Web Data Mining

Albert Weichselbraun

Vienna University of Economics and Business
Department of Information Systems and Operations
Augasse 2-6, 1090 Vienna

albert.weichselbraun@wu.ac.at

May 2011

Agenda

Aufgabenstellung

- Facebook
- Web 2.0 Datenquellen
- Administration

Technologie

- Facebook Application
- Web 2.0 Datenquellen
- Geographische Visualisierungen

Aufgabenstellung

"Web Mining - Web Corpus Challenge"

1. soziale Quellen als Grundlage für sozialwissenschaftliche Forschung und
2. "Games with a Purpose" zur Erstellung und Evaluierung von derartigen Daten

Aufgabenstellung

Ziel dieses Projektes ist es

- Benutzerrezensionen (Rezension + Bewertung) aus dem Web zu extrahieren und diese in einer Datenbank zu speichern.
- Ein "Game with a Purpose" auf einer Social Networking Plattform (Facebook) zu erstellen, welches es dem Benutzer ermöglicht die Top Einflussfaktoren für das gewählte Rating zu identifizieren.
- Eine (minimale) Applikation zur Administration und Visualisierung der Ergebnisse zu erstellen (optional).

Arbeitspakete

- Komponenten zum Spiegeln von Bewertungsportalen (Amazon, Ciao, TripAdvisor, IMDB, RottenTomatoes, ...) und zum Extrahieren der relevanten Informationen
- Facebook “Game with a Purpose” zum Annotieren von Reviews
- Admin Applikation (optional)

Arbeitsaufteilung

Kernaspekte

- Technische Aspekte: Facebook-Applikation, Spider, Visualisierung, ...
- Organisatorische Aspekte: Zeitplan, Organisation von Treffen, Projektdokumentation, ...
- Strategische Aspekte: Planung, Modellierung, Marketing, Aufteilung der Workpackages, ...

Wöchentliches Reporting durch mindestens ein Gruppenmitglied

Achtung: Es wird erwartet, dass *jedes* Gruppenmitglied zumindest einen Programmiertask übernimmt.

Facebook - Anforderungen

- Design, Erstellen und Verteilen einer Facebook Applikation
- Originelle Spielidee
- Verhindern von Cheating
- Empfehlung: Verwendung von vordefinierten Kategorien für die Annotation oder Markieren von Ausdrücken im Text.
- Beispiel: apps.facebook.com/sentiment-quiz/

Integration von Web 2.0 Datenquellen

- Recherche: mögliche nützliche Datenquellen (TripAdvisor, Amazon, Ciao) & der entsprechenden Schnittstellen (API, ..)
- Wenn notwendig: Extraktion der relevanten Information aus dem Datenstrom
- Implementierung von ein bis zwei dieser Quellen und Integration der Quellen ins Datenschema
- Meta Informationen: Benutzer, Ort, Datum, ...

Administration

- Web Applikation zur Kontrolle der im Repository gespeicherten Daten
- Visualisierungen:
  - Anzahl der erfassten Rezensionen
  - Anzahl der von Benutzern annotierten Rezensionen
  - Geographische Verteilung der Benutzer

Facebook Application

Vorgangsweise:

- Facebook:
  - Ausgangspunkt: www.facebook.com/developers/
  - Neue Applikation erstellen
  - Canvas
    → URL: www.facebook.com/sql_example/
    → Callback URL:
    xmdimrill.ai.wu.ac.at/~aweichse/projects/2011s/sql2/GPa/
- xmdimrill:
  - Example Code + Library auf Dimrill kopieren.
  - Applikation mittels Canvas URL oder Canvas Callback URL testen.

Web 2.0 Datenquellen

Stand-alone Applikation (PHP, Python, Java)

- Technologie: anwendungsspezifisch
- Beginn: Recherche und entsprechende Dokumentation
- Integration von mindestens ein bis zwei externen Quellen
- Beachtung der Mirroring Restriktionen (zum Beispiel: maximal ein Request per Sekunde)
- Extraktion der Reviews: Reguläre Ausdrücke, DOM Tree (DOM Inspector, Firebug)

Web 2.0 Datenquellen - Beispiel: IMDB Review einlesen

```php
<?php
  // Beispiel: read Web page with the first
  //           10 comments for the movie
  //           tt0120794 (Prince of Egypt).
  $url='http://www.imdb.com/title/tt0120794'
       . '/usercomments';
  $lines=file($url);
  print_r($lines);
?>
```

Geographische Visualisierungen

Geographische Visualisierungen

- Verwendung der Google MAPS API
- Vorgangsweise:
  - API Key Anfordern
  - KML File erstellen (dynamisch/statisch)
  - JavaScript Code in die Webseite integrieren und KML File angeben.
- Beispiel:
  www.ai.wu.ac.at/~aweichse/projects/2011s/sql2/examples/maps/
  (~aweichse/public_html/projects/2011s/sql2/examples/maps)

Geographische Visualisierungen

```
1  <?xml version="1.0" encoding="UTF-8"?>

3  <kml xmlns="http://earth.google.com/kml/2.1">
4  <Document>
5  <name>Example</name>
6  <description>An Example KML File</description>

8  <Placemark>
9    <name>Perth</name>
10   <description>An Example Placemark</description>
11   <Point><coordinates>
12     115.833297729,-31.9333000183
13   </coordinates></Point>
14 </Placemark>
15 </Document>
16 </kml>
```

Geographische Visualisierungen

```
1  <html xmlns="http://www.w3.org/1999/xhtml"><head>

3  <script
4  src="http://maps.google.com/maps?file=api&amp;v=2&amp;key=XXXX"
5  type="text/javascript"></script>
6  <script type="text/javascript">
7   function load() {
8     if (GBrowserIsCompatible()) {
9       var map = new GMap2(document.getElementById("map"));
10      var geoXml = new GGeoXml("http://t.at/myGeoFile.kml");

12      map.setCenter(new GLatLng(48.2,16.36), 3);
13      map.addControl(new GLargeMapControl());
14      map.addOverlay(geoXml);
15    }
16  }
17 </script>

19  </head>
20  <body onload="load()" onunload="GUnload()">
21    <div id="map" style="width: 680px; height: 400px"></div>
22  </body>
23 </html>
```