In: Hammwöhner, Rainer; Wolff, Christian; Womser-Hacker, Christa (Hg.): Information und Mobilität, Optimierung und Vermeidung von Mobilität durch Information. Proceedings des 8. Internationalen Symposiums für Informationswissenschaft (ISI 2002), Regensburg, 8. – 11. Oktober 2002. Konstanz: UVK Verlagsgesellschaft mbH, 2002. S. 165 – 180

Concept Extractor - Ein flexibler und domänenspezifischer Web Service zur Beschlagwortung von Texten

Lukas C. Faulstich[1], *Uwe Quasthoff*[2], *Fabian Schmidt*[1], *Christian Wolff*[3]

[1]pepper words GmbH,
Ndl. Leipzig
Karl-Heine-Str. 99
04229 Leipzig
{faulstich,fschmidt}@pepper-words.de

[2]Universität Leipzig,
Institut für Informatik,
Augustusplatz 10/11,
04109 Leipzig
quasthoff@informatik.uni-leipzig.de

[3]Universität Regensburg
Institut für Medien-, Informations- und Kulturwissenschaften
93040 Regensburg
christian.wolff@sprachlit.uni-regensburg.de

Zusammenfassung

Der Beitrag beschreibt ein flexibles und modulares System zur automatischen Beschlagwortung von Texten, das auf einer Text Mining-Engine aufbaut. Dabei liegt eine Methode der differentiellen Corpusanalyse zugrunde: Der zu verarbeitende Text wird im Vergleich mit einem unfangreichen Referenzcorpus analysiert und Unterschiede in relativen Häufigkeitsklassen dienen der Auswahl geeigneter Schlagworte. Zusätzlich kommen Datenbanken zum Einsatz, die eine Expansion von Termen hinsichtlich Grundform, Schreibvarianten, Synonymen und Mehrwortbegriffen erlauben. Das System ist als *web service* realisiert und lässt sich problemlos in Content Management-Systeme integrieren.

Abstract

We describe a flexible and modular system for keyword extraction and attribution which operates on top of a text mining engine. Texts are analysed in comparison with a large reference corpus and key words are determined using a frequency based method for determining relative term significance. Additionally, selected terms may be expanded using large knowledge bases on inflected forms, orthographic variants, synonyms and multi word terms. This solution is realised as a web-based service which can easily be integrated into existing content management systems.

1 Einleitung

Die Beschlagwortung redaktioneller Texte, die für verschiedene Medienkanäle aufbereitet und mittels *content syndication* auf unterschiedliche Weise verwertet werden (*cross media publishing*, cf. [Simon 02]), stellt hohe Anforderungen an eine dem jeweiligen Medium angemessene Beschlagwortung. Typische Szenarien umfassen dabei die mehrfache Verwertung medialer Inhalte (hier: redaktionelle Texte) über unterschiedliche Kanäle, z. B. durch

- Publikation auf Websites,
- Lieferung themenspezifischer Inhalte als push-Service per SMS,
- Bereitstellung von Texten für Videotextdienste oder
- die Verwendung in Printmedien.

Für die jeweilige Verwendung gelten jeweils auch unterschiedliche Anforderungen an die inhaltliche Aufbereitung der Texte im Sinne der Informationserschließung, sei es, dass Text für den internen redaktionellen Gebrauch in einem Content Management-System zu beschlagworten ist, sei es, dass eine Beschlagwortung zu erstellen ist, die eine gute Erschließung über Suchmaschinen zulässt (Verwendung von Schlagworten in Meta-Tags der im Web publizierten HTML-Seiten).

Im Folgenden wird ein Beschlagwortungsserver beschrieben, der auf einer mehrschichtigen Analyse von Textcorpora durch Text Mining-Verfahren aufbaut, eine flexible agentenbasierte Beschlagwortungslösung anbietet und als Web Service (cf. [Preece & Decker 02]) verfügbar und damit in unterschiedliche Content Management-Systeme integrierbar ist.

Dabei wird zunächst auf die corpuslinguistischen Grundlagen eingegangen (Kap. 2). Kap. 3 beschreibt Aufbau, Funktionsweise und technische Umsetzung des Beschlagwortungsservers und Kap. 4 gibt praktische Anwendungsbeispiele.

2 Corpuslinguistische Grundlagen

Die Verfügbarkeit elektronischer Texte hat im vergangenen Jahrzehnt zu einer Renaissance der Corpuslinguistik geführt, wie u. a. auch das schnelle Anwachsen internationaler Fachtagungen wie der *Conference on Language Resources and Evaluation* (LREC) zeigt. Mittlerweile liegen nicht nur für viele

Sprachen Referenzcorpora vor, auch die Bemühungen und die Standardisierung von Corpusaufbau und Analyse zeigen Erfolge (cf. [Atkins et al. 02]).

Für die Problematik der automatischen Beschlagwortung von Texten sind corpuslinguistische Verfahren deshalb von Interesse, da sie im Vergleich mit den bekannten Verfahren zur Textindexierung wie dem *vector space model* (cf. [Salton 83], [Baeza-Yates & Ribeiro-Neto 99]) Corpora als zusätzliche Bezugsgröße der Textanalyse einführen und, insofern Corpora nicht nur als Rohdatensammlung, sondern als strukturierter Informationsspeicher zur Verfügung stehen, auch für die Beschlagwortung relevante Informationen bereitstellen können (z. B. Grundformen, Synonyme, Sachgebietsangaben etc.).

2.1 Vergleich von Corpora

Für die dynamische Beschlagwortung von Texten ist dabei der Aspekt des *Vergleichs* unterschiedlicher Corpora von Bedeutung: Text Corpora lassen sich u. a. anhand Faktoren wie Umfang, Art, Anzahl oder Bezugszeitraum der in ihnen enthaltenen Dokumente beschreiben. Durch Vergleich unterschiedlicher Corpora lassen sich dabei Erkenntnisse über die Beschreibungsadäquatheit von Begriffen bezüglich einzelner Dokumente eines Corpus gewinnen.

Mit [Rayson & Garside 00:1] kann man zwei Typen des Corpusvergleichs unterscheiden:

- Vergleich zweier Corpora ähnlicher Größe, die sich hinsichtlich eines Parameters (z. B. Erhebungszeitraum) unterscheiden. Ein solcher Vergleich kann z. B. in der Semiometrie oder Trendforschung eingesetzt werden, um zu bestimmen, inwieweit sich Trends durch geänderten Sprachgebrauch (Wortverwendungshäufigkeiten, unterschiedlicher Vokabularaufbau) nachweisen lassen.
- Vergleich eines kleineren gegen ein größeres (normatives) Corpus, z. B. bei der Differenzierung zwischen Sprachgebrauch in einer Fachdomäne im Vergleich mit einem aus allgemeinsprachlichen Texten aufgebauten Corpus.

Für das nachfolgend beschriebene Beschlagwortungssystem ist der zweite Fall, d. h. der Vergleich von kleinerem Fachcorpus mit einem deutlich größeren normativen allgemeinsprachlichen Corpus der Ausgangspunkt. Als normative Corpusgrundlage dient dabei einerseits

- der im Projekt „Deutscher Wortschatz“ entwickelte Referenzcorpus aus derzeit ca. 300 Millionen laufenden Wortformen, sowie die in ihm enthaltenen zusätzlichen Informationen (cf. [Quasthoff & Wolff 00] [Heyer, Quasthoff, Wolff 02] und http://wortschatz.uni-leipzig.de), andererseits
- die im Rahmen dieses Vorhabens entwickelten Software-Werkzeuge zur Textanalyse, die grundsätzliche auf Textkollektionen beliebigen Umfangs angewandt werden können und als Analyseergebnis eine Datenbank aufbauen, über die (wenigstens) auf Basisdaten wie Wortfrequenzen und Frequenzklassen, Kollokationen oder Grundformrelationen zugegriffen werden kann (cf. [Heyer et al. 01a], [Heyer et al. 01b]).

Fachbegriff	*Häufigkeitsklasse Im Fachcorpus*	*Häufigkeitsklasse im Allgemeinsprachlichen Corpus*	*Differenz*
Hubraum	6	14	8
Nockenwelle	9	18	9
Fahrgeräusch	11	19	8
Zylinder	8	13	5

Tabelle 1: Frequenzvergleich von Fachbegriffen

Der bekannten These folgend, demzufolge sehr seltene Begriffe aufgrund ihrer zu hohen Spezifik für die Beschlagwortung ebenso wenig geeignet sind wie sehr häufige Begriffe (cf. [Salton 83: 62, insb. Abb. 3-2]), spielt bei diesem Ansatz zunächst der Vergleich von Frequenzklassen für Fachbegriffe eine wichtige Rolle bei der Auswahl von Kandidaten für die Textbeschlagwortung: Sowohl in Fach- als auch in Normcorpus hat jeder auftretende Begriff eine absolute sowie eine aus ihr im Verhältnis zur Corpusgröße berechnete relative Frequenzklasse.[1] Über die Voraussetzung einer Mindestfrequenzklasse lassen sich zu häufige bzw. zu seltene Begriffe aus der Analyse ausschließen, über den Vergleich von Frequenzklassen zwischen Fach- und Normcorpus lassen sich geeignete Beschlagwortungskandidaten finden, indem gefordert wird, dass als Kandidaten nur solche Wörter ausgewählt werden, deren Frequenzklasse im Fachcorpus wenigstens um eine Mindestdifferenz verfügen (eine Differenz von 2 besagt dabei eine vierfache relative Häufigkeit im Fachcorpus). Am Beispiel einiger Fachbegriffe aus der Automobiltechnik sei dies verdeutlicht. Grundlage ist dabei ein Fachcorpus, das aus vier Jahrgängen einer bekannten Publikumszeitschrift zur Automobiltechnik besteht und das mit dem Normcorpus „Deutscher Wortschatz“ verglichen wurde (cf. [Wolff 01], [Heyer et al. 01b:81, insb. Tab. 7.7]):

[1] Die Frequenzklasse wird als logarithmisches Maß in Relation zum häufigsten Begriff eines Corpus ermittelt. Eine Klasse 4 besagt daher, dass ein Wort um den Faktor 16 (2^4) seltener gesehen wurde als das jeweils häufigste Wort im Corpus).

Der Vergleich von Häufigkeitsklassen ist ein vergleichsweise einfacher Ansatz, um die für einen Corpus charakteristischen Terme zu extrahieren. Kilgarriff sieht darüber hinausgehend in der Berechnung statistischer Prüftests für die in verschiedenen Corpora auftretenden Begriffe ein wesentliches Merkmal für die bessere Beschreibung von Corpora.[2]

2.2 Mehrschichtiger Corpusvergleich - ein Szenario für die Beschlagwortung von Texten

Für die praktische Anwendung des Corpusvergleichs auf das Problem der Beschlagwortung von Texten sei folgendes Szenario vorausgesetzt:

- Ein hinreichend großes Normcorpus steht als Referenzdatenbank zur Verfügung.
- Die zu beschlagwortende Textkollektion wächst im Vergleich zu ihrer Gesamtgröße relativ langsam.
- Jedes einzelne Dokument kann selbst als ein Textcorpus behandelt werden.

Die Grundlage des Beschlagwortungssystems ist zunächst eine Verallgemeinerung des wortfrequenzbasierten Corpusvergleichs, da hier das jeweils zu beschlagwortende Dokument als dritte Analyseebene hinzukommt: Die statistische Analyse, durchgeführt mit der im Umfeld des Projekts „Deutscher Wortschatz“ entwickelten Text Mining-Engine *Concept Composer* (cf. [Heyer, Quasthoff, Wolff 00], [Quasthoff & Wolff 00] u. unten Abb. 1), erfolgt für den Startbestand der Texte des jeweiligen Anwendungsgebietes sowie gesondert für jedes neue zu beschlagwortende Dokument. Zusammen mit der Corpusdatenbank des „Deutschen Wortschatzes“, die als *allgemeinsprachliche linguistische Datenbank* (ALDB) den Status eines Normcorpus aufweist, ergeben sich, anders als beim einfachen Corpusvergleich, für jedes Dokument zwei Vergleichsebenen:

- Vergleich mit den Analyseergebnissen für die aktuelle Dokumentenkollektion und
- Vergleich mit dem Normcorpus.

---

[2] „Corpus linguistics lacks a vocabulary for talking, quantitatively, about similarities and differences between corpora. […]. One way of describing differences between corpora is by highlighting the words which have consistently been used more in the one corpus than the other“. [Kilgarriff 01: Kap. 10 – Conclusion].

Es ist offensichtlich, dass ein solcher frequenzbasierter Mehrebenenvergleich sprachliche Variation wie Vollformen, Schreibvarianten (Rechtschreibreform!), Gebrauch von Synonymen oder die Erkennung von Mehrwortbegriffen nicht berücksichtigen sollte. Deshalb tritt für die Auswahl von Beschlagwortungstermen eine zweite wesentliche Systemkomponente: Die *Expansion* und *Reduktion* der durch den mehrschichtigen Corpusvergleich ausgewählten Begriffe durch Rückgriff auf in der allgemeinsprachlichen linguistischen Datenbank vorhandenes linguistisches Wissen, z. B. über Vollform-/Grundformbeziehungen, Synonyme oder Eigennamen.

Als Ergebnis dieser Konzeption steht die Entwicklung eines modularen und hochparametrischen Beschlagwortungssystems, dessen Aufbau und Arbeitsweise im folgenden Kapitel näher beschrieben sind.

3 Systemarchitektur und Arbeitsweise

Das System ist als Web Service realisiert und kann über einen Webserver angesprochen werden. Dabei liegt ein einfaches Kommunikationsschema zugrunde: Der zu beschlagwortende Text wird per http an den Beschlagwortungsservice gesandt, dort wird die Textanalyse durchgeführt und das Beschlagwortungssystem liefert Schlagworte als nach Relevanz geordnete Liste an den Client, z. B. ein Content Management-System (CMS), zurück. Die Anzahl gewünschter Schlagworte kann dabei gesteuert werden. Abbildung 1 gibt hierzu einen Überblick. Das Kommunikations- sowie das nachfolgend beschriebene Steuerungsmodul stellen Querschnittkomponenten des Systems dar, die die einzelnen Module zu

- Textsegmentierung,
- Stoppworteliminierung und Frequenzabgleich,
- Expansion durch linguistisches Wissen und
- Ranking.

steuern.

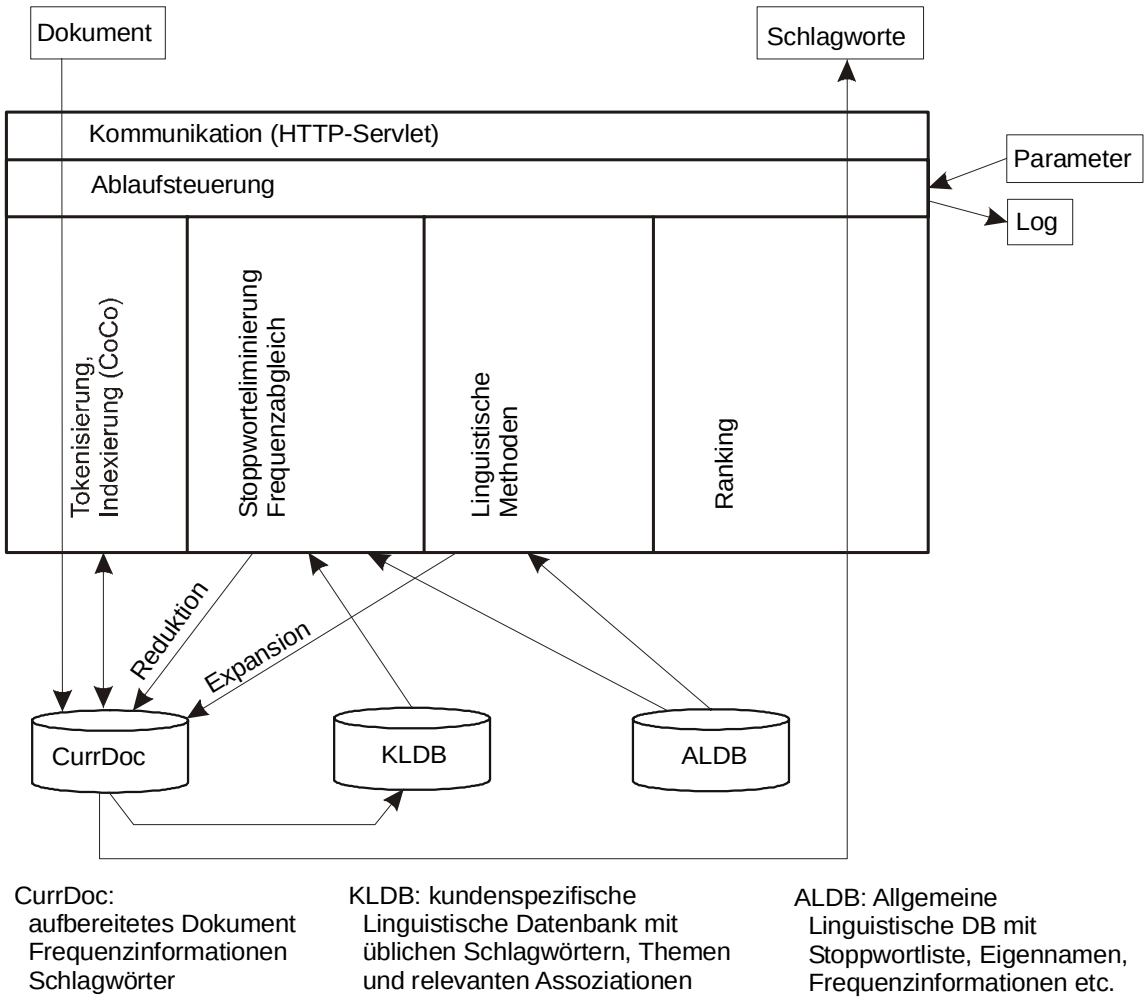

Abbildung 1: Systemarchitektur Concept Extractor

3.1 Steuerung

Das Steuerungsmodul stellt den einzelnen Komponenten des Beschlagwortungssystems einen Zugriff auf

- den aktuell zu verschlagwortenden Text,
- die in den verschiedenen Ebenen der Textanalyse ermittelten linguistischen und statistischen Daten.
- die globalen und komponentenspezifischen Parameter sowie
- ein Logging-System bereit.

Vom Steuerungssystem werden die einzelnen Module zur Bearbeitung der Daten angesteuert und ihre gegenseitigen Abhängigkeiten sichergestellt. Neben der Datenbank *CurrDoc* (Ergebnisse der Textanalyse für das aktuelle Dokument) wird von den Modulen auf eine allgemeinsprachliche Referenz-

datenbank (ALDB) mit umfangreichen linguistischen Daten und eine Datenbank mit dem gesamten Textbestand des Kunden (KLDB) zugegriffen. In dieser sind kundenspezifische Daten (Stoppwörter, Negativ- und Positivliste für Schlagwörter etc.) hinterlegt; sie wird fortlaufend um die neu verschlagworteten Texte ergänzt.

Die Module des Verschlagwortungsprozesses unterteilen sich nach ihren Aufgaben in die Bereiche

- Datenaufbereitung,
- Reduktion und Expansion,
- Bewertung.

Sie werden im Folgenden einzeln beschrieben.

3.2 Segmentierung, Erkennung von Mehrwortbegriffen, Häufigkeitsermittlung, Stoppwörter

Zur Tokenisierung des zu verschlagwortenden Textes wird die Text Mining-Engine *Concept Composer* verwendet. Sie zerlegt den Text in Sätze und Wörter und zählt Wörter aus. Sie enthält zudem eine automatische Erkennung von Mehrwortbegriffen auf der Basis einer umfangreichen, teilautomatisch generierten Mehrwortbegriffsliste des Deutschen, die durch Mehrwortbegriffe aus dem aktuellen Datenbestand der Anwendungsdomäne ergänzt wird (KLDB). In die Analyse gehen auch (HTML-)Strukturmerkmale ein, um eine höhere Gewichtung von Begriffen in Überschriften etc. gewährleisten zu können. Weitere relevante Layoutmerkmale wie etwa Großschreibung können berücksichtigt werden (jeweils durch Ersetzung der häufigsten Schreibvariante, z. B. `ALTERNATIVMEDIZIN` → `Alternativmedizin`).

Die aus dem „Deutschen Wortschatz“ gewonnene frequenzbasierte Stoppwortliste wird erweitert durch Stoppwörter aus der aktuellen Anwendungsdomäne. Sie wird zusätzlich um Großschreibungen und gebeugte Formen von Stoppwörtern ergänzt. Das entsprechende Modul entfernt in der *CurrDoc*-Datenbank alle Begriffe, für die ein Stoppwortflag in der Referenzdatenbank (KLDB) gesetzt ist.

3.3 Expansions- und Reduktionsmodule

Neben den im Text tatsächlich gebrauchten Wörtern kommen weitere Begriffe als Schlagwörter in Frage, die aus den verwendeten Begriffen hergeleitet

werden können. Im Einzelnen kommen dabei die nachfolgend beschriebenen Module zum Einsatz.

3.3.1 Schreibvarianten

Zu allen Wörtern aus dem Ausgangstext wird geprüft, ob es sich um Tippfehler handeln könnte. Dazu wird überprüft, ob das Wort in einer Liste korrekter Wörter enthalten ist. Ist das nicht der Fall, werden für längere Wörter typische Tippfehler mit Levenshtein-Abstand 1 generiert (z.B. Vertauschung zweier Buchstaben, Einfügen und Auslassen von Buchstaben, cf. [Navarro 01, Levenshtein 65]), und diese Varianten auf mögliche Korrektheit hin überprüft. Das Verfahren ist nicht für kurze Wörter anwendbar, da es für diese zu viele mögliche Varianten gibt, die gültige Wörter sind. Für die Wörter des Ausgangstextes wird zudem in der linguistischen Datenbank nachgeschlagen, ob sich die Schreibweise durch die Rechtschreibreform geändert hat. Wenn ja, wird die jeweils neue oder alte Variante ergänzt.

Synonymexpansion

Zu allen Wörtern des Ausgangstextes werden Synonyme aus der Synonymdatenbank der ALDB ergänzt. Diese sind jeweils niedriger bewertet als ihr Ausgangswort. Zusätzliche Informationen zur Stärke der Synonymierelationen können, falls vorhanden, zur differenzierten Gewichtung von Synonymen herangezogen werden.

Grundform- und Wortartbehandlung

Anhand der Informationen in der ALDB werden zu den Wörtern des Ausgangstextes Grundformen ermittelt. Dabei werden

- für gebeugte *Nomina* die Grundform *ergänzt* (Expansion),
- gebeugte *Adjektive* und *Verben* durch ihre Grundform ersetzt, wenn die Grundform nicht deutlich niederfrequenter als die gebeugte Form ist.

Weiterhin erhalten Nomina eine höhere Bewertung für ihre Eignung als Schlagwörter als Adjektive und Verben.

3.4 Frequenzabgleich

Im Anschluss an die zuvor genannten Expansions- und Reduktionsschritte werden alle Wörter einem Frequenzvergleich mit dem in der ALDB gespeicherten Referenzkorpus unterzogen[3]. Hierbei sind zwei Probleme zu berücksichtigen: das zur korrekten Berechnung der Häufigkeitsklasse notwendige

[3] Ein Frequenzvergleich mit der KLDB ist im hier beschriebenen Prototypen noch nicht implementiert.

häufigste Wort des Referenzkorpus kann im zu verschlagwortenden Dokument *CurrDoc* fehlen, und nicht alle Worte aus *CurrDoc* kommen im Referenzkorpus vor. Solche Worte dürfen jedoch nicht a priori ignoriert werden, weil es sich dabei um als Schlagworte relevante Eigennamen handeln kann. Im Prototypen kommt ein Häufigkeits-Quotient zum Einsatz, dessen Logarithmus eine Näherungsformel für die Differenz der Häufigkeitsklassen darstellt, der aber die angesprochenen Probleme vermeidet. Der in der Reduktions- / Expansions-Phase berechnete Score wird dabei gewichtet mit dem Faktor:

```
#(CurrDoc,Wort) / (#(ALDB,Wort) + #(CurrDoc, Wort))
```

wobei *#(CurrDoc,Wort)* bzw. *#(ALDB,Wort)* die absolute Häufigkeit von *Wort* im zu verschlagwortenden Text bzw. im Referenzkorpus *ALDB* ist. Dadurch werden Worte bevorzugt, welche in der ALDB selten vorkommen, insbesondere, wenn sie in *CurrDoc* mehrfach auftreten. Der Score von Worten, welche in der ALDB nicht vorkommen, bleibt nach dieser Formel unverändert. Worte, welche in der ALDB häufig vorkommen, werden dagegen stark abgewertet.

3.5 Heuristiken

Anschließend an den Frequenzabgleich werden verschiedene Heuristiken angewandt, um gute Schlagwortkandidaten mit einem besonderen Bonus zu versehen.

Eigennamen
Gute Kandidaten für Schlagwörter sind Eigennamen (*named entities*). Zu deren Erkennung kann auf einen umfangreichen Bestand geographischer Namen, Namen von Persönlichkeiten und Bezeichnungen von Pflanzen und Tieren in der ALDB zurückgegriffen werden. Außerdem erhalten Begriffe eine höhere Wertung, die als *Sachgebietsbezeichnung* gebräuchlich oder als *Personennamen* bekannt sind. Diese Informationen können um Informationen aus dem Textbestand der Anwendungsdomäne ergänzt werden.

HTML-Markup
Worte, die durch das HTML-Layout besonders hervorgehoben sind (z.B. Link-Texte, Überschriften, fette oder kursive Schrift etc.), werden als wichtig erkannt.

Manuelle Schlagwörter
Besonders gute und einfach zu ermittelnde Kandidaten für Schlagwörter stellen bereits früher verwendete, manuell vergebene, Schlagwörter dar. Eine Liste solcher Schlagwörter wird in der KLDB vorgehalten.

3.6 Bewertung

Zur Erzeugung der Schlagwortliste wird die gewichtete Variante des Bagging-Algorithmus von Breiman [Breiman 94] verwendet. An einer zufällig gewählten, konstanten Menge von Testbeispielen wurden die verschiedenen Expansionsschritte entwickelt. Diese bewerten die vorhandenen Wörter und fügen neue Wörter mit einer Bewertung für ihre Tauglichkeit als Schlagwörter hinzu. Der Bewertungsalgorithmus fasst die einzelnen Resultate gewichtet zusammen. In Abhängigkeit der Bewertung und der Länge der zu erzeugenden Schlagwortliste werden die Schlagwörter nach Relevanz sortiert zurückgegeben.

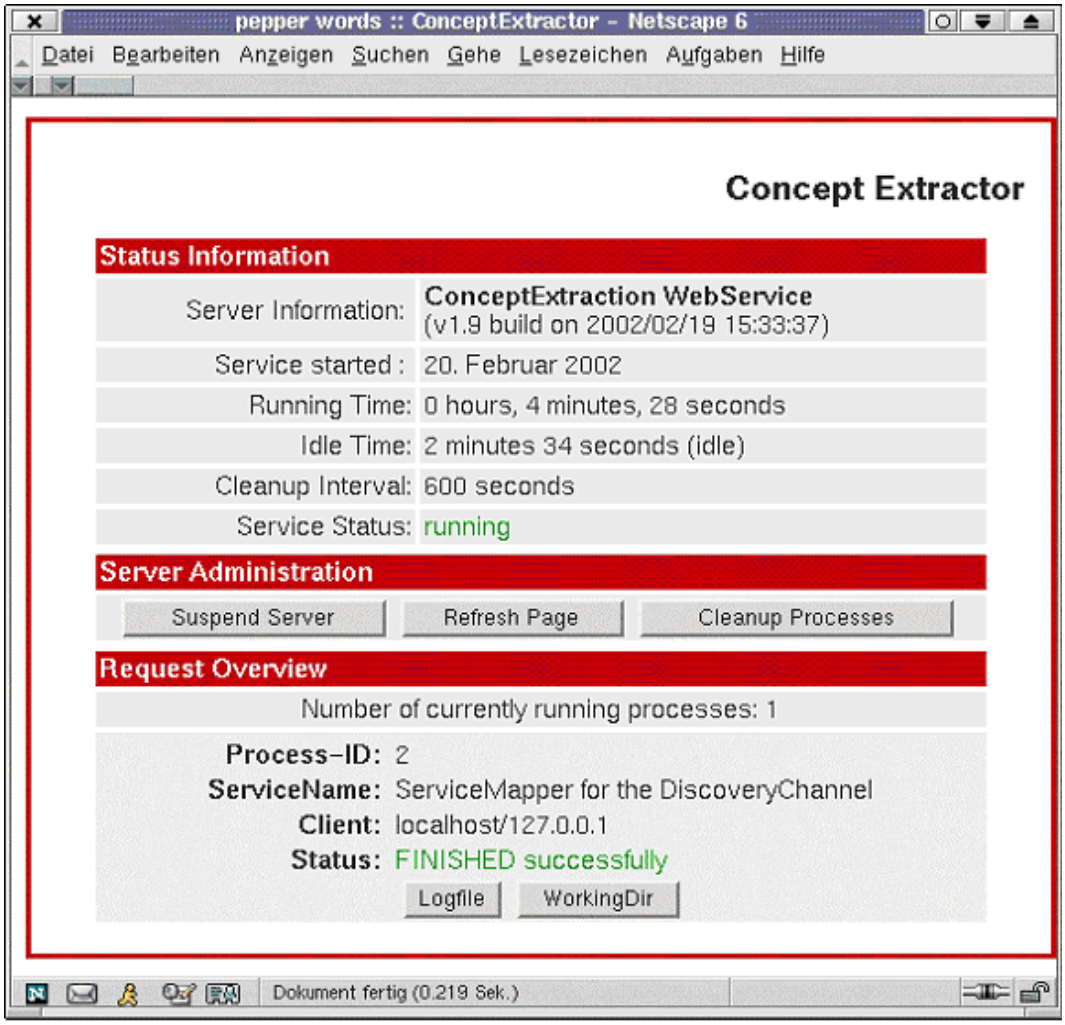

Abbildung 2: Webbasiertes Steuerinterface des Beschlagwortungsservers

3.7 Administration und Steuerung

Um das Beschlagwortungssystem als *web service* bereitstellen und steuern zu können, wurden zusätzlich serverseitige Komponenten entwickelt, die mit Hilfe von *Java Server Pages* (JSP) ein browserbasiertes Administrationsinterface generieren und so die Steuerung und Kontrolle des Beschlagwortungsservers ermöglichen. Die Abbildungen 2 und 3 zeigen jeweils exemplarisch das Steuerinterface sowie die Logging-Ausgabe, die detaillierte Information zum Ablauf der Beschlagwortung enthält.

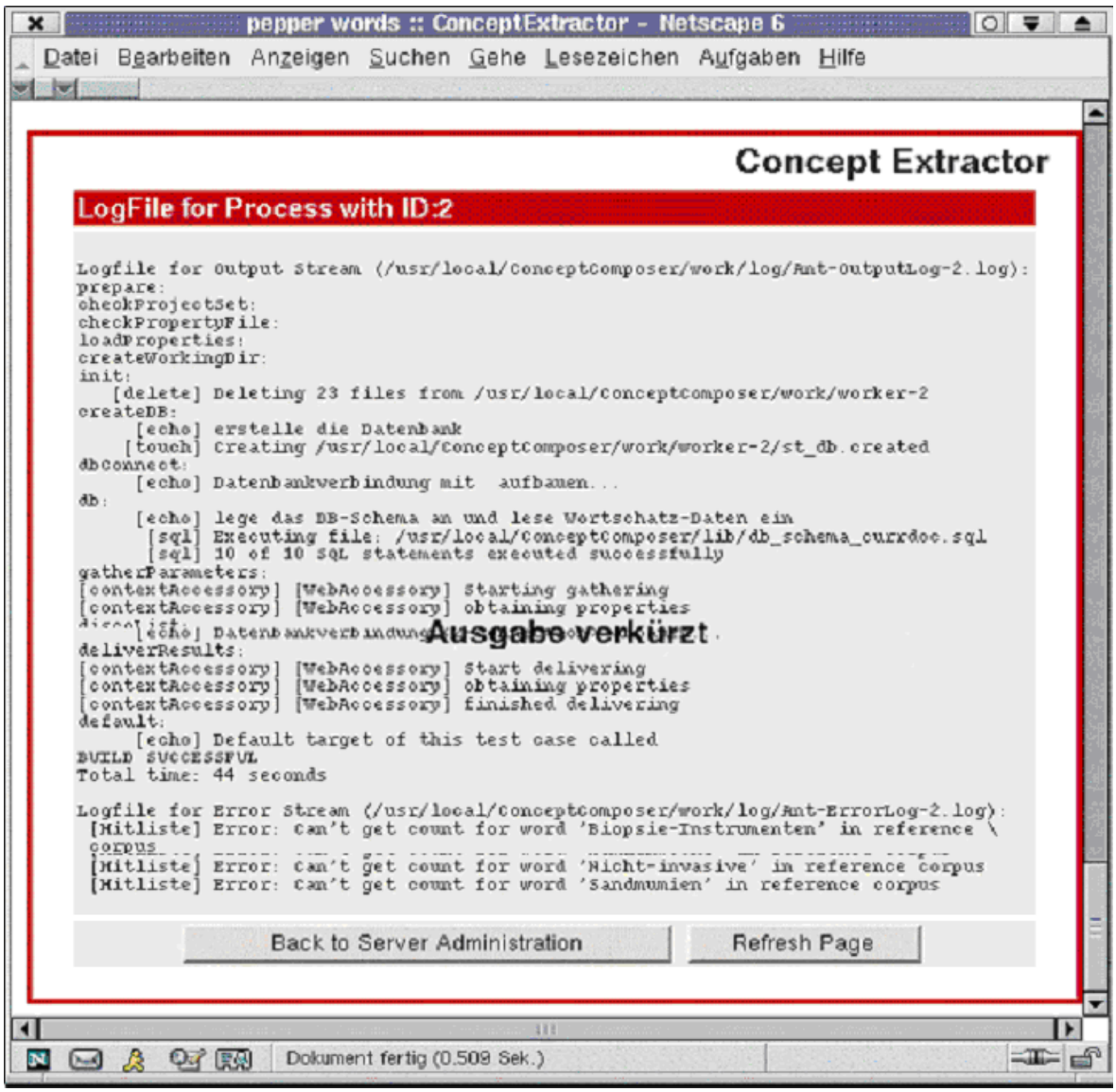

Abbildung 3: Webbasiertes Logging für den Analyseprozess des Concept Extractor

Zusätzlich zu obigen Steuerungsmechanismen existiert ein Testinterface, durch das mit einer webbasierten Upload-Funktion ein beliebiger Text über einen Webbrowser an den Beschlagwortungsserver geschickt werden kann.

4 Ein Beschlagwortungsbeispiel

Abschließend soll an einem Beispiel die Ergebnisqualität der automatischen Beschlagwortung verdeutlicht werden. Als Beispiel wird dabei der in Abb. 4 gezeigte Text über Alternativmedizin verwendet. Von der für den Text verantwortlichen Online-Redaktion wurden dabei durch manuelle Beschlagwortung folgende Begriffe vergeben:

`Medizin, Alternative Heilmethoden, Homöopathie, Traditionelle Chinesische Medizin, Akupunktur, Shiatsu, Ayurveda`

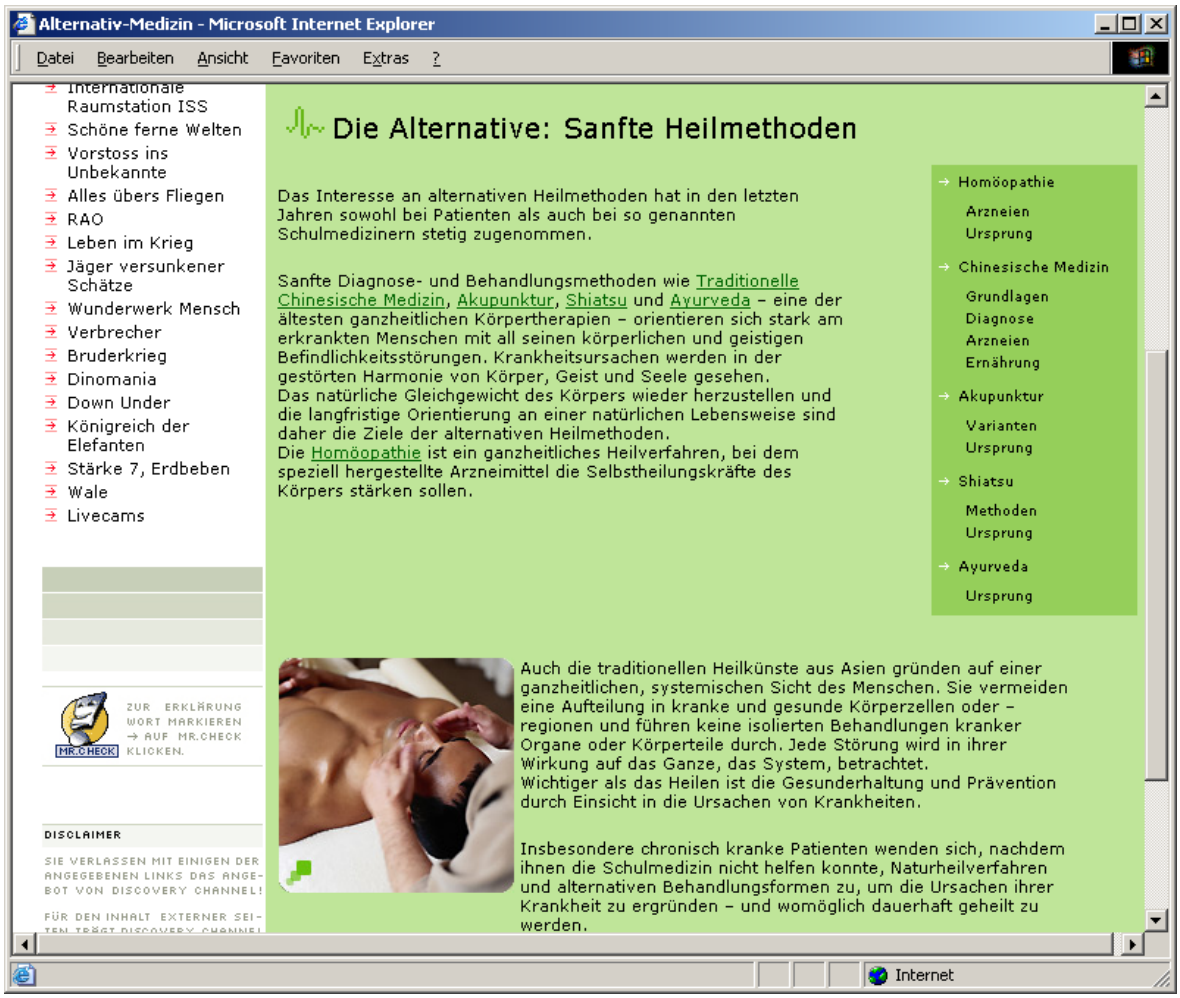

Abbildung 4: Beispieltext Alternativmedizin

Mit Hilfe des Concept Extrator wurden die nachfolgenden Schlagworte ermittelt - die Ausgabe ist dabei nach dem Wert der Zuordnungsfunktion gerankt; die fett gedruckten Begriffe stellen dabei die Untermenge dar, die als erste

256 Zeichen zur Aufnahme in die HTML-Meta Tags der Webseite vorgesehen sind, um die Erfassung durch Suchmaschinen zu verbessern:

Alternativ-Medizin, Shiatsu, Ayurveda, Medizin, Special, Akupunktur, Heilmethoden, Traditionelle, Arzneien, Homöopathie, Chinesische, Heilkunst, Diagnose, Ernährung, Körpertherapie, Grundlagen, Heilmethode, Heilkünste, Körpertherapien, Schulmediziner, Befindlichkeitsstörung, Sanft, Krankheitsursache, Behandlungsform, Störung, Organe, Ursprung, Gleichgewicht, Schulmedizinern, Selbstheilungskraft, Befindlichkeitsstörungen, Orientierung, Gesunderhaltung, Sanfte, Krankheitsursachen, Behandlungsformen, Krankheiten, Seele, Geist, Asien, Selbstheilungskräfte, Behandlungsmethode, Körperzelle, Heilverfahren, Krankheit

Das Beschlagwortungsergebnis zeigt, wie sich mit Hilfe frequenzbasierten Corpusvergleichs und unter Heranziehung zusätzlicher Wissensmodule eine flexible Beschlagwortung erreichen lässt. Dabei sind allerdings auch Schwächen des Systems offensichtlich: Nicht alle Mehrwortbegriffe können automatisch als solche erkannt werden (z. B. Aufnahme des Adjektivs chinesische als Schlagwort, Fehler bei der Groß- und Kleinschreibung).

5 Fazit

Das vorgestellte Beschlagwortungssystem stellt einen Mittelweg zwischen vollautomatischer Volltextindexierung und manueller Vergabe von Schlagworten dar. Durch Corpusvergleich kann die Schlagwortselektion an die Eigenheiten der jeweiligen Dokumentkollektion angepasst werden, zusätzliche Module für die Begriffsexpansion und –reduktion gewährleisten, dass auch sprachliche Varianten in die Beschlagwortung aufgenommen werden. Die Flexibilität des Systems lässt es für sehr unterschiedliche Probleme der Informationserschließung, gerade auch für innovative Informationsdienste aus dem Umfeld des *mobile computing* geeignet erscheinen, für die Restriktionen z. B. bezügliche der möglichen Beschlagwortungstiefe gelten.

Die systematische Evaluierung des Ansatzes steht noch aus. Bislang erfolgte nur eine qualitative Bewertung durch Online-Redakteure in einer ersten Anwendungsdomäne des Systems im Bereich Wissenschaftsjournalismus.

Literaturverzeichnis

[Atkins et al. 02]. Atkins, S. et al. (2002). „From Resources to Applications. Designing the Multilingual ISLE Lexical Entry.“. In: Proc. LREC-2002. Third International Conference on Language Resources and Evaluation. Las Palmas, May 2002, Vol. II, 687-692.

[Baeza-Yates & Ribeiro-Neto 99] Baeza-Yates, R.; Ribeiro-Neto, B. (1999). Modern Information Retrieval. Harlow et al.: The ACM Press/The MIT Press.

[Breiman 94] Breiman, L. (1994). „Bagging Predictors“. Technical Report No. 421, University of California at Berkeley, Department of Statistics, September 1994.

[Heyer et al. 01a] Heyer, G.; Läuter, M.; Quasthoff, U.; Wittig, Th.; Wolff, Ch. (2001). “Learning Relations Using Collocations". In: Proc. IJCAI Workshop on Ontology Learning, Seattle/WA, August 2001, 19-24.

[Heyer et al. 01b] Heyer, G.; Läuter, M.; Quasthoff, U.; Wolff, Ch. (2001): „Wissensextraktion durch linguistisches Postprocessing bei der Corpusanalyse“. In: Lobin, H. (ed.) (2001). Sprach- und Texttechnologie in digitalen Medien. Proc. GLDV-Jahrestagung 2001, Universität Gießen, 71-83

[Heyer, Quasthoff, Wolff 00] Heyer, G.; Quasthoff, U.; Wolff, Ch. (2000). “Aiding Web Searches by Statistical Classification Tools“. In: Proc. Proc. 7. Intern. Symposium f. Informationswissenschaft ISI 2000, UVK, Konstanz, 163-177.

[Heyer, Quasthoff, Wolff 02] Heyer, G.; Quasthoff, U.; Wolff, Ch. (2002). “Knowledge Extraction from Text: Using Filters on Collocation Sets.” In: Proc. LREC-2002. Third International Conference on Language Resources and Evaluation. Las Palmas, May 2002, Vol. III, 241-246.

[Kilgarriff 01] Kilgarriff, Adam (2001). „Comparing Corpora" In: International Journal of Corpus Linguistics 6(1) (2001), 97-133.

[Levenshtein 65] Levenshtein, V. (1965). “Binary Codes Capable of Correcting Spurious Insertions and Deletions of Ones”. In: Probl. Inf. Transmission 1 (1965), 8–17.

[Navarro 01] Navarro, G. (2001). „A Guided Tour to Approximative String Matching“. In: ACM Computing Surveys 33(1) (2991), 33-88.

[Preece & Decker 02] Preece, A.; Decker, M. (2002). “Intelligent Web Services”. In: Intelligent Systems 17(1) (2992), 15-17.

[Quasthoff & Wolff 00] Quasthoff, U.; Wolff, Ch. (2000). “An Infrastructure for Corpus-Based Monolingual Dictionaries”. In: Proc. LREC-2000. Second International Conference on Language Resources and Evaluation. Athens, May / June 2000, Vol. I, 241-246.

[Rayson & Garside 00] Rayson, P.; Garside, R. (2000). „Comparing Corpora Using Frequency Profiling“. In: Proc. Workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000), Hong Kong, October 2000, 1-6.

[Salton & McGill 83] Salton, G.; McGill, M. J. (1983). Introduction to Modern Information Retrieval. New York et al.: McGraw-Hill.

[Simon 02] Simon, M. (2002). „Eine Botschaft auf allen Kanälen“. In: <e>Market, Juni 2002, 10-12.

[Wolff 01]. Wolff, Ch. (2001). „Aspekte des Vergleichs von Fach- und Normcorpora am Beispiel eines Fachcorpus aus der Automobiltechnik“. Arbeitsmaterialie, Universität Leipzig, Institut für Informatik, Abt. Automatische Sprachverarbeitung, Juni 2001.