten die im Abschnitt 4.2 genannten Kriterien zur Topicgenerierung umgesetzt werden. Alle Topics beinhalten einen Informationsdefizit, sind in natürlicher Sprache formuliert und beziehen sich auf die Testkollektion. Die strukturierte Darstellung der Topics

wurde von der Universität Amsterdam mittels der Abgaberoutine auf dem Webinterface gewährleistet. Alle 547 generierten Topics wurden am 15. Mai 2005 als XML Datei den Teilnehmern zur Verfügung gestellt.

4.2.2 Deutsche WebCLEF Topics der Universität Hildesheim

WC0008 LKW Maut Verordnung (named page Topic)

WC0011 Alterssicherung der Landwirte (named page Topic)

WC0012 Ausbildungspakt für Deutschland (named page Topic)

WC0014 Bundeskanzleramt am Spreebogen (homepage Topic)

WC0023 Europäisches Amt für Betrugsbekaempfung OLAF (homepage Topic)

WC0040 Aeltestenrat (homepage Topic)

WC0047 Bundestag und Schule (named page Topic)

WC0057 Sicherheitsmaßnahmen des Atomkraftwerk Biblis (named page Topic)

WC0077 Joschka Fischer auf der Afghanistan Konferenz 2004 (named page Topic)

WC0095 FIFA WM 2006 (homepage Topic)

WC0133 Bundespresseamt Bela Anda (homepage Topic)

WC0180 Agenda 2010 der Bundesregierung (named page Topic)

WC0185 Europa Newsletter (homepage Topic)

WC0197 Welthandelsorganisation (named page Topic)

WC0212 Bildungspolitik der Zentralafrikanischen Republik (named page Topic)

WC0234 PLOTEUS (homepage Topic)

WC0236 Bundesministerium der Verteidigung (homepage Topic)

WC0241 Institutionen der EU (homepage Topic)

WC0257 Diskussionsecke der EU (homepage Topic)

WC0267 Europaeisches Jugendportal (homepage Topic)

WC0270 Der Bundeskanzler für Kinder (homepage Topic)

WC0300 Ihre direkte Verbindung zur Europaeischen Union (named page Topic)

WC0327 Autobahn Suedumfahrung Leipzig (named page Topic)

WC0351 Kanzlerreise nach Washington im Oktober 2001 (named page Topic)

WC0396 Europaeisches Amt für Personalauswahl (homepage Topic)

WC0412 Europaeischer Rechnungshof (homepage Topic)

WC0446 Steuerreform 2000 (named page Topic)

WC0453 Bundeskanzler (homepage Topic)

WC0477 Links zum Thema Klimaschutz (named page Topic)

WC0531 Eine Euro Münze (named page Topic)

4.3 Preprocessing der Testkollektion

Ziel war es, die Erfahrungen und Systeme der Universität Hildesheim aus dem CLEF 2004 Durchgang mit in die WebCLEF Experimente einfließen zu lassen, wie in der Vorgehensweise beschrieben. Aufgrund dieser Entscheidung war die erste Aufgabe, das EuroGOV Korpus in wohlgeformtes XML umzuformen. Der Vorteil von wohlgeformtem XML bei Retrievalexperimenten liegt in der flexiblen Verarbeitung während des Indexierens. Durch das einfache Zugreifen bzw. Bearbeiten einzelner Elemente eines XML Dokuments durch einen SAX- [SAX05] bzw. DOM Parser [DOM05] können ohne großen Aufwand mehrere Indexfelder von einer Quelle erstellt werden. Dies hat für die späteren Retrievalprozesse und ihre Optimierung große Vorteile. Der einfache, aber prägnante Satz über die Dokumentstruktur des EuroGOV Korpus auf der Internetseite der WebCLEF Initiative [Web05] verdeutlicht nur annähernd, welchen Aufwand und Schwierigkeiten die Bearbeitung und Umformung der Kollektion beinhaltete.

> *This might smell like XML, but it will not be XML. Because:*
>
> 1. *We did not escape ampersand in URLs.*
> 2. *Some documents contain the pattern <![CDATA[...]]> in their content, but nested CDATA escaping is not XML. [Web05]*

Wie schon im zweiten Kapitel beschrieben besteht die Kollektion aus 157 Dateien. Jede dieser 157 Dateien musste einzeln umgeformt werden. Mit der Motivation, wohlgeformte XML Dateien für die Indexierung zur Verfügung zu haben, wurden folgende Umformungen der EuroGOV Kollektion geplant und umgesetzt.

XML Deklaration Jeder Datei wurde am Anfang jeweils folgende XML Deklaration eingefügt:
<?xml version=”1.0” encoding=”iso-8859-1”?>

Attribute Einige Attribute wurden als selbstständige Elemente heraus geschnitten. Diese neu generierten Elemente können beim Indexieren als mögliche Indexfelder auftreten. Folgende Attribute wurden zu eigenständigen Elementen umgeformt:

- domain im EuroGOV:bin Element
- url im EuroGOV:doc Element
- id im EuroGOV:doc Element
- title im EuroGOV:content Element (wenn im CDATA Element vorhanden)

CDATA Beim Parsen von XML Dokumenten tritt die Schwierigkeit auf, dass bei dem Vorkommen von illegalen XML Ausdrücken XML Parser sofort den Prozess des Parsens abbrechen, da sie auf eine mögliche, nicht interpretierbare Fehlerquelle gestoßen sind. Um diese häufig auftretende Fehlerquelle bzw. ungültige XML Ausdrücke ignorieren zu können, gibt es den CDATA Bereich. Innerhalb eines CDATA Bereichs sind alle illegalen Ausdrücke erlaubt, da ein XML Parser den gesamten Inhalt eines CDATA Bereichs ignorieren kann und darf. Ein CDATA Bereich beginnt mit **<![CDATA[** und endet mit]]>. Die einzige Zeichenfolge die nicht erlaubt ist, ist das mehrfache Auftreten von schließenden CDATA Tags (]]>), da sonst der folgende Inhalt verloren gehen kann bzw. illegale XML Zeichen für den Parser wieder zugänglich gemacht werden [Beh00]. Die Schwierigkeit des mehrfachen Vorkommens von schließenden CDATA Tags trat häufig in einzelnen Dokumenten des EuroGOV Korpus auf. Diese Konstellation ist während des Crawling Prozesses der EuroGOV Kollektion entstanden. Während des Zusammenführens der Metadaten eines Webdokuments und des eigentlichen Inhalts wurde der komplette Inhalt in einen CDATA Bereich kopiert. Der Hintergedanke dieser Aktion war es, den Parsern das Verarbeiten von Javascript Elementen, nicht wohlgeformten HTML Dokumenten und das Vorkommen von anderen Dateiformaten wie z.B. pdf Dateien zu erleichtern. Im Grundsatz ist diese Zielsetzung auch richtig gewesen und hat beim Vorbereiten der

Kollektion viel Arbeit erspart. Es wurde jedoch nicht bedacht, dass schon in Webdokumenten CDATA Bereiche vorkommen könnten, was verschachtelte CDATA Bereiche zur Folge hätte [Web05]. Verschachtelte CDATA Bereiche würden wiederum das Vorkommen von mindestens 2 schliessenden CDATA Tags bedeuten, womit die Wohlgeformtheit bzw. Verarbeitungsfähigkeit von XML Dokumenten nicht mehr gewährleistet werden kann. Diese verschachtelten CDATA Bereiche mussten in der EuroGOV Kollektion gefunden und korrigiert werden.

Ampersands (&) In dem Attribut url in dem EuroGOV:doc Element der Kollektion lagen vor Bearbeitung der Daten sehr häufig Ampersand (&) Zeichen vor. Das Ampersand Zeichen ist ein ungültiger XML Ausdruck, der zu der Zeichenfolge *&* umgeformt werden musste.

Sprachen der einzelnen Dokumente Von Seiten der Organisatoren wurde zusammen mit der EuroGOV Kollektion eine Liste geliefert, die für jedes einzelne Webdokument die vorkommende Sprache beinhaltete. Diese Information sollte ebenfalls den einzelnen Dokumenten als eigenständiges Element zugefügt werden.

JavaScript Navigationselemente In den meisten Webdokumenten der Kollektion war ein erhöhtes Aufkommen von JavaScript Navigationselementen zu verbuchen; diese sollten entfernt werden, um die schwer verarbeitungsfähige Größe der Dateien zu minimieren.

```
<EuroGOV:bin domain="cy" id="001">
<EuroGOV:doc
 url="http://www.cyprus.gov.cy/"
 id="Ecy-001-35"
 md5="548813557b2c61bb4c42db912ef01247"
 fetchDate="Wed Sep 22 10:57:39 MEST 2004"
 contentType="text/html; charset=ISO-8859-7">
<EuroGOV:content>
<![CDATA[
<HTML>
<HEAD>
<TITLE>Republic of Cyprus / ÊõõñéáéÞ Áçìïêñáôßá</TITLE>
<SCRIPT LANGUAGE="JavaScript">
</SCRIPT>
</HEAD>
<BODY>

Inhalt der Seite

</BODY>
</HTML>
]]>
</EuroGOV:content>
</EuroGOV:doc>
```

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<EuroGOVbin   id="001">
<Domain>cy</Domain>
<EuroGOVdoc
 md5="548813557b2c61bb4c42db912ef01247"
 fetchDate="Wed Sep 22 10:57:39 MEST 2004"
 contentType="text/html; charset=ISO-8859-7">
<id> Ecy-001-35</id>
<language>  greek</language>
<url> http://www.cyprus.gov.cy/</url>
<EuroGOVcontent>
<![CDATA[
<HTML>
<HEAD>
<TITLE>Republic of Cyprus / ÊõõñéáéÞ Áçìïêñáôßá</TITLE>
<SCRIPT LANGUAGE="JavaScript">
<!--
self._domino_name = "_Welcome";
// -->
</SCRIPT>
</HEAD>
<BODY>

Inhalt der Seite

</BODY>
</HTML>
]]>
</EuroGOVcontent>
<title> Republic of Cyprus / ÊõõñéáéÞ Áçìïêñáôßá </title>
</EuroGOVdoc>
```

Abb. 4.4: Ein Webdokument vor und nach der Umformung zu wohlgeformten XML

4.3.1 Umsetzung

Bei der Umsetzung der beschriebenen Ziele stellte sich heraus, dass die Größe der Dateien entscheidend für die mögliche Herangehensweise war. Die Dateigrößen variierten von ca.45MB bis ca.750MB. Der erste Versuch, die Umformung der Kollektion zu realisieren, wurde anhand eines Perlscripts durchgeführt. Hierfür wurde eine eigene Testkollektion aus fünf unterschiedlich großen Dateien (die jeweils ersten Dateien aus den Domänen .at 220MB, .be 598MB, .cy 59MB, .de 396MB und .dk 49MB) zusammengestellt. Anhand von regulären Ausdrücken und dem PERL split (Suchen und Ersetzen) Befehl sollten die Umformungen durchgeführt werden. Schwierigkeiten traten bei der Verarbeitung auf, da bei der Verarbeitung der gesamten Datei als String die Speicherkapzität nicht ausreichte bzw. der Prozess sich aufgrund von Speicherproblemen von selbstständig beendete. Deswegen wurden die Dateien nur zeilenweise ausgelesen und bearbeitet. Nach erfolgreicher Umsetzung wurde der String geleert und für die nächste Zeile zur Verfügung gestellt. Mit dieser Herangehensweise konnte die *eigene* Testkollektion von fünf Dateien erfolgreich bearbeitet werden, sodass die gesamten EuroGOV Kollektion anhand dieses Perlskripts (2XMLv1.pl[1]) umgeformt wurde. Diese in wohlgeformtes XML umgeformte Kollektion sollte anschließend anhand eines SAX Parsers dem Lucene IndexWriter übergeben werden. Bei dieser Übergabe meldete der SAX Parser wiederholte Male *parser exceptions* Meldungen, welche darauf schließen ließen, dass noch kein 100%ig wohlgeformtes XML vorlag. Die Annahme, dass das Perlscript keine vollständige Umformung des Korpus vorgenommen hatte, bestätigte sich, nachdem einige Dateien außerhalb der gewählten Testkollektion kontrolliert wurden. Die Schlußfolgerung war, dass PERL wahrscheinlich für Daten dieser Größe nicht geeignet ist, da alle Dateien zwar bearbeitet wurden, aber eine vollständige Umsetzung des Scripts nicht zu erkennen war. Aufgrund dieser Erfahrung wurde die gesamte Preprocessing Phase der Kollektion - nach schlechten Erfahrungen mit PERL - in der Programmiersprache JAVA umgesetzt.
Für die Bearbeitung der EuroGOV Kollektion wurde die Java Application EuroGOV2XML[2] geschrieben. EuroGOV2XML bedient sich hauptsächlich dem org.apache.regexp Paket. Dieses Paket wird standardmäßig seit dem Java JDK 1.4 mitgeliefert und umfasst sehr effektive Methoden zur Erkennung und Bearbeitung von regulären Ausdrücken. Dieses Programm beinhaltete folgende Funktionen:

Einlesen der EuroGOV Datei Anhand eines Filereaders wird jede Datei zeilenweise eingelesen und in einem String zwischengespeichert. Dieser String wird dann anhand

[1] siehe DVD/EuroGOV2XML
[2] siehe DVD/EuroGOV2XML

der unten aufgeführtet Methoden bearbeitet und durch den FileWriter in die zu generierende XML Datei geschrieben.

Einlesen der Sprachlisten Anhand eines Filereaders wird jede von den Organisatoren zur Verfügung gestellte Sprachliste (pro Domäne eine Datei) zeilenweise eingelesen und in einem String zwischengespeichert. In jeder Zeile sind Dokument-ID und die zugewiesene Sprache bzw. Sprachen enthalten. Anhand der IDs können die Sprachen während des Konvertierungsprozesses den einzelnen Dokumenten zugewiesen werden.

Einfügen der XML Deklaration Der String "<?xml version="1.0" encoding="ISO-8859-1"?>" wird in die erste Zeile jeder EuroGOV Datei eingefügt.

Regulären Ausdrücke In einer *while* Schleife werden alle nötigen Veränderungen der Datei vorgenommen. Alle Ampersands des URL Attributs werden in die nötige XML Form umgeformt. Die oben erwähnten Attribute werden eingelesen und als eigenständige Elemente wieder ausgegeben. Die Dokumentensprache aus der mitgelieferten language Datei wird als Element in die Datei geschrieben und alle verschachtelten CDATA Bereiche werden gefunden und entfernt.

Dollarzeichen ($) In den URLs der EuroGOV Dokumente mussten nachträglich alle Dollarzeichen ($) entfernt werden, da diese in Java Programmen für Referenzen innerhalb von regulären Ausdrücken verwendet werden und somit Schwierigkeiten bei der Umsetzung bzw. beim Ersetzen einzelner Muster verursachten.

Da es sich als schwierig erwies, mit Hilfe des FileReaders den Inhalt des CDATA Bereichs kontrolliert anzusprechen bzw. zu bearbeiten wurden die JavaScript Elemente in dieser Phase noch beibehalten. In der Phase des Indexierens wird es wesentlich unkomplizierter sein, anhand von SAX Parser Funktionen den CDATA Bereich zu bearbeiten. Für zukünftige Bearbeitungen des EuroGOV Korpus bzw. ähnlichen Korpora wurden an dem IFAS sehr gute Erfahrungen mit Java gemacht. Scriptsprachen - wie in unserem Beispiel PERL - scheinen Schwierigkeiten zu bekommen, sobald die Dateigröße und die durchzuführenden Operationen in hohem Maße zunehmen. Das Arbeiten mit Java hat im Vergleich zu PERL in unseren Experimenten deutlich überzeugt. Mit Hilfe der EuroGOV2XML Applikation konnte die EuroGOV Kollektion erfolgreich und vollständig in wohlgeformtes XML umgeformt werden und war somit fertig zur Indexierung mit Lucene 1.4.

4.4 Indexierung des EuroGOV Korpus

Nachdem das Korpus in wohlgeformtem XML vorlag und die Strategie der Indexierung sich so änderte, dass nur noch die gesamte Testkollektion in einem Index verarbeitet werden sollte (Abschnitt 4.1), mussten die zu verwendenden Analyzer an die Anforderungen der WebCLEF Experimente angepasst werden. Eine wichtige Vorarbeit war das Erstellen einer multilingualen Stoppwortliste, die zur Indexierung der EuroGOV Kollektion mit einfließen sollte. Als Basis dieser Liste wurden die zur Verfügung gestellten Stoppwortlisten der Universität Neuchatel [Sav05] und eine an dem IFAS im Rahmen einer Magisterarbeit [Hof05] entwickelten tschechischen Stoppwortliste verwendet. Zusammengeführt wurden die Listen, indem alle 13 zur Verfügung stehenden Listen in eine Gesamtliste kopiert und alle doppelt vorkommenden Worte entfernt wurden. Die Gesamtliste umfasst 4577 Wörter und deckt 63.64% der Topicsprachen und 52,38% des gesamten Korpus ab. Die Bedeutung von gut sortierten Stoppwortlisten konnte aus den Experimenten des CLEF Projektkurses 2004 des IFAS sehr gut bestätigt werden. In den Versuchsreihen des Kurses wurden hauptsächlich verschiedene Fusionsmethoden für den CLEF 2005 ad-hoc multi-lingual Retrieval Track des IFAS [Hac05b] getestet. In diesen Versuchsreihen wurden Indizes verwendet, die teilweise ohne Stoppwortlisten generiert wurden. Runs, die mittels dieser Indizes durchgeführt wurden, fielen im Durchschnitt mit ca. 20% schlechteren Prescision Werten aus. Aufgrund dieser Erkenntnis konnte zum einen die Bedeutung von Stoppwortlisten bestätigt und zum anderen auch der Fokus zukünftiger Experimente auf das ausführliche Zusammensetzten von Stoppwortlisten gesetzt werden [Cla05]. Mit dieser Erkenntnis wurde die Multilinguale Stoppwortliste in die WebCLEF Experimente des IFAS erfolgreich eingebaut.

4.4.1 EuroGOV Indizes der Universität Hildesheim

Um einen multiligualen Index zu erstellen, mussten - wie in der Einleitung näher erklärt - sprachen unabhängige Indexiermethoden gefunden werden. Zur Indexierung des EuroGOV Korpus wurde der N-Gram Ansatz und die Nutzung des Lucene StandardAnalyzers ausgewählt.

Der N-Gram Ansatz wurde auf Basis des Artikels *Character N-Gram Tokenization for European Language Text Retrieval [McN04]* ausgewählt, indem es heißt:

> *Using the CLEF 2002 test set we demonstrated empirically how overlapping character n-gram tokenization can provide retrieval accuracy that rivals the*

> *best current language-specific approaches for European languages. We show that n = 4 is a good choice for those languages, and document the increased storage and time requirements of the technique. ... Our findings demonstrate clearly that accuracy using n-gram indexing rivals or exceeds accuracy using unnormalized words, for both monolingual and bilingual retrieval. [McN04]*

Hier beschreiben McNamee und Mayfield deutlich, dass der N-Gram Ansatz in monolingualen und bilingualen Experimenten überzeugt hat. Diese Aussage sollte anhand eines multilingualen Experiments mit der EuroGOV Kollektion weitergeführt und die Aussage

> *This approach will easily scale to 20 or more languages - a requirement likely to be necessitated by EU enlargement.m[McN04]*

überprüft werden. Zum Einlesen des Korpus wurde ein SAX Parser gewählt. In der Java Applikation EuroGOVSAXHandler[3] werden die EuroGOV XML Dateien eingelesen und anhand der Title, Content und ID Elemente die Indexfelder bestimmt und gefühlt. Wenn in der Parser Methode *StartElement* das Element Content auftritt, wird automatisch die Methode *cleanup* gestartet. Die Methode cleanup reinigt mit Hilfe von regulären Ausdrücken alle nicht notwendigen Zeichen aus den zu indexierenden Elementen. Damit konnte die Korpusgröße um fast 1/3 gesenkt werden. Da - wie in dem Zitat von McNamee und Mayfield schon angedeutet - der N-Gram Ansatz hohe Rechenleistung und Speicherplatz erfordert, einigte man sich auf einen TriGram Index, auch wenn die guten Ergebnisse von McNamee und Mayfield auf einen 4Gram Ansatz beruhten. Diese Entscheidung wurde aufgrund von Zeit- und Hardware Problemen getroffen. Zur Umsetzung der Tri Gram Indexierung wurde der NGramAnalyzer aus dem CLEF 2005 ad-hoc multi-lingual Retrieval Track des IFAS [Hac05b] verwendet und die Stoppwörter entfernt (Abschnitt 4.4). Da der Bedarf an Rechenleistung und Speicherplatz einen unerwarteten Umfang einnahmen, musste nach mehreren Abbrüchen des Indexierungsprozesses - infolge von zu wenig Speicherplatz - die N-Gram Strategie angepasst werden. Man einigte sich darauf, nur das Title und ID Feld der EuroGOV Webdokumente zu indexieren. Dieser Ansatz war erfolgreich und dauerte ca. 90 Minuten. Dieser Prozess wurde auf dem app01 Server durchgeführt (Tab. 4.1). Die Indexgröße betrug 292 MB. Hinsichtlich dieses Indexes war eine sehr gute Performance der Retrieval Prozesse zu erwarten.

Als zweite Indexierungsmethode bzw. Werkzeug wurde der Lucene StandardAnalyzer verwendet. Um die multilinguale Stoppwortliste (Abschnitt 4.4) erneut verwenden zu können, musste der StandardAnalyzer[4] um eine Lucene Wordlistloader Klasse [Luc05] er-

[3] siehe DVD/IndexEuroGOV

[4] siehe DVD/IndexEuroGOV

weitert werden, sodass die selbst zusammen gestellte multilinguale Stoppwortliste, anstelle der im StandardAnalyzer vorhandenen Liste geladen werden konnte. Des Weiteren arbeitet der Lucene StandardAnalyzer ähnlich wie ein Java Tokenizer, indem er alle Ausdrücke in Form von Strings in lower case konvertiert und alle Satzzeichen entfernt. Außerdem entfernt der für WebCLEF angepasste Analyzer alle in der multilingualen Stoppwortliste enthaltenen Wörter. Mit dem StandardAnalyzer wurden zwei unterschiedliche Indizes erstellt. Der erste Index, mit der Erfahrung, dass die Speicher- und Rechenressourcen beim Indexieren stark beansprucht werden, umfasste das Title und ID Feld sowie das Content Feld, wobei der Prozess des Indexierens nach 200 Zeichen im Content Feld abgebrochen wurde. Der gesamte Prozess des Indexierens dauerte mit diesem Ansatz ca. 5 1/2 Stunden. Durchgeführt wurde diese Indexierung auf dem ir01 Server (Tab. 4.1). Der Speicheraufwand für diesen sogenannten Cut Off Index betrug 740 MB. Der zweite und letzte Index der für die offiziellen WebCLEF Runs produziert wurde, war ein Index über das Title und ID Feld sowie das komplette Content Feld. Für diesen Index benötigte der ir01 Server (4.1) ca. 34 Stunden und einen Speicheraufwand von 4,9 GB.

Index	Größe	Dauer	Server	Beschreibung
UHi3Ti	292 MB	1 1/2 h	ir01	TriGram Index auf dem Title Feld
UHiSco	740 MB	5 1/2 h	file01	StandardAnalyzer Index Cut Off im Content Feld (200 Zeichen)
UHiS	4,9 GB	34 h	file01	StandardAnalyzer Index

Tab. 4.2: Alle für die offiziellen WebCLEF Runs generierten Indizes des IFAS

Alle für die offiziellen WebCLEF Runs erstellten Indizes wurden ohne Berechnung der Termvektoren erstellt. Daher kann mit diesen Indizes während der Retrievalprozesse die Blind Relevance Feedback Funktion nicht verwendet werden. Die Termvektoren wurden aufgrund der nicht einzuschätzenden Rechenleistungen vernachlässigt. Des Weiteren wurden alle Indizes, die mit dem StandardAnalyzer produziert wurden, mit 3 Indexfeldern erstellt. Nach der Umformung der EuroGOV Kollektion in XML standen noch die Felder URL und Domain zur Verfügung. Aus Performancegründen wurde hier ebenso auf die Miteinbeziehung dieser Felder verzichtet. Die Title und Content Felder wurden in Form von Ausdrücken invertiert gefüllt (tokenized) und das ID Feld wurde buchstäblich gefüllt (Abschnitt 3.1). Das ID Feld ist im Retrievalprozess nur für die Ergebnisliste von Bedeutung, denn es wird nur auf den Title und Content Feldern gesucht.

4.4.1.1 Herausforderungen beim Indexieren

Am Anfang der dritten Phase zur Indexierung des Korpus traten vermehrt Schwierigkeiten beim Indexieren auf. Trotz der Aussage in der Lucene Dokumentation, dass ein Lucene Index ca. 30 % der Korpusgrösse einnimmt [Luc05], nahm ein generierter Testindex, bestehend aus nur einer EuroGOV Datei mit einer Ausgangsgröße von 49 MB, fast das Zehnfache der Ursprungsgröße ein. Dies hätte zur Folge, dass ein Index der gesamten EuroGOV Kollektion ca. 820 GB an Speicher füllen würde. Dies wäre mit den Hardware Ressourcen, die dem IFAS zur Verfügung standen, nicht umsetzbar gewesen. Es konnte auch nicht richtig sein, dass ein Index an Größe so zunehmen würde, wenn die erwartete Größe eigentlich nur 30 % der Kollektionsgröße einnehmen sollte. Außerdem brachen die Indexierungsprozesse regelmäßig infolge von verständlichen *out of memory* Meldungen ab. Nach langem und ausgiebigem Testen und Überarbeiten des Quellcodes konnte das Problem gelöst werden. Das Ändern von zwei Parametern veranlasste eine reibungslose und erfolgreiche Indexierung des EuroGOV Korpus. Zum Ersten wurde die HTML SAX Parser Klasse, die noch aus alten Testreihen vorhanden war, aus dem Import Block der EuroGOVSAXHandler Klasse entfernt. Zusätzlich wurde die Möglichkeit genutzt Java maximalen Arbeitsspeicher (-Xmx)[5] bzw. Startspeicher (-Xms)[6] zuzuweisen. Auf dem ir01 Server (Tab. 4.1) wurde mit einem maximalen Arbeitsspeicher von 24 GB und einer Startzuweisung von 8 GB gearbeitet. Auf den file01 und app01 Servern (Tab. 4.1) war die Zuweisung als maximaler Wert auf 8 GB und der Startwert bei 4 GB aufgeteilt. Die unterschiedlichen Zuweisungen sind auf die Hardware Gegebenheiten der drei Server zurückzuführen. Bei Betrachtung dieser zwei Maßnahmen erscheint die Entfernung eines fehlerhaften Imports als nicht wirklich relevant für die Fehlerbehebung. Dennoch war dies nach Zuweisung der unterschiedlichen Speicherwerte der ausschlaggebende Punkt, damit das System reibungslos und fehlerfrei funktionierte. Deutlich wurde in dieser Phase, dass die eigentlichen Herausforderungen einer Experimentierreihe nicht vorhersehbar sind und manchmal sogar die entscheidenden Faktoren für Erfolg oder Misserfolg sein können.

4.5 WebCLEF Retrieval Prozess

Nach einer sehr aufwendigen Indexierungsphase konnten mit den drei vorliegenden Indizes die Retrievalprozesse initialisiert werden. Als Basis aller WebCLEF Retrievalexperimente des IFAS wurde - wie in der Indexierungsphase - die Lucene 1.4 Java Bibliothek verwendet.

[5] Java Konsolenbefehl zum Zuweisen von Arbeitsspeicher

[6] Java Konsolenbefehl zum Zuweisen von einem festen Abreitsspeicher am Anfang eines Prozesses

Für den WebCLEF Retrievalprozess wurden die WebCLEF Topics, die generierten EuroGOV Indizes (Abschnitt 4.4.1) und eine für die Experimente adaptierte Lucene Query Engine verwendet.

4.5.0.2 Besonderheiten der WebCLEF Topics für die Query Engine

Die Lucene Suchansätze sind alle anhand der Lucene Bibliothek verwendbar. Für die WebCLEF Retrievalexperimente ist die Variante des Verstärkungsfaktors (Gewichtung) von Interesse. Die Eigenschaft Lucene Queries anhand von Feldern zu entwickeln kommt in den WebCLEF Experimenten des IFAS besonders zur Geltung. Hinsichtlich der Tatsache, dass die WebCLEF Topics in XML vorliegen, können durch die einzelnen Elemente der Queryfelder und damit die gesamte Query (ein WebCLEF Topic) entsprechend zusammengestellt werden. Für die Baseline Runs (Abschnitt 1.6) standen für die monolingualen Runs das *Title* Element und für die multilingualen Runs zusätzlich das *Translation* Element zur Verfügung. Für einen multilingualen Baseline Run wurde die Query mittels einer Feldgruppierung zusammengestellt und dem vorliegenden Index entsprechend indexiert. Das Gewichten von einzelnen Feldern wurde erst in den Postexperimenten durchgeführt.

Ein Kritikpunkt an der WebCLEF Initiative könnten die gut formulierten und strukturierten Queries sein, die keine Herausforderung im Aufbau bzw. in der Verarbeitung in sich bergen. Lucene bietet viel umfassendere Möglichkeiten Queries zu verarbeiten an, als dies bei den WebCLEF Topics vonnöten ist. Da die Priorität der WebCLEF Initiative - und damit auch die des *Cross Lingual Evaluation Forum* - auf dem Evaluieren von crosslingualen Retrievalprozessen liegt, ist das zur Verfügungstellen von gut strukturierten Topics bzw. Testkollektionen von Vorteil, um den eigentlichen Fokus voranzutreiben.

4.5.1 WebCLEF Retrievalexperimente der Universität Hildesheim

Um die Retrievalexperimente durchzuführen, wurde die Java Applikation WebCLEFSearch[7] geschrieben. Diese Anwendung wurde aus den Retrieval Klassen des CLEF 2005 ad-hoc multi-lingual Retrieval Experiments des IFAS [Hac05b] abgeleitet. Mit der WebCLEFSearch.java Anwendung können mono- und multilinguale Runs durchgeführt werden. Außerdem besteht die Möglichkeit, die Suchfelder unterschiedlich zu gewichten und somit den Verstärkungsfaktor der Lucene Query Engine zu verwenden. Der Ablauf der

[7] siehe DVD/WebCLEFSearch

Anwendung basiert auf dem grundsätzlichen Vorgehen der Lucene Query Engine. Als Erstes werden der Index und die WebCLEF Topics eingelesen. Die Topics durchlaufen die generateQueries Methode, in der die Topicfelder Title und Translation indexiert werden. Laut WebCLEF Richtlinien muss die Ergebnisliste einen Cut Off an der 50. Stelle haben. Das bedeutet, dass nach maximal 50 gefundenen Dokumenten pro Topic der Retrievalprozess abgebrochen werden muss. Nachdem der Ablauf des Prozesses kontrolliert nach maximal 50 Dokumenten abbricht, muss die Anwendung die Ergebnisliste in das offizielle WebCLEF Ergebnislisten Format pressen (Tab. 1.5) und dann als Text Datei abspeichern. Die ersten 5 Stellen der Ergebnisliste für den monolingualen Run des UHiS Indexes (Tab. 4.2) für das Topic WC0014 Bundeskanzleramt am Spreebogen sah wie folgt aus:

Topic	Dokument id	Rang	Mean Reciprocal Rank	Name ID
WC0014 Q0	Ede-003-9734011	1	1.0	UHiSMo
WC0014 Q0	Ede-002-250866426	2	0.6610566	UHiSMo
WC0014 Q0	Ede-002-111252603	3	0.54313827	UHiSMo
WC0014 Q0	Ede-002-44967100	4	0.4967923	UHiSMo
WC0014 Q0	Ede-001-462826050	5	0.4956766	UHiSMo

Tab. 4.3: Beispiel einer Ergebnisliste im offiziellen WebCLEF 2005 Format

Die Ergebnislisten mussten vor Abgabe der Runs anhand eines Perlskripts auf Vollständigkeit und Formatkonformität geprüft werden. Das dafür benötigte Skript wurde von den WebCLEF Organisatoren zur Verfügung gestellt. Am 15. Juni 2005 wurden von dem IFAS sechs Runs über das WebCLEF Webinterface eingereicht.

	Run	Beschreibung
1	UHi3TiMo	Monolingualer Run auf TriGram Title Index (UHi3Ti 4.4.1)
2	UHi3TiMu	Multilingualer Run auf TriGram Title Index (UHi3Ti 4.4.1)
3	UHiScoMo	Monolingualer Run auf StandardAnalyzer Cut Off Index (UHiSco 4.4.1)
4	UHiScoMu	Multilingualer Run auf StandardAnalyzer Cut Off Index (UHiSco 4.4.1)
5	UHiSMo	Monolingualer Run auf StandardAnalyzer Index (UHiS 4.4.1)
6	UHiSMu	Multilingualer Run auf StandardAnalyzer Index (UHiS 4.4.1)

Tab. 4.4: Alle offiziell eingereichten WebCLEF Runs der Universität Hildesheim 2005

4.6 QRELS

Am 4. Juli 2005 veröffentlichten die Organisatoren die vollständige Relevanzbewertung der 547 WebCLEF Topics. Die Relevanzbewertung ist die Auswertung aller Dokumente,

die während des Topicentwicklungsprozesses als relevant vermerkt wurden. Zusammen mit den ausgerechneten MRR Werten pro Topic wurden die *QREL* Dateien und ein dazugehöriges Auswertungsskript geliefert. QREL Dateien sind einfache Textdateien, die die relevanten Dokumente pro Topic in Form der Dokumenten-ID auflisten. Es wurden QREL Dateien für named page- und homepage Topics, sowie eine gemeinsame Auswertung aller Topics in den jeweiligen Sprachen als vollständige Sammlung zur Verfügung gestellt. Für das IFAS wurden die Runs anhand der *webclef2005.qrels*[8] Datei ausgewertet, da alle Runs auf einem multilingualen Index basierten. Die QREL Datei webclef2005.qrels (beinhaltet alle relevanten Dokumente aller Topics) gibt für die ersten drei Topics folgende Werte an:

Topic		Topic-ID	relevant
WC0001	0	Eeu-010-180688331	1
WC0002	0	Eeu-009-345357354	1
WC0003	0	Enl-002-229054709	1

Tab. 4.5: Die ersten drei Stellen in der webclef2005.qrels Datei

Das Skript[9] zur Auswertung von WebCLEF Runs ist in PERL geschrieben und bietet die Möglichkeit, eine Auswertung pro Topic und eine Durchschnittsauswertung auf den Positionen 1, 5, 10, 20 und 50 durchzuführen. Mit Hilfe der Position des ersten relevanten Dokuments wird dann der Mean Reciprocal Rank (Abschnitt 1.7) berechnet, sowie ein Durchschnittswert über den gesamten Run. Des Weiteren liefert die Auswertung eine Zusammenfassung, in der die Durchschnittswerte für die fünf oben genannten Positionen berechnet werden. Der Vorteil der ausgehändigten QREL Dateien und des dazu gehörigen Skripts ist die Möglichkeit, Postexperimente durchzuführen und mit den offiziellen Ergebnissen des Tracks zu vergleichen.

4.7 Ergebnisse

Die Ergebnisse der offiziell eingereichten Runs (Tab. 4.4) des IFAS, ergaben folgende Werte zusammengefasst:

Beim Betrachten der Werte fallen die Ergebnisse auf, die anhand des Lucene StandardAnalyzers durchgeführt wurden. Der beste Run ist auf Basis des, mit dem StandardAnalyzer erstellten, gesamten Index produziert worden. Der mit Cut Off bei 200 Zeichen

[8] siehe DVD/QRELS
[9] siehe DVD/QRELS

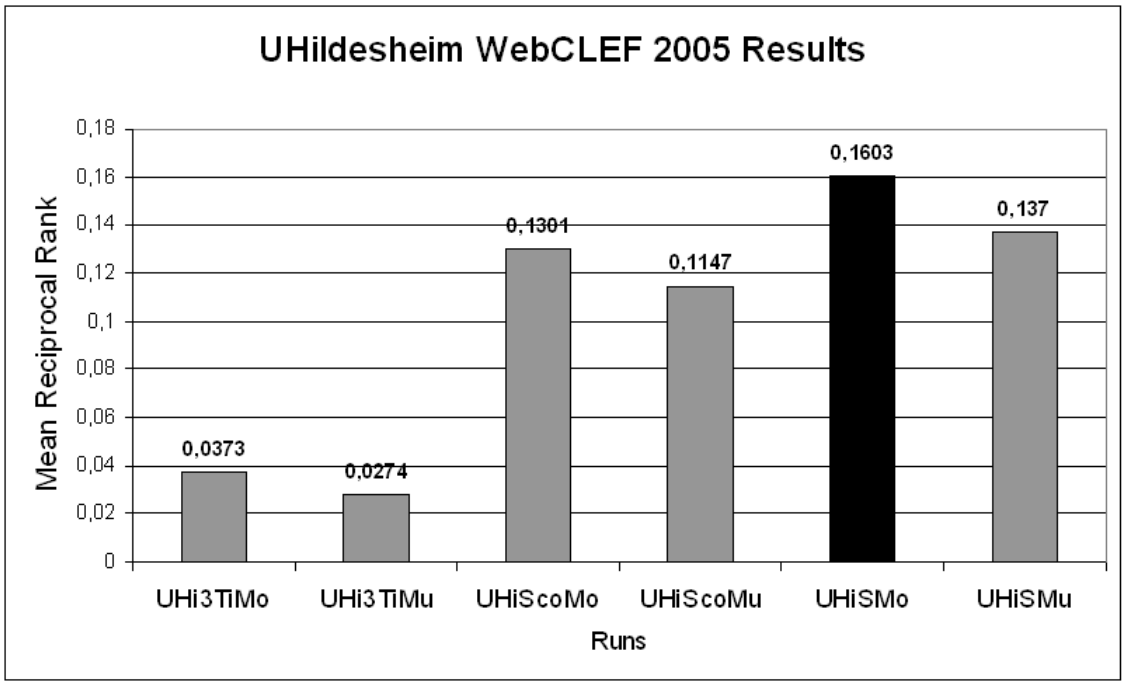


Abb. 4.5: Gesamtdurchschnittlicher MRR aller Runs Tab. 4.4

erstellte Index liegt drei Punkte unterhalb des Gesamtindex. Der StandardAnalyzer erzielt gegenüber dem TriGram Ansatz bessere Ergebnisse, da die Mean Reciprocal Rank Werte (Abschnitt 1.7) deutlich höher liegen als beim TriGram Ansatz. Ein MRR von 0,1603 (Wert des UHiSMo Runs) bedeutet, dass im Durchschnitt jedem Topic an sechster Stelle das erste relevante Dokument zugeordnet werden konnte. Interessant ist, dass alle multilingualen Runs in den Ergebnissen, trotz dem zusätzlichen Benutzen eines weiteren Indexfeldes (Translation), nicht oberhalb der monolingualen Runs landen konnten.

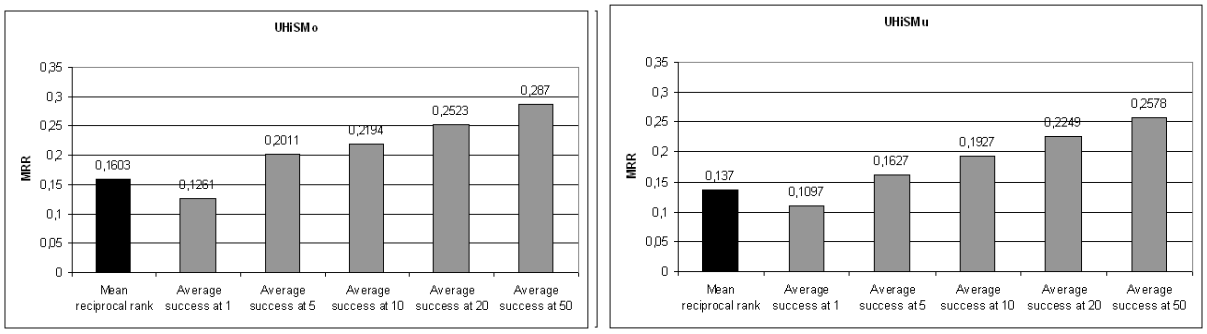


Abb. 4.6: Durchschnittswerte des Mono- und Mulilingualen UHiS Runs (Tab.4.4)

Die weiteren Berechnungen aller Runs stellen die Verteilungen relevanter Dokumente auf der Ergebnisliste dar. *Average success at 5* beinhaltet den Anteil der Topics, die

nach der fünften Position der Ergebnisliste mindestens ein relevantes Dokument erhalten haben. Für die UHiS Runs bedeutet dies, dass im monolingualen Run ca. 13% und im multilingualen Run ca. 11% aller Topics das erste Dokument der Ergebnislisten als relevant bewertet wurde. Der letzte Wert verdeutlicht, dass ca. 29% aller Topics im monolingualen Run ein Ergebnis lieferten und ca. 26% im multilingualen Run. Im Vergleich zu den anderen Runs kann der TriGram Title Index (UH3Ti) nicht überzeugen. Er erzielt die schlechtesten aller offiziell eingereichten Runs. Der StandardAnalyzer Cut Off Index lieferte mittlere Werte, die wahrscheinlich auf die zu kleine Suchbasis (Cut Off bei maximal 200 Zeichen im Content Feld) zurück zu führen sind. Leider konnte die Aussage von McNamee und Mayfield [McN04] bezüglich des erfolgreichen Nutzens von N-Gram Indizes anhand des UHi3Ti Index nicht bestätigt werden. Dies kann an dem stark reduzierten Index liegen, da nur das Title-Feld des Dokuments mit in den Index eingebunden war und nur ein TriGram Index anstelle des von McNamee und Mayfield geforderten 4Gram Index erstellt wurde. Aufgrund von zeitlichen Engpässen konnten nur die in diesem Abschnitt beschriebenen Runs noch vor dem WebCLEF Abgabetermin eingereicht werden. Für die genauere Untersuchung der Aussage von McNamee und Mayfield und dem Nutzen unterschiedlicher Gewichtungsmöglichkeiten werden in den Postexperimenten noch weitere Indizes erstellt.

4.8 Postexperimente

Nach erfolgreicher Abgabe der sechs oben beschriebenen Runs sollte in den Postexperimenten die Aussage von McNamee und Mayfield (zum Thema indexieren mittels des NGram Ansatz [McN04] (Abschnitt 4.4.1)) näher geprüft und die Möglichkeit, Indexfelder unterschiedlich zu gewichten, umgesetzt werden. Für dieses Ziel wurden neue Indizes erstellt und weitere Retrievalprozesse durchgeführt. Bei den neu durchzuführenden Retrievalprozessen wurden alle Indizes monolingual, multilingual und mit unterschiedlichen Gewichtungen der Felder durchgeführt. Hierbei war die Gewichtung jeweils 10:1. Das Title-Feld bzw. das Translation-Feld der WebCLEF Topics wurde jeweils mit dieser Quote gewichtet. Dieses Verhältnis wurde gewählt, um durch das Verwenden von Extremverschiebungen der Gewichtung die unterschiedlichen Einflüsse der Topicfelder zu untersuchen. Des Weiteren war das Ziel, die Ergebnisse der multilingualen Runs zu verbessern. Folgende Indizes wurden erstellt:

Neben den neu erstellten Indizes wurde in der WebCLEFSearch Klasse die Gewichtungsfunktion aktiviert. Diese Funktion wird durch den Verstärkungsfaktor der Lucene Query Engine (Abschnitt 3.2) umgesetzt. In der WebCLEFSearch Klasse wird die Metho-

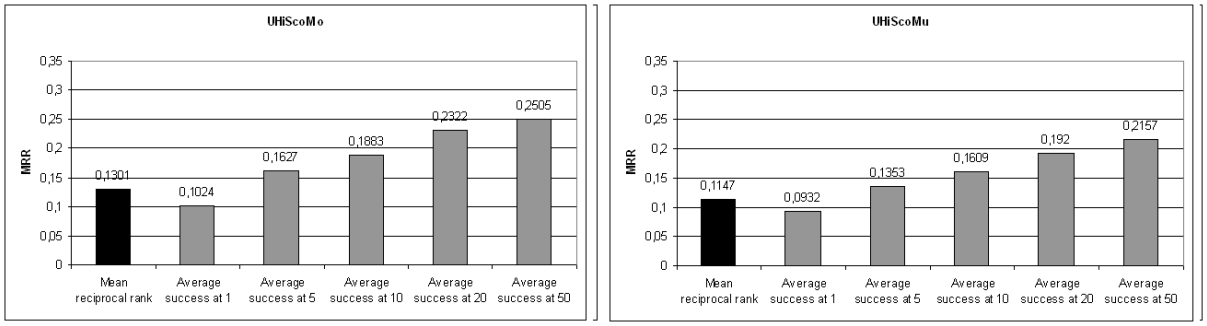


Abb. 4.7: Durchschnittswerte des Mono- und Mulilingualen UHiSco Runs (Tab.4.4)

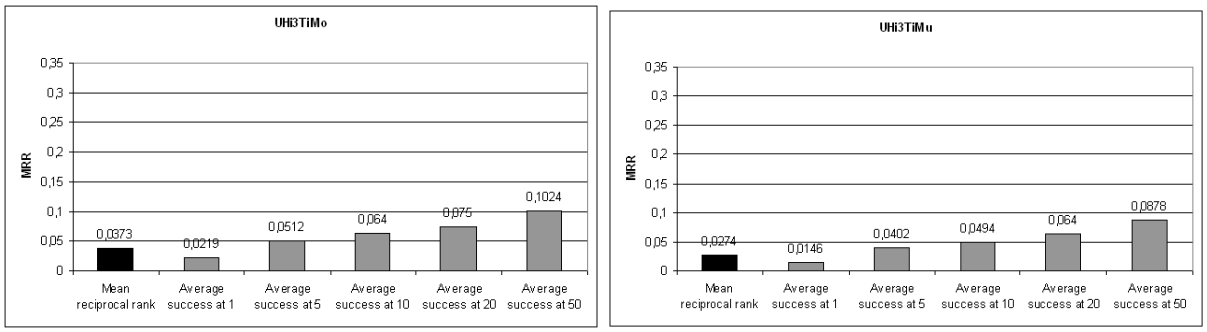


Abb. 4.8: Durchschnittswerte des Mono- und Mulilingualen UHi3Ti Runs (Tab.4.4)

Index	Größe	Dauer	Server	Beschreibung
UHiSTi	343MB	8 1/2h	file01	StandardAnalyzer Index auf Title Feld
UHi3co	1,6GB	40 1/2h	app01	TriGram Index Cut Off im Content Feld (200 Zeichen)
UHi3	909MB	216h	ir01	TriGram Index
UHi4Ti	265MB	6 1/2h	file01	4Gram Index auf Title Feld
UHi4co	884MB	22h	ir01	4Gram index Cut off im Content Feld (200 Zeichen)
UHi5Ti	250MB	6h	file01	5Gram Index auf Title Feld
UHi5co	1,5GB	38 1/2h	file01	5Gram Index Cut Off im Content Feld (200 Zeichen)

Tab. 4.6: Erstellte Indizes der IFAS Postexperimente

de *Boost* aufgerufen, um bestimmten Topicfeldern eine Gewichtung zu vermitteln. Da auf Basis der erstellten Indizes nur Baseline Runs durchgeführt werden können, stehen nur die Topicfelder Title und Translation zur Verfügung. Wenn ein Run geboostet werden soll, kann er damit nur multilingual sein, da sonst nur ein Feld zur Verfügung stehen würde. Die erste Zahl in der Run-ID ist immer das Title Feld. Die zweite Zahl bestimmt immer die Gewichtung des Translation Feldes. Der run UHi4Ti101 steht für einen multilingualen Run auf Basis eines 4Gram Title-Feld Index mit einer Gewichtung der Felder Title und

Translation in einem Verhältnis 10:1. Die Tabelle 4.7 stellt die 38 durchgeführten Postruns des IFAS näher dar. Die Ergebnisse werden in Tabelle 4.8 abgebildet.

	Run	Beschreibung
1	UHiSTiMo	Monolingualer Run StandardAnalyzer Title Index (UHiSTi Tab.4.6)
2	UHiSTiMu	Multilingualer Run StandardAnalyzer Title Index (UHiSTi Tab.4.6)
3	UHiSTi101	Multilingualer Run StandardAnalyzer Title Index Boost 10:1 (UHiSTi Tab.4.6)
4	UHiSTi110	Multilingualer Run StandardAnalyzer Title Index Boost 1:10 (UHiSTi Tab.4.6)
5	UHiS101	Multilingualer Run StandardAnalyzer Index Boost 10:1 (UHiS Tab.4.2)
6	UHiS110	Multilingualer Run StandardAnalyzer Index Boost 1:10 (UHiS Tab.4.2)
7	UHiSco101	Multilingualer Run StandardAnalyzer Cut Off Index Boost 10:1 (UHiSco Tab.4.2)
8	UHiSco110	Multilingualer Run StandardAnalyzer Cut Off Index Boost 1:10 (UHiSco Tab.4.2)
9	UHI3Ti101	Multilingualer Run TriGram Title Index Boost 10:1 (UHi3Ti Tab.4.2)
10	UHi3Ti110	Multilingualer Run TriGram Title Index Boost 1:10 (UHi3Ti Tab.4.2)
11	UHi3coMo	Monolingualer Run TriGram Cut Off Index (UHi3co Tab.4.6)
12	UHi3coMu	Multilingualer Run TriGram Cut Off Index (UHi3co Tab.4.6)
13	UHi3co101	Multilingualer Run TriGram Cut Off Index Boost 10:1 (UHi3co Tab.4.6)
14	UHi3co110	Multilingualer Run TriGram Cut Off Index Boost 1:10 (UHi3co Tab.4.6)
15	UHi3Mo	Monolingualer Run TriGram Index (UHi3 Tab.4.6)
16	UHi3Mu	Multilingualer Run TriGram Index (UHi3 Tab.4.6)
17	UHi3101	Multilingualer Run TriGram Index Boost 10:1 (UHi3 Tab.4.6)
18	UHi3110	Multilingualer Run TriGram Index Boost 1:10 (UHi3 Tab.4.6)
19	UHi4TiMo	Monolingualer Run 4-Gram Title Index (UHi4Ti Tab.4.6)
20	UHi4TiMu	Multilingualer Run 4-Gram Title Index (UHi4Ti Tab.4.6)
21	UHi4Ti101	Multilingualer Run 4-Gram Title Index Boost 10:1 (UHi4Ti Tab.4.6)
22	UHi4Ti110	Multilingualer Run 4-Gram Title Index Boost 1:10 (UHi4Ti Tab.4.6)
23	UHi4coMo	Monolingualer Run 4-Gram Cut Off Index (UHi4co Tab.4.6)
24	UHi4coMu	Multilingualer Run 4-Gram Cut Off Index (UHi4co Tab.4.6)
25	UHi4co101	Multilingualer Run 4-Gram Cut Off Index Boost 10:1 (UHi4co Tab.4.6)
26	UHi4co110	Multilingualer Run 4-Gram Cut Off Index Boost 1:10 (UHi4co Tab.4.6)
27	UHi5TiMo	Monolingualer Run 5-Gram Title Index (UHi5Ti Tab.4.6)
28	UHi5TiMu	Multilingualer Run 5-Gram Title Index (UHi5Ti Tab.4.6)
29	UHi5Ti101	Multilingualer Run 5-Gram Title Index Boost 10:1 (UHi5Ti Tab.4.6)
30	UHi5Ti110	Multilingualer Run 5-Gram Title Index Boost 1:10 (UHi5Ti Tab.4.6)
31	UHi5coMo	Monolingualer Run 5-Gram Cut Off Index (UHi5co Tab.4.6)
32	UHi5coMu	Multilingualer Run 5-Gram Cut Off Index (UHi5co Tab.4.6)
33	UHi5co101	Multilingualer Run 5-Gram Cut Off Index Boost 10:1 (UHi5co Tab.4.6)
34	UHi5co110	Multilingualer Run 5-Gram Cut Off Index Boost 1:10 (UHi5co Tab.4.6)

Tab. 4.7: Postexperimente des IFAS

Beim Betrachten der Ergebnisse wird deutlich, dass alle Runs auf Basis eines StandardAnalyzer Index durchweg die besten Werte erbringen. Zwar zeigen die 4- und 5Gram

Run	MRR	succ. at 1	succ. at 5	succ. at 10	succ. at 20	succ. at 50
Ergebnisse aller Postruns auf Basis von StandardAnalyzer Indizes						
UHiSTiMo	0,2377	0,2267	0,253	0,253	0,253	0,253
UHiSTiMu	0,1939	0,1853	0,2055	0,2055	0,2055	0,2055
UHiSTi101	0,2117	0,2018	0,2257	0,2257	0,2257	0,2257
UHiSTi110	0,1226	0,1193	0,1266	0,1266	0,1266	0,1266
UHiS101	0,1608	0,1298	0,1974	0,2176	0,245	0,2724
UHiS110	0,0811	0,0658	0,0987	0,1133	0,1316	0,1444
UHiSco101	0,1307	0,1042	0,1609	0,1883	0,2285	0,2486
UHiSco110	0,0677	0,053	0,0786	0,1079	0,1207	0,128
Ergebnisse aller Postruns auf Basis von TriGram Indizes						
UHi3Ti101	0,0379	0,0219	0,053	0,0622	0,075	0,1042
UHi3Ti110	0,0139	0,0091	0,0165	0,0238	0,0293	0,0512
UHi3coMo	0,0985	0,0922	0,1062	0,1062	0,1062	0,1062
UHi3coMu	0,0985	0,0922	0,1062	0,1062	0,1062	0,1062
UHi3co101	0,1075	0,1022	0,1142	0,1142	0,1142	0,1142
UHi3co110	0,0985	0,0922	0,1062	0,1062	0,1062	0,1062
UHi3Mo	0,0169	0,0091	0,0238	0,0366	0,042	0,0548
UHi3Mu	0,0099	0,0037	0,0183	0,0238	0,0311	0,0402
UHi3101	0,0172	0,0091	0,0256	0,0329	0,0439	0,053
UHi3110	0,0063	0,0018	0,0073	0,0146	0,0201	0,0402
Ergebnisse aller Postruns auf Basis von 4 Gram Indizes						
UHi4TiMo	0,1116	0,103	0,1232	0,1232	0,1232	0,1232
UHi4TiMu	0,0843	0,0789	0,0917	0,0917	0,0917	0,0917
UHi4Ti101	0,1023	0,0954	0,1119	0,1119	0,1119	0,1119
UHi4Ti110	0,0479	0,044	0,0532	0,0532	0,0532	0,0532
UHi4coMo	0,0607	0,0593	0,0634	0,0654	0,0654	0,0654
UHi4coMu	0,0498	0,0481	0,0536	0,0555	0,0555	0,0555
UHi4co101	0,053	0,0518	0,0555	0,0573	0,0573	0,0573
UHi4co110	0,0363	0,0351	0,0388	0,0407	0,0407	0,0407
Ergebnisse aller Postruns auf Basis von 5 Gram Indizes						
UHi5TiMo	0,121	0,1131	0,1313	0,1313	0,1313	0,1313
UHi5TiMu	0,0953	0,0899	0,1028	0,1028	0,1028	0,1028
UHi5Ti101	0,1127	0,1064	0,1211	0,1211	0,1211	0,1211
UHi5Ti110	0,057	0,0532	0,0624	0,0624	0,0624	0,0624
UHi5coMo	0,121	0,1131	0,1313	0,1313	0,1313	0,1313
UHi5coMu	0,0953	0,0899	0,1028	0,1028	0,1028	0,1028
UHi5co101	0,1127	0,1064	0,1211	0,1211	0,1211	0,1211
UHi5co110	0,057	0,0532	0,0624	0,0624	0,0624	0,0624

Tab. 4.8: Ergebnisse aller IFAS Postruns

Indizes eine leichte Verbesserung zu den TriGram Indizes auf, im Vergleich können die einzelnen Runs allerdings nicht überzeugen. Überraschender Weise fällt der UHiSTiMo Run auf Basis eines sehr kleinen Title Indexes am Besten aus. Mit einem MRR von 0,2377 Punkten erzielt dieser Monolinguale Run mit großem Abstand das beste Ergebnis. Ein MRR von 0,2377 bedeutet, dass im Durchschnitt jedes Topic an der 4,2. Stelle ein relevantes Dokument findet. Dies ist eine Verbesserung von fast zwei Stellen im Vergleich zu dem UHiSMo Run (bester offizieller Run) mit einem MRR von 0,1603 Punkten.

4.9 Auswertung der Ergebnisse

Nach dem Sammeln und Auswerten aller Runs kann als erstes festgehalten werden, dass die Indizes die mit dem Lucene StandardAnalyzer erstellt wurden deutlich bessere Ergebnisse lieferten als alle drei N-Gram Ansätze. Mit dieser Aussage kann die These von McNamee und Mayfield für diese Versuchsreihe nicht bestätigt werden. Beim Betrachten der unterschiedlichen Kurven in der Abbildung 4.9 wird ebenfalls deutlich, dass die MRR Werte für den StandardAnalyzer Cut Off Index im Vergleich zu dem Title Index im Durchschnitt um fast 0,08 Punkte und zum Gesamtindex um 0,02 Punkte abfallen. Die Tendenz, dass der Cut Off Index grundsätzlich schwächere Ergebnisse produziert bestätigt sich bei den N-Gram Indizes nicht. Im Gegenteil. Diese Indizes produzieren die stärksten Ergebnisse mittels des Cut off Index, mit Ausnahme des 4 Gram Index. Mit diesen Ergebnissen kann der Lucene StandardAnalyzer als sprachunabhängiger Indexierungsansatz empfohlen werden. Des Weiteren fiel das Ergebnis des UHiSTi Index unerwartet sehr gut aus, da alle anderen Indizes mit gleichem Ansatz nicht überzeugen konnten. Für alle Runs gilt, dass die monolingualen Runs deutlich bessere Ergebnisse und die multilingualen Runs im Durchschnitt um ca. 0,0345 Punkte schlechtere Ergebnisse ablieferten. Beim Anwenden der Gewichtungen konnten die multilingualen Ergebnisse erheblich verbessert werden, indem das Titlefeld des Topics höher als das Translationfeld gewichtet wurde. In einem Fall (UHi3co101) fiel sogar das multilinguale Ergebnis im Vergleich zum monolingualen Ergebnis um 0.01 Punkte besser aus.

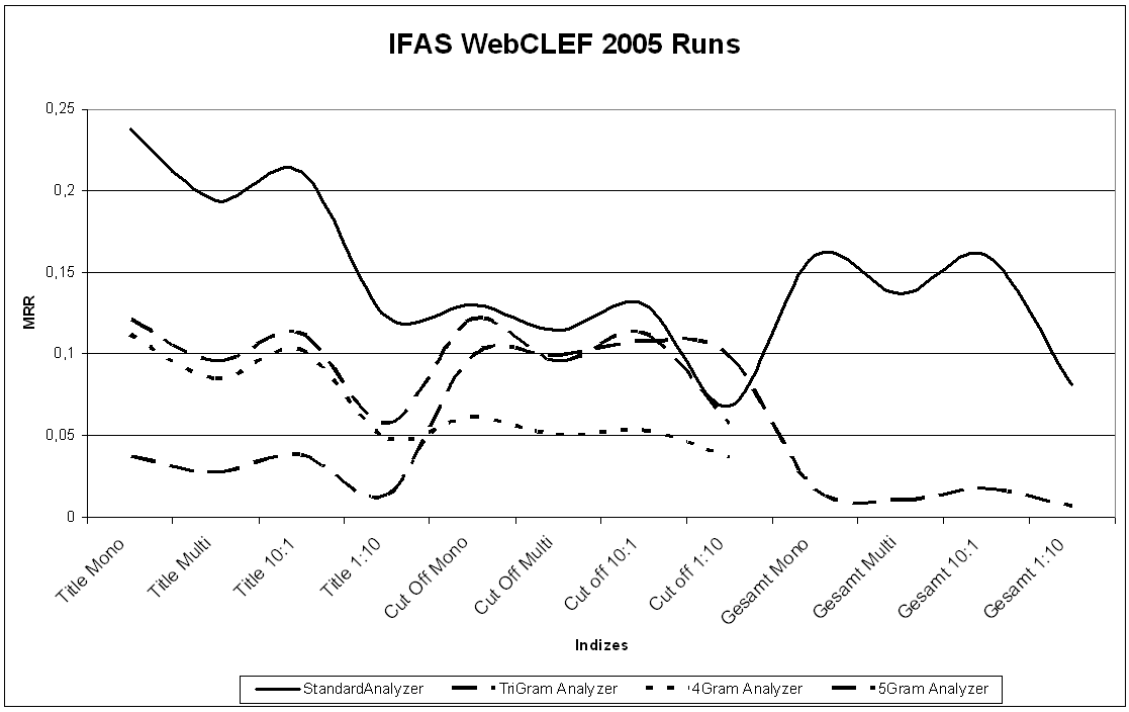


Abb. 4.9: Grafische Darstellung aller IFAS WebCLEF 2005 Runs

Kapitel 5

Ausblick und Fazit zum CLEF Web Track 2005

Beim Arbeiten mit der WebCLEF Testkollektion konnten viele Erfahrungen mit großen Web-Korpora gesammelt werden. Während der Experimente traten einige Schwierigkeiten auf. In diesem Kapitel sollen die Ursachen und Lösungen näher beschrieben werden. Wichtig ist, dass einige Probleme aufgrund von Erfahrungsmangel mit Web Tracks von Seiten des IFAS auftraten; manche Herausforderungen sind auf weniger ausgereifte Vorgaben des WebCLEF Tracks zurück zuführen. Grund ist die erstmalige Durchführung dieses Tracks, da sowohl IFAS als auch die CLEF Initiative erstmals mit einen Web Track betraut war. Dieses Kapitel soll die Thematik dieser Magisterarbeit abschließen und die Erfahrungen im Umgang mit großen multilingualen Korpora sowie Verbesserungsvorschläge für den Web Track 2006 zusammen bringen. Der Inhalt soll zum Wissenstransfer für die weiterführende Gruppe dienen, sodass im nächsten Jahr an der Stelle weiter gemacht werden kann, an der in diesem Jahr aufgehört wurde.
Die allgemeine Herangehensweise und Maxime des IFAS sollte das *Lernen von den Großen* sein. Mit den *Großen* sind zum einen die erfolgreichen Retrievalsysteme anderer in Kapitel 2 beschriebenen Web Tracks und zum anderen die Erfahrungen aus den verschiedenen multilingualen CLEF Tracks gemeint. Wenn man in der Lage ist, die Erkenntnisse aus diesen unterschiedlichen Retrievaldisziplinen zusammen zu bringen, könnte die Effektivität von multilingualen Webretrieval Systemen enorm gesteigert werden.

5.1 Das EuroGOV Korpus

Das Arbeiten mit dem EuroGOV Korpus stellte sich als der zeitaufwendigste Teil im gesamten WebCLEF Experiment des IFAS heraus. Ziel war es, den gesamten Korpus in

der ersten Phase in wohlgeformtes XML zu konvertieren. Dies konnte mittels der EuroGOV2XML Klasse erreicht werden. Während des Indexierungsprozesses meldete der Sax Parser, der die Indexierung der einzelnen Indexfelder steuerte, vermehrt *schwerwiegende Fehler*. Diese Fehler basierten laut *Parser Exception* Meldungen immer auf ungültigen XML Ausdrücken. Dies traf uns sehr unerwartet, da die Konvertierung in der Testphase mit positiven Ergebnissen getestet werden konnte. Zum Testen wurden fünf zufällige EuroGOV Dateien gewählt. Ein Grund für das Auftreten dieser Fehler könnte das nicht vorhanden sein dieser ungültigen XML Ausdrücke in den fünf ausgewählten Testdateien sein. Dann müsste in Zukunft das Testverfahren für große Korpora effektiver gestaltet werden, sodass diese Fehlerquelle zum größten Teil eliminiert wird. Im Grundsatz ist die 100%ige Abdeckung aller Fehlerquellen in einem Korpus dieser Größe kaum möglich. Die vom Crawler heruntergeladenen Dateien wurden in einem nicht konvertierten Zustand in den CDATA Bereich gespeichert. Da nicht konvertierte pdf, ps und Worddokument Dateien in sehr hohem Maße ungültige XML Ausdrücke beinhalten, könnte dies eine weitere Fehlerquelle sein, die während der Konvertierungsphase nicht berücksichtigt worden ist. Für die Zukunft bedeutet dies, dass das Konvertieren noch effektiver kontrolliert und die Rich Document Typen verstärkt gesucht und konvertiert bzw. gelöscht werden müssen. Die Rich-Document-Problematik soll laut WebCLEF Organisatoren in der nächsten Entwicklungsphase eines multilingualen Korpus in den Jahren 2007 und 2008 vermieden werden [Sig05b].

Eine Folge einer verbesserten Konvertierung des EuroGOV Korpus könnte die Verwendung von Blind Relevance Feedback während des Retrievalprozesses sein. Während einiger Indexierungsversuche wurden Termvektoren als Vorbereitung für die Blind Relevance Feedback Funktion berechnet. Aufgrund der häufig auftretenen Parser Exception Meldungen, mussten die Berechnungsprozesse abgebrochen werden. Folge dieser unvollständigen Termvektorenberechnung ist die fehlerhafte Verwendung von Blind Relevance Feedback. Als Ziel für die nächste Teilnahme sollte die Verwendung von Blind Relevance Feedback im Retrievalprozess auf Basis eines Korpus in wohlgeformten XML stehen. Die nötigen Java Klassen zur Nutzung von Blind Relevance Feedback, sind auf der DVD dieser Arbeit zu finden. Erfahrungen aus verschiedenen Teilnahmen am CLEF ad-hoc multi-lingual Retrieval Track [Hac05b, Hac05c, Hac04] haben gezeigt, dass die Verwendung von Blind Relevance Feedback im Retrievalprozess eine erhebliche Steigerung der Ergebnisse erzielen kann.

5.2 Topicentwicklung und Relevanzbewertung

Die Topicentwicklung ist eine zeitintensive Aufgabe und erfordert ein hohes Maß an Geduld und Konzentration von Seiten des Entwicklers. Wie im Abschnitt 4.2 beschrieben, wird der Topicentwicklung eine hohe Bedeutung zugemessen, da sie ausschlaggebend für den guten Verlauf eines Tracks sein kann. Zusätzlich wurde eine Relevanzbewertung parallel zur Topicentwicklung durchgeführt. Die Relevanzbewertung ist der unumgängliche Indikator für die Evaluation eines Runs und im Endeffekt auch für einen gesamten Track. Im Bewertungsprozess der WebCLEF Topics mussten anhand des gleichen Verfahrens zur Topicentwicklung relevante Duplikate zu jedem Topic gefunden werden. In Abschnitt 1.4 ist der genaue Ablauf zur Generierung eines WebCLEF Topics beschrieben. Die Erfahrungen beim Nutzen der zur Verfügung stehenden Tools war nicht sehr befriedigend. Deswegen kann davon ausgegangen werden, dass die Relevanzbewertung aller WebCLEF Topics nicht vollständig ist. Auf der WebCLEF Internetseite wird zu diesem Thema folgendes gesagt:

> *Both duplicate detection and translation detection are likely to be incomplete. I.e. we use the data provided by topic authors at topic development stage. For practical reasons this may be very incomplete assessments. [Web05]*

Die Zahl der relevanten Dokumente, die nicht bekannt sind, kann nur geschätzt werden. In den Richtlinien zur Entwicklung von WebCLEF Topics wurden die Entwickler lediglich aufgefordert, Duplikate und Übersetzungen von Topicdokumenten heraus zu filtern. Mit keinem Wort wurde die Tragweite dieser Aufgabe beschrieben. Aus den Richtlinien geht nicht deutlich hervor, dass der Topicentwickler neben dem Entwickeln von Topics auch für die Relevanzbewertung zuständig ist, bzw. sie parallel mit dem Suchen von Duplikaten und Übersetzungen sogar durchführt. Darauf muss in den nächsten Richtlinien bzw. Aufgabenbeschreibung zur Topicentwicklung eingegangen werden. Die Tatsache, dass die Relevanzbewertung parallel zur Topicentwicklung stattfand, war dem IFAS während des Entwicklungsprozesses nicht bewusst. Da Topicentwicklung in den meisten Fällen durch Studenten der teilnehmenden Universitäten durchgeführt wird, kann davon ausgegangen werden, dass die wenigsten Entwickler die Erfahrung besitzen, um die parallel abgewickelte Bewertung der Topics automatisch zu erkennen. Leider konnte von Seiten des IFAS keine Plausibilitätsprüfung der als relevant eingestuften Dokumente gegenüber den jeweiligen Topics durchgeführt werden.

In Abschnitt 4.2 werden die Schwierigkeiten, die die Tools bei der Generierung von deutschen WebCLEF Topics verursachten haben, näher erklärt. Ob die bereitgestellten

Suchmaschinen den Qualitätsansprüchen der Initiative entsprechen, muss im nachhinein bewertet werden. Grundsätzlich kann man sagen, dass die Qualität bzw. die Relevanzbewertung eines Tracks von den bereitgestellten Tools abhängt. Mit diesem Gedanken sollte die ILPS Lucene: Searchbox überarbeitet werden, da die Ergebnisse während des Entwicklungsprozesses sehr fragwürdig waren. Wie groß die Auswirkung der Entwicklungsphase auf den Verlauf des gesamten Tracks ist, kann erst nach Auswertung der Runs und den dazugehörigen Systemen bestimmt werden. Die Basis der Relevanzbewertung kann aber schon heute als fragwürdig angesehen werden.

5.3 Multilinguale Stoppwortliste

Wie in dem Abschnitt 4.4 näher beschrieben, ist die zur Indexierung verwendete Stoppwortliste den Ansprüchen der Topicsammlung nicht vollständig genug. Für den WebCLEF 2005 Track wurden Topics aus 11 Sprachen generiert. Die Stoppwortliste konnte ca. 64% dieser Sprachen abdecken. Diese nicht ausreichende Abdeckung konnte aus zeitlichen Engpässen nicht mehr behoben werden. Da das Benutzen von gut geführten Stoppwortlisten einen erheblichen Einfluss auf die Retrievalergebnisse hat, würde das vervollständigen der Listen sich empfehlen. Welche Sprachen in den WebCLEF Topics 2006 tatsächlich vorkommen werden bleibt abzuwarten.

5.4 Indizetypen des IFAS

Unabhängig von den Indexierungsmethoden wurden für den WebCLEF Track und in den Postexperimenten drei verschiedene Arten von Indizes erstellt.

- Index auf Basis des HTML Title-Felds jedes Dokuments
- Index auf Basis des gesamten Dokuments (Title- und Content- Feld)
- Index auf Basis des Title- und Content-Felds, wobei nur maximal 200 Zeichen des Content-Felds indexiert wurden

Auf Basis der Abbildung 4.9 konnte sich kein Indextyp deutlich von den anderen absetzen. Das beste Ergebnis wurde mit einem Title Index des StandardAnaylzers erzielt. Dieses Ergebnis konnte von den weiteren Indizes nicht bestätigt werden. Die einzige Aussage, die auf alle Indizetypen zutrifft, ist die Wahrscheinlichkeit, dass an der 50. Stelle noch relevante Dokumente für ein Topic gefunden wird, nimmt mit Größe des Indexes zu. Ein

Title Index kann die Ergebissse über alle getätigten Runs des IFAS an 50. Stelle durchschnittlich um nur 2% verbessern. Die Ergebnisse der Gesamtindizes hingegen steigern ihre Ergebnisse an 50. Stelle um 7% im Durchschnitt über alle Runs. Um konkretere Aussagen zu bestimmten Indextypen machen zu können, müssen noch weitere Runs durchgeführt und ausgewertet werden.

5.5 Evaluierung der Runs

Die Evaluierung der eingereichten Runs wurde anhand der veröffentlichten QREL Dateien durchgeführt. Die QREL Dateien basierten auf den Relevanzbewertungen der einzelnen Topics. Für die Auswertung einzelner Runs wurde der Mean Reciprocal Rank pro Topic und im Durchschnitt über alle 547 Topics errechnet. Zusätzlich wurde die Wahrscheinlichkeit errechnet, dass ein relevantes Dokument an erster, fünfter, zehnter, zwanzigster und fünzigster Stelle auftreten würde. Der Mean Reciprocal Rank ist die Maßeinheit, um known item Search Runs zu bewerten, da der MRR nur das erste relevante Dokument in die Bewertung des Runs mit einfließen lässt. Für den Mixed monolingual Task taugt dieses Maß. Bei einem multilingualen Run stellt sich jedoch die Frage, ob dieses Evaluationsmaß angebracht ist. Die Grundannahme eines multilingualen Runs ist es, alle Dokumente sprachunabhängig zu einem bestimmten Thema / Topic zu finden. Wenn ein System nun vier relevante Dokumente zu einem Topic findet, wird nur das erste Dokument für die Bewertung berücksichtigt. Damit sind die anderen Dokumente und der multilinguale Ansatz dieses Tasks eigentlich hinfällig. Für den nächsten Durchgang muss erörtert werden, ob ein multilingualer Task überhaupt auf Basis des known item Search Prinzips aufgestellt werden sollte.
Im NTCIR-4WEB Information Retrieval Task wird der Weighted Reciprocal Rank als Bewertungsmaß benutzt. Der Hintergedanke in diesem Task ist die multi-grade Bewertung (Abschnitt 2.2) der Topics. Die Bewertungen multilingualer Runs im WebCLEF Track könnten ähnlich durchgeführt werden. Die Basis zur Berechnung des WRR sollten nicht die unterschiedlichen Relevanzen, sondern die möglichen Dokumentsprachen pro Topic sein. Diese könnten auf der Grundlage des Userprofils (Metadaten des WebCLEF Topics) gewichtet werden. Mit vorhandener Gewichtung und den einzelnen MRRs der Dokumente pro Sprache könnte ein angemessener Durchschnittswert (WRR) für multilinguale Runs berechnet werden. Ob diese Variante der Evaluierung von multilingualen Runs sich bewährt, muss anhand von sprachabhängigen Relevanzbewertungen getestet werden. Eine Anpassung der Evaluierungsmethode von multilingualen Runs, muss den multilingualen Ansprüchen entsprechen, um den eigentlichen Sinn dieses Tasks nicht zu verfehlen.

5.6 Teilnahme des IFAS am CLEF Web Track 2006

Die Teilnahme am WebCLEF 2005 Track war für das IFAS der Universität Hildesheim sehr erfolgreich und empfiehlt daher zu einer Weiterentwicklung des verwendeten Prototypen. Aufgrund der Ergebnisse bieten die monolingualen Runs noch ein großes Verbesserungspotential. Beim Betrachten der einzelnen Runs des IFAS waren die Ergebnisse der monolingualen und der multilingualen Runs allerdings sehr eng aneinander gekoppelt. Im Durchschnitt wichen die multilingualen Runs nur um 1,7% Punkte von den monolingualen Runs ab. Dies würde bedeuten, dass Veränderungen des gesamten Prozesses Auswirkungen auf beide WebCLEF Disziplinen haben würde. Deswegen muss abgewogen werden ob ein gekoppelter oder losgelöster Ansatz für beide Disziplinen gewählt wird, bevor die zweite Generation von WebCLEF Experimenten durchgeführt wird. Eine wichtige Erfahrung die während der Experimente gemacht wurde ist die Zeitaufwendigkeit der einzelnen Teilabschnitte. Bei erfolgreicher Umsetzung der Verbesserungsvorschläge ist mit erheblichem ressourcen- und zeittechnischen Aufwand zu rechnen. Daher ist die Umsetzung der folgenden Verbesserungsvorschläge abhängig von dem Beginn der Experimente und dem Know-How der beteiligten Personen.

Index Die vollständige Indexierung des EuroGOV Korpus ist in diesem Jahr, wie in Abschnitt 5.1 näher beschrieben, fraglich. Bei der Verarbeitung des Korpus kann viel verbessert werden. Die Konvertierungsprozesse müssen überarbeitet werden, um die unzähligen ungültigen XML Ausdrücke aus dem Korpus zu eliminieren. Das mögliche Vorgehen wurde im Abschnitt 5.1 erklärt. Ob ein Korpus dieser Größe vollständig konvertiert werden kann oder nicht, bleibt abzuwarten. Bei erfolgreicher Konvertierung wäre die Nutzung der Blind Relevance Feedback Funktion in der WebCLEFSearch Klasse möglich.

Topic Metadaten Während des diesjährigen Tracks wurden von dem IFAS lediglich Baseline Runs eingereicht. Das bedeutet, dass die Metadaten eines WebCLEF Topics noch nicht verwendet wurden. Die Metadaten beinhalten zwei Profile, die zur Verbesserung bzw. Gewichtung der Ergebnisse genutzt werden können. Das erste Profil ist das sogenannte Targetprofile und beinhaltet die Sprache und Domäne des Zieldokuments. Das zweite Profil wird Userprofile genannt und speichert die Sprachfähigkeit und den Geburts- und Wohnort des Users. Beide Profile können zur Gewichtung und somit zur Ordnung der relevanten Dokumente in einer Ergebnisliste beitragen. Das Targetprofile könnte helfen die gesuchten Zieldokumente des Topics in einer multilingualen Ergebnisliste anhand von Sprache und Domäne zu ordnen, sodass

die relevanteren Dokumente im oberen Bereich der Liste aufgeführt werden. Um das Targetprofile nutzen zu können, muss jedes Dokument des EuroGOV Korpus mittels eines Sprachidentifizierers die richtige Sprache zugewiesen bekommen. Die WebCLEF Organisatoren haben der WebCLEF Testkollektion eine Liste mit den identifizierten Sprachen pro Dokument mitgeliefert. Der verwendete Sprachidentifizierer hat im Durchschnitt jedem Dokument 2,2 Sprachen zugewiesen. 15,4% der Dokumente konnte keine Sprache zugeweisen werden. Um das Targetprofile effektiv nutzen zu können müsste dieser Sprachidentifizierungsprozess überarbeitet und den Sprachvorgaben des Topics angepasst werden. Die Informationen über die Sprache in einem EuroGOV Dokument und die Sprachinformationen in einem WebCLEF Topic müssen in ihrem Format übereinstimmen. Da der Sprachidentifizierer der WebCLEF Organisatoren im Ergebnis nicht überzeugt hat, könnte an dieser Stelle eine andere Alternative getestet werden. Die Nutzung des LangIdent Tools, das im Rahmen einer Magisterarbeit [Art05] an der Universität Hildesheim entwickelt wurde, würde sich hier anbieten. Die Nutzung der Userprofile Daten wäre eine weitere Möglichkeit,

```
<topic>
<num>WC0008</num>
<title>LKW Maut Verordnung</title>
<metadata>
<topicprofile>
<language language="DE" />
<translation language="EN">German Truck Toll Regulation</translation>
</topicprofile>
<targetprofile>
<language language="DE" />
<domain domain="de"/>
</targetprofile>
<userprofile>
<native language="DE" />
<active language="EN" />
<passive language="NL" />
<passive language="ES" />
<countryofbirth country="DE" />
<countryofresidence country="DE" />
</userprofile>
</metadata>
</topic>
```

Abb. 5.1: Metadaten eines WebCLEF Topics

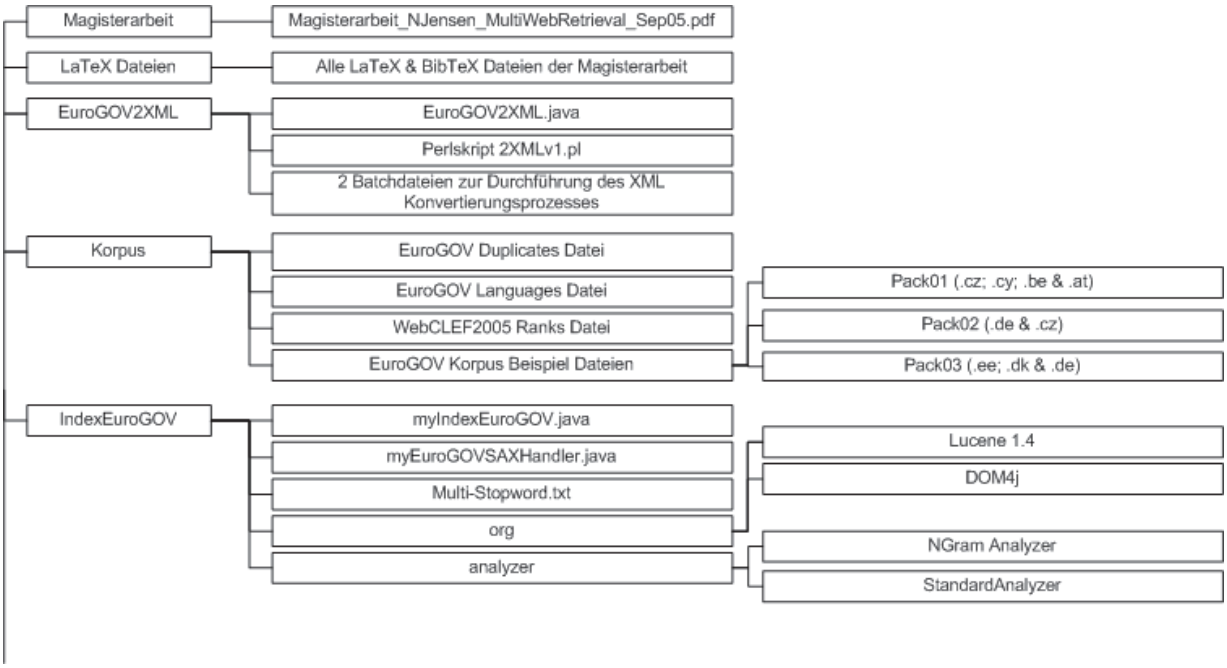
die Ergebnislisten des Retrievalprozesses noch effektiver zu ordnen. Die Userprofile Daten würden den Weg zu einer personalisierten Suche ebnen. Die Dokumente, die durch ein Topic gefunden wurden, könnten so nach den sprachlichen Fähigkeiten des Users geordnet werden. Dokumente in einer Sprache die der User nicht spricht

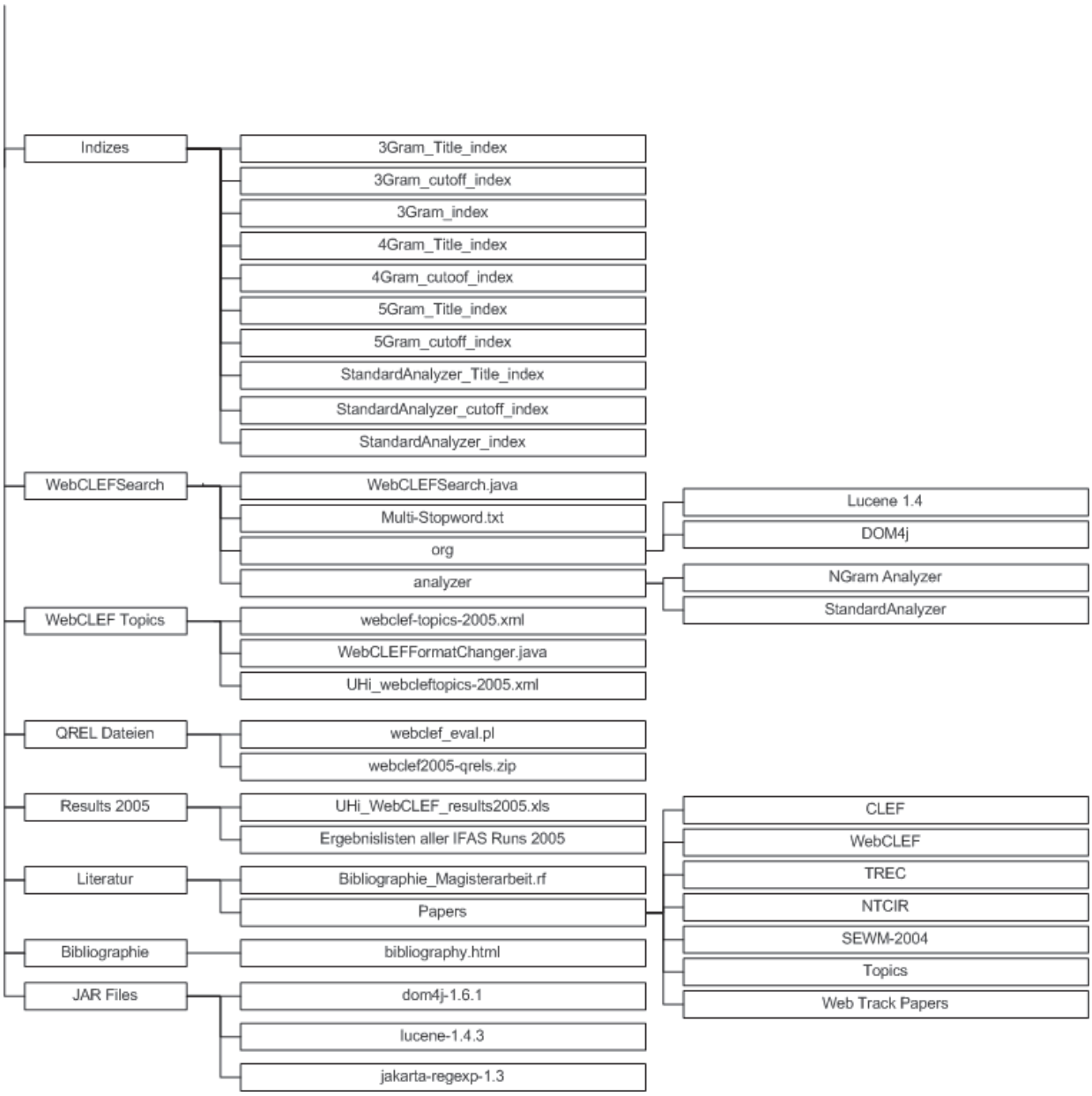
könnten so ans Ende der Ergebnisliste gestellt oder sogar gelöscht werden. Außerdem bieten die Informationen zur aktiven bzw. passiven Sprachfähigkeit des Users eine weitere Möglichkeit des Gewichtens von Dokumenten. Personalisierte Retrievalprozesse würden dem User die Möglichkeit bieten, bequem die ersten Dokumente einer Liste zu evaluieren, da sie keine besondere sprachliche Herausforderung für ihn darstellen. Die erfolgreiche Nutzung der WebCLEF Metadaten kann nur umgesetzt werden, wenn die Sprachidentifizierung der einzelnen Dokumente mit der Sprachidentifizierung der Topics abgleichbar ist.

Die Teilnahme an dem WebCLEF Track 2005 bot viele Herausforderungen fachlicher und technischer Art. Oft funktionierten bekannte Ansätze aufgrund der Korpusgröße nicht und mussten komplett umstrukturiert werden. Die geplante Vorgehensweise, das MIMOR System des CLEF 2004 ad hoc Retrieval Tasks [Hac05c] an die WebCLEF Anforderungen zu adaptieren, konnte nur teilweise umgesetzt werden, da Topics und Korpus zu große Unterschiede zu den vergangenen CLEF Tasks aufwiesen. Insgesamt konnten Erfahrungen zu den Themen Datenkonvertierung von großen Korpora in XML, arbeiten mit dem Lucene Indexierungsmodell und der Lucene Query Engine sowie dem Evaluieren von Web Tracks gesammelt werden. Die genaue Auswertung des Web Tracks kann erst nach Abhalten des CLEF Workshops in Wien vom 21.09. bis zum 23.09.2005 und der Herausgabe der offiziellen Workshop Proceedings erfolgen. Interessant für das Retrievalsystem des IFAS wird der Vergleich der unterschiedlichen Systeme und Herangehensweisen der anderen Teilnehmer. Im Großen und Ganzen kann aber jetzt schon gesagt werden, dass der Ansatz, einen multilingualen Index zu erstellen, die richtige Entscheidung für den multilingual Task war, auch wenn diese Entscheidung aufgrund zeitlicher Engpässe und technischer Probleme getroffen wurde. Für die nächste Teilnahme gilt es also, den Ansatz weiter zu verfolgen und die möglichen Verbesserungen zu implementieren.

Anhang A

IFAS @ WebCLEF2005 DVD Sitemap




Indizes
3Gram_Title_index
3Gram_cutoff_index
3Gram_index
4Gram_Title_index
4Gram_cutoof_index
5Gram_Title_index
5Gram_cutoff_index
StandardAnalyzer_Title_index
StandardAnalyzer_cutoff_index
StandardAnalyzer_index
WebCLEFSearch
WebCLEFSearch.java
Multi-Stopword.txt
org
analyzer
Lucene 1.4
DOM4j
NGram Analyzer
StandardAnalyzer
WebCLEF Topics
webclef-topics-2005.xml
WebCLEFFormatChanger.java
UHi_webcleftopics-2005.xml
QREL Dateien
webclef_eval.pl
webclef2005-qrels.zip
Results 2005
UHi_WebCLEF_results2005.xls
Ergebnislisten aller IFAS Runs 2005
Literatur
Bibliographie_Magisterarbeit.rf
Papers
CLEF
WebCLEF
TREC
NTCIR
SEWM-2004
Topics
Web Track Papers
Bibliographie
bibliography.html
JAR Files
dom4j-1.6.1
lucene-1.4.3
jakarta-regexp-1.3

Literaturverzeichnis

[Aqu05] AQUAINT Corpus http://www.ldc.upenn.edu/Catalog/docs/LDC2002T31/, verifiziert am 07.09.2005.

[Ari04] Arikawa, Masatoshi; Sagara, Takesh; Noaki, Kouzou; Fujita, Hideyuk. Preliminary Workshop of Geographic Information Retrieval Systems for Web Documents. *Working Notes of the 4th NTCIR Workshop Meeting; Tokyo, Japan, June 2004 http://research.nii.ac.jp/ntcir-ws4/NTCIR4-WN/OV/NTCIR4WN-OV-KandoN.pdf*, 2004.

[Art05] Artemenko, Olga; Shramko, Margaryta. Entwicklung eines Werkzeugs zur Sprachidentifikation in mono- und multilingualen Texten. *Magisterarbeit Internationales Informationsmanagement, Universität Hildesheim*, 2005.

[Beh00] Behme, Henning; Mintert, Stefan. XML in der Praxis. *Addison-Wesley*, pages 396 –399, 2000.

[Bia05] Bialecki, Andrzej. Lucene Index Toolbox http://www.getopt.org/luke/, verifiziert am 03.09.2005.

[Bus45] Bush, Vannevar. As We May Think. *Atlantic Monthly*, 176 (1):101–108, 1945.

[Cla05] Claus, Lena; Göpel, Sabine; Jensen, Niels; Spichal, Carsten. Arbeitsbericht der Projektgruppe Fusion. *Projektkurs Cross language Information Retrieval Sommersemester 2004 Universität Hildesheim*, 2005.

[CLE05] Homepage der CLEF Initiative http://clef.isti.cnr.it/, verifiziert am 07.09.2005.

[Cra05] Craswell, Nick; Hawking, David. Overview of the TREC-2004 Web Track. *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004); Gaithersburg, Maryland, November 16-19, 2004 http://trec.nist.gov/pubs/trec13/papers/WEB.OVERVIEW.pdf*, 2005.

[Dim05] Dimmick, D.; OBrien, G.; Over, P.; Rodgers, W.. PRISE Searchengine (z39.50/prise 2.0) Projectguide http://www.nist.gov/itl/div894/894.02/, (1998) verifiziert am 03.09.2005.

[DOM05] Xerces Java Parser (DOM Parser) http://xml.apache.org/xerces-j/, verifiziert am 08.09.2005.

[Egu04a] Eguchi, Koji. Overview of the Topical Classification Task at NTCIR-4WEB. *Working Notes of the 4th NTCIR Workshop Meeting; Tokyo, Japan, June 2004 http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings4/WEB/NTCIR4-OV-WEB-D-EguchiK.pdf*, 2004.

[Egu04b] Eguchi, Koji; Oyama, Keizo; Aizawa, Akiko; Ishikawa, Haruko. Overview of the Information Retrieval Task at NTCIR-4WEB. *Working Notes of the 4th NTCIR Workshop Meeting; Tokyo, Japan, June 2004 http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings4/WEB/NTCIR4-OV-WEB-A-EguchiK.pdf*, 2004.

[Egu04c] Eguchi, Koji; Oyama, Keizo; Aizawa, Akiko; Ishikawa, Haruko. Overview of the NTCIR-4WEB Navigational Retrieval Task 1. *Working Notes of the 4th NTCIR Workshop Meeting; Tokyo, Japan, June 2004 http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings4/WEB/NTCIR4-OV-WEB-B-OyamaK.pdf*, 2004.

[Egu04d] Eguchi, Koji; Oyama, Keizo; Aizawa, Akiko; Ishikawa, Haruko. Overview of WEB Task of the Fourth NTCIR Workshop. *Working Notes of the 4th NTCIR Workshop Meeting; Tokyo, Japan, June 2004 http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings4/WEB/NTCIR4-OV-WEB-EguchiK.pdf*, 2004.

[Ent05] TREC Enterprise Track Homepage http://www.ins.cwi.nl/projects/trec-ent/, verifiziert am 31.08.2005.

[Hac03] Hackl, René; Kölle, Ralph; Mandl, Thomas; Womser-Hacker, Christa. Domain Specific Retrieval Experiments with MIMOR at the University of Hildesheim. In Peters, Carol; Braschler, Martin; Gonzalo, Julio; Kluck, Michael, editor, *Advances in Cross-Language Information Retrieval: Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002, Rome, Italy, September 2002. Berlin et al.: Springer [Lecture Notes in Computer Science 2785] S.*

343-348. Vorab in: Working Notes CLEF-Workshop. 19.-20.9.2002. Rom. S. 241-244. http://clef.iei.pi.cnr.it:2002/workshop2002/WN/32.pdf, 2003.

[Hac04] Hackl, René; Kölle, Ralph; Mandl, Thomas; Ploedt, Alexandra; Scheufen, Jan-Hendrik; Womser-Hacker, Christa. Multilingual Retrieval Experiments with MIMOR at the University of Hildesheim. In Peters, Carol; Braschler, Martin; Gonzalo, Julio; Kluck, Michael, editor, *Comparative Evaluation of Multilingual Information Access Systems: 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Trondheim, Norway, August 21-22, 2003, Revised Selected Papers. Berlin et al.: Springer [Lecture Notes in Computer Science 3237] S. 166-173. http://www.springerlink.com/index/96JHXR0W6229UW5H Vorab in: Working Notes CLEF-Workshop. 21.-22.8.2003. S. 79-82. http://clef.iei.pi.cnr.it:2002/2003/WN_web/10.pdf*, 2004.

[Hac05a] Hackl, René; Mandl, Thomas. Bilingual Retrieval Experiments with Social Science Documents. *Working Notes of the 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005. Sept. 2005, Wien. http://www.clef-campaign.org/2005/working_notes/workingnotes2005/hackl-mandl05.pdf*, 2005.

[Hac05b] Hackl, René; Mandl, Thomas; Womser-Hacker, Christa. Ad-hoc Mono- and Multilingual Retrieval Experiments at the University of Hildesheim. *Working Notes of the 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005. Sept. 2005, Wien. http://www.clef-campaign.org/2005/working_notes/workingnotes2005/hackl05.pdf*, 2005.

[Hac05c] Hackl, René; Mandl, Thomas; Womser-Hacker, Christa. Mono- and Crosslingual Retrieval Experiments at the University of Hildesheim. In Peters, Carol; Clough, Paul; Gonzalo, Julio; Kluck, Michael; Jones, Gareth; Magnini, Bernard, editor, *Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign. Berlin et al.: Springer [Lecture Notes in Computer Science 3491] S. 165-169. http://dx.doi.org/10.1007/11519645_17. Vorab in: Working Notes CLEF-Workshop.15.-17.8.2004. Bath, England. S. 123-125. http://clef.iei.pi.cnr.it:2002/2004/working_notes/WorkingNotes2004/16.pdf*, 2005.

[Hof05] Hofman Miquel, Laura. Informationslinguistische Ressourcen für das Information Retrieval in der tschechischen Sprache im Rahmen des Cross Language

Evaluation Forums (CLEF). *Magisterarbeit Internationales Informationsmanagement, Universität Hildesheim*, 2005.

[Jen05a] Jensen, Niels. Mehrsprachiges Information Retrieval mit einem WEB-Korpus. In Mandl, Thomas; Womser-Hacker, Christa, editor, *Proceedings des Vierter Hildesheimer Information Evaluierungs und Retrieval Workshop (HIER) Hildesheim, 20.7.2005. Univ.-verlag Konstanz*, 2005.

[Jen05b] Jensen, Niels; Hackl, René; Mandl, Thomas; Strötgen, Robert. Web Retrieval Experiments with the EuroGOV Corpus at the University of Hildesheim. *Working Notes of the 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005. Sept. 2005, Wien. http://www.clef-campaign.org/2005/working_notes/workingnotes2005/jensen05.pdf*, 2005.

[Kan04] Kando, Noriko. Overview of the Fourth NTCIR Workshop. *Working Notes of the 4th NTCIR Workshop Meeting; Tokyo, Japan, June 2004*, 2004.

[Klu02] Kluck, Michael; Womser-Hacker, Christa. Inside the Evaluation Process of the Cross-Language Evaluation Forum (CLEF): Issues of Multilingual Topic Creation and Multilingual Relevance Assessment. In Rodriguez, M. G.; Araujo, C.P.S., editor, *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002, Las Palmas de Gran Canaria, 29-31 May 2002. Paris: ELRA*, pages 573–576, 2002.

[Les96] Lesk, Michael. The Seven Ages of Information Retrieval. *International Federation of Library Associations and Institutions Universal Dataflow and Telecommunications Core Programme http://lesk.com/mlesk/ages/ages.html*, Occasional Paper 5, 1996.

[Lev65] Levenshtein, Vladimir I.. Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR*, 163 (4):S. 845–848, 1965.

[Luc05] Lucene Project Homepage http://lucene.apache.org, verifiziert am 16.08.2005.

[Man03] Mandl, Thomas; Womser-Hacker, Christa. Linguistic an Statistical Analysis of the CLEF Topics. In Peters, Carol; Kluck, Michael, editor, *Advances in Cross-Language Information Retrieval: Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002, Rome, Italy, September 2002.*

Berlin et al.: Springer [Lecture Notes in Computer Science 2785]. Vorab in: Working Notes CLEF-Workshop. 19.-20.9.2002. Rom. S. 317-321. http://clef.iei.pi.cnr.it:2002/workshop2002/WN/40.pdf, pages 505–511, 2003.

[Man04] Mandl, Thomas; Womser-Hacker, Christa. Proper Names in the Multilingual CLEF Topic Set. In Peters, Carol; Braschler, Martin; Gonzalo, Julio; Kluck, Michael, editor, *Comparative Evaluation of Multilingual Information Access Systems: 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Trondheim, Norway, August 21-22, 2003, Revised Selected Papers. Berlin et al.: Springer [Lecture Notes in Computer Science 3237] S. 21-28. http://www.springerlink.com/index/W80EYD5RWB7T579E Vorab in: Working Notes CLEF-Workshop. 21.-22.8.2003. S. 439-443. http://clef.iei.pi.cnr.it:2002/2003/WN_web/53.pdf*, 2004.

[Man05] Mandl, Thomas; Womser-Hacker, Christa. How do Named Entities Contribute to Retrieval Effectiveness? In Peters, Carol; Clough, Paul; Jones, Gareth; Gonzalo, Julio; Kluck, Michael; Magnini, Bernard, editor, *Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign. Berlin et al.: Springer [Lecture Notes in Computer Science 3491] S. 833-842. http://dx.doi.org/10.1007/11519645_81 Vorab in: Working Notes CLEF-Workshop.15.-17.8.2004. Bath, England. S. 649-654. http://clef.iei.pi.cnr.it:2002/2004/working_notes/WorkingNotes2004/77.pdf*, 2005.

[McN04] McNammee, Paul; Mayfield, James. Character N-Gram Tokenization for European Language Text Retrieval. *Information Retrieval*, 7:73–97, 2004.

[Med05] MEDLINE Datenbank Homepage http://medline.cos.com/, verifiziert am 03.09.2005.

[NTC05] NTCIR (NII-NACSIS Test Collection for IR Systems) Project Homepage http://research.nii.ac.jp/ntcir/, verifiziert am 03.09.2005.

[Oar99] Oard, Douglas W.. Topic Tracking with the PRISE Information Retrieval System. *Proceedings of the DARPA Broadcast News Workshop; Herndon, Virgina, 28.02.-03.03., 1999 ; http://www.itl.nist.gov/iaui/894.01/publications/darpa99/pdf/tdt350.pdf*, 1999.

[Por05] Porter, Martin. Porter Stemmer http://www.tartarus.org/m̃artin/PorterStemmer/, verifiziert am 16.09.2005.

[Sav05] Savoy, Jacques. Homepage der Stoppwortlisten der Universität NeuChatel http://www.unine.ch/info/clef, verifiziert am 24.08.2005.

[SAX05] SAX Parser http://sax.sourceforge.net/, verifiziert am 18.08.2005.

[SEW05a] Chinesische Homepage des SEWM 2004 Workshops http://www.scut.edu.cn/sewm2004/, verifiziert am 03.09.2005.

[SEW05b] Englische Homepage des SEWM 2004 Workshops http://net.pku.edu.cn/w̃ebg/cwt/en_index.html, verifiziert am 03.09.2005.

[Sha99] Shakespeare, William. As You Like It, Act 2, Scene 7, lines 143-166. 1599.

[Sig05a] Sigurbjönsson, Bökur. WebCLEF: Participations Guidelines. *http://ilps.science.uva.nl/WebCLEF/participants-guidelines-DRAFT-20050510.pdf*, verifiziert am 17.09.2005.

[Sig05b] Sigurbjörnsson, Bökur. EuroGOV: Engineering a Multilingual Web Corpus. *Working Notes of the 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005. Sept. 2005, Wien. http://www.clef-campaign.org/2005/working_notes/workingnotes2005/derijke05.pdf*, 2005.

[Sig05c] Sigurbjörnsson, Bökur. Overview of WebCLEF 2005. *Working Notes of the 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005. Sept. 2005, Wien. http://www.clef-campaign.org/2005/working_notes/workingnotes2005/sigurbjornsson05.pdf*, 2005.

[Sig05d] Sigurbjörnsson, Bökur. Topic Creation Guidelines WebCLEF2005. *http://ilps.science.uva.nl/WebCLEF/webclef-topics-04042005.pdf*, 2005.

[Sig05e] Sigurbjörnsson, Börkur; Kamps, Jaap; de Rijke, Maarten. Blueprint of a cross-lingual web retrieval collection. *Digital Information Management*, 3 (9-13), 2005.

[Ter05] Terrier Homepage http://ir.dcs.gla.ac.uk/terrier/, verifiziert am 16.08.2005.

[TRE05] TREC Homepage http://trec.nist.gov/, verifiziert am 17.09.2005.

[Voo05] Voorhees, Ellen M.. Overview of TREC 2004. *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004); Gaithersburg, Maryland, November 16-19, 2004 http://trec.nist.gov/pubs/trec13/papers/OVERVIEW13.pdf*, 2005.

[Web05] WebCLEF Homepage http://ilps.science.uva.nl/WebCLEF/index.html, verifiziert am 05.08.2005.

Tabellenverzeichnis

Abbildungsverzeichnis

Danksagung

Ich möchte mich ganz herzlich bei René Hackl, Lars Jensen, Rolf Maichel, Thomas Mandl, Thorsten Trippel und Robert Strötgen für Ihre Unterstützung und Hilfsbereitschaft bedanken. Des Weiteren danke ich Charlotte für die bettelnden Blicke neben dem Schreibtisch, doch endlich fertig zu werden und ganz besonders meiner Frau Ulrike, die in den letzten Monaten große Geduld gezeigt hat.

Eigenständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und ohne fremde Hilfe verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Außerdem versichere ich, dass die Arbeit noch nicht veröffentlicht oder in einem anderen Prüfungsverfahren als Prüfungsleistung vorgelegt wurde.

Hildesheim, im September 2005