Wege zur Entdeckung von Communities in Folksonomies

Robert Jäschke, Andreas Hotho, Christoph Schmitz, Gerd Stumme

20. April 2006

Zusammenfassung

Ein wichtiger Baustein des neu entdeckten World Wide Web – des „Web 2.0“ – stellen Folksonomies dar. In diesen Systemen können Benutzer gemeinsam Ressourcen verwalten und mit Schlagwörtern versehen. Die dadurch entstehenden begrifflichen Strukturen stellen ein interessantes Forschungsfeld dar. Dieser Artikel untersucht Ansätze und Wege zur Entdeckung und Strukturierung von Nutzergruppen („Communities“) in Folksonomies.

1 Einleitung

Beginnend mit Blogs und Wikis sowie Techniken wie AJAX[1], RSS[2] oder REST[4] hat sich das World Wide Web in den letzten Jahren weiterentwickelt zum sogenannten „Web 2.0“. Ein wesentlicher Baustein dessen ist die soziale Komponente – die Benutzer können und sollen sich stärker am Inhalt des WWW beteiligen. Dieses Paradigma ist in Form von Folksonomies erfolgreich umgesetzt worden.

In Folksonomy-Systemen können Nutzer Ressourcen wie z. B. Bilder oder Web-Lesezeichen mit Schlagwörtern („Tags”) versehen. Die Ressourcen und Schlagwörter sind öffentlich und letztere dienen einerseits der Navigation im System, andererseits dem schnellen Wiederfinden der Ressourcen. Dabei ist es möglich, alle Benutzer zu sehen, die eine bestimmte Ressource getaggt oder ein bestimmtes Tag verwendet haben.

Weil innerhalb einer Folksonomy eine gute Vernetzung zwischen den Benutzern durch Tags und Ressourcen gegeben ist, ist ein spannendes Forschungsfeld, wie sich Gruppen von Nutzern („Communities“) in diesen Systemen beschreiben lassen, wie sie sich etablieren und ändern und wie man dieses Wissen nutzen kann, um den Benutzern Mehrwert zu bieten.

Im folgenden Abschnitt wird der Begriff Folksonomy erklärt und daraufhin vorhandene Ansätze zur Untersuchung dieser aufgezeigt. Teil 4 formuliert zentrale Fragen zur Entdeckung von Communities und Abschnitt 5 beschreibt eigene Forschung in dieser Richtung. Schließlich werden im Teil 6 weitere Bestandteile unseres Folksonomy-Systems BibSonomy[3] vorgestellt.

2 Folksonomies

Das Wort „Folksonomy” ist eine Zusammensetzung der Wörter „Folk” und „Taxonomy” und bezeichnet das gemeinsame Annotieren von Ressourcen im Web. Dabei schreiben Benutzer Schlagwörter an Ressourcen, wobei ein solcher Baustein (ein Benutzer, eine Ressource und die zugewiesenen Tags) als „Post” bezeichnet wird. Der Vorgang des Annotierens bringt den Nutzern ohne großen Aufwand einen Mehrwert, denn einerseits erleichtert er das Auffinden der eigenen Ressourcen, andererseits ist es dadurch leicht möglich, ähnliche neue Ressourcen zu finden, die von Interesse sein können. Im Gegensatz zu herkömmlichen Suchmaschinen funktioniert dies für Textinhalte gleichermaßen gut wie für Bilder, Videos oder andere nicht-textuelle Inhalte.

[1] `http://adaptivepath.com/publications/essays/archives/000385.php`
[2] `http://web.resource.org/rss/1.0/`
[3] `http://www.bibsonomy.org`

Ein allen diesen Systemen zu Grunde liegendes Modell beschreiben wir in [9]. Dieses formalisiert eine Folksonomy als tripartiten Graphen mit den Nutzern, Tags und Ressourcen als Knoten und einer Kante für jedes Tag, das ein Nutzer an eine Ressource geschrieben hat. Zusätzlich schlagen wir eine Benutzer-spezifische binäre Relation auf den Tags vor, welche eine Strukturierung der Tags erlaubt. Dies deckt sich einerseits mit Wünschen zahlreicher Nutzer (auch anderer Dienste, wie auf Mailinglisten ersichtlich) und ist in unserem BibSonomy-System umgesetzt. Andererseits sehen wir dies als einen ersten Schritt, die in Folksonomies nur schwach ausgebildete Ontologie zu erweitern.

Allgemein kann man Folksonomies als schwache Ontologien auffassen, wobei Tags durch Benutzer und Ressourcen miteinander verbundene Konzepte sind. Benutzer und Ressourcen können als Instanzen der Konzepte betrachtet werden. Folksonomies stellen ein spannendes neues Feld für die Forschung dar, denn erstmals erzeugt eine sehr große Zahl von Menschen gemeinsam eine schwache Form von Metadaten und annotiert Inhalte. Es gilt, dieses Potential zu nutzen, um ausgehend von einfachen Strukturen, den Aufbau des Semantic Web zu unterstützen. Ein besseres Verständnis der sozialen Struktur der Folksonomy kann dabei zum Stärken der begrifflichen Struktur beitragen.

3 Vorhandene Ansätze

Einen guten Überblick über Folksonomy-Systeme bieten [8] und [13], [14] analysiert Stärken und Schwächen dieser Systeme, [7] beschäftigt sich mit der Struktur – insbesondere von del.icio.us – und identifiziert sieben Arten von Tags. Die Visualisierung von Tags über einen gewissen Zeitraum steht im Vordergrund von [3], worin ausführlich Algorithmen für diese Aufgabe beschrieben werden. Eine verwandte Arbeit auf dem Gebiet der Blogs ist [17], worin Hauptkomponentenanalyse und Clustering-Verfahren auf einem FOAF-Netzwerk[4] angewendet werden. Insbesondere die zeitliche Änderung der Benutzer-Struktur steht dabei im Mittelpunkt der Betrachtung.

In [16] wird das „Actor-Concept-Instance“-Modell einer Folksonomy als tripartiter Graph modelliert, durch Projektion erzeugte 1-mode-Netzwerke mit typischen Graph-Maßen analysiert und Co-Occurence Techniken zum Clustern der Netzwerke angewendet. Dabei liegt der Schwerpunkt mehr auf der Extraktion schwacher Ontologien als auf dem Finden von Communities.

Das Programm Ontocopi [1] stellt einen Algorithmus zur Netzwerkanalyse von Ontologien bereit, um ein organisationales Gedächtnis erstmalig zu füllen. Dazu werden Methoden der Netzwerkanalyse auf eine bereits bestehende Ontologie angewandt, um innerhalb dieser wichtige Objekte zu erkennen. Insbesondere können durch ein dem PageRank [2] ähnliches Verfahren „Communities of Practice“ entdeckt werden.

Auf der Webseite der „Three-Mode-Company“[5] gibt es eine umfangreiche Literaturliste zu Analyseverfahren für 3-mode-Netzwerke, wobei zu prüfen ist, welche dieser Verfahren für die Untersuchung von Folksonomies anwendbar sind.

Für einige Folksonomy-Systeme existieren externe Systeme, die durch Analysemöglichkeiten Mehrwert für die Nutzer bereitstellen. Zu diesen sogenannten „Mashups” gehören Cloudalicious[6], welches die Struktur der Tags von del.icio.us graphisch aufbereitet, „Clusty”[7] mit einem ähnlichen Service sowie der Collaborative Rank [15] – ein System, um mit Hilfe der Folksonomy-Daten aus del.icio.us das Ranking für Suchanfragen zu verbessern. Des weiteren finden wichtige Diskussionen zu Folksonomies auf Mailinglisten[8] und Blogs statt.

[4] `http://www.foaf-project.org`

[5] `http://three-mode.leidenuniv.nl`

[6] `http://cloudalicio.us`

[7] `http://laurie.informatik.uni-bremen.de/clusty/`

[8] beispielsweise auf TagDB: `http://lists.tagschema.com/mailman/listinfo/tagdb` oder auch auf Connotea: `http://lists.sourceforge.net/lists/listinfo/connotea-discuss`

4 Zentrale Fragen

Neu und interessant an Folksonomies ist insbesondere, dass eine breite Masse von Menschen beginnt, Ontologien (wenn auch schwache) selbständig zu erstellen und zu pflegen, wobei jeder Nutzer dies bis zu einem gewissen Grade autonom tut. Wird dies auf eine große (Welt-)Ontologie hinauslaufen oder wird jeder Nutzer seine eigene kleine Ontologie besitzen? Gibt es etwas was dazwischen liegt – Ontologien beschrieben durch Communities? Werden sich Gruppen von Nutzern finden, die ein gemeinsames Vokabular benutzen, eine gemeinsame Ontologie bauen? Dem zugrunde liegt besonders die Frage: Wie ist die Struktur der Communities in einer Folksonomy?

Die Beschäftigung mit dieser Frage stellt einen wesentlichen Schritt zum Verständnis der begrifflichen Struktur einer Folksonomy dar. Möchte man die Community-Struktur einer Folksonomy erfassen, wirft dies weitere Fragen auf:

- Was ist eine Community und welche Arten von Communities gibt es (vergleiche Sorten von Tags in [7])?
- Welche Communities gibt es und wie kann man sie beschreiben?
- Zu welchen Communities gehört ein Benutzer?
- Wie entwickeln sich Communities über die Zeit?
- Wie lassen sich Communities in Folksonomies entdecken?

Zusammenfassend kann man diese reduzieren auf eine zentrale Frage: *Was macht eine Community in einer Folksonomy aus?*

Nach Beantwortung dieser Fragen stellt sich die nach dem Nutzen des gewonnenen Wissens. Aus Sicht der Benutzer stellen zielführende und nutzerspezifische Vorschläge eine Bereicherung des Systems dar, ebenso ein intelligentes Ranking von Ressourcen (oder auch Tags/Benutzern) oder eine erweiterte Darstellung der Folksonomy. Schließlich kann der weitere Ausbau der Ontologie damit beeinflusst werden, falls dies sinnvoll erscheint.

5 Eigene Vorarbeit

Zu den von uns bereits untersuchten Verfahren gehören der FolkRank, die triadische formale Begriffsanalyse sowie Assoziationsregeln.

Der FolkRank[9] ist eine Anpassung des PageRank [2] an den tripartiten Graphen einer Folksonomy und ermöglicht unter anderem die Erstellung von Ranglisten für Benutzer, Tags oder Ressourcen. Damit ist es insbesondere möglich, relevante Nutzer zu einem Tag zu finden oder für eine Menge von Nutzern die für sie relevanten Tags und Ressourcen zu entdecken. Daher ist der FolkRank eine Methode zur Entdeckung von Communities. Jedoch stellt auch bei diesem Algorithmus die enorme Datenmenge eines typischen Folksonomy-Systems ein Problem dar, da die Berechnungen zu aufwendig sind, um sie in Echtzeit durchführen zu können.

Mit der formalen Begriffsanalyse [6] in ihrer triadischen Form [12] existiert ein hierarchisches Clusterverfahren für tripartite Graphen, die dabei als triadische Kontexte aufgefasst werden. Mit Hilfe des Next Closure Algorithmus [5, 11] gelang es uns, Cluster von Benutzern zu berechnen, welche die gleiche Ressource mit den gleichen Tags versehen haben. Auf einem Folksonomy-Datensatz mit 3301 Nutzern, 30461 Tags, 220366 Ressourcen und 616819 Tripeln erhielten wir mehrere tausend Cluster. Besonders interessant erwiesen sich jene, deren Kantenlänge in jeder Dimension größer als Eins war. Da dieser Ansatz stets „exakte“ Cluster berechnet, bleibt zu untersuchen, wie das gewonnene Wissen interpretiert und genutzt werden kann.

Schließlich haben sich Assoziationsregeln [18] als mögliches Mittel zum Entdecken von Nutzerverhalten in Folksonomies erwiesen.

Oftmals problematisch ist die Größe der Daten (z. B. hat del.icio.us mehr als 300.000 Nutzer und mehrere Millionen Lesezeichen), denn viele der genannten Verfahren skalieren schlecht.

Daher stellt die Anwendbarkeit neuer Verfahren auch auf großen Datensätzen eine besondere Herausforderung dar, die Berücksichtigung finden muss.

6 Weitere Strukturen zur Unterstützung von Communities

Die bisher betrachteten Ansätze zur Analyse von Communities in Folksonomies stützen sich im wesentlichen auf die Grundstruktur des tripartiten Graphen. In einem typischen Folksonomy-System sind jedoch weitere Informationen vorhanden, die genutzt werden können.

Als erstes wäre das *Datum* eines Posts zu nennen, denn erst dieses ermöglicht die zeitliche Einordnung und damit das Erkennen von Änderungen der Community-Struktur. Des weiteren kann es auch zum Klassifizieren der Nützlichkeit von Ressourcen bzw. deren Annotation dienen, wie beim Collaborative Rank [15] umgesetzt.

Ein spezieller Punkt unseres BibSonomy-Systems sind *Gruppen*, die mehrere Nutzungsvarianten aufweisen. Einerseits als Menge von Anwendern (dem trivialen Fall einer formalen Community) und damit als aggregiertem Blick auf alle Posts einer Nutzergruppe. Des weiteren als Möglichkeit, nur für Gruppenmitglieder sichtbare Posts oder private Posts zu erstellen. Und schließlich die spezielle Gruppe „friends", die für jeden Benutzer existiert und von ihm verwaltet wird. Damit lässt sich eine Art FOAF-Netzwerk[9] innerhalb der Folksonomy abbilden und User können Posts nur für ihre Freunde sichtbar machen.

Das Verhalten der Menschen beim Benutzen des Systems stellt eine weitere Quelle zur Analyse dar. Methoden des Web Usage Mining [10] können dafür verwandt werden, wobei die Nutzung durch Logdateien des Webservers nachvollzogen werden kann. Zusätzlich enthält BibSonomy eigene Datenbanktabellen, die Änderungen an Einträgen protokollieren.

Des weiteren enthalten viele Systeme eine *Kopierfunktion*, mittels der sich vorhandene Einträge kopieren lassen. Die Auswertung dieser Aktion kann ebenfalls sinnvoll sein. Eine Spezialität unseres Systems ist der *Download*-Bereich, eine Art Warenkorb für Literaturreferenzen. Dort kann der Benutzer eigene und fremde Posts einsammeln und sich die Sammlung dann in verschiedenen Formaten ausgeben lassen. Der Inhalt des Warenkorbs verspricht eine interessante Quelle zur Community-Analyse zu sein, da sich damit Rückschlüsse auf den Nutzer interessierende Themen ziehen lassen.

7 Ausblick

Das Ziel der Beantwortung der gestellten Fragen und die Bereitstellung der Erkenntnisse zum Nutzen des Anwenders, verlangt neben der Untersuchung vorhandener Techniken und ihrer Adaption auf die Problemstellung insbesondere die Entwicklung neuer Verfahren zur effizienten Entdeckung von Communities in Folksonomies. Unsere existierende Forschung ist ein wichtiger Baustein, ebenso wie die in BibSonomy vorhandene Funktionalität, die den Entdeckungs- und Strukturierungsvorgang unterstützen kann. All dies trägt bei zum Verständnis von Folksonomies, insbesondere im größeren Kontext des „Web 2.0", in welchem diese eine wichtige Rolle spielen.

Literatur

[1] Harith Alani, Srinandan Dasmahapatra, Kieron O'Hara, and Nigel Shadbolt. Identifying Communities of Practice through Ontology Network Analysis. *IEEE Intelligent Systems*, 18(2):18–25, March/April 2003.

[2] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[9] `http://www.foaf-project.org`

[3] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing tags over time. In *Proceedings of the 15th International WWW Conference*, May 2006.

[4] Roy T. Fielding. *Architectural Styles and the Design of Network-based Software Architectures*. PhD thesis, University of California, Irvine, 2000.

[5] B. Ganter. Algorithmen zur Formalen Begriffsanalyse. In B. Ganter, R. Wille, and K. E. Wolff, editors, *Beiträge zur Begriffsanalyse*, pages 241–254. B.I. Wissenschaftsverlag, 1987.

[6] B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer, 1999.

[7] Scott Golder and Bernardo A. Huberman. The structure of collaborative tagging systems. Technical report, Information Dynamics Lab, HP Labs, Aug 2005.

[8] Tony Hammond, Timo Hannay, Ben Lund, and Joanna Scott. Social Bookmarking Tools (I): A General Review. *D-Lib Magazine*, 11(4), April 2005.

[9] Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Information retrieval in folksonomies: Search and ranking. In *Proceedings of the 3rd European Semantic Web Conference*, Lecture Notes in Computer Science. Springer, 2006. (to appear).

[10] R. Kosala and H. Blockeel. Web mining research: A survey. *SIGKDD Explorations*, 2(1):1–15, 2000.

[11] S. Krolak-Schwerdt, P. Orlik, and B. Ganter. TRIPAT: a model for analyzing three-mode binary data. In H. H. Bock, W. Lenski, and M. M. Richter, editors, *Studies in Classification, Data Analysis, and Knowledge Organization*, volume 4 of *Information systems and data analysis*, pages 298–307. Springer, Berlin, 1994.

[12] F. Lehmann and R. Wille. A triadic approach to formal concept analysis. In G. Ellis, R. Levinson, W. Rich, and J. F. Sowa, editors, *Conceptual structures: applications, implementation and theory*, volume 954 of *Lecture Notes in Artificial Intelligence*, pages 32–43. Springer Verlag, 1995.

[13] Ben Lund, Tony Hammond, Martin Flack, and Timo Hannay. Social Bookmarking Tools (II): A Case Study - Connotea. *D-Lib Magazine*, 11(4), April 2005.

[14] Adam Mathes. Folksonomies – Cooperative Classification and Communication Through Shared Metadata, December 2004. `http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html`.

[15] Amir Michail. CollaborativeRank: Motivating People to Give Helpful and Timely Ranking Suggestions. `http://collabrank.web.cse.unsw.edu.au/collabrank.pdf`, April 2005. (work in progress).

[16] Peter Mika. Ontologies are us: A unified model of social networks and semantics. In *Proceedings of the 4th International Semantic Web Conference*, volume 3729 of *Lecture Notes in Computer Science*, pages 522–536. Springer, 2005.

[17] John C. Paolillo, Sarah Mercure, and Elijah Wright. The social semantics of livejournal foaf: Structure and change from 2004 to 2005. In Gerd Stumme, Bettina Hoser, Christoph Schmitz, and Harith Alani, editors, *Proceedings of the 1st Workshop on Semantic Network Analysis at the ISWC 2005 Conference*, pages 69 – 80, Galway, Ireland, November 2005.

[18] Christoph Schmitz, Andreas Hotho, Robert Jäschke, and Gerd Stumme. Mining association rules in folksonomies. In *Proceedings of the IFCS 2006 Conference*, Lecture Notes in Computer Science. Springer, 2006. (to appear).