Web 2.0 - Begriffsklärung

Der Begriff „Web 2.0“ bezieht sich primär auf eine veränderte Nutzung und Wahrnehmung des Internets: Die Benutzer erstellen und bearbeiten Inhalte selbst.

Er bezeichnet aus technischer Sicht auch eine Anzahl von Methoden wie
- Web-Service-APIs,
- Ajax (Asynchronous Javascript und XML)
- und Abonnement-Dienste wie RSS.

(Siehe http://de.wikipedia.org/wiki/Web_2.0)

Typen von Web 2.0- Anwendungen

- Wikis (z.B.: Wikipedia)
- Blogs (z.B.: irgendein journalistisches Blog?)
- Photo- und Videoplattformen (z.B.: Youtube, Flickr)
- Social Bookmarking (z.B.: del.icio.us, BibSonomy)
- soziale Online-Netzwerke (z.B.: Xing, Myspace, Facebook, StudiVZ)
- virtuelle Welten (z.B. Second Life, Bailamo)
- Mikroblogs (z.B.: Twitter)

Tagging / Folksonomies

tagging is a distributed process
tagging has a small cognitive overhead
system contents can be browsed by tag
the system evolves in time: new resources, new users, new tags
there may be an underlying social network, explicitly exposed or not
the behavior of users is “selfish”
users are exposed to each other’s activity
users share implicit knowledge (language, cultural background)
Social Bookmarking Systems

- Collaborative annotation of web resources
- Easy to use, open for everyone
- Joint use leads to converging vocabularies and emergent semantics.

There are many popular folksonomy systems on the web, eg:
- flickr (photos)
- YouTube (videos)
- del.icio.us (bookmarks)

Our system: BibSonomy

Folksonomies allow users to assign tags to resources.

A folksonomy is a tuple $F := (U, T, R, Y, ~)$ where
- $U$, $T$, and $R$ are finite sets, whose elements are called users, tags and resources,
- $Y \subseteq U \times T \times R$, called set of tag assignments,
- $\prec \subseteq U \times T \times T$ is a user-specific sub-tag/super-tag relation.

The personomy $P_u$ of user $u$ is the restriction of $F$ to $u$.
Types of Tags

- content/topic of resource (nouns, proper nouns, ...)
- category of resource
- opinion about resource (adjectives)
- ownership of resource (user names)
- self-reference, relation between resource and user (mystuff, myown, citingme)
- task organization (toread, tobuy)
- social coordination (for:andrea)

[ see Golder & Huberman '06 ]

Probleme und Vorteile des Web2.0 (insbes. Folksonomies)

Probleme:
- keine formale Semantik
- viele Mehrdeutigkeiten, Tippfehler, etc.

Vorteile:
- Viele Beitragende tragen große Mengen an Wissen zusammen
- Hilft gegen den Wissensakquisitions-Flaschenhals

Ziel ist es, die Lücke zwischen dem Semantic Web und dem Web 2.0 zu schliessen. („Bridging the Gap“)

(Dies wird gelegentlich schon als „Web3.0“ bezeichnet.)

Wenn dies (semi-)automatisch gelingt, kann man das Wissen der Vielen („Wisdom of the Crowd“) in eine formale Sprache überführen und somit maschinell verarbeitbar machen.

EU Project: TAGora - Emergent Semantics in Social Online Communities

Motivation

- Final Goal: Understand “tag semantics” in a folksonomy, i.e.,
  - Which tags describe the same / a more specific / a more general concept?
- Two basic approaches:
  1. Look up tags in external thesaurus:
     + semantically grounded metrics
     - “folksonomy jargon” (misspellings, neologisms etc.) not present
  2. Apply measures directly to folksonomy structure (e.g., cooccurrence statistics, ...)
     + inclusion of complete vocabulary
     - semantic interpretation of measures is not clear

→ Understand characteristics of (distributional) measures
→ assess their applicability for tag recommendation, ontology learning, ...

Semantic Grounding

Relatedness Measures

- Take Co-occurrence frequency as similarity measure (freq).
- Use FolkRank to find related tags (folkrank).

Describe each tag as a vector, whereby each dimension of the vector space corresponds to another tag. Compute similar tags by cosine similarity (cosine).
(The same can be done in the user space or the resource space and with TF-IDF.)

Dataset

- Del.icio.us crawl 2006
  - \(|U| = 667,128\) \(|T| = 2,454,546\) \(|R| = 18,782,132\)
  - \(|Y| = 140,333,714\)

- Excerpt: 10,000 most popular tags
  - \(|U| = 476,378\) \(|T| = 10,000\) \(|R| = 12,660,470\)
  - \(|Y| = 101,491,722\)

- In the following: tag rank = position in most-popular list:
  - 1: design
  - 2: software
  - 3: blog
  - 4: web
  - ...

Example for cosine measure
Examples of most related tags

<table>
<thead>
<tr>
<th>rank</th>
<th>tag</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>web2.0</td>
<td>ajax</td>
<td>web</td>
<td>tools</td>
<td>blog</td>
<td>webdesign</td>
</tr>
<tr>
<td>15</td>
<td>howto</td>
<td>tutorial</td>
<td>reference</td>
<td>tips</td>
<td>linux</td>
<td>programming</td>
</tr>
<tr>
<td>28</td>
<td>games</td>
<td>fun</td>
<td>flash</td>
<td>game</td>
<td>free</td>
<td>software</td>
</tr>
<tr>
<td>30</td>
<td>java</td>
<td>programming</td>
<td>development</td>
<td>opensource</td>
<td>software</td>
<td>web</td>
</tr>
<tr>
<td>39</td>
<td>opensource</td>
<td>software</td>
<td>linux</td>
<td>programming</td>
<td>tools</td>
<td>free</td>
</tr>
<tr>
<td>1152</td>
<td>tobuy</td>
<td>shopping</td>
<td>books</td>
<td>book</td>
<td>design</td>
<td>torread</td>
</tr>
</tbody>
</table>

First insights

- **Freq / FolkRank** show bias to high-frequency tags, i.e., to hyperonyms.
- **Cosine** seems to yield more synonymy and “siblings”.

→ Now: grounding of these observations in WordNet.

Semantic Grounding in WordNet

- **WordNet** is a large lexical database for English.
- **Words** with the same meaning are grouped in **synsets**, which are ordered by an *is-a* relation.
- **Introduction of single artificial root node** enables application of graph-based similarity metrics between pairs of nouns / pairs of verbs.
- **Inclusion of top n del.icio.us tags in WordNet:**
  - 100: 82%
  - 1,000: 79%
  - 5,000: 69%
  - 10,000: 61%

Shortest paths between original tag and most closely related one
Edge composition of shortest paths (for lengths 1 and 2)

Similar tags live on www.bibsonomy.org

Learning Ontologies from Folksonomies

Idea:
- automatically induce a concept hierarchy
- semantics of the relations resembles closely the one of taxonomic relations

Data:
- The tag-tag co-occurrence network of the delicious dataset forms the basis of the experiments (UTC = user-based tag-tag-co-occurrence, RTC = resource based tag-tag-co-occurrence)

Possible approaches:
- Social network analysis
- Set theoretic approaches (association rules, TRIAS)
- Statistical approaches (clustering, similarity measure)

Main steps of an Ontology Learning Algorithm

Filter the tags by an occurrence threshold

Order the tags in descending order by generality (measured by degree centrality in the UTC network)

Starting from the most general tag, add all tags subsequently to an evolving tree structure:
- identify the most similar existing tag
- (decide whether the tags are synonyms or form a compound expression and expand the tree accordingly)

Results for delicious (dataset 2005, 320 tags, used by > 2000 users)

Results for delicious together with similarity pruning
Conclusion

- Folksonomies overcome the knowledge acquisition bottleneck
  - due to ease of use
  - and therefore of fastly increasing amounts of users.

- Cosine measure seems most suitable to discover synonyms and siblings.

- Similarity measures can be used for Ontology Learning.

Try it yourself: www.bibsonomy.org