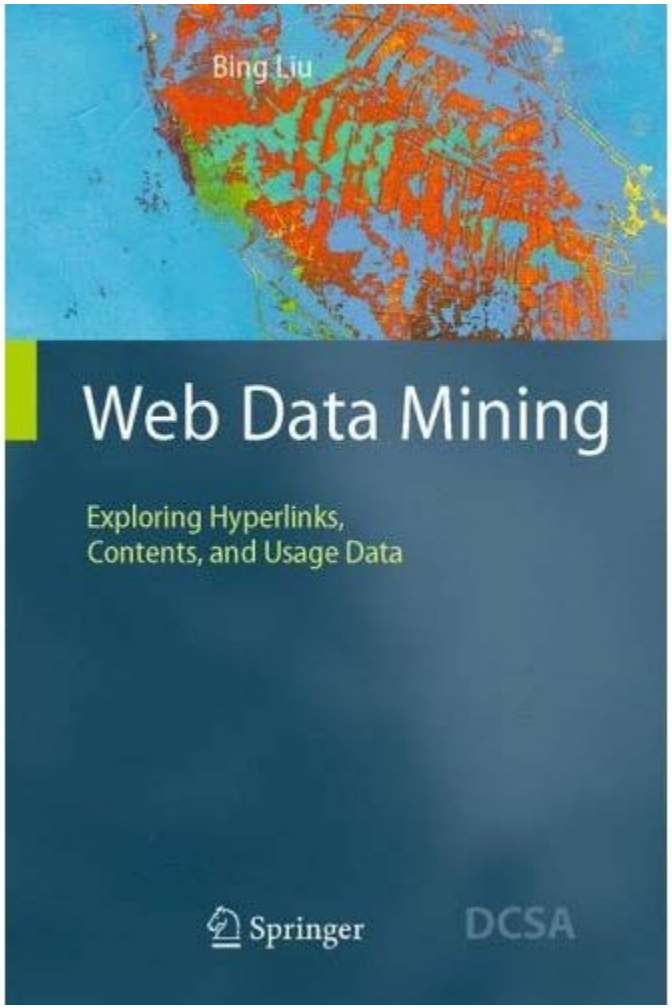
Web Data Mining

Alexander Hinneburg

Sommersemester 2007

Termine

- Vorlesung
  - Mi. 10:00-11:30 Raum ??
- Übung
  - Mi. 11:45-13:15 Raum ??
- Klausuren
  - Mittwoch, 23. Mai
  - Donnerstag, 12. Juli
- Buch
  - Bing Liu: Web Data Mining Exploring Hyperlinks, Contents, and Usage Data, Springer 2007



Einführung

- Das World Wide Web (kurz Web) hat in den letzten 10 Jahren viele Aspekte unseres Lebens beeinflußt.
  - besteht aus Milliarden von verknüpften Dokumenten
  - ist leicht zugänglich und durchsuchbar
  - neue Weise um Informationen zu beschaffen und zu verteilen
- Neue Plattform für Geschäfte
- Menschen drücken ihre Meinungen in Foren, Blogs, usw... aus => das Web ist ein Teil unserer Gesellschaft

Was ist das Web?

- Lexikon
  - WWW ist ein verteiltes Hypermedium das Menschen Zugriff auf eine große Dokumentmenge gibt
  - Zugriff auf Dokumente wird über das Internet gewährleistet
- Implementierung des Web
  - Client-Server Modell
  - Navigieren mittels Browser
  - Browser schicken Anfragen an Web-Server, interpretieren die Anwort und stellen die empfangen Dokumente dar
  - Hypertext erlaubt Verknüpfungen (Links) zu anderen Dokumenten
- Hypertext wurde 1965 von Ted Nelson entwickelt
  - Hypertext in Verbindung mit anderen Medien (Bilder, Ton,..) ist Hypermedium

Kurze Geschichte des Web (1/2)

- Tim Berners-Lee, CERN
  - Vorschlag über ein verteiltes Hypertextsystem wurde 1989 abgelehnt
  - Nach Wiedereinreichung 1990 angenommen
  - Entwicklung des HTTP Protokolls, HTML, URL und erste Implementierungen eines Servers und Browsers.
- Mosaik (1993) und Netscape (1994), IE (1995)
  - erste grafische Darstellung mit Mausunterstützung
- Internet
  - Grundlage für Web
  - begann als ARPANET, 1969-72
  - Vinton Cerf und Bob Kahn entwickeln 1973 TCP/IP

Kurze Geschichte des Web (2/2)

- Suchmaschinen
  - Excite, Stanford, 1993
  - Yahoo, 1994, Hierarchischer Verzeichnisdienst von Webseiten
  - Weitere Systeme
    - Lycos, Infoseek, AltaVista, Inktomi, AskJeeves, Northernlight
  - Google, Stanford, 1998
  - Microsoft, MSN, 2003
  - Yahoo, bietet seit 2004 allg. Suche an (nach Kauf von Inktomi)
- W3C
  - Gremium entwickelt Standards für das Web
- ACM WWW Konferenz gibt es seit 1994

Web Data Mining (1/2)

- 8 Eigenschaften des Web, die Web Mining interessant machen

1. Datenmenge im Web ist sehr groß und immer noch wachsend.
2. Sogut wie alle vorstellbaren Datentypen kommen im Web vor, strukt. Tabellen, semistrukt. Web-Seiten, unstrukt. Text, Multimedia Dateien
3. Informationen sind sehr heterogen, viele Seiten enthalten die gleiche Information in verschiedener Repräsentation
4. Viele Dokumente im Web sind verlinkt. Links haben verschiedene Semantik, navigierend, referenzierend,...

Web Data Mining (2/2)

1. Informationen im Web sind verrauscht, (a) Webseiten enthalten neben Hauptinhalt viel Beiwerk, (b) keine consistente Prüfung des Inhalts, widersprüchliche Aussagen möglich
2. Web ist eine Plattform für Geschäfte
3. Web ist dynamisch und ändert sich schnell
4. Web ist Teil der Gesellschaft, enthält Meinungsäußerungen, erlaubt Interaktion zwischen Menschen

Was ist Data Mining ?

- Finden von nützlichen Mustern in Datenquellen
- Gebiete
  - überwachtes Lernen, Klassifikation
  - unüberwachtes Lernen
    - Assoziationsregeln, Sequenzielle Muster, Clusteranalyse
- Schritte
  - Vorverarbeitung
  - Data Mining
  - Nachbearbeitung
- Datenquellen
  - Relationale Datenbanken
  - Texte und Dokumente

Was ist Web Data Mining ?

- Finden von nützlichen Mustern in
  - dem Web Graphen, der durch die Links entsteht
  - den Inhalten der Webseiten
  - den Logdaten der Webserver
- Web Struktur
  - finden von Mustern im Webgraphen
  - wichtige Web-Seiten
  - Kommunities entdecken
- Web Inhalte
  - Automatisch Webseiten nach Inhalt/Thema clustern
  - Meinungen aus Foren und Blogs extrahieren und zusammenfassen
- Web Benutzung
  - Finden von typischen Zugriffspfaden

Vorlesungsüberblick (1/2)

- Assoziationsregeln und sequenzielle Muster
  - Objekte, die häufig gemeinsam auftreten
  - Ereignisse, die in gleicher Reihenfolge ablaufen
- Überwachtes Lernen, Klassifikation
  - Grundlagen, Entscheidungsbäume, Regeln, Evaluation
  - Textklassifikation, SVM, Ensembles
- Unüberwachtes Lernen, Clusteranalyse
  - k-Means, EM Algorithmus, Evaluation
  - Teilweise überwachtes Lernen

Vorlesungsüberblick (2/2)

- Information Retrieval, Textsuche
  - Boolschen und Vektorraummodell, Finden von Duplikaten
- Linkanalyse
  - PageRank, HITS, Kommunities entdecken
- Web Crawling
- Erzeugen von Wrappern zur Informationsextraktion
- Extrahieren und Zusammenfassen von Meinungen
- Auswerten von Log-Daten