18. Social Bookmarking am Beispiel BibSonomy

Andreas Hotho[1], Robert Jäschke[1,2], Dominik Benz[1], Miranda Grahl[1], Beate Krause[1,2], Christoph Schmitz[1] und Gerd Stumme[1,2]

[1] Hertie-Lehrstuhl Wissensverarbeitung, FB Elektrotechnik/Informatik, Universität Kassel; nachname@ cs.uni-kassel.de
[2] Forschungszentrum L3S, Hannover

Zusammenfassung: BibSonomy ist ein kooperatives Verschlagwortungssystem (Social Bookmarking System), betrieben vom Fachgebiet Wissensverarbeitung der Universität Kassel. Es erlaubt das Speichern und Organisieren von Web-Lesezeichen und Metadaten für wissenschaftliche Publikationen. In diesem Beitrag beschreiben wir die von BibSonomy bereitgestellte Funktionalität, die dahinter stehende Architektur sowie das zugrunde liegende Datenmodell. Ferner erläutern wir Anwendungsbeispiele und gehen auf Methoden zur Analyse der in BibSonomy und ähnlichen Systemen enthaltenen Daten ein.

Einleitung

BibSonomy ist ein kooperatives Verschlagwortungssystem (Social Bookmarking System) zur Verwaltung von Web-Lesezeichen und Publikationen. Das System ist seit Dezember 2005 unter der Adresse `http://www.bibsonomy.org/` online und wird vom Fachgebiet Wissensverarbeitung der Universität Kassel betrieben. Dieses für jeden frei nutzbare System erlaubt es, Lesezeichen (Favoriten, Bookmarks) für Webseiten zentral auf dem BibSonomy-Server abzuspeichern und zu verschlagworten. Die vom Nutzer frei wählbaren Schlagwörter, auch ‚Tags' genannt, erlauben es ihm, seine Lesezeichen- und Literatursammlung zu strukturieren und mit Hilfe der Schlagwörter einen Eintrag später einfach wieder zu finden. Die Sammlungen aller Benutzer sind in einer ‚Folksonomie' zusammengefasst. Das System bietet jedem Benutzer die Möglichkeit, in dieser Folksonomie auch nach Lesezeichen und Publikationsreferenzen anderer Benutzer mit verwandten Interessen zu suchen. Diese soziale Komponente erzeugt also personalisierte Empfehlungen, die globale Suchmaschinen wie Google nicht leisten können, da sie die Interessen

des Anfragenden nicht kennen. Durch die zentrale Speicherung hat der Benutzer außerdem jederzeit von jedem Rechner Zugriff auf seine Lesezeichen und Publikationsdaten.

Ein Grund für die Entwicklung von BibSonomy war die Tatsache, dass Literaturrecherche und -verwaltung eine zentrale Rolle in der täglichen Arbeit eines Wissenschaftlers spielt, es für die systematische Katalogisierung der gefundenen Publikationen jedoch wenig Unterstützung gibt – jeder Forscher entwickelt typischerweise sein eigenes Verwaltungs- und Ablageschema, und jedes Institut führt darüber hinaus mit hohem Aufwand jeweils eigene Literaturlisten. Da diese Schwierigkeiten auch in unserem Fachgebiet auftraten, und wir die Möglichkeit sahen, diese in Kooperation mit anderen Forschern effektiver und effizienter anzugehen, wurde die kollektive Literaturverwaltung in BibSonomy als zweite Kern-Komponente neben der Lesezeichenverwaltung eingebaut. Für interessierte Gruppen (z. B. Universitätsinstitute oder Projektkonsortien) bietet das Fachgebiet die Einrichtung von Benutzergruppen im System an, mit denen sowohl der interne als auch der externe Literaturaustausch organisiert werden kann.

BibSonomy basiert auf dem `BibTEX`-Format [19] zur Speicherung von Publikationsdaten. Seine Publikationsverwaltung ist somit leicht in das Satz-System `LATEX` [14] integrierbar, mit dem Forscher – insbesondere aus den Naturwissenschaften – ihre wissenschaftlichen Veröffentlichungen druckfertig gestalten. Das System erzeugt automatisch Literaturlisten in weiteren Formaten und verringert so den Gesamtaufwand, der Wissenschaftlern zur Erstellung ihrer Publikationslisten für verschiedene Zwecke (Web-Publikationslisten der einzelnen Wissenschaftler und der ganzen Forschungsgruppe, Referenzlisten in Publikationen, Projekt-Berichte, Berichte an die Hochschulleitung, etc.) – vielfach durch die Verwendung unterschiedlicher Systeme – aufgezwungen wird. In BibSonomy wird die gesamte Literaturliste einmal zentral verwaltet. Aus ihr können dann die gewünschten Teillisten automatisch in die gewünschten Formate exportiert werden. Dafür werden neben `BibTEX` auch die für Anwender von Microsoft Word interessanten Formate RTF und EndNote erzeugt, aber auch Formate wie XML, RSS-Feeds und formatiertes HTML. Ausgewählte Anwendungsszenarien werden in den Kapiteln 4 und 5 beschrieben.

BibSonomy ist eines von mehreren kooperativen Verschlagwortungssystemen, die im Zuge des Web 2.0 – der nächsten Generation des World Wide Webs – entstanden sind. Andere Verschlagwortungssysteme dienen der Verwaltung von Photos und Musik – und selbst dem Austausch von guten Vorsätzen. BibSonomy ist derzeit das einzige System, das die Ver-

waltung von Lesezeichen und Publikationen verbindet.[1] Im Zentrum der Entwicklung stand die Anwendbarkeit im akademischen Bereich; die Rückmeldungen von vielen Wissenschaftlern sind in die Entwicklung des Systems mit eingeflossen. Zurzeit (Stand 7.3.2008) hat das System 2.632 aktive Benutzer, die sich 199.849 Web-Lesezeichen und 76.984 Publikationen teilen. Hinzu kommen 18.193 Web-Lesezeichen und 979.033 Publikationen, die automatisch von der Computer Science Library der Universität Trier übernommen werden. Das System verzeichnet derzeit im Schnitt 810.487 Seitenzugriffe pro Tag; die Tendenz ist steigend. Ein besonderer Schwerpunkt bei der Implementierung wurde auf die effiziente Beantwortung der Anfragen gelegt. Details zur Implementierung sowie zur Funktionalität des Systems sind in den Kapiteln 2 und 3 beschrieben.

Die frei wählbaren Schlagwörter ergänzen die üblichen Publikationsdaten wie Autor, Titel, Verlag, Jahr etc. Ähnlich gehen Bibliotheken vor, wenn sie Bücher in ihren Bestand aufnehmen und diese zum Zwecke des Wiederfindens verschlagworten. Während in einer Bibliothek jedoch – wie in Semantic-Web-Anwendungen – die Schlagwörter aus einem fest vorgegebenen Katalog ausgewählt werden, kann in BibSonomy jeder Nutzer seine Schlagwörter frei wählen und so auch aktuelle Themen beschreiben, die in einem vorgegebenen Katalog (noch) nicht erfasst oder die nicht offensichtlich mit dem Werk verbunden sind. Die Bibliothekslösung hat dagegen den Vorteil, dass die Literatursuche nicht durch mehrdeutige Schlagwörter erschwert wird. Bibliothekare werden allerdings für die systematische Verschlagwortung geschult, was für einen breiteren Benutzerkreis nicht vorausgesetzt werden kann.

Die zentrale Frage, um die Vorteile von Web 2.0 (Nähe zum Anwender und daraus resultierende umfangreiche Datensammlungen) und Semantic Web (Wissensrepräsentation mit formaler Semantik und daraus resultierender Unterstützung von strukturierter Navigation und Wissensableitung durch logisches Schließen) zu verbinden – eine Kombination, die mancherorts schon als ‚Web 3.0' vermarktet wird – ist daher, wie untrainierte Anwender effektiv unterstützt werden können, ohne durch eine komplexe Interaktion mit den zugrunde liegenden Wissensverarbeitungsverfahren irritiert zu werden. Wir haben in diese Richtung mehrere Ansätze in den Bereichen Datenanalyse, Information Retrieval und Wissensentdeckung auf die neue Datenstruktur von Folksonomien erweitert. Diese Ansätze werden wir in Kapitel 6 kurz beschreiben.

[1] Bekannte Systeme zur kooperativen Publikationsverwaltung sind http://www.citeulike.org und http://www.connotea.org sowie http://del.icio.us zur Lesezeichenverwaltung. Ein Vergleich von BibSonomy mit diesen Systemen wurde in [23] durchgeführt.

Aufbau des Systems

In diesem Abschnitt beschreiben wir das `BibSonomy`-System. Nach einer Einführung in die Benutzeroberfläche, der Vorstellung des zentralen formalen Folksonomie-Modelles, der Semantik und der Architektur von BibSonomy erklären wir weitere Eigenschaften sowie zukünftige Verbesserungen. Eine Beschreibung der Funktionalität und der internen Datenstrukturen der ersten Version von BibSonomy findet man in [7]. Das aktuell laufende System basiert auf einem neuen internen Modell. Es bietet eine in vielen Punkten erweiterte Funktionalität an. Der neue modulare Aufbau vereinfacht die Wartung und Weiterentwicklung des Systems erheblich.

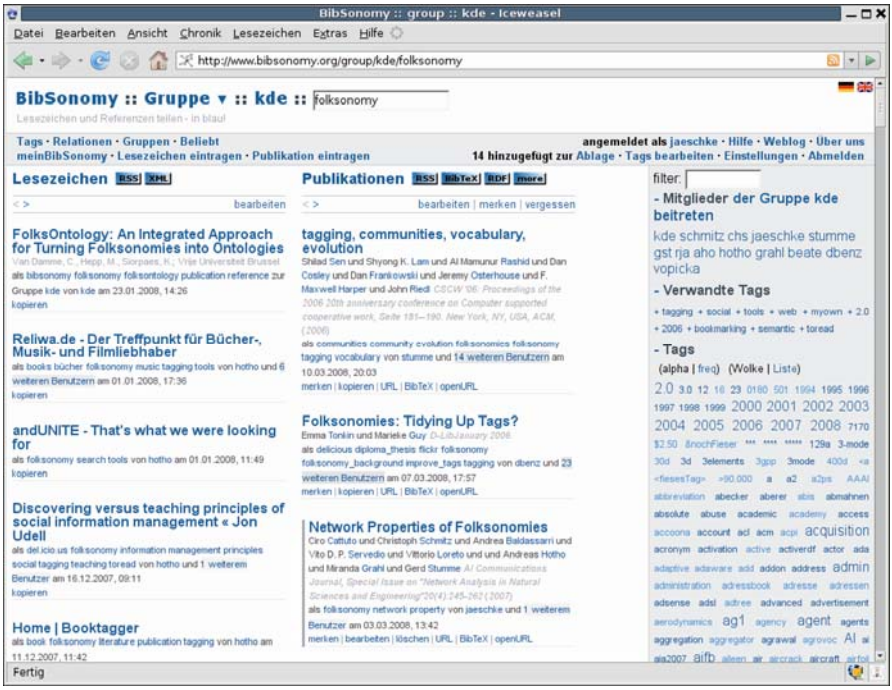

Abb. 1. BibSonomy zeigt gleichzeitig Lesezeichen und auf BibTEX basierende Literaturverweise an

Benutzeroberfläche

Abbildung 1 zeigt eine typische Liste von Einträgen bestehend aus Lesezeichen und Literaturreferenzen, die mit dem Schlagwort *web* annotiert wurden. Die Seite besteht aus vier Abschnitten: dem Kopf (mit Informa-

tionen zur aktuellen Seite, dem aktuellen Pfad, Navigationselementen und Suchfeldern), zwei Listen von Einträgen – eine für Lesezeichen und eine für Literaturreferenzen – die jeweils absteigend nach Datum sortiert sind, und einer Liste von Schlagwörtern, die mit den Einträgen verwandt sind. Dieses Schema gilt für alle Seiten, die Einträge anzeigen, und ermöglicht das Navigieren in allen Dimensionen der Folksonomie. Die darunterliegende Semantik dieser Seiten wird in Abschn. 3 erklärt.

Abb. 2. Detaillierte Ansicht eines einzelnen Lesezeicheneintrags

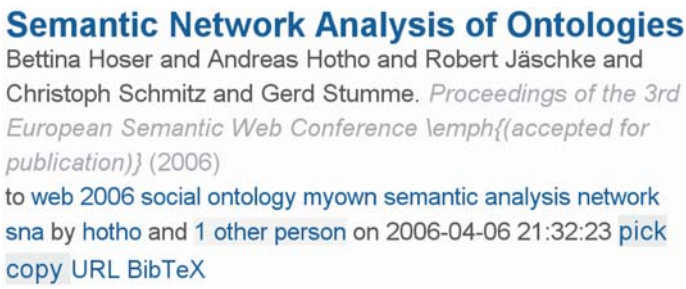

Abb. 3. Detaillierte Ansicht eines einzelnen Literatureintrags

Abbildung 2 zeigt eine detaillierte Ansicht eines Lesezeichen-Eintrags aus der Liste in Abb. 1. Die erste Zeile zeigt fett gedruckt den Titel des Lesezeichens in Form eines Hyperlinks. Die zweite Zeile zeigt eine optionale Beschreibung, die der Benutzer jedem Eintrag zuordnen kann. Die letzten zwei Zeilen gehören zusammen und zeigen detaillierte Informationen: Erstens alle Schlagwörter, die der Benutzer diesem Eintrag zugeordnet hat (*web, service, tutorial, guidelines* und *api*), zweitens den Benutzernamen dieses Benutzers (*hotho*) gefolgt von einem Hinweis, wie viele Benutzer insgesamt diese spezifische Ressource verschlagwortet haben. Diese Informationen fungieren als Hyperlinks zu den entsprechenden Schlagwort-Seiten des Benutzers (`/user/hotho/web, /user/hotho/service,` …), der Seite des Benutzers selbst (`/user/hotho`) und einer Seite, die alle vier Einträge (d. h. den einen des Benutzers *hotho* und die der drei anderen) dieser Ressource zeigt (`/url/`r). Abschn. 3 erklärt die Pfade, die bis auf

weiteres in Klammern angegeben sind. Der letzte Teil zeigt Datum und Uhrzeit des Eintrags, gefolgt von Links auf Aktionen, die der Benutzer ausführen kann – je nachdem, ob es sein eigener Eintrag ist (*bearbeiten*, *löschen*) oder der eines anderen Benutzers (*kopieren*).

Die Struktur eines Literatureintrags in BibSonomy ist ähnlich (siehe Abb. 3). Die erste Zeile zeigt wieder den Titel des Eintrags, der derselbe ist wie der Titel der Publikation in `BibTEX`. Er hat einen darunter liegenden Hyperlink, der zu einer Seite führt, die detaillierte Informationen zu diesem Eintrag präsentiert. Nach dieser Zeile kommen die Autoren oder Herausgeber der Publikation sowie der Titel der Zeitschrift oder des Buchs mit dem Erscheinungsjahr. Die nächsten Zeilen zeigen die Schlagwörter, die diesem Eintrag vom Benutzer zugeordnet wurden, danach dessen Benutzername, gefolgt von einem Hinweis, wieviele Personen diese Publikation verschlagwortet haben. So wie schon für die Lesezeicheneinträge beschrieben, verweisen diese Teile auf entsprechende weiterführende Seiten. Nach der Angabe von Datum und Zeit, zu denen der Benutzer diesen Eintrag gemacht hat, folgen die Aktionen, die der Benutzer ausführen kann (in diesem Fall Auswahl des Eintrags für einen späteren Download, Kopieren, Folgen eines Hyperlinks zu einer externen Referenz des Eintrags und Ansicht des `BibTEX`-Quellcodes).

Das System enthält eine Reihe spezieller Funktionen, die an dieser Stelle kurz vorgestellt werden.

Gruppen

BibSonomy verfügt über Benutzergruppen, die das gemeinsame Sammeln von Einträgen in offenen und geschlossenen Gruppen unterstützen. Eine Liste aller im System registrierten Gruppen findet man unter `/groups`. Die Gruppen fungieren auf der einen Seite als organisatorische Einheiten und erlauben zusammengefasste Sichten über die Einträge aller Gruppenmitglieder. Auf der anderen Seite dienen sie zur Verwaltung der Zugriffsrechte. Auf diesem Weg können Einträge auch gruppenintern geteilt und gemeinsam bearbeitet werden. Jeder Nutzer kann zwischen mindestens drei Gruppen wählen: `public`, `private` und `friends`. In `public` werden alle für jedermann sichtbaren Einträge gesammelt und in `private` die ausschließlich privaten Einträge. `Friends` ist eine Gruppe, in der jeder Benutzer seine befreundeten Benutzer festlegen und mit ihnen Einträge gemeinsam sammeln kann. Die Liste alle Freunde und der dazugehörenden Einträge findet man unter `/friends`.

Warenkorb

Eine weitere Funktion zur Unterstützung der Nutzer bei der Arbeit mit BibSonomy ist der `basket`, ein Warenkorb für Literatureinträge. Er dient dazu, Listen mit Literaturverweisen zusammenstellen zu können. Jeder Literaturverweis im System, unabhängig davon ob es ein eigener oder ein fremder Eintrag ist, kann mittels der `pick`-Funktion in den Warenkorb aufgenommen werden. Für den Warenkorb stehen alle typischen Funktionen, wie das Editieren von Schlagwörtern oder Exporte in andere Formate, zur Verfügung.

Dublettenerkennung

Insbesondere bei Literaturverweisen besteht das Problem doppelter Einträge, da die Art und Weise, wie Benutzer Felder wie ‚Name der Zeitschrift' oder ‚Autoren' ausfüllen, sehr unterschiedlich ist. Einerseits ist es wünschenswert, dass Benutzer verschiedene Einträge haben können, die nur wenig variieren. Andererseits möchte man die Einträge anderer Benutzer, die sich auf denselben Artikel beziehen, auch dann finden können, wenn die Einträge leicht voneinander abweichen.

Jeder Nutzer erhält auf der Seite `/myDuplicates` einen Überblick über die vom System als Duplikate erachteten Einträge in seiner Literaturliste. Um diese zu erkennen, werden spezielle Hashes aus den Literaturreferenzen des Benutzers errechnet. Sind zwei Hashes identisch, so deutet dies auf ein Duplikat hin und es wird auf der `myDuplicates`-Seite angezeigt. Die gegenwärtige Dublettenerkennung ist sehr einfach; Fehler in der Schreibweise, Unterschiede bei der Eingabe besonderer Zeichen (wie der deutschen Umlaute), und zusätzliche $^{L}_{A}{}^{T}_{E}{}^{X}$-Befehle werden derzeit nicht erkannt. Diese Themen stehen auf unserer Agenda und sind leicht zu ergänzen, denn unsere Implementierung erlaubt die einfache Integration neuer Hashes in das System.

Import- und Export-Funktionen

Um Benutzer zum Wechsel von anderen Systemen hin zu BibSonomy zu ermutigen, haben wir Importfunktionen eingebaut. Lesezeichen können von del.icio.us oder aus dem Webbrowser Firefox automatisch übernommen werden. Bei der Literatur ist der Import aus `BibTEX`-Dateien Standard; man kann auch Daten im EndNote-Format importieren. Zusätzlich sind im System mehr als 25 so genannte Screenscraper[2] integriert. Sie ermöglichen die automatische Übernahme der Publikationsdaten aus Digi-

[2] http://www.bibsonomy.org/scraperinfo

talen Bibliotheken (beispielsweise SpringerLink und CiteSeer). Literaturdaten von weniger strukturierten Webseiten werden durch Informationsextraktion [20] in BibSonomy übernommen. Unsere Implementierung basiert auf dem MALLET-System [17].

Verschiedene Export-Formate vereinfachen die Interaktion von BibSonomy mit anderen Systemen. RSS-Feeds ermöglichen die einfache Integration von Listen in Webseiten oder RSS-Aggregatoren, und die `BibTEX`- bzw. EndNote-Ausgabe kann dazu verwendet werden, automatisch Publikationslisten für Veröffentlichungen zu generieren. Weiterhin stehen spezielle HTML-Export-Formate zur Verfügung. Einen Überblick über alle unterstützten Exportformate findet man unter `/export/`. Besonders leicht kann der Nutzer die Ausgabe von BibSonomy mit dem aus dem Open-Source-Projekt JabRef übernommenen Layout-Filter[3] erzeugen. Der Layout-Filter bietet die Möglichkeit, in einer einfachen Sprache ein eigenes Ausgabeformat für Literaturreferenzen festzulegen. Der Nutzer kann den selbst definierten Filter ins System hochladen und dann alle BibSonomy -Seiten passend ausgeben.

mySearch

Zur Exploration der eigenen Literaturreferenzen kann man die Funktion `/mySearch` verwenden. Sie bietet ein übersichtliches, Javascript-basiertes Benutzer-Interface, über das man die Referenzen entlang der Schlagwörter und Autoren durchforsten kann.

Formales Folksonomie-Modell

Benutzer, Ressourcen (Webseiten, `BibTEX`-Einträge o. ä.) und Schlagwörter bilden die Basis einer Folksonomie, des zentralen Datenmodells von kooperativen Verschlagwortungssystemen. Im Folgenden präsentieren wir die formale Definition einer Folksonomie, wie sie auch dem BibSonomy-System zugrunde liegt.

Definition 1: Eine *Folksonomie* ist ein Tupel $F := (U, T, R, Y, \prec)$ wobei

- U, T und R endliche Mengen sind, deren Elemente man *Benutzer (users)*, *Schlagwörter (tags)* und *Ressourcen (resources)* nennt,
- Y eine ternäre Beziehung zwischen diesen ist, d. h. $Y \subseteq U \times T \times R$ gilt, und

[3] http://jabref.sourceforge.net/help/CustomExports.php (Dieser, sowie alle folgenden Links auf Webseiten, wurde am 3.3.2008 aufgerufen.)

- $\prec$ eine benutzerspezifische *Unterschlagwort/Oberschlagwort-Beziehung* ist, d. h. $\prec \subseteq U \times T \times T$ gilt.

Die Elemente von Y heißen *Schlagwort-Zuweisungen* (tag assignments). Ein *Eintrag* (post) von F ist ein Tripel (u, S, r) mit $u \in U$, $r \in R$, und $S := \{t \in T \mid (u,t,r) \in Y\}$ mit $S \neq \emptyset$. Die Menge aller Einträge einer Folksonomie wird mit P bezeichnet.

Die *Personomy* P_u eines Users $u \in U$ ist die Einschränkung F auf u, d. h. $P_u := (T_u, R_u, I_u, \prec_u)$ mit $I_u := \{(t,r) \in T \times R \mid (u,t,r) \in Y\}$, $T_u := \pi_1(I_u)$, $R_u := \pi_2(I_u)$ und $\prec_u := \{(t_1,t_2) \in T \times T \mid (u,t_1,t_2) \in \prec\}$.

Benutzer werden typischerweise durch eine User-ID und Schlagwörter durch beliebige Zeichenketten beschrieben. Was man als Ressource betrachtet, hängt vom Systemtyp ab. Ressourcen sind in del.icio.us zum Beispiel Bookmarks und in Flickr Bilder. Unser System BibSonomy enthält zwei Arten von Ressourcen: Lesezeichen und BIBTEX-Einträge. Verschiedene Typen von Ressourcen unterscheiden sich strukturell nicht, da sie unabhängig vom Typ intern durch eine ID repräsentiert werden. Nur die Art der Bildschirmanzeige hängt von der Art der Ressource ab.

Semantik des URL-Schemas

Alle hier beschriebenen relativen URLs beziehen sich auf `http://www.bibsonomy.org`. Wir beschränken die Beschreibung auf die für das System zentralen URLs; organisatorische Seiten wie `/help`, `/settings` oder `/post_bookmark` lassen wir in dieser Übersicht aus. Am Ende dieses Abschnittes gehen wir auf einige interessante Erweiterungen ein. Die folgende Liste gibt zu jeder URL die Menge C der Einträge (Lesezeichen oder Literaturreferenzen) an, die von BibSonomy angezeigt werden.

- **`/tag/`** t_1 … t_n zeigt jeden Eintrag, dem alle Schlagwörter $t_1,...,t_n$ zugeordnet sind:
  $$C_{t_1,..,t_n} := \{(u,S,r) \in P \mid \{t_1,...,t_n\} \subseteq S\} \quad (1)$$
- **`/user/`** u zeigt alle Einträge des Benutzers u:
  $$C_u := \{(\hat{u},S,r) \in P \mid \hat{u} = u\} \quad (2)$$
- **`/user/`** u / t_1 … t_n zeigt jeden Eintrag des Benutzers u, dem alle Schlagwörter $t_1,...,t_n$ zugeordnet sind:
  $$C_{u,t_1,..,t_n} := \{(\hat{u},S,r) \in P \mid \hat{u} = u, \{t_1,...,t_n\} \subseteq S\} \quad (3)$$
- **`/concept/`** t_1 … t_n zeigt jeden Eintrag, dem für jedes Schlagwort $t \in \{t_1,...,t_n\}$ mindestens eines seiner Unterschlagwörter oder t selbst zugeordnet sind (siehe auch Abschn. 1):

$$C_{t_1,..,t_n} := \{(u,S,r) \in P \mid \forall t_i (i = 1,...,n) \exists t \in S : (u,t,t_i) \in \prec \vee t = t_i\} \quad (4)$$

- **/concept/user/** u / t_1 … t_n zeigt jeden Eintrag des Benutzers u, dem für jedes Schlagwort $t \in \{t_1,...,t_n\}$ mindestens eines seiner Unterschlagwörter oder t selbst zugeordnet ist (siehe auch Abschn. 1):

$$C_{u,t_1,..,t_n} := \{(\hat{u},S,r) \in P \mid \hat{u} = u, \forall t_i (i = 1,...,n) \exists t \in S : (\hat{u},t,t_i) \in \prec \vee t = t_i\} \quad (5)$$

- **/url/** r zeigt alle Einträge zu Ressource r, wenn sie ein Lesezeichen ist:

$$C_r := \{(u,S,\hat{r}) \in P \mid \hat{r} = r\} \quad (6)$$

- **/url/** r / u zeigt den Eintrag von Benutzer u zu Ressource r, wenn sie ein Lesezeichen ist:

$$C_{r,u} := \{(\hat{u},S,\hat{r}) \in P \mid \hat{r} = r, \hat{u} = u\} \quad (7)$$

- **/bibtex/** r zeigt alle Einträge zu Ressource r, wenn sie ein Literaturverweis ist:

$$C_r := \{(u,S,\hat{r}) \in P \mid \hat{r} = r\} \quad (8)$$

- **/bibtex/** r / u zeigt den Eintrag von Benutzer u der Ressource r wenn sie ein Literaturverweis ist:

$$C_{r,u} := \{(\hat{u},S,\hat{r}) \in P \mid \hat{r} = r, \hat{u} = u\} \quad (9)$$

- **/groups** zeigt alle Gruppen des Systems. Mehr zu Gruppen siehe Abschn. 1.
- **/group/** g zeigt alle Einträge aller Benutzer, die zu Gruppe g gehören:

$$C_g := \{(u,S,r) \in P \mid u \in g\} \quad (10)$$

- **/group/** g / t_1 … t_n zeigt jeden Eintrag, dem alle Schlagwörter $t_1,...,t_n$ zugeordnet sind, und dessen Benutzer zu Gruppe g gehört:

$$C_{g,t_1,..,t_n} := \{(u,S,r) \in P \mid u \in g, \{t_1,...,t_n\} \subseteq S\} \quad (11)$$

- **/viewable/** g zeigt alle Einträge, die für Mitglieder der Gruppe g sichtbar eingestellt sind.
- **/viewable/** g / t_1 … t_n zeigt alle Einträge, die für Mitglieder der Gruppe g als sichtbar eingestellt sind und denen alle Schlagwörter $t_1,...,t_n$ zugeordnet sind.
- **/search/** s zeigt (basierend auf der MySQL-Volltext-Suchfunktion[4]) alle Einträge, deren Volltext oder Schlagwörter dem Suchausdruck s entsprechen.

[4] http://dev.mysql.com/doc/refman/5.0/en/fulltext-boolean.html

- **/basket** zeigt alle Literatureinträge, die der Benutzer in den Warenkorb gelegt hat, wie in Abschn. 1 beschrieben.
- **/popular** zeigt die 100 am häufigsten abgespeicherten Ressourcen innerhalb der letzten 1000 Einträge. (Diese Liste unterliegt einem stetigen Wechsel.)
- **/** ist die Homepage von BibSonomy; sie zeigt die aktuellsten Einträge.
- **/author/** a_1 … a_n zeigt jeden Eintrag, der alle Namen aus a_1 … a_n als Nachnamen enthält. Dabei spielt es keine Rolle, ob der Name im Autor- oder Herausgeber-Feld vorkommt. Die übergebenen Namen werden auf Nachnamen normalisiert. Ist ein Name von Anführungszeichen umgeben, so werden nur Einträge ausgegeben, die im Autor- oder Herausgeber-Feld exakt diesen Namen enthalten.
- **/uri/…** ist eine für den Zusammenschluss mit anderen Semantic-Web-Diensten notwendige URL-Erweiterung. Sie erlaubt die Auswahl des auszugebenden Datenformats durch den anfragenden Dienst unabhängig von der URL und leitet auf die Seite mit dem gewünschten Ausgabeformat weiter (siehe Abschn. 2).
- **/bibtexkey/** k zeigt die Einträge an, die den BIBTEX-Key k haben.
- **/api/** ist die Basis-URL für den Zugriff auf die REST-API von BibSonomy. Details findet man in Abschn. 3.
- **/export/** zeigt die Liste aller zur Verfügung stehenden Exportformate für die aktuelle Seite an. Einige Formate werden in den Abschnitten 1 und 2 beschrieben.
- **/my…** umfasst eine Reihe von URLs, die dem Benutzer einen einfachen Zugriff auf seine gespeicherten Daten erlauben: `/myBibSonomy`, `/mySearch`, `/myPDF`, `/myRelations` und `/myDuplicates`.

Allen oben beschriebenen URL-Pfaden kann eine Zeichenkette vorangestellt werden, die das Ausgabeformat ändert (siehe Abschn. 4). Im Allgemeinen werden Einträge als HTML-Listen angezeigt, die von Navigationselementen und einer Schlagwort-Wolke umgeben sind (siehe Abb. 1), aber diese Eigenschaft ermöglicht es dem Benutzer, aufgabenspezifische Ausgabeformate wie BIBTEX, RSS oder passend formatierte HTML-Seiten (nach eigenen Layoutvorgaben) zu erhalten.

Architektur

Die Grundbausteine von BibSonomy sind ein Apache Tomcat[5] Servlet Container, der Java Server Pages[6] und Java Servlet[7]-Technologie verwen-

[5] http://tomcat.apache.org
[6] http://java.sun.com/products/jsp

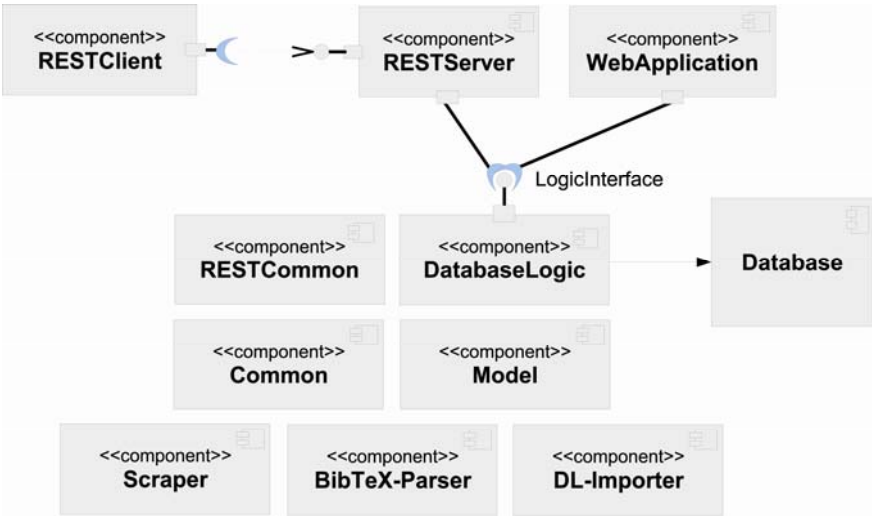

Abb. 4. UML-Diagramm der Komponenten von BibSonomy

det, sowie eine MySQL[8]-Datenbank. Gegenwärtig beinhaltet das Projekt mehrere zehntausend Codezeilen und verwendet das Model View Controller (MVC) Programmier-Paradigma [12], um die logische Verarbeitung der Daten von deren Präsentation zu trennen. Dies ermöglicht es, Ausgaben in verschiedenen Formaten zu erzeugen (siehe Abschn. 4), da das Hinzufügen eines neuen Ausgabeformats lediglich den Austausch einer JSP als Modellansicht erfordert. Weiterhin wird das Spring-MVC-Framework[9] verwendet, um die unterschiedlichen Teile einer HTML-Seite unabhängig voneinander zu erzeugen und so das System zusätzlich zu modularisieren.

Einen Überblick über die Komponenten des Systems gibt Abb. 4. Man erkennt die zentralen Komponenten `DatabaseLogic` und `Model` des Systems, um die sich auf der einen Seite `RESTServer` und `WebApplication` als Elemente zur Kommunikation mit der Außenwelt, auf der anderen Seite die Hilfskomponenten `RESTCommon`, `Common`, `Scraper`, `DL-Importer` und `BibTeX-Parser` gruppieren.

- **DatabaseLogic.** Diese Komponente kapselt alle Datenbankanfragen und stellt ein internes Interface für den Datenzugriff bereit. Zur Kodierung der SQL-Anfragen wird IBATIS[10] eingesetzt. Ein Hauptgrund für die Auswahl dieses Persistenz-Frameworks war die Möglichkeit, direkte

[7] http://java.sun.com/products/servlets
[8] http://www.mysql.com
[9] http://static.springframework.org/spring/docs/2.0.x/reference/mvc.html
[10] http://ibatis.apache.org/

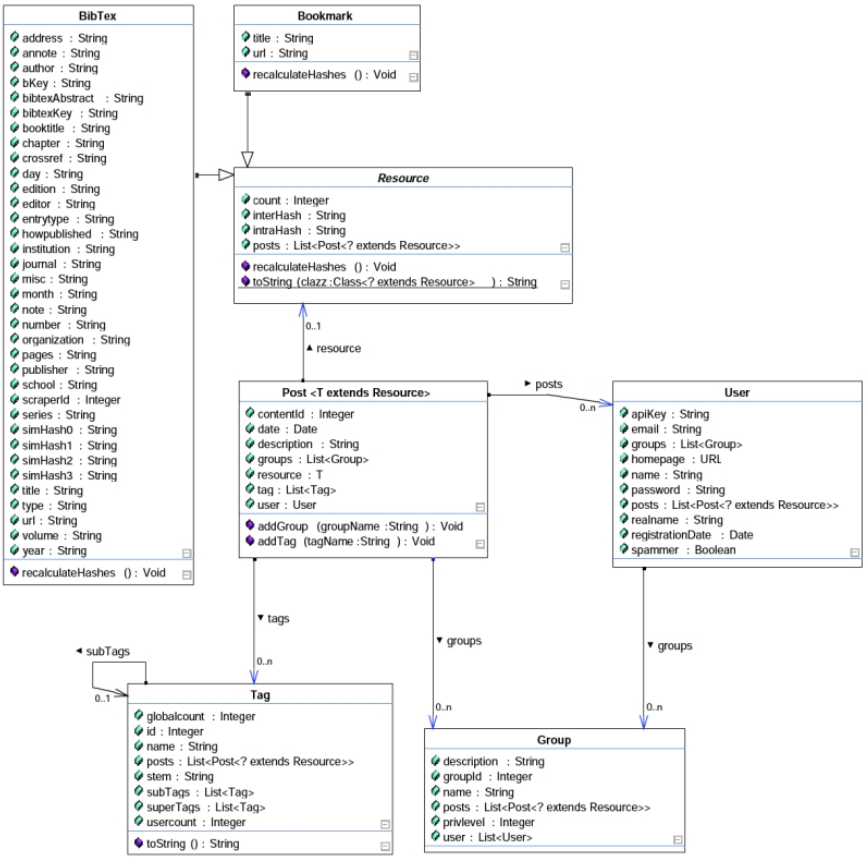

Abb. 5. UML-Diagramm des BibSonomy zugrunde liegenden Datenmodells

Kontrolle über die SQL-Anfragen beizubehalten und sie somit ‚von Hand' optimieren zu können – was bei anderen Frameworks wie z. B. Hibernate [11] nicht in diesem Maß möglich ist.

- **Model.** Das Modell bildet alle zentralen Elemente einer Folksonomie in Javaobjekten ab (siehe Abb. 5). Es bildet die Grundlage für alle weiteren Datenverarbeitungs- und Austauschprozesse innerhalb des Systems sowie für die Kommunikation mit der Außenwelt. Ein wichtiger Bestandteil ist die Definition eines XML-Schemas, welches das Format für den Datenaustausch über die REST-API festlegt. Die zentralen Elemente des Modells entsprechen den Elementen der formalen Definition einer Folksonomie aus Abschn. 2.

[11] http://www.hibernate.org/

- **RESTServer.** Implementiert die REST-API (siehe Abschn. 3). Das zentrale Element hierbei ist das RESTServlet, das alle eingehenden API-Anfragen entgegennimmt und je nach Typ der Anfrage (z. B. GET, POST,...) sowie nach der Struktur der angefragten URL eine passende Strategie zur Bearbeitung der Anfrage auswählt. Nachdem die passenden Daten über `DatabaseLogic` geholt beziehungsweise geschrieben wurden, sendet das Servlet die Antwort an den Client zurück.
- **WebApplication.** Diese Komponente ist für die BibSonomy-Webseiten zuständig. Wie oben erwähnt, kommt an dieser Stelle das Spring-MVC-Framework zum Einsatz, das sich nahtlos in die Model-View-Controller-Architektur des gesamten Systems einfügt. Für eingehende Anfragen sind verschiedene Controller zuständig, die die angeforderten Daten (wiederum basierend auf dem definierten Datenmodell) über die `DatabaseLogic` anfragen. Diese werden dann mittels einer entsprechenden View-Komponente gerendert.
- **RESTCommon.** Enthält Funktionen, die sowohl auf Seiten des REST-Servers wie auch auf Seiten des REST-Clients benötigt werden. Hierzu gehört beispielsweise das Parsen und Serialisieren von XML-Dateien. Diese Funktion wurde mittels JAXB[12] implementiert.
- **Common.** Enthält allgemeine Hilfsfunktionen, die im gesamten System benötigt werden. Hier befinden sich unter anderem typische ‚Utility'-Funktionen, zum Beispiel um mit `BibTEX`-Objekten oder Zeichenketten zu arbeiten. Des Weiteren sind an dieser Stelle mehrere Konstanten definiert, auf die von allen anderen Modulen aus zugegriffen werden kann.
- **Scraper.** Extrahiert automatisch die Metadaten aus bekannten elektronischen Bibliotheksseiten. Enthält auch ein Modul zur automatischen Extraktion von Metadaten aus Text, welcher z. B. per Copy-and-Paste aus einer Webseite eingefügt werden kann. Diese Funktionalität nimmt dem Benutzer die (etwas mühsame) Aufgabe ab, alle Metadaten wie Autor, Titel, Seitenzahl, etc. zu einer Publikation manuell über ein Formular einzugeben. Stattdessen reicht ein Klick auf der entsprechenden Bibliotheksseite und das Scraper-Modul führt den jeweiligen Extraktionsvorgang automatisch aus.
- **DL-Importer.** Dient dazu, größere Metadaten-Bestände aus bekannten Bibliotheken zu importieren. Zur Zeit wird über diese Komponente der Import von DBLP[13] abgewickelt. Um die Aktualität der Metadaten zu gewährleisten, wird dieser Import regelmäßig durchgeführt.
- **`BibTEX`-Parser.** `BibTEX` stellt ein zentrales Import- und Export-Format des Systems dar. Diese Komponente ist dafür zuständig, `BibTEX`-

[12] Java Architecture for XML Binding: https://jaxb.dev.java.net/
[13] http://www.informatik.uni-trier.de/ley/db/

Einträge oder ganze BibTEX-Dateien in Java-Objekte zu parsen, um sie dann weiterverarbeiten zu können. Darüber hinaus wird dieses Modul dazu verwendet, BibTEX-Einträge auf syntaktische Korrektheit zu überprüfen.

Das Datenbankschema (Abb. 6) von BibSonomy basiert im Wesentlichen auf vier Tabellen: je eine für Lesezeichen- und Literaturbeiträge (posts), eine für die Schlagwort-Zuordnungen (tas) und eine für die Relationen (relations). Zwei weitere Tabellen speichern Informationen über Benutzer (users) und Gruppen (groups). Die beiden posts-Tabellen für Lesezeichen- beziehungsweise Literatureinträge verhalten sich sehr ähnlich (und sind daher in Abb. 6 zusammengefasst), die Literatureintrags-Tabelle hat lediglich einige zusätzliche Spalten, um die BIBTEX-Felder aufzunehmen. In der Datenbank sind sie aus Gründen der Effizienz getrennt, da diese zusätzlichen Spalten nur für Publikationen gespeichert werden müssen.

Die posts-Tabellen sind mit der tas-Tabelle durch den Schlüssel post_id verbunden. Das Schema ist nicht normalisiert, es wurde im Gegenteil sogar viel Redundanz hinzugefügt, um Abfragen zu beschleunigen. Zum Beispiel speichern wir Gruppe, Benutzername und Datum nicht nur

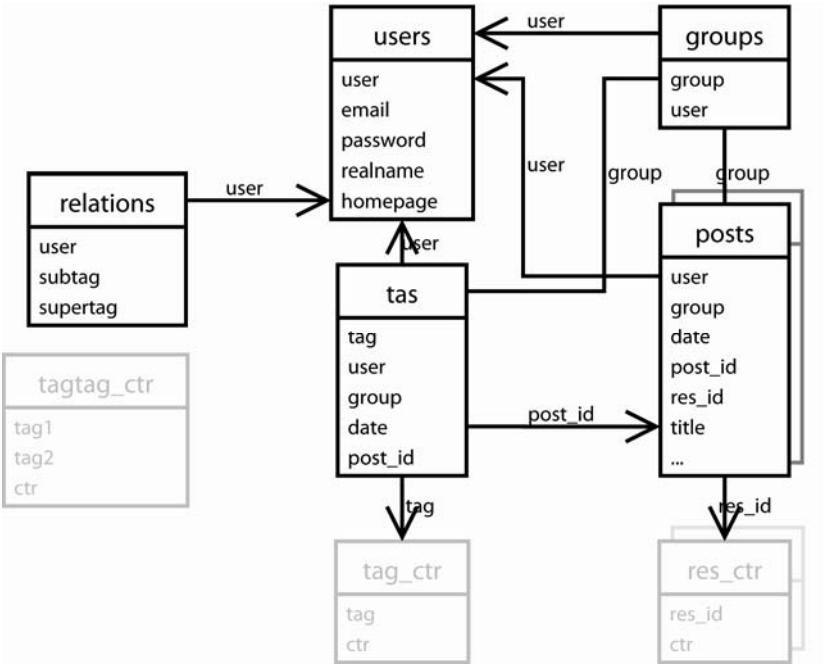

Abb. 6. Beziehungsschema der wichtigsten Tabellen

in den `posts`-Tabellen, sondern auch in der `tas`-Tabelle, um bei der Anfrageabarbeitung die Anzahl der benötigten Joins klein zu halten. Außerdem wurden viele Zähler materialisiert (beispielsweise die Anzahl der Einträge für eine Ressource oder die Häufigkeit der Verwendung eines Schlagwortes), und sehr viele Indizes (allein zwölf in der `tas`-Tabelle) eingefügt, um die Antwortzeiten klein zu halten.

Insgesamt verbrachten wir viel Zeit damit, das Datenbank-Schema und die SQL-Anfragen zu optimieren, und testeten beides mit Folksonomie-Daten mit bis zu 8.000.000 Einträgen. Da das System gut skaliert, benötigen wir momentan keine besondere Pufferung oder physische Verteilung der Datenbank, um angemessene Antwortzeiten zu erhalten. Eine Verteilung von Anfragen über synchronisierte Datenbanken ist mit MySQL möglich.

Ein zentraler Bestandteil des aktuellen BibSonomy-Systems ist die REST-API (Details in Abschn. 3). Ihre wesentlichen Komponenten sind der REST-Server und der REST-Client (siehe Abb. 4). Der REST-Server stellt dabei über definierte URLs den Zugriff auf die Daten von BibSonomy bereit. Die Daten werden in XML ausgegeben. Das Schema entspricht dem internen Modell und erlaubt das automatische Erzeugen von Java Objekten aus XML mittels der ‚Java Architecture for XML Binding', kurz JAXB.[14]

Semantische Funktionalitäten und Interoperabilität

Dieser Abschnitt beschreibt einige semantische Erweiterungen und Funktionen zur Integration von BibSonomy in andere Systeme, die nicht Teil des Grundsystems sind, sich aber als notwendig für die tägliche Nutzung von BibSonomy erwiesen haben.

Tag-Hierarchie

Das ‚Tagging' hat in den letzten zwei Jahren eine solche Popularität erlangt, weil es einfach ist und keine besonderen Fertigkeiten voraussetzt. Bei wachsender Datenmenge kommt dann jedoch schnell das Bedürfnis auf, die (eigene) Schlagwortsammlung stärker strukturieren zu können. Eine Schlagwort-Hierarchie ($\prec$ in unserem Folksonomie-Modell in Abschn. 2) ist eine einfache Möglichkeit, Tags zu ordnen.

Um das Hinzufügen von Elementen zur Relation bereits während des Verschlagwortens zu ermöglichen, wurden die Zeichenfolgen `<-` und `->` reserviert. Wenn der Benutzer u die Folge $t_1 -> t_2$ eingibt, ordnen wir

[14] https://jaxb.dev.java.net/

die Tags t_1 und t_2 dem entsprechenden Eintrag zu und fügen das Tripel (u,t_1,t_2) zur Relation $\prec$ hinzu. Die Eingabe $t_2 < -t_1$ wird als $t_1 -> t_2$ ausgewertet. Dabei wird $t_1 -> t_2$ gelesen als „t_1 *ist ein* t_2" oder „t_1 ist ein Unterschlagwort des Schlagworts t_2". Es gibt auch andere Möglichkeiten, Elemente zu $\prec$ hinzuzufügen, insbesondere mit einem Relationen-Editor.

Relationen werden in verschiedenen Situationen genutzt. Erstens kann der Benutzer seine Schlagwort-Wolke strukturieren, indem er alle Unterschlagwörter eines bestimmten Schlagworts anzeigen lässt und dadurch die Schlagwörter in einer Hierarchie sehen kann. Zweitens bietet BibSonomy die Möglichkeit, auf der Schlagwort-Seite eines Benutzers nicht nur die einem Schlagwort zugeordneten Einträge anzuzeigen, sondern auch die Einträge, die einem seiner Unterschlagwörter zugeordnet wurden. Dies funktioniert auch bei der Kombination von Schlagwörtern: Für die Schlagwörter $t_1,...,t_n$ und einen Benutzer $u \in U$ erhält man eine Seite mit den in Gl. 5 in Abschn. 3 beschriebenen Einträgen. Weiterhin ist es möglich, die Aggregation über alle Unterschlagwörter nicht nur benutzerspezifisch, sondern über alle Benutzer im System anzeigen zu lassen. Eine Übersicht der am häufigsten verwendeten Relationen findet man auf der Seite `/relations`. Verglichen mit den in del.icio.us erhältlichen *tag-bundles* sind Relationen in BibSonomy allgemeiner und ausdrucksstärker.

Interoperabilität mit anderen Semantic-Web-Diensten

Eines der Ziele des Semantic Web ist die Nutzung von verteilten, maschinenlesbaren und -nutzbaren Daten. Ein aktueller Schritt in diese Richtung besteht im Erstellen eines großen, verteilten Datenrepositories (Linked Data on the Web)[15], welches seinerseits die Daten in verschiedensten Formaten zur Verfügung stellt. Die unterschiedlichen Datenquellen sind miteinander verlinkt und können als ein gemeinsamer großer Datensatz angesehen werden. Dieses Ziel verfolgt die W3C-Gruppe „Semantic Web Education and Outreach". Ende Oktober 2007 bestand der Datensatz aus mehr als zwei Milliarden RDF-Tripeln.

Für den Zusammenschluss dieser heterogenen, verteilten Daten müssen die Daten in verschiedenen Formaten vom Inhaltsanbieter zur Verfügung gestellt werden, damit sowohl Menschen als auch Maschinen mit den Daten arbeiten können. Weiterhin muss der Inhaltsanbieter zwischen Menschen und Maschinen als Anfragern unterscheiden und zur Anfragezeit die Daten passend ausliefern. In BibSonomy wurden erste Schritte umgesetzt, um an der Linked Data Initiative teilnehmen zu können.

[15] http://www.w3.org/DesignIssues/LinkedData.html und ein Tutorial unter: http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/

BibSonomy stellt die Daten nicht nur in verschiedenen HTML-Formaten und in `BibTEX` zur Verfügung, sondern implementiert auch einen Reihe von RDF-Ausgabe-Formaten. Die bekanntesten Formate sind RSS, SWRC und BURST. Bei SWRC handelt es sich um eine RDF-Ausgabe passend zur SWRC-Ontologie, die in [27] definiert wurde. BURST ist eine Ontologie, die SWRC, DC und RSS zusammenfasst.[16] Ausgabeformate geben nur die passenden Literaturreferenzen aus. Sie lassen sich mit allen inhaltlichen URLs kombinieren (siehe Abschn. 3). Diese Seiten können direkt zum Linked Data Repository hinzugefügt werden. Sie sind allerdings für den Menschen sehr schwer verständlich.

Die Auswahl des passenden Ausgabe-Formates ermöglicht die URL-Erweiterung `/uri/`. Sie kann mit ausgewählten BibSonomy -URLs kombiniert werden. Beim Aufruf dieser erweiterten URL wird ein Header erwartet, der das Ausgabeformat spezifiziert. Als Antwort bekommt man nicht den eigentlichen Inhalt, sondern einen Verweis auf die BibSonomy -Seite, die den Inhalt im passenden Format ausgibt, beispielsweise für den Web-Browser HTML und für eine Semantic-Web-Applikation RDF. Dieser Vorgang ist in der Literatur unter dem Begriff „Content Negotiation“ bekannt und wird über den Statuscode 303 des Web-Servers, der eine Weiterleitung auf eine andere Seite anzeigt, realisiert.

REST-API

Die Erfahrung hat gezeigt, dass eine Anwendungs-Programmier-Schnittstelle (API) entscheidend für den Erfolg eines Systems ist. Sie wurde von vielen BibSonomy-Anwendern gefordert und ermöglicht ein gutes Zusammenspiel zwischen BibSonomy und anderen Systemen. Die meisten Systeme verwenden leichtgewichtige APIs, die der Idee von REST (Representational State Transfer) [3] ähneln. Diese können auch von nicht so erfahrenen Programmierern verwendet werden, da ihre Funktionalität leicht verständlich ist. Daher wurde REST dem Standard SOAP[17] vorgezogen. Für BibSonomy wurde eine REST-API implementiert. Sie ist für alle Nutzer unter `/api/` zugreifbar.

Die prinzipiellen Eigenschaften der REST-API fußen auf dem internen Datenmodell von BibSonomy, welches mittels JAXB in XML serialisiert werden kann. Fragt man Daten über die API an oder möchte neue Daten über die API in das System laden, so werden die Daten gemäß des Modells in XML abgelegt.

[16] http://www.cs.vu.nl/pmika/research/burst/BuRST.html
[17] http://www.w3.org/TR/soap/

Das Konzept der REST-API nutzt die typischen HTTP-Verben GET, PUT, POST und DELETE, um Aktionen auf den URLs durchzuführen. Dabei wurden für alle API-URLs entsprechende Namen gewählt. So kann man z. B. mit einem GET-Befehl unter `/api/tags` die Liste aller Schlagwörter des Systems erfragen oder mittels POST über `/api/users/[username]/posts` einen neuen Eintrag ins System einstellen. Details zur API findet man auf den Hilfeseiten.[18]

Für den einfachen Zugriff auf die API steht ein Java-Client zur Verfügung, der einen einfachen Zugriff auf die BibSonomy -Daten über Java-Objekte gemäß dem definierten Modell erlaubt. Als weitere Beispielanwendung der API wurde BibSonomy in die Literaturverwaltungs-Software JabRef integriert, siehe Kap. 3.

Anwendungen

Die Möglichkeiten, Literatur-Referenzen in BibSonomy zu verwalten, thematisch zu sortieren und mit Kollegen auszutauschen, werden ergänzt durch einige nützliche Erweiterungen, die wir hier kurz an Beispielen vorstellen.

Erzeugung von Publikationslisten für das Web

Die in den Abschnitten 1 und 2 diskutierten Ausgabeformate können genutzt werden, um Publikationslisten für verschiedene Zwecke zu generieren. So verschlagworten beispielsweise alle Mitglieder unserer Arbeitsgruppe ihre eigenen Publikationen mit *myown*, um ihre privaten Publikationslisten[19] und die des Fachgebiets[20] aktuell und synchron zu halten. Auch eine persönliche Schlagwort-Wolke[21] kann so aus BibSonomy übernommen werden.

Publikationslisten für Projekte können auf die gleiche Weise aus demselben Datenbestand generiert werden. So werden beispielsweise die Publikationslisten der Europäischen Projekte ‚Nepomuk – The Social Semantic Desktop'[22] und ‚TAGora – Emergent Semiotics in Online Social Communities'[23] aus BibSonomy generiert. Für das Nepomuk-Projekt wird

[18] http://www.bibsonomy.org/help/doc/api.html
[19] Zum Beispiel http://www.kde.cs.uni-kassel.de/stumme/publications.html
[20] http://www.kde.cs.uni-kassel.de/pub
[21] Siehe beispielsweise http://www.kde.cs.uni-kassel.de/stumme
[22] http://nepomuk.semanticdesktop.org/xwiki/bin/view/Main1/Publications
[23] http://www.tagora-project.eu/publications/

zusätzlich der Teil des für die Europäische Kommission erforderlichen Berichtswesens, der publikationsbezogen ist, über ein zusätzlich implementiertes, projektintern erreichbares Formular abgewickelt.

Auch für Lehrveranstaltungen kann BibSonomy genutzt werden. So wurden beispielsweise für das Seminar ‚Online-Communities und Web 2.0'[24] die Literaturreferenzen unter dem Schlagwort *seminar2006* online bereit gestellt; und die Studierenden wurden aufgefordert, diese durch weitere Referenzen zu ergänzen. Auch die Publikationslisten für Vorlesungen werden auf diese Weise erstellt.[25]

Je nach verwendetem Webserver und Content-Management-System bieten sich unterschiedliche Formen des Zugriffs auf BibSonomy an. Die oben genannten Beispiele für private und Fachgebietswebseiten laufen etwa auf einem ZOPE-System[26], das die Publikationslisten direkt im HTML-Format von BibSonomy importiert. Die Nepomuk-Publikations-Webseite wird durch XWiki[27] generiert; die Daten werden hierzu von BibSonomy als RSS-Feed übertragen. Die TAGora-Seite wird mit WordPress erstellt; auch hierfür werden die Publikationsdaten per RSS übertragen.

BibSonomy kann auch verwendet werden, um für Tagungen die Sammlung der Publikationen zu den Vorträgen navigierbar zu machen. So wurden beispielsweise für die 23. IUPAP International Conference on Statistical Physics[28] und für die 6. International Semantic Web Conference[29] alle Einträge der Tagungsbände bereits vor der Veranstaltung in BibSonomy eingeladen, auf die über speziell für die Tagung erzeugte Schlagwort-Wolken zugegriffen werden kann. Je nach Publikations-Strategie der Tagung ist auf diese Weise auch der direkte Zugriff auf Online-Proceedings denkbar.

Merklisten für Bibliotheksrecherchen und E-Learning-Anwendungen

Für die regelmäßige Arbeit mit einem Online-Bibliothekskatalog ist ein ‚Merkzettel' eine hilfreiche Funktion. Der Kölner Universitäts-Gesamtkatalog[30] hat daher auf allen Seiten, auf denen die Treffer zu den Kataloganfragen angezeigt werden, einen Link eingerichtet, der es dem Benutzer

[24] http://www.kde.cs.uni-kassel.de/lehre/ss2006/online_communities
[25] Siehe beispielsweise http://www.kde.cs.uni-kassel.de/lehre/ws2007-08/IR
[26] http://zope.org/
[27] http://www.xwiki.org/
[28] http://www.bibsonomy.org/events/statphys23
[29] http://www.bibsonomy.org/events/iswc2007
[30] http://kug.ub.uni-koeln.de/

erlaubt, die gefundenen Publikationsdaten direkt in BibSonomy abzuspeichern. Dieselbe Funktionalität bietet die E-Learning-Plattform eduCampus[31] der Universität Kassel. Umgekehrt kann man in BibSonomy durch Anklicken des Knopfes ‚OpenURL' hinter einem Eintrag direkt nach der entsprechenden Publikation in seiner Heimat-Bibliothek suchen, wenn diese einen OpenURL-Server betreibt und dieser auf der Seite der persönlichen Einstellungen in BibSonomy vermerkt ist.

Lokale Verwaltung von Literaturdaten mit JabRef

JabRef[32] ist ein Java-Tool zum Verwalten von Literaturdaten. Es wurde eine spezielle BibSonomy-Version[33] entwickelt, die den Im- und Export von Referenzen in und aus JabRef über die API-Schnittstelle ermöglicht. Damit steht allen JabRef-Nutzern eine einfache Schnittstelle zu BibSonomy zur Verfügung, so dass sie ihre Daten sowohl lokal als auch auf dem BibSonomy-Server speichern und verwalten können. Die Literatureinträge können somit auch im Offline-Modus genutzt werden.

Wissensmanagement in Unternehmen

Die leichte Bedienbarkeit von kooperativen Verschlagwortungssystemen macht sie zu viel versprechenden Kandidaten für Wissensmanagement-Anwendungen in einem kommerziellen Umfeld, insbesondere in Bereichen, in denen sich strukturierte Intranetlösungen noch nicht etablieren konnten, oder die so schnelllebig sind, dass die regelmäßige Pflege der Bestände durch Wissensingenieure zu kostspielig ist. Dies trifft insbesondere auf Anwendungen zu, die von Wissensmodellierungslaien bedient werden müssen. Da Intranet-Inhalte häufig aus Office-Dokumenten bestehen, die in der Regel keine Navigation zu verwandten Inhalten ermöglichen, können Folksonomien hier als Mittel zur Strukturierung helfen. Diese Vorteile in Bezug auf die Erstellung und insbesondere Wartung von Wissensmanagement-Systemen im Intranet haben auch große Firmen erkannt. So erwägt z. B. IBM die Benutzung von Folksonomien im Intranet, und auch [18] diskutiert eine solche Anwendung.

Aus demselben Grund haben wir in einem Industrieprojekt einen auf BibSonomy basierenden Prototyp entwickelt. Die gemeinsame Anmeldung an BibSonomy und das im Unternehmen eingesetzte SAP-Portal erfolgt

[31] http://educampus.uni-kassel.de/
[32] http://jabref.sourceforge.net/
[33] http://www.bibsonomy.org/help/doc/jabrefclient.html

per Single-Sign-On. Für den internationalen Einsatz wurde BibSonomy um die Möglichkeit erweitert, die Inhalte in verschiedenen Sprachen anzubieten. Zur Erhöhung der Akzeptanz wurde darüber hinaus eine automatische Validierung der Links aller vorhandenen Lesezeichen eines Benutzers implementiert und es wurden integrierte Bookmarklets zur Unterstützung der Navigation über die Systemgrenzen hinweg erstellt. Um Nutzer des Intranets auf den neuen Dienst aufmerksam zu machen, wurde in jede Seite des Intranets ein Link zum automatischen Speichern der Seite in BibSonomy integriert. Mit Hilfe der iViews-Funktionalität des SAP-Portals werden darüber hinaus die Top-5-Schlagwörter eines Nutzers sowie die letzten fünf von ihm eingestellten Einträge angezeigt. Ziel ist auch hier, die Sichtbarkeit und den Austausch von Informationen im Portal zu erhöhen.

Datenanalyse, Information Retrieval und Wissensentdeckung

Ein wesentlicher Grund für uns, BibSonomy zu entwickeln und zu betreiben, ist, dass das System praxisrelevante Fragestellungen der Wissensverarbeitung im Web 2.0 aufzeigt und gleichzeitig Möglichkeiten zu deren Evaluierung bietet. Wir haben in den letzten beiden Jahren verschiedene Fragestellungen aus den Bereichen Datenanalyse, Information Retrieval und Wissensentdeckung auf kooperative Verschlagwortungssysteme übertragen und bearbeitet. In diesem Kapitel stellen wir in Kürze die Ansätze vor, die auf Daten von BibSonomy angewandt wurden oder bereits in das System integriert sind. Für Details verweisen wir auf die jeweiligen Referenzen.

Zu allen Ansätzen gibt es verwandte – aber nicht auf BibSonomy bezogene – Forschungsarbeiten. Es würde jedoch zu weit führen, diese hier zu diskutieren. Die jeweils zitierten Arbeiten beinhalten entsprechende weiterführende Übersichten. Publikationen über kooperative Verschlagwortungssysteme sind auch in BibSonomy unter dem Schlagwort *folksonomies*[34] zu finden.

Analyse der Folksonomie-Struktur

Analysen der Netzwerkeigenschaften der Folksonomie von BibSonomy [2, 24] haben gezeigt, dass sie die so genannten *Small-World*-Eigenschaften aufweist: ein kleiner Graph-Durchmesser (d. h. die Pfade zwischen je zwei Einträgen sind im Mittel kurz) und ein hoher Clustering-Koeffizient. Dies bedeutet, dass es zum einen thematische Gruppierungen

[34] http://www.bibsonomy.org/tag/folksonomies

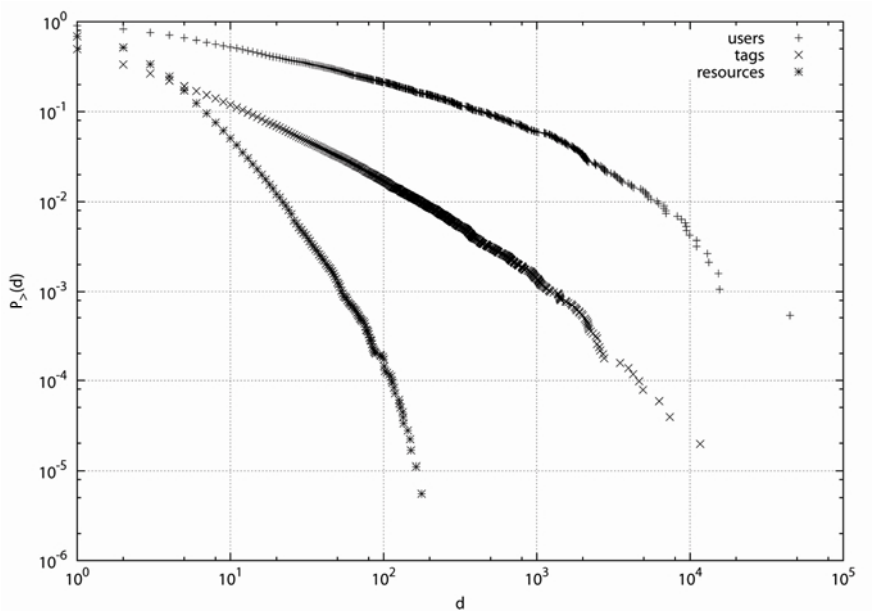

Abb. 7. Kumulierte Häufigkeiten der Nutzer, Schlagwörter und Ressourcen in BibSonomy. Auf der *x*-Achse ist die Anzahl *d* der Schlagwortzuweisungen (tag assignments) abgetragen, und auf der *y*-Achse die relative Häufigkeit der Nutzer resp. Schlagwörter resp. Ressourcen, die in mehr als *d* Zuweisungen auftreten. Beide Achsen sind logarithmisch skaliert

der Einträge gibt, dass zum anderen aber eine erkundende Navigation durch das System sehr effizient ist, da man mit nur wenigen Clicks von einem Thema zu jedem anderen kommt. Dass die implizite Semantik der Schlagwörter einen strukturierenden Einfluss auf die Folksonomie hat, konnte durch eine Analyse der Gewichtsverteilungen im Schlagwort-Nachbarschaftsgraphen nachgewiesen werden.

Abbildung 7 zeigt darüber hinaus, dass die Häufigkeiten der Nutzer, Schlagwörter und Ressourcen in BibSonomy extrem schief verteilt sind: Einige wenige Elemente (Nutzer/Schlagwörter/Ressourcen) sind sehr stark vernetzt, fast alle Elemente jedoch sehr schwach. Damit hat BibSonomy (wie auch andere Folksonomies, siehe etwa [8, 2]) eine Struktur, die auch typisch ist für andere Netze, die aus sozialen Aktivitäten entstehen.

Ranking von Suchergebnissen

Die meisten kooperativen Verschlagwortungssysteme sortieren ihre Trefferlisten in umgekehrt chronologischer Reihenfolge: Einträge, die zuletzt

eingestellt wurden, werden am weitesten oben gelistet. Del.icio.us bietet darüber hinaus ein Ranking an, bei dem die Einträge, die in der letzten Zeit am häufigsten eingestellt wurden, die Liste anführen.

In [8] haben wir den relevanzbasierten FolkRank-Algorithmus vorgestellt. Ein Eintrag wird als relevant betrachtet, wenn er von vielen relevanten Nutzern mit vielen relevanten Schlagwörtern versehen wurde. Die Definition der Relevanz von Nutzern und Schlagwörtern erfolgt rekursiv auf die gleiche Weise. Diese Anpassung des von Google bekannten, auf Eigenvektor-Zentralität basierenden PageRank-Algorithmus [1] auf Folksonomies erlaubt es, für eine gegebene Anwenderpräferenz in einem Schritt ein Ranking für Schlagwörter, Anwender und Ressourcen zu erstellen. Das Verfahren ist in BibSonomy für die Schlagwort-Seiten implementiert.[35]

Entdeckung von Communities

Neben organisatorischen Gruppen entstehen in kooperativen Verschlagwortungssystemen thematische Gruppen (Communities of Interest). Eine Identifikation solcher Gruppen böte die Möglichkeit, themenspezifisch auf andere Nutzer zu verweisen und so die Navigation durch das System zu lenken oder auf spannende Inhalte aufmerksam zu machen. Gruppen werden in der Regel automatisch mittels Clusterverfahren identifiziert, die aber in diesem Fall auf die spezielle Struktur der Daten angepasst werden müssen.

Ein Ansatz hierzu ist eine triadische Variante [15] der Formalen Begriffsanalyse [5]. In [9] haben wir TRIAS, eine triadische Version des Next-Closure-Algorithmus [4, 13], vorgestellt, der es ermöglicht, Gruppen von Benutzern zu bestimmen, die alle dieselben Ressourcen mit denselben Schlagwörtern verschlagwortet haben. Wir haben die BibSonomy-Publikationsdaten mit dem TRIAS-Algorithmus analysiert [10]. Der Datensatz enthält alle Publikationen, die bis zum 26. November 2006 eingegeben wurden. Ausgenommen wurden die Einträge der DBLP Computer Science Bibliography,[36] die von BibSonomy gespiegelt wird. Analysiert wurden so Daten von 262 Nutzern, die 11.101 Publikationen mit 5.954 verschiedenen Schlagwörtern versehen haben.[37]

Ein *Tri-Begriff* besteht aus einer Menge A von Benutzern, einer Menge B von Schlagwörtern und einer Menge C von Ressourcen, so dass jeder Benutzer aus A jede Ressource aus C mit jedem Schlagwort aus B versehen hat und keine der drei Mengen vergrößert werden kann, ohne diese

[35] Siehe beispielsweise http://www.bibsonomy.org/tag/ontology?order=folkrank
[36] http://www.informatik.uni-trier.de/ley/db/
[37] BibSonomy-Benchmark-Datensätze sind für wissenschaftliche Zwecke auf Anfrage erhältlich, siehe http://www.bibsonomy.org/faq

Bedingung zu verletzen. In den Publikationsdaten finden sich insgesamt 13.992 solche Tri-Begriffe. Abbildung 8 zeigt alle 21 Tri-Begriffe der BibSonomy-Publikationsdaten mit mindestens drei Nutzern, zwei Schlagwörtern und zwei Ressourcen. Die Publikationstitel wurden aus Platzgründen durch Nummern ersetzt und können unter `http://www.bibsonomy.org/group/kde/trias_example` nachgeschlagen werden; sie sind in lexikographischer Ordnung der Titel nummeriert.

Die 21 Punkte in der Mitte des Diagramms repräsentieren die 21 häufigen Tri-Begriffe. Die Mengen von Nutzern, Schlagwörtern und Ressourcen, die jeweils einen Tri-Begriff konstituieren, können an den drei Seiten des Diagramms abgelesen werden. Ein Knoten in einer der drei Hierar-

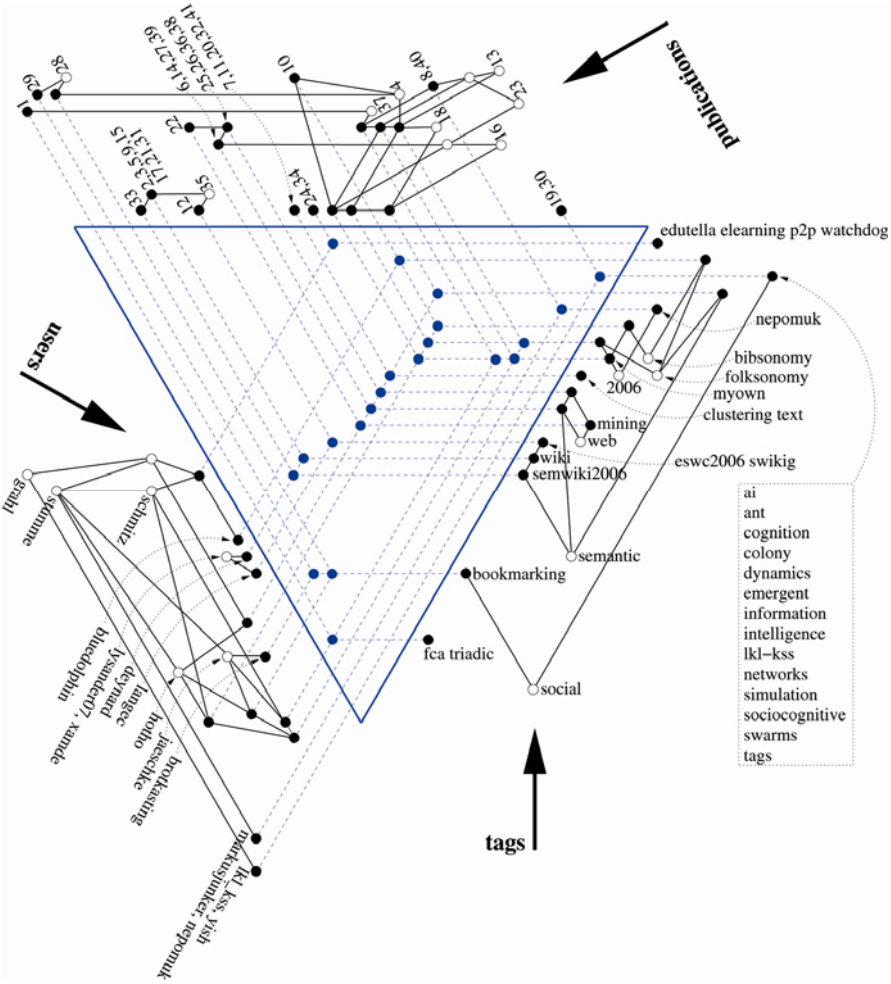

Abb. 8. Die häufigen Tri-Begriffe in den BibSonomy-Publikationsdaten

chien beinhaltet immer alle Nutzer, Schlagwörter bzw. Ressourcen, die direkt an ihm dran stehen, sowie die, die im Sinne der Pfeilrichtung unter ihm liegen. Dies bedeutet beispielsweise in der Schlagworthierarchie, dass der äußerste rechte Knoten neben den Schlagwörtern *ai*, …, *tags* auch das Schlagwort *social* enthält. Ein Tri-Begriff ist durch seine drei Projektionen eindeutig identifiziert. So repräsentiert zum Beispiel der unterste Punkt im Inneren des Diagramms den Tri-Begriff, der aus der Menge {*jaeschke, schmitz, stumme*} von Nutzern, der Menge {*fca, triadic*} von Schlagwörtern und der Menge {*1, 37*} von Ressourcen (d. h. [26, 9]) besteht.

Die Schlagworthierarchie gibt einen Hinweis auf die zentralen Themen, die im November 2006 durch die Publikationssammlung in BibSonomy abgedeckt wurde. (Inzwischen hat sich die Anzahl der Einträge versechsfacht und thematisch verbreitert.) Das Schlagwort *social* tritt am häufigsten mit anderen Schlagwörtern auf. Dieses Schlagwort wurde zusammen mit den Schlagwörtern *ai* [= Artificial Intelligence], …, *tags* durch die Benutzer *lkl_kss* und *yish* den Publikationen *19* und *30* [22, 21] zugeordnet, zusammen mit dem Schlagwort *bookmarking* durch die Benutzer *hotho, jaeschke, stumme* den Publikationen *4* und *28* [7, 6] zugeordnet, und wieder zusammen mit dem Schlagwort *bookmarking* durch die Benutzer *brotkasting, jaeschke, stumme* den Publikationen *28* und *29* [6, 16] zugewiesen. Dies deutet darauf hin, dass {*lkl_kss, yish*} und {*brotkasting, hotho, jaeschke, stumme*} zwei Communities of Interest bilden, die sich beide mit sozialen Phänomenen im Web 2.0 beschäftigen, aber mit unterschiedlicher Ausrichtung. Eine zweite Gruppe von Themen wird durch das Schlagwort *semantic* aufgespannt, welches in drei verschiedenen Zusammenhängen auftritt: semantische Wikis, Semantic Web Mining und im Zusammenhang mit Folksonomien.

Eine detaillierte Analyse dieses Datensatzes ist in [10] beschrieben.

Empfehlungen von Schlagwörtern

Die häufigste Anwendung von Recommendern im Social Bookmarking ist die Empfehlung von Schlagwörtern für aktuell abzuspeichernde Webseiten. Solche Systeme schlagen dem Benutzer meistens mehrere Schlagwörter vor. In BibSonomy ist derzeit ein erster Collaborative-Filtering-Ansatz eingebaut, basierend auf [25]. In [11] haben wir den FolkRank-Algorithmus für die Vorhersage von Schlagwörtern analysiert. Wir konnten empirisch eine deutliche Verbesserung der Performanz (gemessen in Precision und Recall) gegenüber inhaltsbasierten und CF-Ansätzen nachweisen. Dieselbe Güte wird außerdem (fast) durch eine Kombination von Nutzer- und Ressourcen-basierten Vorschlägen erreicht, die gegenüber dem Folk-

Rank-Algorithmus den Vorteil einer geringeren Komplexität zur Laufzeit hat. Die Implementierung dieses Ansatzes in BibSonomy ist geplant.

Ausblick

Mit der Kombination von Lesezeichen und Literaturreferenzen ist BibSonomy als Unterstützung für die tägliche Arbeit von Wissenschaftlern konzipiert, da diese typischerweise eine große Menge von Informationsquellen organisieren müssen. Wenn kooperative Verschlagwortungssysteme wie BibSonomy wachsen, muss die Benutzerunterstützung über einfache Funktionalitäten wie beispielsweise Volltextsuche hinausgehen. Dies bedingt eine verbesserte Organisation der Daten. Ein offensichtlicher Ansatz hierfür sind Technologien des Semantic Web. Die entscheidende Frage bleibt jedoch, wie man deren Vorteile nutzt, ohne ungeübte Benutzer mit dem Formalismus zu behelligen. Wir sind sicher, dass dieses Thema ein fruchtbares Forschungsgebiet für die Semantic Web-Community in den kommenden Jahren ist.

Danksagung

Die Entwicklung von BibSonomy wurde teilweise durch die Projekte „Nepomuk – the Social Semantic Desktop“ (FP6-027705) und „TAGora – Emergent Semiotics in Online Social Communities“ (FP6-IST5-34721) der Europäischen Kommission sowie dem durch Microsoft Research finanzierten Projekt „Social Search: Bringing the Social Component to the Web“ finanziert.

Literatur

1. 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, April 1998.
2. Ciro Catutto, Christoph Schmitz, Andrea Baldassarri, Vito D. P. Servedio, Vittorio Loreto, Andreas Hotho, Miranda Grahl, and Gerd Stumme. Network properties of folksonomies. AI Communications Journal, Special Issue on ”Network Analysis in Natural Sciences and Engineering” (im Druck), 2007.
3. Roy T. Fielding. Architectural Styles and the Design of Network-based Software Architectures. PhD thesis, University of California, Irvine, 2000.
4. 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.

5. Bernhard Ganter and Rudolf Wille. Formale Begriffsanalyse: Mathematische Grundlagen. Springer, Heidelberg, 1996.
6. Tony Hammond, Timo Hannay, Ben Lund, and Joanna Scott. Social Bookmarking Tools (I): A General Review. D-Lib Magazine, 11(4), April 2005.
7. Andreas Hotho, Robert J¨aschke, Christoph Schmitz, and Gerd Stumme. BibSonomy: A social bookmark and publication sharing system. In Proceedings of the Conceptual Structures Tool Interoperability Workshop at the 14th International Conference on Conceptual Structures, pages 87–102, 2006.
8. Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Information retrieval in folksonomies: Search and ranking. In York Sure and John Domingue, editors, The Semantic Web: Research and Applications, volume 4011 of LNAI, pages 411–426, Heidelberg, June 2006. Springer.
9. Robert Jäschke, Andreas Hotho, Christoph Schmitz, Bernhard Ganter, and Gerd Stumme. Trias – an algorithm for mining iceberg tri-lattices. In Proceedings of the 6th IEEE International Conference on Data Mining (ICDM 06), pages 907–911, Hong Kong, December 2006. IEEE Computer Society.
10. Robert Jäschke, Andreas Hotho, Christoph Schmitz, Bernhard Ganter, and Gerd Stumme. Discovering shared conceptualizations in folksonomies. Journal of Web Semantics, 2008 (to appear).
11. Robert Jäschke, Leandro Balby Marinho, Andreas Hotho, Lars Schmidt-Thieme, and Gerd Stumme. Tag recommendations in folksonomies. In Joost N. Kok, Jacek Koronacki, Ramon López de Mántaras, Stan Matwin, Dunja Mladenic, and Andrzej Skowron, editors, Knowledge Discovery in Databases: PKDD 2007, 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, volume 4702 of Lecture Notes in Computer Science, pages 506–514, Berlin, Heidelberg, 2007. Springer.
12. G. E. Krasner and S. T. Pope. A cookbook for using the model-view controller user interface paradigm in Smalltalk-80. Journal of Object Oriented Programming, 1(3):26–49, 1988.
13. 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.
14. Leslie Lamport. LaTeX: A Document Preparation System. Addison-Wesley, 1986.
15. 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.
16. Ben Lund, Tony Hammond, Martin Flack, and Timo Hannay. Social Bookmarking Tools (II): A Case Study – Connotea. D-Lib Magazine, 11(4), April 2005.
17. Andrew Kachites McCallum. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu, 2002.
18. David R. Millen, Jonathan Feinberg, and Bernard Kerr. Dogear: Social bookmarking in the enterprise. In CHI ’06: Proceedings of the SIGCHI conference

on Human Factors in computing systems, pages 111–120, New York, NY, USA, 2006. ACM Press.
19. Oren Patashnik. BibTeXing, 1988. (Included in the BIBTEX distribution).
20. Fuchun Peng and Andrew McCallum. Accurate information extraction from research papers using conditional random fields. In HLT-NAACL, pages 329–336, 2004.
21. Vitorino Ramos, Carlos Fernandes, and Agostinho C. Rosa. Social cognitive maps, swarm collective perception and distributed search on dynamic landscapes. 2005.
22. Vitorino Ramos, Carlos Fernandes, and Agostinho C. Rosa. On self-regulated swarms, societal memory, speed and dynamics. 2006.
23. Katharina Regulski. Aufwand und Nutzen beim Einsatz von Social Bookmarking Services als Nachweisinstrument für wissenschaftliche Forschungsartikel am Beispiel von BibSonomy. http://www.bibliothek-saur.de/preprint/2007/ar2460regulski.pdf, 2007.
24. Elizeu Santos-Neto, Matei Ripeanu, and Adriana Iamnitchi. Tracking user attention in collaborative tagging communities, 2007.
25. Badrul M. Sarwar, George Karypis, Joseph A. Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In Proceedings of the 10th International WWW Conference, pages 285–295, 2001.
26. Gerd Stumme. A finite state model for on-line analytical processing in triadic contexts. In Bernhard Ganter and Robert Godin, editors, Proceedings of the 3rd International Conference on Formal Concept Analysis, volume 3403 of Lecture Notes in Computer Science, pages 315–328. Springer, 2005.
27. York Sure, Stephan Bloehdorn, Peter Haase, Jens Hartmann, and Daniel Oberle. The SWRC ontology – semantic web for research communities. In Carlos Bento, Amilcar Cardoso, and Gael Dias, editors, Proceedings of the 12th Portuguese Conference on Artificial Intelligence – Progress in Artificial Intelligence (EPIA 2005), volume 3803 of LNCS, pages 218–231, Covilha, Portugal, DEC 2005. Springer.