Exposé zur Diplomarbeit

Identifikation von Erdbebenmeldungen über Focused Crawling und Textsegmentierung

Institut für Informatik
Humboldt-Universität zu Berlin

Mario Lehmann

3. September 2012

Betreuer: Prof. Dr. Ulf Leser, Prof. Dr. Felix Naumann

1 Einleitung

Die Fülle der frei verfügbaren Informationen im Internet effizient zugänglich zu machen ist ein Problem, welches mit dem Wachstum des Internets immer mehr an Relevanz gewinnt. Neben der Suche geeigneter Informationen werden zunehmend Methoden relevant themenspezifische Webtexte in großem Umfang mit dem Ziel der weiteren Analyse zu sammeln. Hierfür werden sogenannte *Focused Crawler* eingesetzt, die im Gegensatz zu herkömmlichen *Webcrawlern* das Netz themenspezifisch durchsuchen. Die Grundannahme dabei ist, dass themenrelevante Seiten mit größerer Wahrscheinlichkeit auf ebenfalls themenrelevante Seiten verlinken, während hingegen themenirrelevante Seiten eher zu weiteren irrelevanten Seiten führen und somit ein Endpunkt der Suche darstellen sollten [Nov04]. In Abbildung 1 sind die beiden Suchverfahren gegenübergestellt[1].

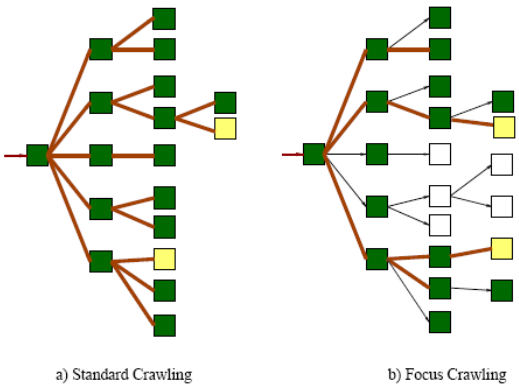

Abbildung 1: Vergleich des Standard Crawl-Verfahrens (a) mit dem *Focused Crawling* (b). Um eine bestimmte Zielseite (gelbes Quadtrat) zu besuchen, wird bei der Standard-Suche (a) jeder Link einer Seite verfolgt. Die Suche ist nicht optimal, da irrelevante Seiten ebenfalls besucht werden. In (b) wird nur eine Teilmenge aller verlinkten Dokumente durchsucht, wodurch "nicht zielführende" Seiten ausgelassen werden (weiße Quadrate). Dies reduziert den Suchraum, wodurch das Auffinden relevanter Seiten beschleunigt wird.

[1] Quelle: [DCL00]

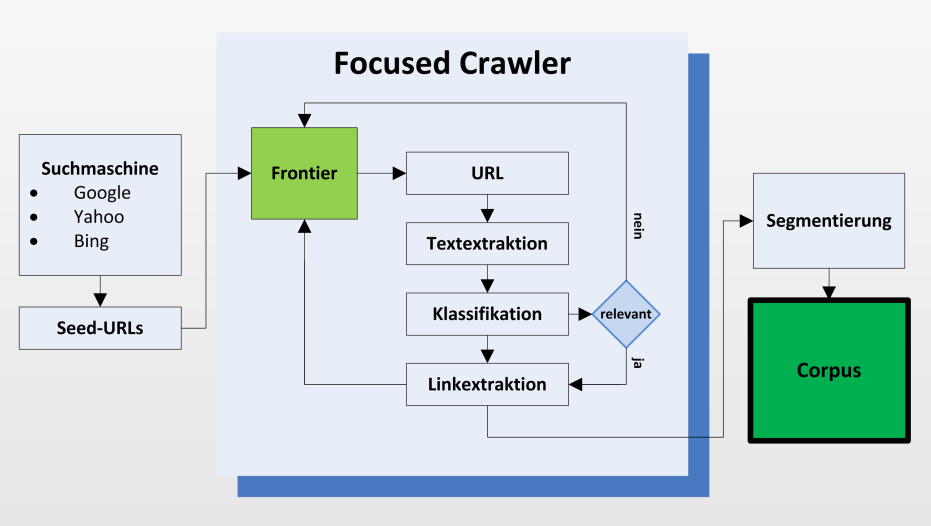

Abbildung 2: Darstellung der Arbeitsschritte als Blockdiagramm

2 Zielstellung

In der Arbeit soll ein *Focused Crawler* entstehen, der aus dem Internet Texte extrahiert, die von Erdbeben handeln. In Anlehnung an die Arbeit [Leh12] soll für das *EquatorNLP*-Projekt [DL11] ein domänenspezifisches Korpus [Das08] generiert werden.
Zudem sollen innerhalb der Korpustexte erdbebenrelevante Textstellen detektiert und raumzeitlich eingeordnet werden. Die Suche nach relevanten Textstellen wird auch als Segmentierungsproblem bezeichnet. Ein Segment ist im Sinne der Arbeit ein Textabschnitt, der von einem konkreten Erdbebenereignis handelt.

3 Vorgehen

In Abbildung 2 ist der Ablauf der Arbeit skizziert. Im ersten Schritt werden per Suchmaschine über den Begriff ”earthquake” *seed-URLs* generiert, welche die Start-URLs für den *Crawler* bilden. Aus diesen wählt der *Crawler* eine URL aus und extrahiert von der entsprechenden Seite den Haupttext. Dieser wird anhand eines zuvor antrainierten Modells (Siehe 4.1.3) hinsichtlich seiner Erdbebenrelevanz klassifiziert. Enthält der extrahierte Text keine

erdbebenrelevanten Inhalte, so wird im *Frontier* die entsprechende URL gelöscht (bzw. als besucht gekenntzeichnet) und eine neue URL zur Extraktion herausgesucht. Im anderen Fall werden die auf der Seite enthaltenden Links extrahiert und an den *Frontier* weitergegeben. In diesem Fall wird zusätzlich der Haupttext segmentiert und inklusive Metadaten im Korpus abgespeichert.

4 Methoden

4.1 Focused Crawler

4.1.1 Webcrawler

Der *Focused Crawler* soll auf Basis einer bereits bestehenden Implementierung eines *Webcrawlers* entstehen. Da in der Arbeit eine große Textmenge extrahiert werden soll, muss der *Crawler* bestimmte Anforderungen erfüllen. In [Hei10] sind die Anforderungen zusammengefasst. Auszugsweise seien drei der wichtigsten Punkte erwähnt:

Robustness: Ein *Crawler* muss mit dynamisch erzeugten Links, sogenannten *spider traps*, umgehen können.

Politeness: Ein *Crawler* muss die Regeln, die auf einer bestimmten *Domain* herrschen (z.B. max. Anzahl der Requests / min) berücksichtigen.

Distributed und Scalable: Die Bandbreite der Crawlvorgangs muss durch *multithreading* oder verteilter Datenverarbeitung einfach adaptierbar sein.

Zudem muss der *Crawler* in seinem Quelltext frei verfügbar sein. Es kommen die beiden *Webcrawler Heritrix*[1] und *Apache Nutch*[2] genauer in Frage, wobei *Heritrix* nicht über die Möglichkeit der verteilten Datenverarbeitung verfügt [Hei10].

[1] `https://webarchive.jira.com/wiki/display/Heritrix/Heritrix`
[2] `http://nutch.apache.org/`

4.1.2 Textextraktion

Analog zur Studienarbeit [Leh12] ist für die Textextraktion angedacht die *Boilerpipe*-Bibliothek zu verwenden [KFN10]. *Boilerpipe* identifiziert für die Extraktion des Haupttextes einer HTML-Seite diejenigen Blöcke, die nicht zum Haupttext einer Seite gehören („boilerplates“) und entfernt sie. In Abhängigkeit der Extraktionsqualität bei beliebigen Seiten ist jedoch nicht ausgeschlossen, dass noch andere Extraktionsverfahren zur Anwendung kommen.

4.1.3 Klassifikation

Für die Klassifikation der Erdbebentexte steht bereits ein Klassifikationsmodell aus der Arbeit [Leh12] zur Verfügung. Ein Klassifikationsmodell ist ein statistisches Modell, das von einem Klassifikator im Lernschritt für den Klassifikationsvorgang erstellt wird.
Für die Erstellung des Modells wurde eine *Support Vector Maschine* [HCL10] auf Basis eines Korpus mit 453 Nachrichtenartikeln angelernt. Das Modell erreichte auf einer 10-Feld-Kreuzvalidierung [WIK] eine hohe *Precision* [MRS08] von 97.5-100% bei einem moderaten Recall von 55-75%, welcher vom relativen Anteil der Positivklasse (erdbebenrelevante Artikel) abhängig war. Da die Klassenverteilung beim *focussed crawling* nicht abzuschätzen ist, bietet es sich an die Texte ähnlich wie in der Studienarbeit über das Wort ”earthquake” zu präselektieren. Da der Recall bei dieser Vorklassifikation nahezu bei 100% ist (Siehe Baseline [Leh12]), könnte man nahezu verlustfrei die Klassenverteilung zwischen den gecrawlten Texten und den Texten, mit denen das Modell antrainiert wurde, angleichen. Konkret hieße das, dass man das Modell nur mit Texten anlernt, die das Wort ”earthquake” enthalten und zudem den Präselektionsschritt im *Focused Crawler* implementiert.
Eine weitere Überlegung wäre, dass man alle Texte zuerst segmentiert und die Segmente anschließend vom Modell klassifiziert. Da die Segmentierung auch eine Merkmalsreduktion nach sich zieht, könnte möglicherweise die Klassifikationsleistung dadurch positiv beeinflusst werden.

4.1.4 Linkextraktion

Grundsätzlich sollen die Links in Abhängigkeit ihrer Themenrelevanz in den *Frontier* einsortiert werden, so dass die relevantesten URLs vom Crawler prioritär weiter verfolgt werden.

Für die Abschätzung der Relevanz eines Links bietet sich der Abstand eines Links zum Wort ”earthquake” an. Unter Umständen wäre es auch möglich das Segmentierungsergebnis zu nutzen, um bspw. die Relevanz eines Links, der sich nicht in einem Erdbebensegment befindet, zusätzlich abzumildern.

4.2 Textsegmentierung

Ein wichtiges Ergebnis aus der Studienarbeit [Leh12] ist, dass Ereignisse, auch wenn sie nur als Nebenthema vorkommen, über raumzeitliche Attribute recht gut detektiert werden können. Besonders interessant ist, dass die Detektion auch bei Artikeln funktioniert hat, die mehrere Erdbebenereignisse enthalten. Da in der geplanten Arbeit ein Erdbebenereignis einem Segment bzw. mehreren Segmenten entsprechen soll, könnten daher raumzeitliche Attribute u.U. auch für die Textsegmentierung nützlich sein.
Um dies zu überprüfen, müssen die extrahierten Texte analog zur Studienarbeit zuerst raumzeitlich annotiert werden. Hierfür gibt es bereits verfügbare Werkzeuge, welche für die automatische Detektion genutzt werden können. Für die Extraktion der räumlichen Attribute kommt *MetaCarta* [RBBA03] in Frage. Für die Detektion der zeitlichen Attribute bietet sich *HeidelTime* [SG10] an.
Sind die raumzeitlichen Attribute detektiert, kann im Text nach zusammenhängenden ”raumzeitlich homogenen” Abschnitten gesucht werden. Enthält einer dieser Abschnitte das Wort ”earthquake”, so bildet er ein Erdbebensegment. Diese Zuordnung wäre ebenfalls über das oben beschriebene Klassifikationsmodell denkbar.
Sind zwei Textsegmente ”raumzeitlich homogen”, beschreiben sie womöglich das gleiche Ereignis, so dass sie dem gleichen Ereignis zugeordnet werden können. Auf diese Weise ließe sich anschließend die Gesamtanzahl der detektierten Ereignisse innerhalb eines Artikels ermitteln. Sind die Textsegmente erzeugt, kann anschließend ermittelt werden, welches genaue Ereignis ein Textsegment bzw. eine Menge gleichartiger Textsegmente enthält. Hierbei wird für jedes gefundene Ereignis in den dazugehörigen Textsegmenten nach demjenigen Raum-Zeit-Paar gesucht, das das Ereignis (mit größter Genauigkeit) repräsentiert. Somit erhält man für jeden Text eine Auflistung aller enthaltenen Erdbebenereignisse, sowie deren Position bzw. Segmente innerhalb des Textes.
Die entsprechenden Segmentgrenzen besitzen hierbei eine duale Funktion. Zum einen gren-

zen sie erdbebenrelevante von nicht erdbebenrelevanten Textstellen ab. Zum anderen fungieren sie innerhalb erdbebenrelevanter Abschnitte als Grenzen zwischen mehreren Erdbebenereignissen.
Besonders im ersten Fall ist unklar, ob auf alleiniger Basis der raumzeitlichen Attribute die Segmentierung ausreichen wird. Daher soll das Verfahren im Verlauf der Arbeit um weitere Segmentierungsmethoden erweitert werden. Im Abschnitt 5 werden Methoden, die zum Einsatz kommen könnten, genauer vorgestellt.

4.3 Evaluation

4.3.1 Focused Crawler

Grundsäzlich kann man bei *Webcrawlern* den *Recall* nicht abschätzen, da es ungewiss ist, wie viele Texte im Internet zu einem speziellen Thema existieren. Daher beschränkt sich die Evaluation des Crawlverfahrens auf die *Precision*. Hierfür wird aus dem generierten Korpus eine Stichprobe entnommen und manuell überprüft, ob die Texte aus der Stichprobe richtig klassifiziert wurden.
Die Fokussierungsgüte des *Crawlers* kann durch das Verhältnis der Anzahl von besuchten Seiten durch die Gesamtanzahl der extrahierten Seiten im Korpus (*"harvest ratio"* [CvdBD99]) ermittelt werden. Zum Vergleich wird als Baseline ein unfokussierter Crawler eingesetzt, der die gleichen *seed-URLs* verwendet. Die Abhängigkeit der *harvest ratio* beider Crawler von ihrer *seed*-Größe könnte man ebenfalls untersuchen.

4.3.2 Segmentierung

Die Segmentierungsgüte wird ebenfalls anhand einer befundeten Stichprobe abgeschätzt. Als Gütemaß schlagen Beefer und Kollegen in [BBL99] das P_k-Maß vor. Das in [Cho00] auch als *error metric* bezeichnete Maß gibt die Wahrscheinlichkeit dafür an, dass ein im Text zufällig gezogenes Wortpaar mit einem Abstand von k vom Segmenter inkonsistent klassifiziert wird. Ob eine Inkonsistenz vorliegt, entscheiden die Segmentgrenzen. Liegt bspw. in der Referenzsegmentierung zwischen den Worten des Paares eine Segmentgrenze jedoch bei der angenommenen Segmentierung keine, so gilt die Situation als "miss" und im umgekehrten Fall als "false alarm".

Da die detektierten Segmente konkreten Ereignissen zugeordnet werden, muss ggf. die P_k-Berechnung um "identifizierte" Segmentgrenzen angepasst werden. Wechselt bspw. ein Textabschnitt von keinem Erdbebenereignis zu einem bestimmten Erdbebenereignis, der Segmenter erkennt jedoch einen Wechsel zu einem anderen Ereignis, so liegt ebenfalls ein "miss" vor, obwohl eine Segmentgrenze richtig erkannt wurde.

5 Verwandte Arbeiten

Wegbereitend für das *Focussed Crawling* ist u.a. die Arbeit von Davison [Dav00], in der topologische Charakteristiken des Internets untersucht wurden. Auf Grundlage eines Subgraphen von 100000 Knoten wurde über Similaritätsmaße (*z.B. TFIDF cosine similarity*) u.a. untersucht, inwieweit sich lokal benachbarte Knoten thematisch ähneln. Ein wichtiges Ergebnis war, dass die Ähnlichkeit verlinkter Seiten im Vergleich zu randomisiert gezogenen signifikant höher ist. Zudem zeigte sich, dass zwei Geschwisterseiten ähnlicher zueinander sind, je geringer ihre Links auf der Elternseite voneinander entfernt sind.
In der Arbeit [CvdBD99] wurde ein *Focussed Crawler* auf Basis eines probabilistischen Klassifikationsmodells realisiert. Im Vergleich zu einem unfokussierten Crawler, der bereits nach einigen besuchten URLs keinerlei relevante Seiten mehr besuchte, erreichte der Crawler eine *harvest rate*[1] von ca. 0.5. In einer späteren Arbeit [CvdBD99] erweiterte Chakrabarti den Ansatz um einen zweiten Klassifikator, wodurch die Anzahl der falsch besuchten Seiten um weitere 30-90% reduziert werden konnte.

Textsegmentierungsalgorithmen findet man in der Literatur häufig im Kontext der *text summarization*, bei der es darum geht, irrelevante Bestandteile eines Textes zu eliminieren. In [KKM98] ist dargestellt, dass Texte hierarchisch oder linear segmentiert werden können. Die Arbeit selber beschäftigt sich mit der linearen Segmentierung, bei der ein Eingabetext in lineare Sequenzen adjazenter Segmente aufgeteilt wird. Hierfür wurde ein regelbasiertes Verfahren entwickelt, welches jedem Absatz innerhalb eines Textes einen gewichteten Wert (*score*) auf Basis von zuvor extrahierten Termketten zuweist. Kommt bspw. eine bestimmte Nominalphrase (ein Term) über mehrere Absätze hinweg in einem Text vor, so ist es unwahr-

[1] ist in 4.3.1 als *harvest ratio* bezeichnet

scheinlich, dass die jeweiligen Absatzgrenzen auch eine Segmentgrenze darstellen. In diesem Fall werden mehrere Absätze zu einem Segment zusammengefasst. Das Verfahren erreichte auf einem getesteten Korpus von 20 Nachrichtenartikeln (*Wall Street Journal, Economist*) eine mittlere Precision von 42.6% bei einem Recall von 39.6%.
In [Cho00] wird von Choi ein Verfahren beschrieben, welches die Segmentgrenzen eines Textes mit Hilfe einer Satzähnlichkeitsmatrix berechnet. Hierfür wird über ein Rangverfahren der Kontrast der Matrix zunächst erhöht und anschließend eine Clusterung durchgeführt. Das als C99 bezeichnete Verfahren erreichte auf einem Corpus von 700 Texten im Mittel eine Fehlerrate von ca. 10%. In [RB12] wurde das C99-Verfahren um ein probabilistisches Modell erweitert, welches jedem Wort das wahrscheinlichste Thema zugeordnet. Diese *"Topic IDs"* ersetzen anschließend die eigentlichen Worte. Mit diesem Verfahren ließ sich auf dem Originalkorpus von Choi die Fehlerrate des C99-Verfahren zusätzlich halbieren.

Zuletzt sei im Kontext der Ereigniserkennung die Arbeit [LNF+11] erwähnt, in welcher der Einfluss konkreter Ereignisse (z.B. politische Ereignisse, Naturkatastrophen) auf statistische Daten (z.B. Einkommensentwicklung in Deutschland) untersucht wurde. Zwei Ziele standen hierbei im Vordergrund. Zum einen sollten statistische Ausreißer (sogenannte *"black swans"*) automatisch detektiert und mit zeitlich und kausal entsprechenden Ereignissen annotiert werden. Zum anderen sollten bestimmte Muster erfasst werden, die die Wechselwirkung von Ereignistypen und statistischen Daten beschreiben. Für die Mustererkennung wurden Assoziationsregeln generiert, deren Qualität anhand probabilistischer Metriken abgeschätzt wurde. Als Datengrundlage wurden mehrere Quellen verwendet (z.B. *World Bank, IMF, DBpedia, EMDat*). Das Ergebnis der Arbeit ist im Internet frei verfügbar[1].

[1] http://www.blackswanevents.org/

Literatur

[BBL99] D. Beeferman, A. Berger, and J. Lafferty. Statistical models for text segmentation. *Machine learning*, 34:177–210, 1999.

[Cho00] F.Y.Y. Choi. Advances in domain independent linear text segmentation. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 26–33. 2000.

[CvdBD99] Soumen Chakrabarti, Martin van den Berg, and Byron Dom. Focused crawling: a new approach to topic-specific Web resource discovery. *Computer Networks*, 31(11-16):1623–1640, May 1999.

[Das08] Niladri Sekhar Dash. *Corpus Linguistics: A General Introduction.* Pearson Education-Longman, 2008.

[Dav00] BD Davison. Topical locality in the Web. *Proceedings of the 23rd annual international ACM ...*, 1(212), 2000.

[DCL00] M Diligenti, F Coetzee, and S Lawrence. Focused crawling using context graphs. In *Proceedings of the 26th international conference on very large data bases*, pages 527–534. 2000.

[DL11] Lars Döhling and Ulf Leser. EquatorNLP: Pattern-based Information Extraction for Disaster Response. In *Proceedings of Terra Cognita 2011 Workshop*, 2011.

[HCL10] Chih-wei Hsu, Chih-chung Chang, and Chih-jen Lin. A Practical Guide to Support Vector Classification. *Bioinformatics*, 1(1):1–16, 2010.

[Hei10] Arvid Heise. *Extraction of Management Concepts from Web Sites for Sentiment Analysis.* Master's thesis, Humboldt-Universität zu Berlin, 2010.

[KFN10] Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. Boilerplate Detection using Shallow Text Features. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 441–450, 2010.

[KKM98] M.Y. Kan, J.L. Klavans, and K.R. McKeown. Linear segmentation and segment significance. In *Proceeding of the 6th Workshop on Very Large Corpora (WVLC-98)*, pages 197–205. 1998.

[Leh12] Mario Lehmann. *Textklassifikation von Erdbebenmeldungen*. Studienarbeit, Humboldt-Universität zu Berlin, 2012.

[LNF+11] Johannes Lorey, Felix Naumann, Benedikt Forchhammer, Andrina Mascher, Peter Retzlaff, and Armin ZamaniFarahani. Black swan: augmenting statistics with event data. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management*. 2011.

[MRS08] C.D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*, volume 1. Cambridge University Press, 2008.

[Nov04] Blaž Novak. A survey of focused web crawling algorithms. *unpublished*, 2004.

[RB12] Martin Riedl and Chris Biemann. How Text Segmentation Algorithms Gain from Topic Models. (2009):553–557, 2012.

[RBBA03] Erik Rauch, Michael Bukatin, Kenneth Baker, and Massachusetts Avenue. A confidence-based framework for disambiguating geographic terms. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references-Volume 1*, pages 50–54. 2003.

[SG10] Jannik Strötgen and Michael Gertz. HeidelTime : High Quality Rule-based Extraction and Normalization of Temporal Expressions. *Computational Linguistics*, (July):321–324, 2010.

[WIK] WIKIPEDIA. Cross-validation (statistics). `http://en.wikipedia.org/wiki/Cross-validation_%28statistics%29`.