

Wikipedia-Analysen

Technologien des Wissensmanagements

Daniel Fay, Bernd Hermes, Shimal Ibrahim

02. Juli 2013

- Wikipedia Informationen
- Unser Konzept
- Ergebnisse
 - Analyse des Artikels
 - Analyse des Users (Benutzers)
 - Analyse der Daten
 - Sonstige Datenanalysen
- Fazit
- Quellen

- Artikel
 - Titel
 - Mehrere Revisionen
 - ID
 - User
 - Name
 - ~~Content~~
 - Timestamp
 - Änderungsart
 - Size
 - Kategorie

- Folgende Infos, die von Wiki zur Verfügung gestellt werden, werden in unserem Konzept weiter betrachtet
 - Artikeltitel
 - Revisionen eines Artikels
 - Liste von Usern
 - Kategorien
 - Änderungsarten

- Wie werden die Wiki Infos analysiert?
 - Artikel und User
 - Nutzung einer API-Schnittstelle
 - Speichern der Daten in die Datenbank
 - Absetzung von SQL-Befehle zur Erhaltung von relevanten Informationen
 - Änderungen
 - Durch Analyse der einzelnen Revisionen werden die Änderungen festgestellt

- Ein Artikel besitzt n-Revisionen
- Revisionen sind nicht identisch
- Letzte Revision wird bearbeitet



- Für Analyse wichtig
 - Maximale Anzahl an Zeilen
 - Maximale Anzahl an Wörter
 - Liste von einzigartigen Zeilen
 - Liste von einzigartigen Wörtern

- Vergleich von zwei Revisionen ($n-1$ & n) in 3 Schritte
 - Schritt 1: Zeilenvergleich
 - Schritt 2: Wörtervergleich
 - Schritt 3: Zählung der einzigartigen Zeilen und Wörtern

- Vergleich von zwei Revisionen in 3 Schritte
 - **Schritt 1: Zeilenvergleich**
 - Prüfung auf Ähnlichkeit
 - Identische Zeilen werden aus beide betrachteten Revisionen entfernt

Artikel n	Revision n-1	Zeilen	Wort 1	Wort 2	Wort 3	Wort 4	Wort 5
			Wort 1	Wort 2	Wort 3	Wort 4	Wort 5
		Zeilen	Wort 1	Wort 2	Wort 3	Wort 4	Wort 5
			Wort 1	Wort 2	Wort 3	Wort 4	Wort 5
	Revision n	Zeilen	Wort 1	Wort 2	Wort 3	Wort 4	Wort 5
			Wort 1	Wort 2	Wort 3	Wort 4	Wort 5
		Zeilen	Wort 1	Wort 2	Wort 3	Wort 4	Wort 5
			Wort 1	Wort 2	Wort 3	Wort 4	Wort 5

- Vergleich von zwei Revisionen in 3 Schritte

- **Schritt 2: Wörtervergleich**

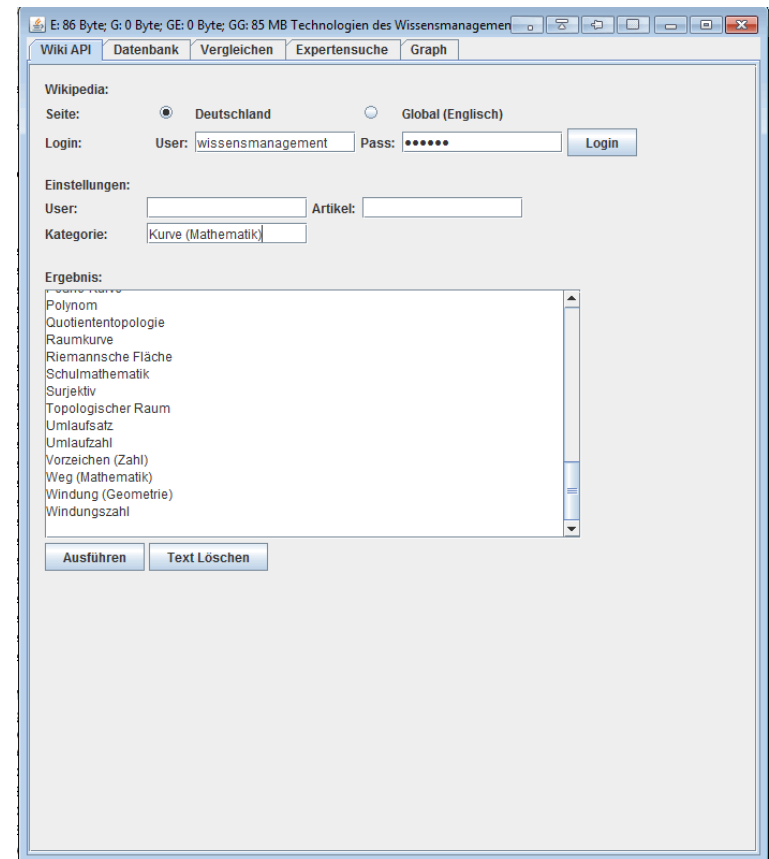
- Prüfung auf Ähnlichkeit der Wörter in den nicht übereinstimmenden Zeilen
 - Identische Wörter werden aus beiden betrachteten Revisionen entfernt

Artikel n	Revision n-1		Wort 1						Wort 2						Wort 3						Wort 4						Wort 5						Wort 6					
	Revision n		Wort 1						Wort 2						Wort 3						Wort 4						Wort 5						Wort 6					
			Wort 8						Wort 9						Wort 10						Wort 11						Wort 12						Wort 13					

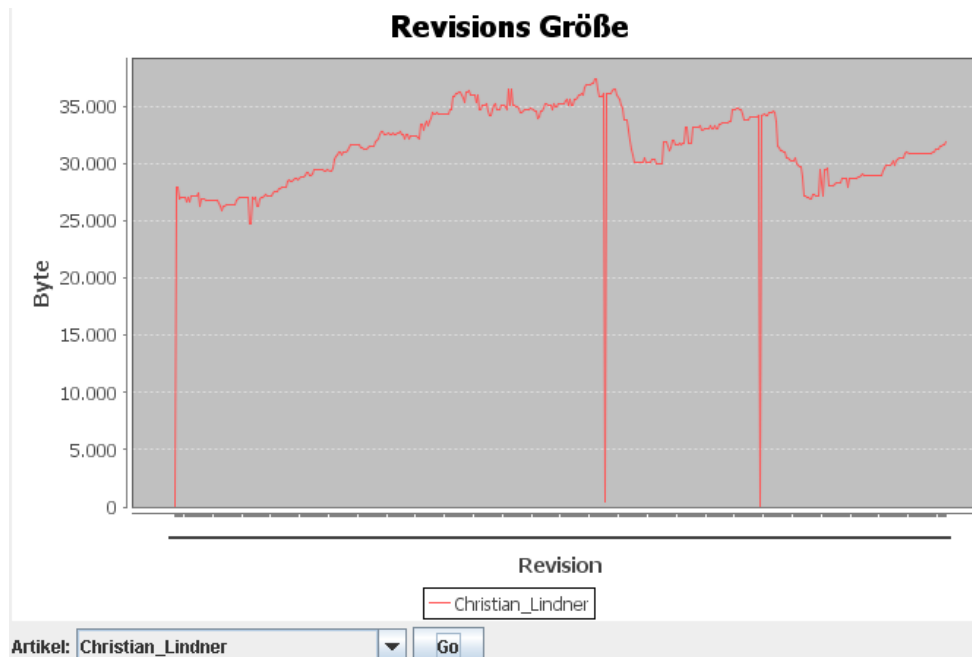
- Vergleich von zwei Revisionen in 3 Schritte
 - **Schritt 3: Zählung der einzigartigen Zeilen und Wörtern**
 - Nicht identische Zeilen werden gezählt
 - Nicht identische Wörter werden gezählt
 - Vergleich über Anzahl der nicht identischen Zeilen & Wörter
 - Klassifizierung von Änderungen

Artikel n	Revision n-1		Wort 1	Wort 2	Wort 3	Wort 4	Wort 5	Wort 6
	Revision n		Wort 1	Wort 2	Wort 3	Wort 4	Wort 5	Wort 6
			Wort 8	Wort 9	Wort 10	Wort 11	Wort 12	Wort 13

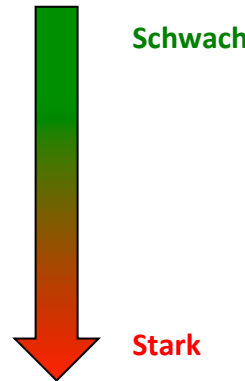
- Wie werden die Wiki Infos analysiert?
 - GUI



- Analyse des Artikels
 - Revisionsgröße über die Zeit
 - Screenshots vom Graphen



- Analyse des Artikels
 - Änderungsarten eines Artikels
 - Klassifizierung in
 - Keine Änderung
 - Korrektur
 - Formatierung
 - Verbesserung
 - Wissensproduktion
 - Überarbeitung



- Analyse des Users (Benutzers)
 - Betrachter User
 - Aka
 - Artikel des Users
 - SQL-Befehl: Select Artikel from revision where User=„Aka“ group by Artikel;
 - Christian_Lindner, Datenschutz, Deutschland
 - Homöopathie, PRISM_(Überwachungsprogramm), U2_(Band)
 - Wie gehören diese Artikel des Users zusammen?
 - Betrachtung von Kategorien und Artikel zeigt Zusammengehörigkeit

- Analyse des Users (Benutzers)
 - Betrachter User
 - Aka
 - Wie gehören diese Artikel des Users zusammen?
 - SQL-Befehl: `Select Kategorie from kategorie where Artikel=„Christian_Lindner“;`
 - Kategorien sind u.a. Praktische Informatik, Persönlichkeitsrecht, Datenschutz, Mitgliedstatt der OECD, Staat in Europa, Deutschland
 - Deutschland und Homoöpathie (Zusammengehörigkeit: ja)
 - Datenschutz, Deutschland und PRISM_(Überwachungsprogramm) gehören zu Kategorie Deutschland (Zusammengehörigkeit: ja)
 - Christian_Lindner und U2_(Band) haben keine Zusammengehörigkeit

- Analyse des Users (Benutzers)
 - Wie Arten von Änderungen führen User durch?
 - Kleine und große Änderungen
 - Durch Benutzung unserer Software können weitere Änderungsarten festgestellt werden, z.B.

Artikel	User	Änderung
Christian_Lindner	Kleiner Stampfi	Korrektur
Datenschutz	Mikered	Keine Änderung, da er/sie die erste Revision erstellt hat
Datenschutz	134.91.6.204	Formatierung, und Korrektur (User hat zwei Revisionen bearbeitet)

- Analyse des Users (Benutzers)
 - Sind Typen von Nutzern erkennbar?
 - Stammesmitglied: verfasst mehrere Artikel, regelmäßige bei Änderungen aktualisiert
 - Administratorstatus: User ist über längeren Zeitraum aktiv beteiligt
 - Durch Benutzung unserer Software können weitere Typen von Nutzern erkannt werden, z.B.

Artikel	User	Änderung
Christian_Lindner	Miltrak	Korrektur (3x)
Christian_Lindner	Bluemangenius	Keine Änderung (1x) und Korrektur(1x) (Kein Typ erkennbar)
Deutschland	Horst-schlaemma	Formatierung (2x), Korrektur (1x) und Wissensproduktion (1x) (Kein Typ erkennbar)

- Analyse der Daten
 - Welche Nutzer sind sich ähnlich?
 - USER X Y sind sich ähnlich, weil sie beide die gleiche Kategorien überarbeiten
 - Weil sie die selben Typen von Änderungen haben
 - Korrektur
 - Formatierung
 - Verbesserung
 - Wissensproduktion
 - Überarbeitung

- Analyse der Daten
 - Wie kann man Kooperationsfähigkeit zweier Nutzer messen?
 - Nutzer müssen am selben Artikel gearbeitet haben
 - Auf große Änderung folgt kleine Änderung (Rechtschreibkontrolle)
 - Artikelgröße nimmt stetig zu (Separate Erweiterungen)
 - Wiederholende Folge von wachsende und schrumpfende Anzahl der Artikelgröße (Editwar)

- Analyse der Daten
 - Wie kann man die Reputation von Nutzern bewerten?
 - Mehrere Kriterien
 - Nur Gast
 - Bann (Fehlverhalten)
 - Gruppen (Bürokrat, Bot, Administrator)
 - Rechte (Gruppenabhängig)
 - Registrierungsdatum
 - Kommentare

- Sonstige Datenanalysen
 - Expertensuche
 - Hat an Artikel gearbeitet
 - Geordnet nach Menge und Typ der Veränderung
 - Hat Artikel einer Kategorie bearbeitet
 - Zeitgeist
 - In welcher Reihenfolge wurden die Homonyme hinzugefügt
 - Welches Homonym wurde wann, wie oft verändert
 - Es steht im Artikel

- Spannendes Projekt
- Die Funktionsweise von Wikipedia kennengelernt
- Funktion und Hintergründe von Wikipedia
- Texte vergleichen
- Lange Einarbeitungszeit in die Wiki-API

- Wikipedia (Deutschland)

DANKE FÜR IHRE AUFMERKSAMKEIT!