Referatsthemen: WPF 45 Web Data Mining

1. Qualitätsmessung bei Suchmaschinen

Um die Qualität einer Suchmaschine zu bewerten, wird zumeist auf die Maße Recall und Precision (siehe http://wikis.gm.fh-koeln.de/wiki_ir/InformationRetrieval/Precision) abgestellt. Dies ist allerdings eine sehr einseitige Sichtweise. In der unten angegebenen Literatur werden verschiedene Aspekte der Suchmaschinenqualität untersucht. Dabei werden auch die Suchmaschinennutzer in die Betrachtung mit einbezogen.

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. Web-Site Ranking

Bei Internet Suchanfragen besteht das Problem, dass 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 die 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.

Sergey Brin , Lawrence, Page (1998): The anatomy of a large-scale hypertextual Web search engine, In: Computer Networks and ISDN Systems 1998. http://infolab.stanford.edu/~backrub/google.html

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

*Soumen Chakrabarti (2003): Mining the Web, Morgan Kaufman: Amsterdam et al. pp. 209-224.

*Bing Liu (2007): Web Data Mining, Springer: Berlin, Heidelberg, New York. pp. 246-254.

3. Modelle und Maßzahlen für das Internet

Für die Entstehung der Linkstruktur des Internets gibt es verschiedene Modelle, aus denen unterschiedliche Verteilungseigenschaften der Links resultieren. Die Linkverteilungen, die man im Internet empirisch beobachten kann, stimmen mit älteren Resultaten von Lotka (1926) über die Verteilung von Autoren in Publikationsmedien überein.

Kleinberg, Jon M. & Kumar, Ravi & Raghavan, Prabhakar & Rajagopalan, Sridhar & Tomkins, Andrew S.: (1999): The Web as a Graph: Measurements, Models, and Methods. COCOON 1999: 1-17.

http://en.wikipedia.org/wiki/Erd%C5%91s%E2%80%93R%C3%A9nyi_model

http://de.wikipedia.org/wiki/Lotkas_Gesetz

*Potter, William (1981): Lotka's Law Revisited, Bibliometrics 30 (1), 21-40.

4. 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, deren Eigenschaften in den folgenden beiden Aufsätzen untersucht werden

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).

*Bing Liu (2007): Web Data Mining, Springer: Berlin, Heidelberg, New York. pp. 261-271.

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

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

5. Web Crawling

Web-Crawler (spider , robots) sind Programme, die automatisch WEB-Seiten laden und indizieren. Sie werden 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 diesen Referaten 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)
- Ethische Problem 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,http://infolab.stanford.edu/~backrub/google.html
* Segaran, T. (2008), Kollektive Intelligenz analysieren, programmieren und nutzen, O'Reilly

6. SEO: Suchmaschinenoptimierung (SEO)

Suchmaschinenoptimierung gibt es seit dem Start der ersten Suchmaschinen. Doch erst in den letzten Jahren gewinnt SEO deutlich an Bedeutung. Daraus ist sogar eine eigene Berufsbezeichnung entstanden. Natürlich liegt dies primär an der steigenden Anzahl der Internetnutzer und an ihrem Verhalten. Vor einigen Jahren kannten die meisten „Surfer“ die Adressen der bevorzugten Domains auswendig oder hatten sie notiert. Doch dieses Verhalten hat sich mit dem stark steigenden Angebot an Webseiten geändert. Für viele ist der Einstiegspunkt ins Internet eine Suchmaschine, mit knapp 98% Marktanteil ist bevorzugt natürlich Google zu nennen. Dadurch bekommt die Positionierung der eigenen Webpräsenz in Suchmaschinen für die Erfolgschancen eine bedeutende Rolle. Schätzungsweise 200 Faktoren nehmen im Algorithmus Einfluss auf die Rangordnung pro Suchphrase. Dies lässt erahnen welcher Aufwand und welche Kosten auf ein Unternehmen für die

Optimierung der Webpräsenzen zukommen können. Das Referat soll sich mit diesen Faktoren praxisnah beschäftigen.

Literatur:

www.suchmaschinenoptimierung.de

Patrick Klingberg, “Google Algorithmus”, www.seo-monster.de/google-algorithmus/ (

Florian Boxberg: Suchmaschinenoptimierung, Master-Thesis, Gummersbach 2010

7. Wrapper und strukturierte Datenextraktion

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
- Automaische 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 Multiple Pages

  RoadRunner System als praktische Umsetzung

Literatur:

Bin Liu, Web Data Mining, Kapitel 9, S, 330-373
* http://www.dia.uniroma3.it/db/roadRunner/ und * http://rtw.ml.cmu.edu/readtheweb.html

8. Opinion Mining

Hier geht es um die 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 bestimmten 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 S. 411-444

9. 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, S. 449-482

*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:

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

10. Empfehlungssysteme
Empfehlungssysteme (Recommendersystem) werden oft in E-Commerce-Anwendungen genutzt, um den Nutzern sinnvolle Empfehlungen für interessante Produkte zu geben, in dem sie ihm ein bisher ungekanntes Objekt vorschlagen. Ein bekanntes Beispiel ist www.amazon.de. Dabei unterscheidet man Content-based filtering, basierend auf Objekteigenschaften und Collaborative filtering, basierend auf Nutzung der Objekte, sowie Hybridverfahren. Das Referat soll eine Einführung in die grundlegenden Methoden dieser Systeme geben.

Literatur:

Linden, G., Smith, B., and York, J. 2003. Amazon.com Recommendations: Item-to-Item Collaborative Filtering. IEEE Internet Computing 7, 1 (Jan. 2003), 76-80

Adomavicius & A.Tuzhilin (2005), Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Trans. Knowl. Data Eng. 17(6):734-749.

*Segaran, T. (2008), Kollektive Intelligenz analysieren, programmieren und nutzen, O'Reilly

Legende: * = optionale Zusatzliteratur

Literatur kann zu Verfügung gestellt werden bzw. ist in der Bücherei vorhanden

Eigene Literaturrecherche oder auch eigene Themenvorschläge sehr erwünscht!

Es können jeweils 1-2 Personen ein Thema bearbeiten.