Ontologies are us: A unified model of social networks and semantics

Christopher Sonneborn

30. Juni 2009

Inhaltsverzeichnis

Ontologie

- Ontologien dienen der Verbesserung der Kommunikation zwischen menschlichen und maschinellen Akteuren.
- Bestehen aus:
  - Lexikon: enthält eine Menge von Worten (lexikalischen Einträgen, Symbolen), mit denen Begriffe und semantische Relationen bezeichnet werden
  - Begriffen: charakterisieren, welche Begrifflichkeiten für einen Anwendungsbereich als relevant erachtet werden
  - semantische Relationen: setzen Ontologioen zueinander in Beziehung
  - regelhafte Zusammenhänge: erfasst zusätzliche Bedeutungsinhalte von Begriffen und Relationen

Standard für Ontologien

- Beschreibung von Ontologien durch OWL von W3C
- Mit OWL (Web Ontology Language) kann man mit einer formalen Beschreibungssprache Ontologien erstellen.
- OWL basiert auf RDF-Syntax und DAML+OIL

Emergente Semantiken

- Problem im (Semantic) Web:
  - Ontologien sind zu starr für den Einsatz zur Bildung von Web Communities
  - Lexikon müsste permanent mitgepflegt werden
- Lösung: Emergente Semantiken
  - Semantiken werden durch Interaktionen von vielen Agenten erzeugt
  - Daraus entstehende Ontologien werden dynamisch

Folksonomie

- Folksonomie kommt von Folk und Taxonomie
- Teilen von Wissen oder Objekten durch freiwählbare Schlüsselwörter
  - Schlüsselwörter sind nicht festgelegt
  - Schlüsselwörter unterliegen keinerlei Bedingungen, Einschränkungen oder Semantiken
  - Beschreibung von Objekten auch durch mehrere Schlüsselwörter

Definition: Folksonomie Netzwerk

- Beschreibung eines Netzwerkes von Folksonomien durch einen dreigeteilten Graphen mit Hyperkanten
- Syntax:
  - T Folsonomie
  - $A = \{a_1, a_2, ..., a_k\}$ Menge von Actors (User)
  - $C = \{c_2, c_2, ..., c_l\}$ Menge von Concepts (Tags bzw. Begriffe)
  - $I = \{i_1, i_2, ..., i_m\}$ Menge von Instances (Objekte)

Definition: Folksonomie Netzwerk (forts.)

- Solch ein Netzwerk beschreibt Beziehungen zwischen Usern, Begriffen und Objekten.
- Der Graph ist definiert als:
  - $T \subseteq A \times C \times I$
  - $H(T) = \langle V, E \rangle$
  - wobei $V = A \cup U \cup I$
  - und $E = \{\{a, c, i\} | (a, c, i) \in T\}$

Reduzierung des Folksonomie Netzwerks

- Solch ein dreigeteilter Graph ist schwer zu verarbeiten.
- Lösung: Reduzierung auf drei zweigeteilte Graphen:
  - AC Graph: Graph für Beziehungen zwischen User und Begriffen
  - CI Graph: Graph für Beziehungen zwischen Begriffen und Objekten
  - AI Graph: Graph für Beziehungen zwischen User und Objekten

AC-Graphen

- AC Graphen sind definiert als:
  - $AC = (A \times C, E_{ac})$
  - $E_{ac} = \{(a, c) | \exists i \in I : (a, c, i) \in E\}$
  - $w : E \rightarrow \mathbb{N}$
  - $\forall e = (a, c) \in E_{ac}$
  - $w(e) := |\{i : (a, c, i) \in E\}|$
- In Worten: Der Graph verbindet User mit Begriffen, die sie zusammen für ein Objekt genutzt haben.
- Die Kanten des Graphen sind durch die Häufigkeit der Benutzung eines Begriffs durch einen User gewichtet.

Matrix Darstellung

- Verarbeitung des Graphen in Matrizen:
- $B = \{b_{ij}\}$ wobei $b_{ij} = 1$ wenn User a_i mit dem Begriff c_j verbunden ist.
- $S = s_{ij} = \sum_{x=1}^{k} b_{ix} b_{xj} = BB'$ wird Affiliation Matrix genannt
  - Sie verbindet User, die gleiche Begriffe nutzen
  - gewichtet nach Anzahl der Begriffe, die sie zusammen nutzen
- $O = B'B$ ist eine zu S ähnlich Matrix
  - Sie verbindet Begriffe die von Benutzern zusammen verwendet wurden
  - Gewichtet nach der Anzahl der User

Reduzierung des AC-Graphen

- Aus dem Affiliation Netzwerk kann man zwei weitere Graphen extrahieren:
  - ein soziales Netzwerk von Usern, basierend auf überlappenden Mengen von Objekten
  - eine leichtgewichtige Ontologie von Begriffen, basierend auf überlappenden Mengen von Communities
- Im weiteren Verlauf geht es nur noch um die leichtgewichtigen Ontologien basierend auf überlappenden Communities (O_{ac}) und basierend auf überlappenden Mengen von Objekten (O_{ci})

Arten von Inhalten

- Ontologien beinhalten verschiedene Kombinationen von Inhalten.
  - allgemeine Inhalte
    Eigenen sich zur Bildung von Clustern von Inhalten
  - bestimmte Inhalte
    Haben eine enge Bindung zu ihrer Umgebung
- Zur Abgrenzung der beiden Kategorien kann der *clustering coefficient*, der *(local) betweenness centrality* oder der *network constraint* genutzt werden, die teilweise in Pajek und UCINET implementiert sind.

broader/narrow Relationen

- Extrahierung von broader/narrow Relationen
  - Konzept A ist Oberbegriff von Begriff B, wenn gilt: $B \subseteq A \rightleftharpoons A \cap B = B$
  - A ist signifikant größer als B wenn gilt: $|B|/|A| < k$ für eine Variable k.
  - Definition von Untermengenrelation durch das Finden von fast perfekten Überlappungen: $|A \cup B|/|B| < n$
- In CI-Graphen kommen Items aus narrow auch in broader vor. Extrahierung von Klassifizierungs Hierarchie
- In AC-Graphen extrahieren von Hierarchie basierent auf Subcommunitie Relationen.

Was ist del.icio.us?

- del.icio.us ist ein social bookmarking tool
- Man kann online Bookmarks speichern und verwalten
- Bookmarks können mit Schlüsselwörtern versehen werden
- Teilen von Bookmarks mit Freunden
- Finden von Bookmarks durch die Schlüsselwörter

Was wurde gemacht?

- del.icio.us stellt Daten durch RSS Feeds dar
- Informationen wurden durch einen RDF Crawler gesammelt
  - Initialisierung des Crawler mit dem Tag „web“
  - Durchlaufen des entstehenden RSS Netzwerkes mit einer Art Breadth-First-Suche
- Ergebnis: 51852 Kommentare zu 30790 URLs von 10198 Benutzern, die 29476 eindeutige Tags verwendeten

Was wurde gemacht? (forts.)

- Verkleinerung der Daten durch Ausfiltern von Einträgen:
  - Tags, die weniger als zehn Einträge klassifizieren
  - User mit weniger als fünf Begriffen
- Erstellung von zwei Ontologien mit Hilfe von Pajek:
  - O_{ac}
  - O_{ci}

Darstellung der Graphen

- Nach weiterem Ausfiltern erhält man die zwei Graphen:

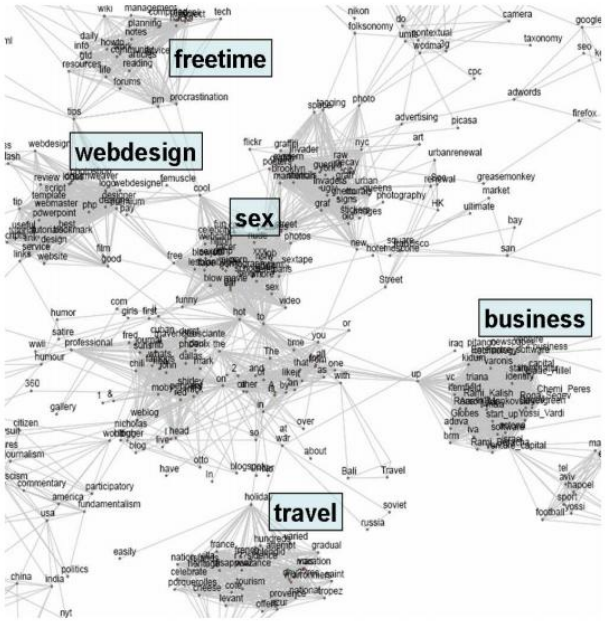

Abbildung: O_{ci}-Graph

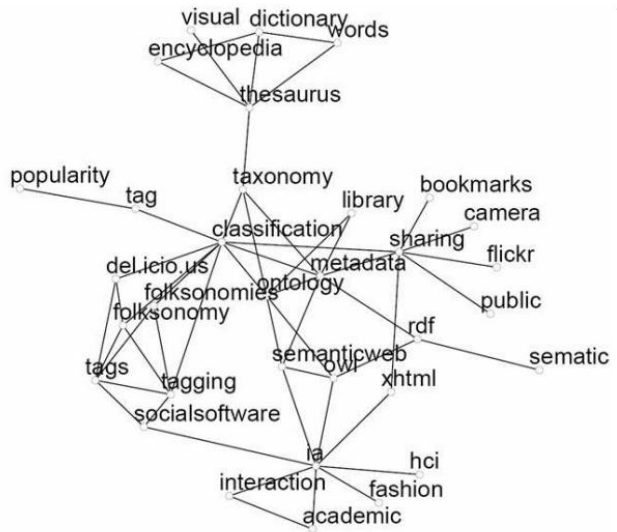

Abbildung: O_{ac}-Graph

Unterschiede der beiden Netzwerke

- Unterschiedliche Auswahl von Inhalten:

O_{ci}	*/GoogleHacks, _0, 04, 1, 2, 2005, 3g, a, A, a9, Aaron_Mankovski, actona, actors, adult, aduva, advice, ajax, all, Allegrini, america, an, and, angeles, apparel, Apple, as, assembly, attempt, attention, attention.xml, aviv, axml, azur
O_{ac}	.net, 3d, 43folders, academic, accessibility, acronym, actionscript, activism, ad, ads, adsense, advertising, advice, advisories, adwords, agile, ajax, amazon, america, analysis, and, Apache, apache, api, app, apple, application, architecture, archive, Art, art, articles, asia, astronomy, atlas, Audio

Abbildung: Inhalte beider Netzwerke, die mit „A“ bzw. „a“ beginnen

Unterschiede der beiden Netzwerke (forts.)

- Clusterbildung beider Netzwerke:
  - O_{ci}: Hauptsächlich spezielle Inhalte, z.B.: up, cool, hot, in, to
  - O_{ac}: Wenig Clustering: Häufiges Fehlen von speziellen und allgemeinen Inhalten

Ergebnisse der Netzwerke

- O_{ci} Netzwerk:
  - Starke Verbindung zwischen Begriffen, bei der Teilung von einem großen Prozentsatz von Inhalten, unabhängig von der Anzahl von User, die sich dafür interessieren.
  - Die durchschnittliche Gewichtung ist recht hoch.
- O_{ac} Netzwerk:
  - Starke Verbindung, wenn zwei Begriffen mit einen großen Anteil von Usern verbunden sind, unabhängig von der Anzahl von Instanzen.
  - Die durchschnittliche Gewichtung ist recht niedrig.

λ-set Analyse mit UCINET

- Ergebins:
  - größeres Netzwerk mit 751 Begriffen (vorher 64)
  - fünf geschlossene Gruppen von Begriffen mit Möglichen Interssen in Reisen, Geschäft, Freizeit, Porno und Web Design

Fazit

- O_{ci} Netzwerke ignorieren die Relevanz von Begriffen für User
- O_{ci} geben nur ungenau Communities wieder
- O_{ac} geben Communities besser weider
- Begiffe in einem O_{ac} Netzwerk sind wichtig für User
- O_{ac} ergeben bessere Ergebnisse für die bildung von Sub-Communities, z.B.:

broader	ajax	linux	xml	mac	flash
narrow	json	ubuntu, gnome	xslt	iPhoto	actionscript

Tabelle: Broader/Narrow Relation basierend auf Sub-Communities

- Die vorgestellten Methoden zur Netzwerkanalyse sind wenig zur Extrahierung von Taxonomie Relationen geeignet

Vorbereitung

- Fiktives Modell für Ontologien:
  - Mitglieder einer Community sind User
  - Vorbereitet Liste von Inhalten
  - Web Seiten sind Objekte
  - Web Seiten sind beschrieben durch ein Begriff wenn das Begriff auf der Seite auftritt

Generierung der Netzwerke

- O_{ci}: Nutzen einer Suchmaschine um für alle Paare von Begriffen die Anzahl von Seiten zu bekommen, wo diese auftreten. Diese werden dann mit Hilfe ihrer eigenen Seitenzahlen normalisiert.
- O_{ac}: Es existiert eine Verbindung zwischen einer Person, einem Begriff und einer Web Seite, wenn der Name einer Person zusammen mit einem Begriff auf einer Web Seite auftritt.
- Zur Erzeugung dieser Netzwerk wurde das Flink System genutzt.

Ergebnis der Netzwerke

- Ergebnis des O_{ci} Netzwerkes:
  - Starke Assoziation zwischen allgemeinen Inhalten
  - Bestimmte Inhalte fehlen fast vollständig

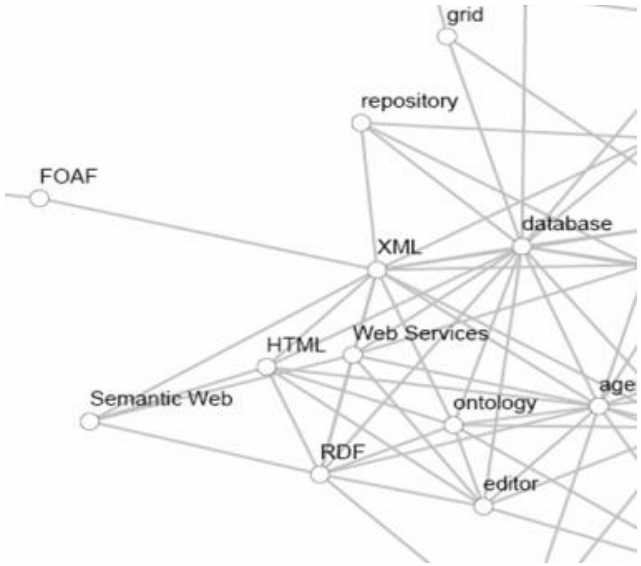

Abbildung: Ausschnitt des O_{ci} Netzwerkes

Ergebnis der Netzwerke (forts.)

- Ergebnis des O_{ac} Netzwerkes:
  - Starke Assoziation zwischen gebietsspeziefischen Begriffen
  - Korrektes Clustering

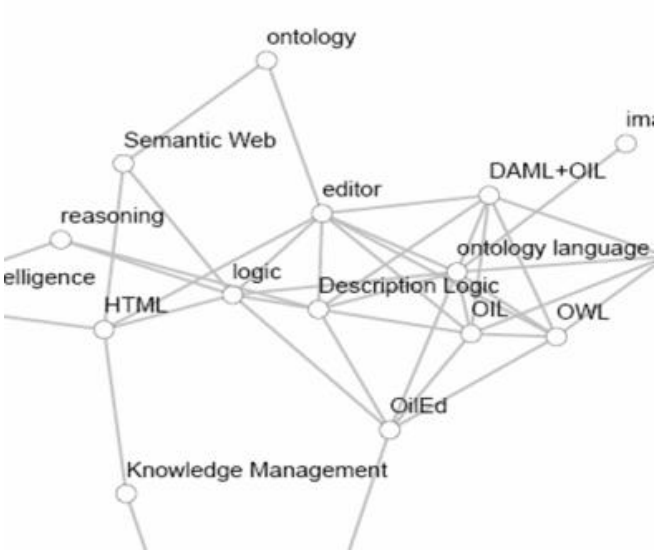

Abbildung: Ausschnitt des O_{ci} Netzwerkes

Woher kommen die Unterschiede in den Netzen?

- O_{ci} Netzwerke ignorieren die Relevance für die Communities
- Für das O_{ac} Netzwerk wurde erst nach der Person und dann nach dem Begriff gefragt.

Bewertung der Ontologien

- 61 Leute wurden per E-Mail gefragt, welche Ontologie sie auf Grund der Ergebnisse für aussagekräftiger halten
- Ergebnis:

	N	O_{ac}	O_{ci}	Ratio	Sign.
Alle	30	22	8	73,3%	0,0055
ISWC	23	18	5	78,3%	0,004
ISWC-Kern	15	13	2	86,7%	0,0032

Tabelle: Verteilung der Antworten

Fragen?

Fragen?