

A.1 Tagging als soziales Bindeglied für Communities

Walter Christian Kammergruber

Technische Universität München

Manfred Langen

*Siemens AG, Corporate Technology, Fachzentrum für
Wissensmanagement*

Social Tagging und soziale Netzwerke sind zentrale Bausteine des Web 2.0 und Enterprise 2.0. In diesem Beitrag werden die sozialen Aspekte von Social Tagging beleuchtet und ein Ansatz aufgeführt, um in Folksonomies Personen mit ähnlichen Interessen zu finden. Ferner wird ein Tagging-Framework beschrieben, das im Use Case Alexandria im Rahmen des BMWi-Projekts Theseus¹ entstanden ist.

1 Einleitung

Social Tagging erfreut sich seit einigen Jahren vor allem im Umfeld von Web 2.0 Applikationen enormer Beliebtheit, wenn es darum geht, einfach und nutzerzentriert Artefakte zu annotieren. Einer der Pioniere in diesem Kontext ist Joshua Schachter, der Gründer von Delicious², einem Social Bookmarking Service, welcher später von Yahoo! übernommen wurde.

Clay Shirky, Professor an der New York University, wird vom Guardian folgendermaßen zitiert:

*„Clay Shirky [...] studied tagging and advised Delicious. He describes Schachter as, the first person to figure out the **social value** of labeling. Any one person's labels are **messy, inconsistent and partial**, and are therefore much less valuable than formal classification systems. However, if there is a way to **aggregate** those*

¹ <http://theseus-programm.de/home/>

² <http://delicious.com/>

*labels, and therefore their value, they become more valuable than formal systems, because they are **robust, socially accurate and cheap.*** ' " [13]

In dieser Beschreibung stecken komprimiert einige wesentliche Aussagen. Tags sind chaotisch, inkonsistent und subjektiv. Das bedeutet, dass sie im Vergleich zu kontrollierten Vokabularen (z.B. Glossaren, Thesauri, Ontologien) einen erheblichen Nachteil haben, und zwar ein Mangel an Struktur und Konsistenz.

Allerdings haben Tags einen oft unterschätzten sozialen Wert, der sich ergibt, sobald die Tags geeignet aggregiert werden. Glückt diese Aggregation, so sind Tags in Summe von größerem Wert als formelle Annotationssysteme. Tags spiegeln nämlich die Sprache der Nutzer wider und sind in der Masse robust (es ergibt sich zumeist eine Verteilung, die dem Potenzgesetz folgt – vgl. dazu [2]).

Ein weiteres Problem kontrollierter Vokabulare, die Anpassung an die Dynamik des Sprachgebrauchs, wird gelöst. Die gelebte Sprache ist einem stetigen Wandel ausgesetzt. Es entstehen und veralten Termini, bzw. Wortgebräuche (z.B. „Web 2.0“ taucht erst seit der betreffenden O'Reilly Konferenz als Wendung auf und wird sicherlich die nächsten Jahre von einem anderen Modewort abgelöst) oder auch so manche Gruppe von Personen hat ihre eigene Sprache. Man denke nur an kontextspezifische Abkürzungen, wie z.B. AI steht unter anderem für Activity Item, Artificial Intelligence oder auch Angewandte Informatik³. Der Wandel der Sprache stellt generell ein beachtliches Problem für kontrollierte Vokabulare dar (vgl. speziell für Thesauri z.B. [14]).

Die Einfachheit und die Tatsache, dass Tags kostengünstig erstellt werden, ist ein Grund für die breite Akzeptanz von Tagging zum Annotieren und somit Kategorisieren von Objekten. Zu dem bereits erwähnten Delicious kann als weiteres Parade-Beispiel Flickr⁴ dienen, eine beliebte Community-Plattform um Fotos und mittlerweile auch (kurze) Filme zu veröffentlichen und zu diskutieren. Aufzuführen sind aber auch andere Dienste wie Librarything⁵, eine Buch-

³ <http://de.wikipedia.org/wiki/AI>

⁴ <http://www.flickr.com/>

⁵ <http://www.librarything.com/>

Community, Last.fm⁶, ein Musik-Portal, oder auch 43Things⁷, eine Gemeinschaft, die sich um Lebensziele dreht.

2 Tagging als Web 2.0 Phänomen

Im Grunde kann Tagging für alles eingesetzt werden, was im philosophischen Sinne existiert. Tagging funktioniert selbst in den Fällen, wo automatisierte Verfahren, die typischerweise bei textbasierten Dokumenten (Bücher, Webseiten, Berichte, etc.) ihre Anwendung finden, an ihre Grenzen stoßen.

Tags erfüllen verschiedenste Rollen bei der Organisation von Objekten. Golder et al. [6] machen sieben gebräuchliche Muster aus, in der Tags eingesetzt werden – angefangen mit Tags, die das entsprechende Objekt beschreiben, bis hinzu Tags, die zur Selbstorganisation dienen, z.B. „toDo“ oder „toRead“.

Für welche Zwecke Tags eingesetzt werden beschreiben Marlow et al. [12]. Diese Zwecke beschränken sich nicht nur auf persönliche, private Aspekte – man tagged ein Objekt, damit man es später wiederfindet – sondern es treten auch soziale, interaktive Tendenzen auf, z.B. bei Flickr werden von manchen Nutzern Fotos getagged, damit Freunde oder Kollegen sie besser finden können.

An der großen Verbreitung und Beliebtheit kann zum einen die Benutzerfreundlichkeit und Bedienbarkeit erkannt werden, aber auch der Unterschied zwischen Tagging und *Social Tagging*: Der Nutzer ist ein zentrales Element.

Hendler et al. [7] schreiben in einem Artikel mit dem Titel: „Metcalf's law, Web 2.0, and the Semantic Web“ u.a. über die Netzwerk-Effekte von Web 2.0 Anwendungen, insbesondere Tagging als eine Kerntechnologie dieses Phänomens. In dem Artikel beklagen sie den Mangel an Struktur, den das Tagging in Web 2.0 Anwendungen mit sich bringt und vertreten die These, dass dieser durch Semantic Web Technologien möglicherweise beseitigt oder zumindest abgeschwächt werden kann. Gleichwohl erörtern sie die sozialen Aspekte vom Web 2.0 und das Potential, das darin verborgen liegt. Metcalf's Law folgend argumentieren Hendler

⁶ <http://www.last.fm/>

⁷ <http://www.43things.com/>

et al., dass der Nutzen eines Netzwerks exponentiell zur Anzahl der teilnehmenden Nutzer steigt. Allerdings kann die Kommunikation zwischen Benutzern nur in begrenztem Maß erfolgen. Es kann sich nicht jeder mit jedem beschäftigen. Zeit ist bekanntlich ein endlicher Faktor und allein darin liegt eine Einschränkung für die Aufmerksamkeit, die man generell Angelegenheiten – und Personen insbesondere – schenken kann. Folglich ist es von besonderem Nutzen, algorithmische Verfahren zur Verfügung zu haben, um interessante Artefakte und Personen zu finden und potentielle Communities aufzudecken. Tags bieten sich als Bindeglied zwischen einzelnen Benutzern an.

In diesem Artikel fließen Ergebnisse aus dem Alexandria Use Case des BMWi-Projekts Theseus ein. Im Rahmen von Alexandria werden verschiedene Ansätze erdacht und erprobt, um aus den zunächst losen und unstrukturierten Tags weitaus wertvollere Strukturen zu generieren und zusätzliches „Wissen“ zu extrahieren. Zusätzlich zu den Relationen zwischen Tags spielen Nutzer und Strukturen zwischen Nutzern (sei es mögliche oder tatsächliche) eine entscheidende Rolle. In diesem Beitrag soll auf das Finden von ähnlichen Nutzern und somit das Clustern von Nutzergruppen näher eingegangen werden.

3 Tagging und Soziale Netzwerke

Die Interessen eines Benutzers können potentiell über seine Tags beschrieben werden. Die häufige Verwendung bestimmter Tags lässt darauf schließen, dass der Benutzer ein Experte auf dem Gebiet ist oder sich zumindest mit einer bestimmten Thematik intensiv beschäftigt (hat). Tags können dazu herangezogen werden, eine gewisse Form von *implizitem* Benutzerprofil zu erstellen. Diese Benutzerprofile können dazu benutzt werden, um für den jeweiligen Nutzer passend zu seinen Tags interessante Artefakte, also z.B. Blogposts, Wiki-Artikel oder Bookmarks, vorzuschlagen. Im Sinne von Social Networking ist eine weitere Verwendung der Tagprofile bedeutsam: Es können Personen mit ähnlichen Interessen aufgezeigt werden. Insbesondere bei global agierenden Firmen mit weltweit verteilten Mitarbeitern ist dies ein wichtiger Aspekt.

Teilweise beschäftigen sich Kollegen mit ähnlichen Problemen, wissen aber nicht voneinander, so dass potentielle Synergien nicht genutzt werden. Bisher waren

z.B. sogenannte Yellow Pages ein Mittel der Wahl. Allerdings sind diese aufwändig bei der Einführung und werden selten aktuell gehalten. Tags stellen eine kostengünstige Alternative dar, um Nutzer-Profile zu generieren.

Mit der wachsenden Verbreitung von Social-Software im Enterprise-Umfeld entstehen immer wieder neue Web 2.0 Applikationen als Quelle für Tagging-Daten. Das damit verbundene Potential zur Unterstützung einer firmeninternen Vernetzung gilt es zu nutzen. Nachfolgend wird genauer darauf eingegangen, wie mittels Tags ähnlich interessierte Nutzer ermittelt werden können.

4 Tag-Aggregation und Verarbeitung

Innerhalb des bereits erwähnten Alexandria Use Case des Theseus Projekts wird prototypisch ein Tagging-Framework zum Aggregieren von Tag-Daten aus verschiedenen Applikationen entwickelt.

Die Tags einer Applikation für sich genommen stellen bereits eine Folksonomie (zu dem Begriff siehe [15]) dar. Um allerdings die Grenzen zwischen Applikationen zu überwinden, und damit einen Mehrwert auf Unternehmensebene zu schaffen, werden beim Tagging-Framework Folksonomien aus verschiedenen Anwendungen aggregiert.

Zurzeit werden die Tag-Daten aus der *Siemens Blogosphere* (vgl. [3]), der *Siemens Wikisphere* (vgl. [11], [9]) und einem Projekt-Management-Tool exportiert und in einem RDF-Repository gehalten. RDF wurde aus technischer Sicht deshalb gewählt, da es sich um ein äußerst flexibles Format handelt, um Graphen, bzw. semi-strukturierte Daten zu modellieren. Auf eine zusätzliche explizite Modellierung mittels z.B. OWL wird (derzeit) verzichtet, da keine Reasoning Methoden eingesetzt werden.

Langfristig sollen sich die Vorteile von Semantic Web Technologien herauskristallisieren, ähnlich wie sie von der Linking Open Data Community⁸

8

vertreten werden, z.B. dass Daten offen (zumindest firmenintern und unter notwendigen Einschränkungen) zugänglich sind.

Bereits jetzt sind verschiedene Datenquellen über Standard-Internet-Protokolle erreichbar (z.B. Blog-Posts, Wiki-Seiten, Sharepoint, etc.). Allerdings zumeist (noch) nicht in einem standardisiertem maschinen-interpretierbaren Format. Die Zukunft wird zeigen, welche Protokolle/ Datenformate sich in diesem Bereich etablieren. Für die Erprobung von verschiedenen Ansätzen wurden für die jeweiligen Applikationen entsprechende Plugins, respektive Exporter geschrieben.

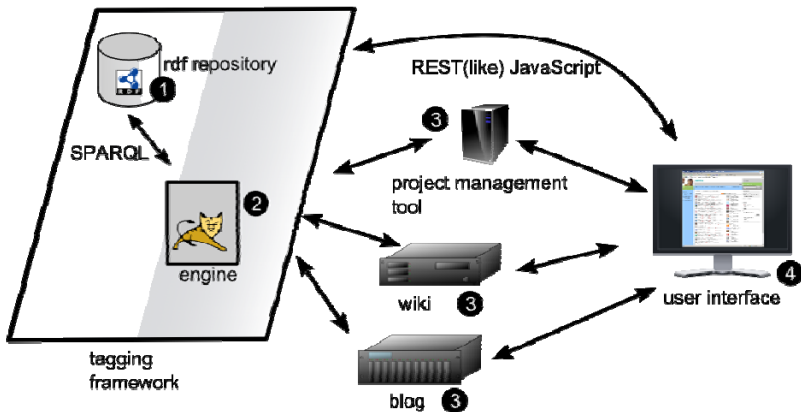


Abbildung 1: Allgemeine Architektur des Tagging Frameworks.

Abbildung 1 zeigt eine schematische Architekturbeschreibung unseres Tagging-Frameworks. Die Daten werden aus den verschiedenen Anwendungen (3) je nach Möglichkeit über Push- oder Pull-Mechanismen im Tagging-Framework (1+2) gesammelt. Dieses stellt Services für externe Applikationen zur Verfügung, mit denen z.B. dem Benutzer Tag-Vorschläge beim Anlegen von zu taggenden Objekten unterbreitet werden. Zugleich werden kleinere Widgets (Webelemente, die z.B. über so genannte IFrames in die jeweiligen Oberflächen mit eingebunden werden) angeboten (4). Ein exemplarisches Widget ist eine Tag-Cloud, die Tags über alle angeschlossenen Anwendungen hinweg bezogen auf einen Nutzer darstellt.

Ein weiteres Beispiel ist ein Such-Widget. Mit diesen kann z.B. nach Artefakten zu einem Tag gesucht werden, aber auch nach Personen, die ein gewisses Tag verwendet haben. Weitere Funktionalitäten werden sukzessive ausgebaut.

Ein beträchtlicher Vorteil des Tagging-Frameworks liegt darin, dass die einzelnen angeschlossenen Applikationen nicht jeweils selbst eigene Implementierungen erstellen müssen. Eine service-orientierte Implementierung ermöglicht direkte Anfragen an das Tagging-Framework. Als Rückgabewert werden verschiedene Formate, wie JSON oder RDF, angeboten. Zusätzlich wird eine grafische Darstellung mittels entsprechender Widgets unterstützt.

5 Finden von ähnlichen Benutzern

Kammergruber et al. [8] beschreiben in einem Paper, wie sich über die Verwendung von Tags Benutzer mit korrespondierenden Interessen finden lassen. Im Folgenden wird der Ansatz kurz beschrieben. Für Details sei auf den Artikel verwiesen.

Um die Ähnlichkeit von zwei Benutzerprofilen zu bestimmen, wird die Cosinus-Distanz verwendet, was dem Winkel zwischen zwei Vektoren entspricht.

Die Cosinus-Distanz, bzw. umgekehrt die Ähnlichkeit, zwischen zwei Vektoren v_1 und v_2 wird folgendermaßen berechnet:

$$\cos(v_1, v_2) = \frac{v_1 \cdot v_2}{\|v_1\| \|v_2\|}$$

Die Cosinus-Distanz kann Werte zwischen -1 und 1 annehmen. -1 steht für genau gegenteilig, 0 für unabhängig und 1 exakt gleich.

Zwei Nutzern $User_1$ und $User_2$ ist eine Menge an Tags respektive der Häufigkeit ihrer Verwendung zugeordnet. Somit kann die Ähnlichkeit von zwei Benutzerprofilen über die Cosinus-Distanz zweier individueller Tag-Häufigkeitsvektoren ermittelt werden.

Ein einfaches Beispiel: Angenommen man hat für $User_1$ die Tags $t_1 = [\text{blog: 5, km: 2, rss: 10}]$ und für $User_2$ die Tags $t_2 = [\text{ajax: 10, blog: 3, rss: 5}]$. Dann werden wechselseitig für jeden Nutzer die Tags, die der andere Nutzer verwendet hat, er

selber aber nicht mit Häufigkeit Null aufgefüllt und es resultieren die beiden Vektoren $t_1 = (0, 5, 2, 10)$ und $t_2 = (10, 3, 0, 5)$.

Für das Beispiel beträgt die Cosinus-Distanz zwischen den Tag-Vektoren der beiden Nutzer ca. 0,5. Da die Häufigkeiten der Tags immer positive Werte annehmen, ergeben sich für die Cosinus-Ähnlichkeit in diesem Kontext Werte zwischen 0 und 1. Der Wert von 0,5 aus dem Beispiel entspricht einer gewissen Übereinstimmung, aber man sieht, dass z.B. die Häufigkeit von „ajax“ beim ersten Nutzer hoch ist und beim zweiten gar nicht auftaucht.

Dieses Ähnlichkeitsmaß kann dazu eingesetzt werden, um eine sortierte Liste von ähnlichen Benutzern zu einem bestimmten Nutzer zu erzeugen. Aus dieser Liste wird für einen Nutzer ersichtlich, welche Personen sich mit ähnlichen Themen beschäftigen. In diesem Zusammenhang wird auch der Begriff „social search“ verwendet.

Eine weitere Anwendung liegt darin, ausgehend von dem Ähnlichkeitsmaß ein Clustering von Personen durchzuführen. Dies erfolgt mit geeigneten Clustering-Algorithmen, wie z.B. DBSCAN (siehe [4]) oder auch agglomerative hierarchische Verfahren.

Im anfangs zitierten Paper wurde beispielsweise mittels DBSCAN ein Cluster von Personen, die sich mit Photographie beschäftigen, gefunden. Diese Cluster stellen entweder bereits existierende oder potentielle Communities dar. Es ergibt sich somit die Möglichkeit die Bildung von neuen Gruppen mit ähnlichen Interessen anzustoßen.

6 Andere Ansätze

Die (semi-)automatische Extraktion von Mustern in Folksonomies hat in den letzten Jahren viele wissenschaftliche Arbeiten geprägt. Zumeist wird dabei versucht, semantische Relationen zwischen Tags zu entdecken – z.B. hierarchische Beziehungen, wie es bei „Mensch“ als Unterbegriff zu „Säugetier“ der Fall ist. Es sind dabei mannigfaltige Relationen denkbar. Allerdings beschränken sich die meisten Ansätze auf die Standard-Beziehungen, welche in der Thesaurus-Norm DIN 1463-1 bzw. dem internationalen Äquivalent ISO 2788 beschrieben werden.

Das zumeist gewählte Mittel um Relationen zwischen Tags statistisch zu bestimmen, ist die Ausnützung der Co-Occurrence-Beziehung (z.B. bei [1], [5]). Eine Co-Occurrence-Beziehung zwischen zwei Tags bedeutet, dass Benutzer beide Tags zugleich für dieselbe Ressource verwendet haben.

Li et al. [10] haben ein so genanntes Internet Social Interest Discovery System (ISID) entwickelt, welches darauf ausgerichtet ist, Nutzer mit gemeinsamen Interessen zu finden. Sie verfolgen einen zweistufigen Ansatz. Zunächst setzen sie Frequent Itemset Mining ein (welches normalerweise beim Association Rule Mining eingesetzt wird), um häufige Tag-Pattern zu finden, also Tags, die häufig miteinander vorkommen. In einem zweiten Schritt versuchen sie Benutzer zu finden, die diese Pattern verwendet haben. Die Nutzer-Mengen stellen die Cluster von Personen mit ähnlichen Interessen dar. Erprobt wurde der Ansatz an einem exzessiven Auszug des Delicious Bookmarking Service.

Zarnadi et al. [16] benutzen ein Cosinus-Distanzmaß angewandt auf Tag-, Nutzer- oder Ressourcen-Vektoren, um ein Ranking für Empfehlungen in einem Recommender-System zu berechnen.

7 Ausblick

Das Tagging-Framework wurde mit realen Daten aus Social-Software-Anwendungen im Forschungsprojekt Theseus Alexandria erprobt. Im nächsten Schritt werden Services aus dem Tagging-Framework in interne, global im Intranet erreichbare Knowledge-Management Tools der *Siemens AG* integriert, insbesondere (aber nicht ausschließlich) der Wikisphere (einer corporate Wiki-Applikation) und Blogosphere (einer corporate Blog-Plattform).

Von weiterem Interesse bezogen auf das Finden von Gruppen ähnlicher Benutzer ist die Berücksichtigung von (Thesauri-)Relationen zwischen Tags. Diese Relationen zwischen den Tags stammen dann entweder vom Nutzer selbst – jemand legt beispielsweise fest, dass „knowledge mangement“ ein Unterbegriff von „knowledge“ ist – oder aus externen Quellen, wie beispielsweise WordNet (bei englischen Begriffen), DBpedia (indem z.B. Kategorieinformationen genutzt werden) oder auch anderen Quellen mit strukturierten Informationen.

Literatur

- [1] Grigory Begelman, Philipp Keller und Frank Smadja. Automated tag clustering: Improving search and exploration in the tag space. In *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland*, 2006.
- [2] Ciro Cattuto, Andrea Baldassarri, Vito D. P. Servedio und Vittorio Loreto. Vocabulary growth in collaborative tagging systems, Apr 2007.
- [3] Karsten Ehms. *Globale Mitarbeiter-Weblogs bei der Siemens AG.*, S. 199–209. Oldenbourg, München, 2008.
- [4] Martin Ester, Hans-Peter Kriegel, Jörg Sander und Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proc. of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, S. 226–231, 1996.
- [5] Eirini Giannakidou, Vassiliki Koutsonikola, Athena Vakali und Yiannis Kompatsiaris. Co-clustering tags and social data sources. In *WAIM '08: Proceedings of the 2008 The Ninth International Conference on Web-Age Information Management*, S. 317–324, Washington, DC, USA, 2008. IEEE Computer Society.
- [6] Scott A. Golder und Bernardo A. Huberman. Usage patterns of collaborative tagging systems. *J. Inf. Sci.*, 32(2):198–208, 2006.
- [7] J. Hendler und J. Golbeck. Metcalfe's law, Web 2.0 und the Semantic Web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(1):14–20, 2008.
- [8] Walter Christian Kammergruber, Maximilian Viermetz und Cai-Nicolas Ziegler. Discovering communities of interest in a tagged on-line environment. In *CASoN2009: Proceedings of the 1st International Conference on Computational Aspects of Social Networks*, 2009.
- [9] Manfred Langen und Karsten Ehms. Social Software als Ansatz für dezentrales Wissensmanagement im Unternehmen. In *Virtuelle Organisation und Neue Medien 2006*, S. 75–83, 2006.
- [10] Xin Li, Lei Guo und Yihong E. Zhao. Tag-based social interest discovery. In *Proceedings of the 17th International World Wide Web Conference*, S. 675–684. ACM, 2008.

-
- [11] Bernd Lindner. Der Einsatz von Wikis in der Siemens AG. I-KNOW, 2008.
- [12] Cameron Marlow, Mor Naaman, Danah Boyd und Marc Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HYPertext '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, S. 31–40, New York, NY, USA, 2006. ACM Press.
- [13] Quinn Norton. 'I want to build something that grows'. <http://www.guardian.co.uk/media/2006/jan/26/newmedia.technology1>, 2006.
- [14] Jiri Panyr. *Automatische Klassifikation und Information Retrieval*. Niemeyer Max Verlag GmbH, 1986.
- [15] Thomas Vander Wal. Folksonomy. Folksonomy Coinage and Definition, 2007. <http://vanderwal.net/folksonomy.html>.
- [16] Valentina Zanardi und Licia Capra. Social Ranking: Uncovering Relevant Content Using Tag-based Recommender Systems. In *RecSys '08: Proceedings of the 2008 ACM conference on Recommender systems*, S. 51–58, New York, NY, USA, 2008. ACM.