

Referatsthemen: WPF 45 Web Data Mining

Edda Leopold und Heide Faeskorn-Woyke

1. Qualitätsmessung bei Suchmaschinen

Hier sind einige Arbeiten zu referieren, die die gebräuchlichsten Suchmaschinen hinsichtlich der Qualität und Aktualität ihrer Ergebnisse evaluieren.

Dirk Lewandowski & Nadine Höchstötter (2007): Qualitätsmessung bei Suchmaschinen. System- und nutzerbezogene Evaluationsmaße, in: Informatik Spektrum 30 (3), 159-169.

Dirk Lewandowski (2008): Search engine user behaviour: How can users be guided to quality content? In: Information Services & Use 28 (3-4), 261-268.

Dirk Lewandowski (2008): A three-year study on the freshness of Web search engine databases, in: Journal of Information Science 34 (6), 817-831.

2. PageRank

Bei Internet Suchanfragen besteht das Problem, dass die es in der Regel zehntausende von Dokumenten gibt, die zu einer Suchanfrage passen, von denen aber nur wenige wirklich relevant sind. Empirische Studien haben gezeigt, dass verschiedene Suchmaschinen (Google, Bing, Yahoo, Ask) sich vor allem darin unterscheiden, in welcher Reihenfolge die Suchergebnisse Präsentiert werden, und dass dieser Unterschied für Beliebtheit einer Suchmaschine entscheidend ist, weil Nutzer sich in der Regel nur die ersten 10 Resultate anschauen. Der Page-Rank Algorithmus und der Hits Algorithmus dienen gerade dazu die Dokumente entsprechend ihrer Relevanz in eine Reihenfolge zu bringen.

Soumen Chakrabarti (2003): Mining the Web, Morgan Kaufman. pp. 209-212.

Sergey Brin , Lawrence Page (1998): The anatomy of a large-scale hypertextual Web search engine, Computer Networks, 30(1-7), 107-117.

3. Hits-Algorithmus

Hits steht für Hyperlink-Induced Topic Search und bezeichnet einen Algorithmus, der Internetseiten anhand der Linkstruktur, in eine Relevanzreihenfolge bringt. Im Gegensatz zum Page-Rank Algorithmus berechnet der Hits Algorithmus für jede Suchanfrage einen entsprechenden Subgraphen des Internets der der Relevanzberechnung zugrunde gelegt wird.

Soumen Chakrabarti (2003): Mining the Web, Morgan Kaufman. pp. 212-224.

Jon M.Kleinberg (1998):Authorative Sources in a hyperlinked environment, in: Journal of the ACM 46 (5), 604-632.

4. Web-Structure Mining (Social Network Analysis)

Das untenstehende Paper von Kleinberg et al. ist eine nachträgliche Analyse zum Hits Algorithmus. Sein wesentliches Resultat besteht in einem Modell für das Wachstum von Seiten und Links im Internet.

Soumen Chakrabarti (2003): Mining the Web, Morgan Kaufman. pp. 243-253.

Jon M. Kleinberg & Ravi Kumar & Prabhakar Raghavan & Sridhar Rajagopalan & Andrew S. Tomkins (1999): The Web as a graph: measurements, models, and methods, in: Takao Asano & Hiroshi Imai & D. T. Lee & Shin-Ichi Nakano & Takeshi Tokuyama (eds.): Proceedings of the 5th International Computing and Combinatorics Conference (COCOON), Tokyo July 26-28, 1999, Springer LNCS 1627, pp. 1-17

5. Community Discovering.

Sowohl für die Terrorismusbekämpfung als auch für die Entdeckung von potentiellen Kundengruppen ist es interessant zu wissen welche Individuen innerhalb eines sozialen Netzwerkes eine Gemeinschaft bilden. Hierzu gibt es verschiedene Algorithmen.

Soumen Chakrabarti (2003): Mining the Web, Morgan Kaufman. pp. 203-209.

Ronen Feldman & James Sanger (2007): Textmining Handbook, CUP: Cambridge, S. 242-272.

Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto and Domenico Parisi (2004) Defining and identifying communities in networks, in: PNAS, 101(9): 2658-2663.

Mark E. J. Newman and Michelle Girvan (2004): Finding and evaluating community structure in networks, in: Physical Review, E 69 (026113).

6. Web Crawling 1

Web-Crawler (spider , robots) sind Programme, die automatisch WEB-Seiten laden und indizieren. Sie werden in hauptsächlich Suchmaschinen, aber auch in vielen anderen Bereichen, wie z.B. Business Intelligence –Anwendungen (Konkurrenzanalyse), zum Suchen von E-Mail-Adressen und persönlichen Informationen, Monitoring von interessanten Web-Seiten etc. verwendet. Es werden verschiedene Arten von Crawlern unterschieden: Allgemeine (Universal)Crawler, thematische (topical) Crawler und zielgerichtete (focudes) Crawler. In diesem Referat sollen folgende Themen besprochen werden:

- Grundlegende Techniken und Implementierungsaspekte von Crawlern
 - Wie gehen crawler mit anderen Formaten als HTML um?
 - Stemming und Stopwortentfernung
 - Linkextraktion und kanonische URLs
 - Tiefensuche
 - Breitensuche
- Spezielle Crawler(universal, focudes, topical)
- SEO (Suchmaschinenoptimierung)
- Ethische Problem beim Crawlern
- Neuere Entwicklungen beim Crawlern

Literatur:

Bin Liu, Web Data Mining, Kapitel 8

Sergey Brin , Lawrence, Page, The anatomy of a large-scale hypertextual Web search engine, Computer Networks, 30(1-7,), pp. 107-117, 1998 bzw.

<http://infolab.stanford.edu/~backrub/google.html>

P. Baldi, P. Frasconi, P. Smyth: "Modeling the Internet and the Web", Wiley, 2002, Kapitel 6.

7. Wrapper und strukturierte Datenextraktion (1-2-Vorträge)

Ein Wrapper ist ein Programm, das automatisch (semi-)strukturierten Daten aus einer bestimmten Datenquelle (Text, WEB) extrahiert. Es gibt drei Ansätze:

- Manuelle Extraktion mittels menschlicher Unterstützung
- Halbautomatische Extraktion mit überwachten Lernmethoden
- Automatische Extraktion mittels nicht überwachtem Lernen

Die Referate sollen sich mit den beiden letzten Methoden auseinandersetzen. Dazu gehören folgende Punkte:

- Klassifizierung von Web-Seiten, die strukturierte Daten enthalten
- Ein Datenmodell für die Wrapper-Generierung
- Wrapper Induction
- Wrapper Extraction
 - Tree-Matching-Algorithmus
 - Bildung DOM-Trees
 - Extraktion von List Pages und Mutiple Pages

RoadRunner System als praktische Umsetzung

Literatur:

Bin Liu, Web Data Mining, Kapitel 9 mit Bibliographie

<http://www.dia.uniroma3.it/db/roadRunner/>

<http://rtw.ml.cmu.edu/readtheweb.html>

8. Opinion Mining

Hier geht es um die Analyse von **unstrukturierten** Texten, wie sie im WEB sehr häufig vorkommen, also insbesondere um die Meinungsforschung im WEB. Es gibt drei Grundaufgaben:

- Klassifizierung von Text, ob er einer bestimmten Meinung oder einem Produkt positiv oder negativ gegenübersteht
- Automatische Suche von Eigenschaften, die ein bestimmtes Objekt betreffen (z.B. Produkteigenschaften, die oft kommentiert werden)
- Automatischer Vergleich von unterschiedlichen Objekten oder Produkten, die vorgegeben sind

Literatur:

Bin Liu, Web Data Mining, Kapitel 11 mit Bibliographie
Verschiedene Arbeiten von Bing Liu selber [245], [246]

10. Web Usage Mining

Web Usage Mining meint die automatische Suche und das Finden in Mustern, die Internetbenutzer beim Besuch von Web-Seiten erzeugen. Grundlage sind Web Server Log-Dateien, Clickstream-Protokolle, Seiteninhalte und Daten, die über Benutzerverhalten auf WEB-Seiten generell gesammelt werden. Diese Daten werden mit Data Warehouse bzw. Data Mining Tools aufgearbeitet.

- Datensammlung, insbesondere aus Web Server-Log-Dateien
- Ein Datenmodell für das Web Usage Mining
- Mustersuche und Analyse von Web Usage Pattern

Literatur:

Bin Liu, Web Data Mining, Kapitel 12 mit Bibliographie

From Web to Social Web: Discovering and Deploying User and Content Profiles

Workshop on Web Mining, WebMine 2006, Berlin, Germany, September 18, 2006. Revised Selected and Invited Papers, Lecture Notes in Computer Science, Springer Berlin / Heidelberg, darin:

Conor Hayes, Paolo Avesani and Uldis Bojars: An Analysis of Bloggers, Topics and Tags for a Blog Recommender System

Federico Michele Facca :Combining Web Usage Mining and XML Mining in a Real Case Study

Bettina Berendt and Anett Kralisch: From World-Wide-Web Mining to Worldwide Webmining: Understanding People's Diversity for Effective Knowledge Discovery

<http://www.cs.kuleuven.be/~berendt/teaching/2009-101stsemester/adb/Lecture/Session10/truemper.html>