Bringing Semantics into Folksonomies – Semantische Analyse nutzergenerierter Indexierungen

Steffen Lohmann, Jürgen Ziegler

Abteilung für Informatik und Angewandte Kognitionswissenschaft
Universität Duisburg-Essen
Lotharstrasse 65
47057 Duisburg
{lohmann, ziegler}@interactivesystems.info

Abstract: Die zunehmende Popularität des gemeinschaftlichen Indexierens (Social Tagging) führt zu umfangreichen Sammlungen an nutzergenerierten Metadaten. Eine automatisierte Verwertung und interoperable Nutzung dieser Daten ist jedoch schwierig, da sie nur eine geringe formale Semantik aufweisen. In diesem Beitrag wird ein konzeptioneller Rahmen für die Abbildung von Tags auf Ontologiekonzepte aufgestellt und anhand eines Beispielszenarios veranschaulicht.

1 Motivation

Die nutzerseitige Verschlagwortung von Inhalten durch frei gewählte Bezeichner (sog. Tagging) hat sich in letzter Zeit zu einer populären Form der Indexierung entwickelt. In „sozialen Umgebungen" entstehen durch die Verknüpfung der Tagging-Daten mehrerer Nutzer umfangreiche Sammlungen von Metadaten (sog. Folksonomies [Va07]), die sich effizient zur Strukturierung und Exploration von Inhalten einsetzen lassen [Ma06, GH06]. Da beliebig gewählte Tags im Gegensatz zu Deskriptoren eines kontrollierten Vokabulars in der Regel keine eindeutige oder vereinbarte Semantik besitzen, gestaltet sich die automatisierte Verwertung und interoperable Nutzung von Tagging-Daten schwierig. Probleme bereiten Polysemie und Synonymie, aber auch Tipp- und Rechtschreibfehler sowie der Gebrauch verschiedener Flexionsformen.

In diesem Beitrag wird die Abbildung von Tags auf Ontologiekonzepte systematisiert. Die Tagging-Daten sollen im Nachhinein automatisiert mit Semantik angereichert werden. Damit unterscheidet sich dieses Vorgehen von Ansätzen wie „Semantic Collaborative Tagging" [Ma07] oder „Machine Tags" [Fl07], die semantische Informationen durch ergänzende Nutzereingaben bereits während des Prozesses der Verschlagwortung zu erfassen versuchen. Erfahrungen haben gezeigt, dass der durchschnittliche Nutzer solche zusätzlichen Angaben ungern tätigt, insbesondere, wenn sich hierdurch nicht unmittelbar eine Qualitätsverbesserung der persönlichen Indexierung beobachten lässt. Im Folgenden wird anhand eines Beispielszenarios die Normalisierung und Kontextualisierung von Tagging-Daten thematisiert. Anschließend werden die Vorteile und Grenzen dieses Vorgehens diskutiert und ein Ausblick auf zukünftige Entwicklungen in diesem Bereich gegeben.

2 Ausgangssituation

Die Ausgangsbasis des hier betrachteten Vorgehens bildet die elementarste Informationseinheit eines Systems zum gemeinschaftlichen Tagging, ein Tripel der Form `(Ressource, Nutzer, {Tags})`, wobei `Ressource` die indexierte Informationsressource, `Nutzer` den indexierenden Nutzer und `{Tags}` die bei der Indexierung verwendeten Tags bezeichnet.

Als exemplarisches Szenario soll angenommen werden, ein solcher Tagging-Datensatz wurde in einer Social Bookmarking-Anwendung [Ha05] erstellt, indem Nutzer A die Startseite der Webpräsenz des Welterbes Zeche Zollverein in Essen – wie in Abbildung 1a dargestellt – mit beliebig gewählten Tags zu seiner persönlichen Bookmark-Sammlung hinzugefügt hat.

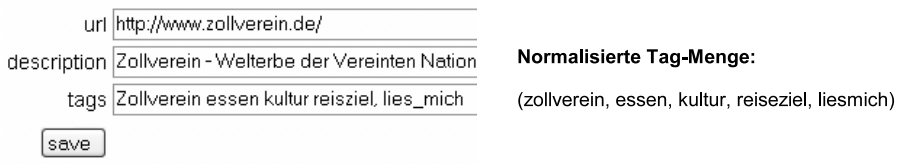

Abbildung 1: (a) Tagging einer Webseite; (b) Tag-Menge nach Normalisierung.

3 Normalisierung

Trotz der zunehmenden Etablierung von Tagging-Konventionen ist unpräzises, flüchtiges Tagging eher Regel als Ausnahme [GT06]. Da gewöhnliche Nutzer keine Indexierungsexperten sind, lassen sich zudem über einen kleinen Satz an Konventionen hinausgehende Tagging-Regeln nicht realisieren. Die von den Nutzern eingegebenen Tags werden deshalb in vielen Anwendungen vor ihrer weiteren Verarbeitung zu einem gewissen Grad normalisiert. Üblich ist eine Konvertierung in Kleinbuchstaben; mögliche weitere Verfahren sind die Entfernung von Satz- und Sonderzeichen, die Dekomposition, verschiedene Varianten der Grundformreduktion (z.B. Lemmatisierung) sowie die Korrektur von Tipp- und Rechtschreibfehlern. Abbildung 1b zeigt die beispielhaft normalisierte Tag-Menge aus dem Szenario.

In Tagging-Systemen wird im Allgemeinen eine defensive Normalisierungsstrategie verfolgt, da Nutzer individuelle Indexierungsstile pflegen, bei denen spezifische Schreibweisen und Sonderzeichen eine wichtige Bedeutung einnehmen können. Für die Abbildung von Tags auf Ontologiekonzepte ist hingegen eine möglichst weitgehende Normalisierung erforderlich, um ein exakteres Konzept-Matching zu erzielen. Das Tag „Reisziel“ aus dem oben beschriebenen Szenario lässt sich beispielsweise erst nach Korrektur des Tippfehlers einem entsprechenden Ontologiekonzept zuordnen.

4 Kontextualisierung

Im Fall von monosemen Tags ist eine Abbildung auf Ontologiekonzepte bereits im Anschluss an die Normalisierung möglich. Nach Korrektur des Tippfehlers sind alternative Auslegungen der Semantik von „Reiseziel“ vernachlässigbar. Für polyseme Tags hingegen ist zusätzlich eine Analyse des Kontextes notwendig. Hierbei können alle drei Elemente des Tripels `(Ressource, Nutzer, {Tags})` als Kontextualisierungsdimensionen dienen.

Bei der Kontextualisierung werden Kookurrenzen zwischen Tags analysiert und mit existierendem Kollokationswissen abgeglichen. Statistisch ermittelte Kollokationswerte sind für diesen Zweck nicht ausreichend, da sich hierüber die Semantik von Tags nicht eindeutig identifizieren lässt. Abbildung 2a zeigt den aus dem deutschen Wortschatz der Universität Leipzig [HQW06] statistisch ermittelten Kollokationsgraphen für den Begriff „Zollverein“; Abbildung 2b zeigt einen Ausschnitt aus einer beispielhaften Domänenontologie zum Thema Ruhrgebiet. In beiden Graphen existiert zwischen den Knoten „Zollverein“ und „Essen“ eine semantische Beziehung, jedoch ist nur in den Ontologiekonzepten explizit repräsentiert, welche Bedeutungen von „Zollverein“ und „Essen“ in diesem Zusammenhang gemeint sind.

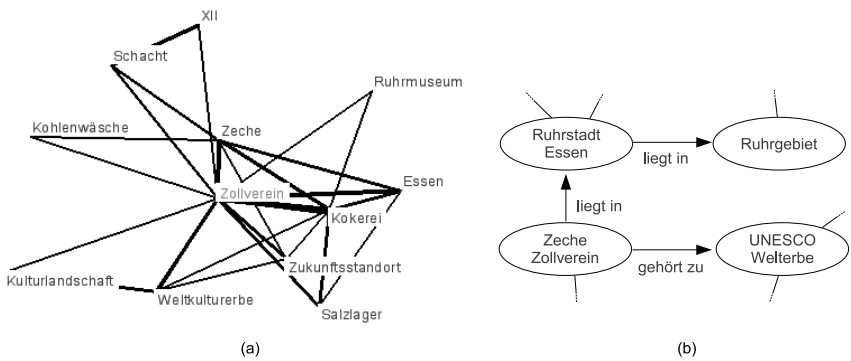

Abbildung 2: (a) Statistisch ermittelter Kollokationsgraph; (b) Ontologieausschnitt.

4.1 Nutzerbezogene Kollokationsanalyse

In den meisten Fällen indexiert ein Nutzer eine Ressource mit mehreren Tags [MC07]. Kollokationen innerhalb dieser Tag-Menge können erste Aufschlüsse über die Semantik einzelner Tags geben. Im Beispielszenario wird durch Abgleich mit der Ruhrgebietsontologie zwischen den Tags „Zollverein“ und „Essen“ eine semantische Beziehung ermittelt, die darauf hindeutet, dass mit dem Tag „Zollverein“ das Konzept „Zeche Zollverein“ und mit dem Tag „Essen“ das Konzept „Ruhrstadt Essen“ gemeint ist. In einer Gegenprobe könnten die alternativen Konzepte „Deutscher Zollverein“ und „Essen (Nahrungsaufnahme)“ durch den Abgleich mit entsprechenden Ontologien auf semantische Relationen überprüft werden. Da hierbei vermutlich keinerlei Beziehungen ermittelt werden, bekräftigt das die genannte semantische Ausrichtung der Tags.

4.2 Nutzerübergreifende Kollokationsanalyse

Das Grundprinzip der gemeinschaftlichen Indexierung ist, dass mehrere Nutzer dieselben Ressourcen mit Tags auszeichnen. Hierüber ergibt sich eine weitere Dimension für die Kollokationsanalyse. Angenommen, Nutzer A hätte das Tag „Zollverein“ nicht verwendet, so dass die Ermittlung semantischer Relationen für das Tag „Essen“ über die nutzerbezogene Kollokationsanalyse zu keinem Ergebnis geführt hätte. Dafür hat jedoch Nutzer B die Ressource mit den Tags „Zollverein“ und „Ruhrgebiet“ versehen. In diesem Fall ließen sich semantische Relationen zwischen den Tags „Essen“ (Nutzer A) und „Zollverein“ sowie „Ruhrgebiet“ (beide Nutzer B) identifizieren – dies sind Indikatoren für die Abbildung des Tags „Essen“ auf das Ontologiekonzept „Ruhrstadt Essen“. Allgemein erhält eine semantische Relation umso größeres Gewicht, je häufiger sie nutzerbezogen oder nutzerübergreifend identifiziert wird.

4.3 Ressourcenbezogene Kollokationsanalyse

Bei der freien Indexierung von Texten übernehmen die Nutzer häufig Schlüsselbegriffe aus dem Text direkt als Schlagworte. Dadurch liefert auch die Inhaltsdimension einen Kontext für die Kollokationsanalyse, sofern die indexierten Ressourcen textbasiert sind und verarbeitet werden können. Steht beispielsweise im Quelltext der indexierten Webseite der Begriff „Zollverein“ im Zusammenhang mit dem Begriff „Welterbe“ (s. Abbildung 1), deutet dies auf eine Interpretation des Tags „Zollverein“ in Richtung des Ontologiekonzepts „Zeche Zollverein“ hin.

5 Diskussion und Ausblick

Die Normalisierung und Kontextualisierung anhand der aufgezeigten Dimensionen ermöglicht die Abbildung von Tagging-Daten auf Ontologiekonzepte und bildet damit die Voraussetzung für semantische Interoperabilität. Die Effektivität dieses Vorgehens ist jedoch insbesondere durch die Verfügbarkeit entsprechender Ontologien determiniert. Für den breiten Einsatz ist die notwendige ontologische Grundlage derzeit nicht gegeben, doch versprechen aktuelle Fortschritte in Bereichen wie dem Semantic Web und Ontological Engineering [TB06, ES07] und insbesondere die Verknüpfung von ontologischen und lexikalischen Ansätzen [GNV03] in Zukunft eine umfassendere Verfügbarkeit von allgemein zugänglichen, formalen Wissensbasen für verschiedene Themenbereiche.

Durch das beschrieben Vorgehen nicht einwandfrei erfasst werden neben unbekannten Begriffen und Abkürzungen insbesondere Tags, die in erster Linie für den indexierenden Nutzer Sinn ergeben. Hierzu zählen beispielsweise aus der Situation oder Aufgabe erwachsene Tags (wie „zu_erledigen“, „projektrelevant“) oder persönliche Wertungen (wie „lustig“, „cool“). Die semantische Analyse solch funktionaler Tags [Sa06] müsste von einer erweiterten Kontextualisierung begleitet werden, die Rückschlüsse auf die Nutzerintentionen bei der Verwendung dieser Tags zulässt. Die große Mehrheit nutzergenerierter Tags besitzt jedoch deskriptiven Charakter [GH06], so dass die Vernachlässigung funktionaler Tags kein größeres Problem darstellt.

Mit dem hier beschriebenen Vorgehen wurde die semantische Interoperabilität von Tagging-Daten auf die Ebene des Ontology Alignment und Mapping [ES07, Eh06] verlagert und damit in ein besser lösbares Problem gewandelt. Neben einer verringerten Anzahl polysemer Tags führt das Vorgehen zu einer Konvergenz synonymer Tags, wenn diese auf gleiche Ontologiekonzepte verweisen. Durch Multilingualität entstehende Probleme werden reduziert, da sich Konzepte leichter in verschiedene Sprachen übersetzen lassen als semantisch mehrdeutige Tags. Konzept-basierte Suchanfragen lassen darüber hinaus verbesserte Recall- und Precision-Werte erwarten [Mo07, HL07].

Literaturverzeichnis

[Eh06] Ehrig, M.: Ontology Alignment – Bridging the Semantic Gap. Springer, Heidelberg, Berlin, 2006.

[ES07] Euzenat, J.; Shvaiko, P.: Ontology Matching. Springer, Heidelberg, Berlin, 2007.

[Fl07] Flickr Machine Tags, http://www.flickr.com/groups/api/discuss/72157594497877875/, 2007 (28.Apr 2007).

[GNV03] Gangemi, S.; Navigli, R.; Velardi, P.: The OntoWordNet Project: Extension and Axiomatisation of Conceptual Relations in WordNet. In: Proc. of Int. Conf. on Ontologies, Databases and Applications of Semantics, 2003, LNCS 2888, Springer; S. 820-838.

[GH06] Golder, S.; Huberman, B.A.: The Structure of Collaborative Tagging Systems. In: Journal of Information Science, 32(2), 2006; S. 198-208.

[GT06] Guy, M.; Tonkin, E.: Folksonomies – Tidying up Tags? In: D-Lib Magazine 12,1, 2006.

[Ha05] Hammond, T.; Hannay, T.; Lund, B.; Scott, J.: Social Bookmarking Tools (I): A General Review. In: D-Lib Magazine, 11, 4, 2005.

[HL07] Haav, H.M., Lubi, T.-L.: A Survey of Concept-based Information Retrieval Tools on the Web. In: Proceedings of 5th East-European Conference ADBIS, 2, 2001; S. 29-41.

[HQW06] Heyer, G.; Quasthoff, U.; Wittig, T.: Text Mining: Wissensrohstoff Text – Konzepte, Algorithmen, Ergebnisse. W3L-Verlag, Herdecke, Bochum, 2006.

[Le06] Lee, K.J.: What Goes Around Comes Around: An Analysis of del.icio.us as Social Space. In: Proceedings of the 20th Anniversary Conference on Computer Supported Cooperative Work, 2006; S. 191-194.

[Ma06] Marlow, C.; Naaman, M.; Boyd, D.; Davis, M.: HT06, tagging paper, taxonomy, Flickr, academic article, to read. In: Proceedings of the 17th Conference on Hypertext and Hypermedia, 2006; S. 31-40.

[Ma07] Marchetti, A.; Tesconi, M.; Ronzano, F.; Rosella, M.; Minutoli, S.: SemKey: A Semantic Collaborative Tagging System. In: Proceedings of WWW 2007 Workshop on Tagging and Metadata for Social Information Organization, 2007.

[MC07] Michlmayr, E.; Cayzer, S.: Learning User Profiles from Tagging Data and Leveraging them for Personal(ized) Information Access. In: Proceedings of WWW 2007 Workshop on Tagging and Metadata for Social Information Organization, 2007.

[Mo07] Moskovitch, R.; Martins, S.B.; Behiri, E.; Weiss, A.; Shahar, Y.: A Comparative Evaluation of Full-text, Concept-based, and Context-sensitive Search. In: Journal of American Medical Informatics Association, 14, 2, 2007; S. 164 - 174.

[Sa06] H. Sack: Kollaborative Indexierung und die Emergenz neuer sozialer Netzwerke. In: Workshop Social Software in der Wertschöpfungskette, 2006.

[TB06] Pellegrini, T.; Blumauer, A. (Hrsg.): Semantic Web. Wege zur vernetzten Wissensgesellschaft. Springer, Heidelberg, Berlin, 2006.

[Va07] Vander Wal, T.: Folksonomy Coinage and Definition, http://www.vanderwal.net/folksonomy.html, 2007 (28. Apr 2007).