Open-Access-Statistik

Wie funktioniert der zentrale Service Provider?

Workshop OA-Statistik:
„Was zählt? - Nutzungsstatistiken als alternative Impact-Messung“
Göttingen, 21. Januar 2010

Hans-Werner Hilse

Niedersächsische Staats- und Universitätsbibliothek Göttingen

<hilse@sub.uni-goettingen.de>

Initiiert durch:

DINI – Deutsche Initiative für Netzwerkinformation e.V.

Gefördert durch:

Deutsche Forschungsgemeinschaft DFG

Rolle des zentralen Service Provider:

- Garantiert einheitliches Berechnungsmodell:
  - Zeitspannen bei Mehrfachzugriffen
  - Crawler-/Robot-Listen
- Kann – mit zunehmender Zahl teilnehmender Data Provider – Service-übergreifende Effekte berücksichtigen
- Zentrale Implementation notwendiger Schnittstellen/Formate (z.B. SUSHI)

Funktionsweise

- Aggregation:
  - tägliches Einsammeln der neuen Nutzungsdaten von jedem Data Provider
  - Einstellen der Daten in RDBMS
- Verarbeitung: Auswertungslauf
  - Für auszuwertende Dokumente (definiert durch Identifier) dann Berechnungsschritt
- Bereinigung des Datenbestands
  - „Ausmisten“ von Alt-Daten
- Bereitstellung
  - Schnittstellen zur Nutzung der generierten Daten

Aggregation

- Gemäß OAI-PMH-Standard
- Verarbeiten der Context Objects:
  - Parsen des XML, Kontrolle der Validität
  - Neue bzw. geänderte Daten werden in eine relationale Datenbank gespeist
  - Herausforderungen:
    - Komplexes Datenformat, keine einfache Speicherung in einer einzigen Tabelle
    - Anzahl der Datensätze: Beispiel „Webdoc“-Repository der SUB Göttingen (ca. 3000 Dokumente, überwiegend Dissertationen): ca. 50.000-70.000 Einzelzugriffe (HTTP) pro Tag – betrifft nötigen Rechenaufwand und Speicherplatz
    - Geschwindigkeit/Durchsatz der Datenübertragung

Verarbeitung

- Für relevante Identifier (solche von Dokumenten, für die gemessen werden soll) wird 1x täglich Auswertungslauf gemacht
- Vgl. folgenden Vortrag von T. Schäfer
- Herausforderungen:
  - Rechenaufwand: Umsetzung der Zählalgorithmen ist rechenintensiv

Bereitstellung

- Verfügbarmachung der Statistiken
  - Fertige Grafiken oder Teil-Webseiten, die einfach in Repositories o.ä. eingebunden werden können
  - Reines Zahlenmaterial
- Vgl. ebenfalls folgenden Vortrag

Technik

- Fokus: Portabilität, Erweiterbarkeit
- Ausschließlich (X)HTML/CSS, PHP und JavaScript
- Nutzung freier Software-Bibliotheken
- Einfache Einbindung in bestehende Repositorien-Software
- Programm-Code auf einfache Anpassung und Erweiterbarkeit angelegt

Technik: Datenmodell

- Eingesammelte Nutzungsdaten verteilen sich wegen detaillierter Strukturen auf viele Tabellen:
  - Allein 8 Tabellen für Daten in den ContextObjects,
  - weitere Tabellen für Management der OAI-PMH-Aggregation
- Auswertung durch Verknüpfung der Daten in den verschiedenen Tabellen
- Folge: Hohe Rechenlast bei Datenbank

Randnotiz: Ein paar Zahlen

- Verarbeitungsgeschwindigkeit bei Datenbank-Aufnahme: ca. 30 Nutzungsereignisse pro Sekunde, skaliert aber mit verfügbarer Rechenleistung
- Bis jetzt ca. 50.000 verschiedene „User Agents“, also Browsertypen bzw. Crawler/Robots
- Verhältnis Datenmenge einer Webserver-Logzeile zu einem ContextObject: ca. 1:10 bis 1:15
- Datenmenge Log-Daten eines einzigen Werktags für den GoeDoc-Server (3000 Dokumente): ca. 55 MByte

Open-Access-Statistik

Vielen Dank für Ihre Aufmerksamkeit!

Workshop OA-Statistik:
„Was zählt? - Nutzungsstatistiken als alternative Impact-Messung“
Göttingen, 21. Januar 2010

Hans-Werner Hilse

Niedersächsische Staats- und Universitätsbibliothek Göttingen

<hilse@sub.uni-goettingen.de>

Initiiert durch:

DINI – Bibliotheken, Medienzentren, Rechenzentren, Wissenschaft – DEUTSCHE INITIATIVE FÜR NETZWERKINFORMATION E.V.

Gefördert durch:

Deutsche Forschungsgemeinschaft DFG