FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Mechanismen und Methoden für die Generierung und Klassifizierung von User Generated Content

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur/in

im Rahmen des Studiums

Software Engineering & Internet Computing

eingereicht von

Martin Trenkwalder
Matrikelnummer 0728267

an der
Fakultät für Informatik der Technischen Universität Wien

Betreuung
Betreuer/in: Prof. Dr. Schahram Dustdar
Mitwirkung: Di. Martin Treiber

Wien, 10.06.2012

(Unterschrift Verfasser/in)

(Unterschrift Betreuer/in)

Abstract

This master thesis focuses on the mechanism and methods for generating and classifying user generated content. First of all different semantic approaches like semantic web, ontologies, topic maps and tags were analysed in detail. In this context, a comparison of Evernote, Delicious, Fickr, Picasa, Zootool and Mister Wong was performed to show their strenghts and weaknesses. After gathering these informations a prototype (Web Content Maps), which is based on folksonomies was developed. The collection of different tags are namend 'Folksonomies'. In this thesis one whole chapter focuses on the different kinds of folksonomies, broad and narrow folksonomies and their advantages and disadvantages. Additionaly, attention to the phenomenon of the long tail curve at the assignment of tags in the broad folksonomy will be drawn. An analysis for information retrieval with the FolkRank-algorithm, which is a customised PageRank-algorithm was done. Whereas the criteria tags, collaboration and prosumer determine the importance of a resource in a folksonomy. Additionally, the same procedure was made for retrieval tools like Paper.li, Flipboard, Tweeted Times and Journal+.

For a better understanding of the process of tagging it will be explained in further detail with an expample, which starts from brainstorming to finally adding the tag. Additionally, the different types of motivation of the user will be shown.

As mentioned above a prototype was developed, which includes a tagging system and a information retrieval tool. For the tagging procedure a Google Chrome Extension was created. A dynamic webpage, which was developed with PHP enables the information retrieval. Furthermore the implementation will be explained from the point of view of the developer and also of the user. Finally, in a forecast new approaches are shown which should eliminate the disadvantages of folksonomies.

Kurzfassung

Diese Arbeit befasst sich mit Mechanismen und Methoden für die Generierung und Klassifizierung von User Generated Content. Basierend auf einer Analyse existenter Konzepten wie dem Semantic Web, Ontologien, Topic Maps und Tags wurde der Tagging Prozess detailliert untersucht. In diesem Zusammenhang wurden existente Systeme, wie Evernote, Delicious, Fickr, Picasa, Zootool und Mister Wong verglichen und deren Stärken und Schwächen aufgezeigt. Ausgehend von der Analyse wurde ein Software Prototyp (Web Content Maps), basierend auf Folksonomies, entwickelt. Folksonomy, eine Sammlung von mehreren Schlagwörtern, wird im Zuge der Arbeit in einem eigenen Kapitel genauer betrachtet um somit die unterschiedlichen Arten von Folksonomies, Broad und Narrow Folksonomy, zu erklären. Dabei werden Vor- und Nachteile aufgezeigt. Zusätzlich wir das Phänomen der Long Tail Kurve bei der Vergabe von Tags in der Broad Folksonomy angeschnitten. Zur Wiederauffindung der getaggten Ressourcen wird der FolkRank-Algorithmus, ein auf Folksonomies angepasster PageRank-Algorithmus, analysiert. Dabei bestimmen die einfließenden Faktoren Tags, Kollaboration und Prosumer die Wichtigkeit einer Ressource in einer Folskonomy. Diesbezüglich wurden bereits existente Information Retrieval Tools für Folskonomies, wie Paper.li, Flipboard, Tweeted Times und Journal+ untersucht und gegenübergestellt.

Zur besseren Verständnis des Vorganges eines Nutzer bei einer Verschlagwortung, wird der Tagging-Prozess anhand eines Beispiels genauer beschrieben, vom Gedanken sammeln bis zum Hinzufügen des Schlagwortes. Zudem wird dabei auf die unterschiedlichen Motivationen der Nutzer eingegangen.

Der oben genannte Prototyp beinhaltet ein Tagging-System und ein Information Retrieval Tool. Für den Taggingvorgang wurde eine Google Chrome Extension erstellt. Das Wiederauffinden der Ressourcen in der Folksonomy erfolgt durch eine mit PHP entwickelte dynamische Webseite. Die Realisierung wird sowohl technisch als auch aus der Anwendersicht erklärt. In dem Ausblick werden neue Lösungsansätze angeführt, die versuchen, die Nachteile von Folksnomies zu beseitigen.

Erklärung

Ich erkläre ehrenwörtlich, dass ich die vorliegende wissenschaftliche Arbeit selbstständig angefertigt und die mit ihr unmittelbar verbundenen Tätigkeiten selbst erbracht habe. Ich erkläre weiters, dass ich keine anderen als die angegebenen Hilfsmittel benutzt habe. Alle ausgedruckten, ungedruckten oder dem Internet im Wortlaut oder im wesentlichen Inhalt übernommenen Formulierungen und Konzepte sind gemäß den Regeln für wissenschaftliche Arbeiten zitiert und durch Fußnoten bzw. durch andere genaue Quellenangaben gekennzeichnet.

(Ort, Datum) (Unterschrift)

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

KAPITEL 1

Einleitung

Das digitale Zeitalter, welches einhergeht mit einer ständig wachsenden Informationsflut von digitalen Daten, stellt die gemeinsame Nutzung und Nachnutzung von Ressourcen immer mehr in den Vordergrund. Die Problematik wie man die großen Mengen an Informationen nach ihrer Wichtigkeit filtert wird in Zukunft an Komplexität gewinnen. Sogar mit der Unterstützung von Suchmaschinen können die nicht textuellen Objekte wie Videos und Musikdateien nur begrenzt aufgefunden werden. Somit müssen die digitalen Objekte auf eine Art und Weise strukturiert werden, um den Zugang zu diesen zu erleichtern und Wichtiges von Unwichtigem zu unterscheiden.

Durch die semantische Anreicherung der Dokumente kann dem entgegengewirkt werden. Dabei werden Detailinformationen bzw. Metainformationen zu einer Ressource bzw. Webseite hinzugefügt, durch welche die Ressource und ihr Inhalt beschrieben werden. Zu den Ressource-Eigenschaften gehören Informationen wie zum Beispiel Titel, Name des Erstellers, Thema, Schlagwörter, persönliche Meinungen, Ortsangaben, usw. Diese Eigenschaften müssen eine definierte Struktur aufweisen, damit sie von einem Computer verarbeitet werden können. Die semantische Anreicherung der Ressourcen dient in erster Linie zur maschinellen Identifizierung, um somit den Inhalt leichter auffindbar und interpretierbar zu machen. Bei einer Suche im Internet sollten die gewünschten Daten gezielt aus der Informationsflut herausgefiltert werden. Dies erweist sich bei der Fülle an Daten meist als schwieriges Unterfangen.

Detailinformationen wurden vor der Einführung von eXtensible Markup Language (XML)[1] in Flat File- und Datenbanken abgespeichert. Diese hatten meist ein proprietäres Format und waren untereinander nicht interoperabel. Sie konnten gewöhnlich nur von den Anwendungen des jeweiligen Herstellers interpretiert und verwendet werden. Durch den Gebrauch von XML werden die Informationen hierarchisch strukturiert und für den Menschen sowie für die Computer lesbar gemacht. Jedoch sagt XML nichts über die Syntaktik und Semantik der Beschreibung der Ressourcen aus.

Für dies wurde das Konzept 'Semantic Web'[2] eingeführt. Das eine Weiterentwicklung des

[1] http://de.wikipedia.org/wiki/Extensible_Markup_Language - Zugegriffen am 21.07.2012

[2] http://www.w3.org/2001/sw/ - Zugegriffen am 20.03.2012

Internets, zum Internet der Dinge darstellt. Für die Verarbeitung der Detailinformationen wird vorausgesetzt, dass die von den Menschen erstellten Informationen automatisch von den Maschinen verstanden und interpretiert werden können. Das Konzept versucht dadurch die fehlende Syntaktik und Semantik in die Webseiten einzubringen. Der Inhalt der Webseiten soll anhand von expliziten Annotationen mit Metadaten für die Maschinen leichter verständlich gemacht werden. Dies bringt den Vorteil, dass sie von den Suchmaschinen sinnvoller indexiert werden und sie somit schneller und leichter gefunden werden können.

Die nötige Syntaktik wird der XML Struktur mittels dem Ressource Description Framework (RDF)[3] hinzugefügt. Es definiert somit die Anordnung der Metadaten. RDF liefert ein einfaches Datenmodell, das Aussagen über Ressourcen anhand RDF-Triples[4] modelliert. Ein Tripel, eine Sequenz aus drei Elementen, besteht jeweils aus einer Ressource (Subjekt), ihrer Eigenschaften (Prädikat) und die Werte der Eigenschaften (Objekt).

Nach der Syntaktik muss den Metainformationen noch die nötige Semantik, die Bedeutung der einzelnen Informationen, hinzugefügt werden. Dies kann durch ein Ressource Description Framework Schema (RDFS)[5] oder einer Web Ontology Language (OWL)[6] erfolgen. Sie definieren ein Vokabular, das bestimmt welche Eigenschaften für eine bestimmte Domäne verwendet werden dürfen. Die Metadaten werden somit in hierarchische Beziehungen gebracht.

OWL basiert technisch auf die RDF-Syntax ist aber weit ausdrucksmächtiger als RDFS. Es ermöglicht mit abstrakten Konzepten das akzeptierte Wissen unserer Welt zu beschreiben. Anhand dieser Konzepte kann Wissen für ein bestimmtes Gebiet bzw. Domäne repräsentiert werden. Ontologien bilden somit die Grundlage von Wissensbasen für die künstliche Intelligenz.

Ein anderes Konzept XML Daten mit Semantik anzureichern sind Topic Maps[7]. Es ist ein internationaler Industriestandard, der bei Informationspräsentation und Informationsintegration angewendet wird. Sie reichern ebenso XML mit Semantik an, aber verfolgen im Gegensatz zu RDF und OWL unterschiedliche Ziele. Topic Maps dienen primär zur Visualisierung von Wissen und die Navigation durch Themengebieten für den Menschen.

Die bereits erwähnten Konzepte (RDF, OWL, Topic Maps) setzen alle eine hierarchische Struktur voraus, die die Beziehungen der Metadaten untereinander klar beschreiben. Sie können wegen ihrer Komplexität meist nur durch Experten erstellt werden. Dies erweist sich als aufwendig, starr und unflexibel. Für die Erstellung und Änderung muss sich ein fachkundiges Personal beschäftigen. Aus diesem Grund gibt es für sehr viele Wissensgebiete bzw. Objekte, die noch durch keine Ontologie oder Topic Map dargestellt wurden.

Dieser große Nachteil existiert bei Tags hingegen nicht. Tags sind einfache Schlagwörter die als Metadaten einer Ressource hinzugefügt werden können. Sie besitzen, zum großen Unterschied der oben genannten Konzepte, keine Struktur. Dadurch kann keine Beziehung zwischen den Metainformationen und der Ressource aufgebaut werden. Alle hinzugefügten Schlagwörter stehen auf derselben Ebene. Das Fehlen dieser gewöhnlich komplexen Strukturen erlaubt es, dass auch Laien Tags zu den Ressourcen hinzufügen können.

[3] http://www.w3.org/RDF/ - Zugegriffen am 21.07.2012
[4] http://www.w3.org/TR/rdf-concepts/ - Zugegriffen am 21.07.2012
[5] http://de.wikipedia.org/wiki/RDF-Schema - Zugegriffen am 15.03.2012
[6] http://de.wikipedia.org/wiki/Web_Ontology_Language - Zugegriffen am 15.03.2012
[7] http://www.topicmapslab.de/introduction - Zugegriffen am 07.02.2012

Die Sammlung von mehreren Tags wird als Folksonomy bezeichnet. Die erstellte Klassifikation können sowohl von Experten als auch von Laien erstellt werden. Jeder Internetnutzer kann ein oder mehrere Metainformationen zu einer Ressource hinzufügen und somit eine Folksonomy erweitern. Diese Offenheit birgt einige Probleme in sich. Nutzer können zum Beispiel wahllos Detailinformationen hinzufügen, die überhaupt keinen Bezug zu der Ressource haben.

Ein sehr interessantes Muster in Folskonomies zeichnet sich bei der Verteilung der vergebenen Schlagwörter ab. Eine geringe Anzahl von Schlagwörtern wird von den meisten Nutzern hinzugefügt. Die Häufigkeit der meisten Schlagwörter ist aber sehr gering. Dieses Phänomen wird als 'The Long Tail' bezeichnet und findet sich auch in dem Online-Verkauf von Musik und Büchern wieder.

In dieser Masterarbeit wurden die verschiedenen Konzepte detailliert erklärt und paarweise gegenüber gestellt. Der Fokus liegt jedoch bei der ausführlichen Beschreibung der Tags bzw. Taggingsystemen. Da das, im Zuge der Masterarbeit, erstellte System auf Tags bzw. Folksonomies basiert. Sie werden ausführlicher erklärt, Probleme, Systeme, Taggingvorgänge sowie Motivationen aufgelistet und erläutert.

Für die bereits existierenden und bekanntesten Tagging-Systeme wurde eine Übersicht erarbeitet. In denen die unterschiedlichen Systeme ausführlich erklärt und in einer Tabelle verglichen werden. Die verschiedenen graphischen Darstellungsmöglichkeiten für gesammelte bzw. getaggten Inhalte werden in einem eigenen Kapitel genauer erläutert.

Nach der Sammlung und Anreicherung der Inhalte mit Metadaten werden die unterschiedlichen Methoden zur Informationswiedergewinnung beschrieben. Dabei wird genauer auf den FolkRank Algorithmus[8] eingegangen der einen auf Folksonomies angepassten PageRank Algorithmus[9] darstellt. Zudem werden die bekanntesten 'Information retrieval'-Tools genauer analysiert und verglichen.

Der praktische Teil dieser Masterarbeit bestand in der Entwicklung eines Prototyps, das die Sammlung von Tags sowie die Informationsbeschaffung beinhaltet. Für das Hinzufügen von Tags wurde eine Google Chrome Extension[10] erstellt, die den Nutzer bei dem Taggingprozess unterstützt. Für die Informationsrückgewinnung wurde ein Portal erstellt, über dies die getaggten Objekte gefunden werden können. Bei der Darstellung der auf eine Suchanfrage resultierenden Informationen kann zwischen einer Listen- und Zeitungs-Ansicht gewählt werden. Das Portal beinhaltet eine soziale Komponente, dies es ermöglicht einen Freundeskreis aufzubauen und Ressourcen nur mit diesen zu teilen. Für die angezeigten Ressourcen werden automatisch ähnliche Informationen geladen und angezeigt.

[8] http://www.bibsonomy.org/tag/www?order=folkrank – Zugegriffen am 09.02.2012

[9] http://de.wikipedia.org/wiki/PageRank - Zugegriffen am 09.02.2012

[10] https://chrome.google.com/webstore/category/extensions?hl=de – Zugegriffen am 03.06.2012

KAPITEL 2

Metadaten für elektronische Objekte

Metadaten werden in der bibliothekarischen Praxis schon seit Jahrhunderten eingesetzt. Sie dienen dazu, Ressourcen zu beschreiben, damit diese zu einem späteren Zeitpunkt leichter aufgefunden werden können. Metadaten sind somit Informationen über andere Daten, die eine gewisse Struktur aufweisen müssen, damit sie für die Recherche, Informationswiedergewinnung und Nutzung des Dokumentes verwendet werden können. Die inhaltlichen und formalen Angaben eines Dokumentes werden dadurch strukturiert und für einen Computer leichter lesbar gemacht.

Metadaten sollten laut Schön (vgl. [Schön2005]) folgende Anforderungen erfüllen:

- Müssen eine bestimmte Struktur aufweisen, dass sie durch Maschinen verarbeitet werden können
- Sie sollten auch für den Menschen leserlich und verständlich sein
- Metadaten müssen für den Austausch über Computernetze geeignet sein
- Die Interoperabilität der Metadaten sollte gewährleistet bleiben, damit sie von verschiedensten Anwendungen genutzt werden können
- Metadaten sollten Domänenneutral und nicht spezifisch für eine Domäne konzipiert werden

Zur Standardisierung der Metadatenbeschreibung wurde durch die Dublin Core Metadata Initiative (DCMI)[1] 1994 eine Konvention eingeführt. Sie beschreibt Metadaten für elektronische Ressourcen. Das Dublin Core Metadata Element Set setzt sich aus 15 verschiedene Elementen zusammen. Sie beinhalten Informationen über den Inhalt (wie zum Beispiel die Beschreibung des Inhaltes, den Titel der Ressource, die verwendete Sprache), Informationen zum Urheber (wie zum Beispiel der Namen des Autors und des Verlages) und formale Informationen (wie zum Beispiel Erscheinungsdatum, Identifikationsnummer, Größe der Ressource).

[1] http://www.dublincore.org/ - Zugegriffen am 20.03.2012

2.1 Semantic Web

Semantic Web ist ein aufbauendes Konzept das zur Weiterentwicklung des Internets zum Internet der Dinge führt. Für diesen Ansatz ist es erforderlich, dass Maschinen die von den Menschen erzeugten Informationen automatisch verarbeiten können. Dazu müssen die Informationen mit einer eindeutigen Beschreibung ihrer Bedeutung, der Semantik, versehen werden. Durch diese Zusatzinformationen kann ein Computer diese Informationen eindeutig zuordnen und diese anschließend effektiv verarbeiten.

Ein Vorbote des Semantic Web ist eXtensible Markup Language (XML) bei dem die Daten anhand von XML-Elemente (XML-Tags) hierarchisch strukturiert werden. Die hinzugefügten Elemente beschreiben, für den Menschen lesbar, den jeweiligen Inhalt der Elemente. So können XML-Dokumente zum Beispiel Informationen über den Ersteller einer Webseite, Themenbereich und relevante Schlagwörter beinhalten. Diese Metadaten dienen den Suchmaschinen die versehenen Dokumente leichter auffindbar zu machen.

Vor der Einführung von XML wurden die Metadaten in Flat File- und Datenbanken abgespeichert. Die verwendeten Daten hatten meist proprietäre Formate, die das Austauschen der Daten fast unmöglich machte. Für den Austausch von XML-Daten müssen XML-Schemas definiert werden. Diese Schemas legen die Struktur der XML-Dokumente fest und garantieren die Interoperabilität und den simplen Austausch deren.

Die Interoperabilität der XML Dokumente basiert ausschließlich auf die Syntaktik, jedoch nicht auf die Semantik. Die Syntaktik definiert die Beziehung eines Zeichens zu anderen Zeichen, wobei die Semantik die Bedeutung der Zeichen festlegt. Bei den Austausch von XML Daten müssen beide Seiten die verwendeten Elementnamen kennen und verstehen. Wenn zum Beispiel ein Element mit dem Namen 'Geschäft' (<Geschäft>) und ein anderes Element mit dem Namen 'Laden' (<Laden>) bezeichnet wird. Ist eine Maschine nicht in der Lage, zu verstehen, dass damit das selbe Element gemeint ist. Bei Semantic Web werden die Elemente nicht nur für den Menschen sondern auch für Maschinen verständlich gemacht.

Damit dies möglich ist muss ein einheitliches Format definiert werden, indem zum Beispiel eine 'Wohnadresse' immer dasselbe Format hat. Eine solche Art findet man heutzutage schon bei Hotel-Internetauftritten, auf denen dem Benutzer ermöglicht wird eine Anfrage bzw. eine Buchung abzuschicken. Der Gast muss das Anreise- bzw. Abreisedatum eingeben, die Anzahl der Erwachsenen und Kinder, seinen Namen und seine Emailadresse.

Reine Xml-Daten reichen somit nicht aus um eine Struktur aufzubauen, die für das Semantic Web notwendig ist. Den Xml-Daten muss ein Ressource Description Framework (RDF) aufgesetzt werden, das Metadaten organisiert, strukturiert und ihnen eine Semantik hinzufügt. RDF ist ein Standard des World Wide Web Consortiums (W3C)[2]. Fensel (vgl. [Fensel2004]) definiert RDF wie folgt:

> *'RDF is an infrastructure that enables the encoding, exchange, and reuse of structured metadata. Search engines, intelligent agents, information brokers, browser and human users can make use of the semantic information. RDF is an XML application (i.e., its syntax is defined in XML) customized for adding meta-information to Web documents."*

[2] http://www.w3.org/ - Zugegriffen am 14.03.2012

Das RDF nutzt bestehende URI- (Uniform Resource Identifier) [3] und XML-Technologien. Wobei die XML zur Strukturierung der Daten und die URI zum Identifizieren der einzelnen Ressourcen eingesetzt wird. Das Ressource Description Framework besteht aus RDF-Triples. Sie beschreiben die Ressource (identifiziert durch einem URI), ihre Eigenschaften und die Werte der Eigenschaften. Somit besteht ein Triple aus einem Subjekt (Ressource), Prädikat (Eigenschaft) und Objekt (Objekt). Die RDF-Triples werden mittels XML-Tags bezeichnet.
Ein RDF-Triple in XML verpackt schaut wie folgt aus:

```
<?xml version='1.0' encoding='UTF-8' ?>
<rdf:RDF xmlns:rdf=http://www.w3.org/1999/02/22-rdf-syntax-ns#
xmlns:dc='http://purl.org/dc/elements/1.1/'>
        <rdf:Description rdf:about=' http://www.trenkwalder.info/ '>
                <dc:title>Haus Trenkwalder – Bed and Breakfast</dc:title>
                <dc:publisher>Martin Trenkwalder</dc:publisher>
        </rdf:Description>
</rdf:RDF>
```

In der ersten Zeile wird der XML-Header definiert. Damit eine Maschine erkennt, dass es sich um eine XML-Datei handelt. Mit xmlns:rdf wird der Namensraum[4] für die Elemente definiert, die ein 'rdf' als Präfix haben. Anhand xmlns:dc wird der Namensraum für die Elemente mit einem Präfix 'dc' bestimmt. Das Description-Element wird durch das Attribut 'rdf:about' identifiziert und beinhaltet die Beschreibung. Die Elemente <dc:title> und <dc:publisher> sind die Eigenschaften der Ressource. In Abbildung 2.1 sind RDF-Triples graphisch dargestellt.

Auf Basis der Assoziationen zwischen Subjekt und Prädikat können Maschinen logische Annahmen erstellen und mit diesen arbeiten. Durch die Identifizierung der Ressourcen mittels URI's ist jede Ressource mit einer eindeutigen ID versehen und ihre Definition kann über das Internet abgefragt werden. Mithilfe von RDF ist es zwar möglich ein Model und eine Syntax für die Beschreibung der Ressource zu bestimmen. Jedoch verfügt es nicht über die Funktion einer Ressource eine Semantik, also deren Beschreibung, hinzuzufügen. Um die Semantik zu definieren werden RDFS (Ressource Description Framework Schema) oder OWL (Web Ontology Language) benötigt.

RDFS legt die Syntax für den Aufbau der XML-Daten fest. In einem RDFS-Vokabular wird definiert, welche Eigenschaften den Ressourcen für eine bestimmte Domain zugewiesen werden dürfen. RDFS baut ebenfalls auf den RDF-Triple-Modell auf, wobei ein RDF-Triple aus Klassen, Klasseneigenschaften und Werten besteht. Anhand der Werte wird für eine bestimmte Domäne die Beziehung zwischen den Klassen bestimmt. Die Ressourcen sind als Instanzen von Klassen definiert. Jede Klasse ist wiederum eine Ressource und kann eine Unterklasse einer anderen sein. Durch diesen hierarchischen Klassenaufbau und Dank der Eigenschaften und Klassen von Ressourcen wird die nötige Semantik hinzugefügt. Maschinen können dadurch die Bedeutung der Ressourcen automatisch bestimmen.

[3] http://de.wikipedia.org/wiki/ Uniform_Resource_Identifier - Zugegriffen am 15.03.2012
[4] http://de.wikipedia.org/wiki/Namensraum_(XML) - Zugegriffen am 15.03.2012

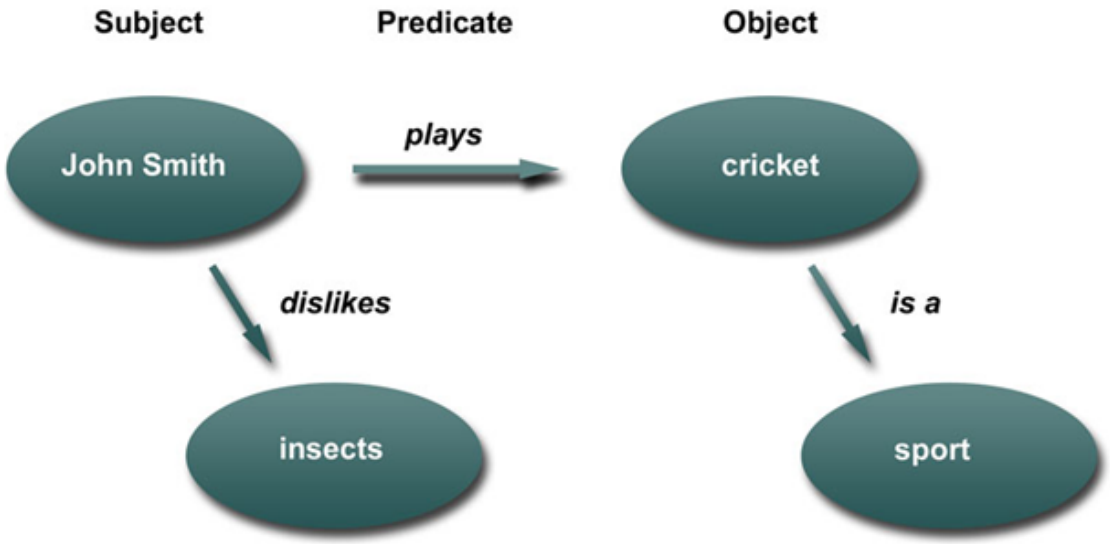

Abbildung 2.1: RDF-Triples

Web Ontology Language(OWL) ist auf RDFS aufbauend und besitzt zur Definition von Semantic Web Ontologien ein viel umfangreicheres Vokabular als RDFS. OWL ist ebenfalls eine Spezifikation des World Wide Web Consortiums. Anhand Welchem Ontologien mit einer formalen Beschreibungssprache erstellt werden können. Somit ist OWL ein Schema, das dazu dient Hierarchien und Beziehungen mittels XML zwischen verschiedenen Ressourcen zu definieren.

2.2 Ontologien

Fensel (vgl. [Fensel2004]) definiert Ontologien wie folgt:

'An Ontology is a formal, explicit specification of a shared conceptualization'

Eine Ontologie hat einen formalen Aufbau, damit er von Maschinen gelesen werden kann. Die darin enthaltenen Informationen sollen explizit spezifiziert werden und mit abstrakten Konzepten das akzeptierte Wissen unserer Welt beschreiben. Laut Lauer (vgl. [Lauer2003]) ist eine Ontologie ein Konstrukt, um Wissen über ein bestimmtes Gebiet, durch den Menschen oder einer Maschine, zu repräsentieren, auszutauschen und wiederzuverwenden. Ontologien bilden somit die Grundlage von Wissensbasen für die künstliche Intelligenz.

Ontologien bestehen aus Konzepten und Relationen. Laut Lauer gibt es zwei Arten von Konzepten.

Bei der ersten Art wird eine Klasse oder eine Menge von individuellen Objekten der Welt beschrieben. Die Objekte müssen gemeinsame Eigenschaften besitzen. Das Konzept 'Fahrzeug' beinhaltet zum Beispiel alle Fahrzeuge die zum Transport von Menschen, Werkzeug und Güter genutzt werden. Egal ob es sich dabei um ein Lastkraftwagen, einem Auto, einem Schiff oder

einem Raumfahrzeug handelt. Die gemeinsame Eigenschaft besteht darin, dass sie sich fortbewegen. Sollen zwischen den oben genannten Fahrzeugen unterschieden werden, so können die Konzepte 'Wasserfahrzeug', 'Landfahrzeug', 'Luftfahrzeug' und 'Raumfahrzeug' eingeführt werden.

Bei der zweiten Art beschreibt ein Konzept nur ein ganz bestimmtes Objekt der Welt. Ein Beispiel für solch ein Objekt wäre das aktuelle Formel-1 Fahrzeug von Sebstian Vettel[5]. Ein Konzept kann nicht nur aus physischen Objekten, sondern auch aus nicht-physischen bestehen.

Um das Wissen der Welt beschreiben zu können, reichen Objekte nicht aus. Die Objekte müssen anhand von Relationen in Beziehungen zueinander gestellt werden. Laut Lauer kann man zwei Klassen an Beziehungen unterscheiden:

1. **Relationen, die Konzepte hierarchisch ordnen**
   Mit dieser Art von Relationen kann beschrieben werden, dass ein Konzept eine Teilmenge eines anderen Konzepts ist. Mit der 'isA-Beziehung', Konzept A isA Konzept B, kann ausgedrückt werden, dass das Konzept A eine Spezialisierung des Konzeptes B ist. Ein Beispiel dafür 'VW Polo' isA 'Fahrzeug'. Anhand einer Äquivalenz-Beziehung kann ausgedrückt werden, dass Konzept A denselben Sachverhalt wie Konzept B hat. Zum Beispiel Konzept 'Bank' $\equiv$ Konzept 'Geldinstitut';

2. **Relationen, die Konzepte untereinander in beliebige Beziehung setzen**
   Diese Kategorie behandelt alle anderen Beziehungen von Konzepten. Durch Rollen können beliebige, nicht hierarchische Beziehungen zwischen Konzepten definiert werden, wie zum Beispiel Franz 'istVerheiratetMit' Stefanie oder Menschen 'trinken' Wasser. Die 'hasA'-Beziehung stellt eine besondere Form der Rollen-Beziehung dar. Die Relation 'Elternteil hat Kind', wobei es sich sowohl bei den Elternteil als auch bei dem Kind um Personen handelt. Sie könnte zum Beispiel durch die Relation 'Elternteil' 'hasA' 'Person' ausgedrückt werden.

Im Zusammenhang mit Semantic Web werden Ontologien wie folgt eingeschränkt:

- Ontologien beinhalten ein Datenmodel (z.B. XML)
- Die Semantik wird anhand eins RDFS- oder OWL-Vokabulars definiert
- Logische Zusammenhänge werden mittels Regeln ausgedrückt

Aus den gegebenen Informationen können weitere Informationen gewonnen werden, dieser Vorgang wird laut Lauer als 'Schließen' bezeichnet. Der Algorithmus der zum 'Schließen' verwendet wird heißt Inferenz-Algorithmus[6]. Dieser verwendet unter anderem die aufgestellten Relationen und Axiome, falls es die eingesetzte Ontologiesprache unterstützt.

Aus einer 'isA'-Relation lassen sich am meisten Informationen gewinnen. Da dadurch die Eigenschaften eines Konzeptes vererbt wird. Ist zum Beispiel in einer Ontologie hinterlegt, dass

[5] http://www.sebastianvettel.de/ - Zugegriffen am 15.03.2012

[6] http://www.informatik.uni-augsburg.de/lehrstuehle/swt/vs/lehre/archiv/WS_07_08/seminar_semantics/ dokumente/Seminarband.pdf - Zugegriffen am 07.02.2012

ein Rennfahrer eine Person ist (Rennfahrer isA Person), so weist ein Rennfahrer auch die Eigenschaften einer Person auf. Die zum Beispiel Größe, Gewicht, Haarfarbe, Geschlecht, Geburtsdatum usw. beinhalten. Eine aufwendige Recherche im Internet nach dem Rennfahrer 'Sebastian Vettel' könnte erspart werden, wenn der Rennfahrer selbst schon seine gesamten Daten in einer Ontologie hinterlegt hätte.

Anhand der Äquivalenz-Relation können Konzepte in verschiedenen Sprachen gleichgesetzt werden. So hat das Konzept 'Auto' und das Konzept 'car' dieselbe Bedeutung und kann in einer Ontologie mit der Äquivalenz-Relation verbunden werden. Stößt nun ein System in der Ontologie auf das Konzept 'car' so sieht es, dass 'Auto' ein äquivalentes Konzept ist und kann die Konzepte die mit dem Konzept 'Auto', anhand von Relationen verbunden sind, in die Suche einbinden und verarbeiten.

Eine Rollen-Relation 'fahren' zwischen dem Konzept 'Sebastian Vettel' und dem Konzept 'Rennauto' in einem Model, könnte dem System dazu dienen, dass es zu den Daten von Vettel seine wichtigsten Rennsiege mit in das Suchergebnis aufnimmt.

2.3 Topic Maps

Topic Maps ist ein internationaler Industriestandard für die semantische Informationsrepräsentation und Informationsintegration (ISO 13250). Ein Topic ist ein Modell aus einer bestimmten Domäne. Es besteht aus einer Menge von Topics und Assoziationen.

> *'In einer Topic Map ist ein Topic der Repräsentant eines Aussagegegenstandes (Subjekt) der realen Welt im Modell. An diesem Repräsentanten werden alle Informationen, die in dem Modell über den zugehörigen Aussagegegenstand repräsentiert werden sollen, geheftet. Die Typen für Topics, Begegnungen, Belegstellen, Beziehungen und Beziehungsrollen werden ebenfalls durch Topics innerhalb der Topic Map definiert.'* (vgl. [Maicher2008])

Ein Topic kann also als ein leerer Container angesehen werden, der mit Inhalt gefüllt werden kann. Diese Inhalte füllen das Topic mit Eigenschaften. Ein Topic, zum Beispiel ein Projekt, kann anhand der Eigenschaften die Ziele, die Probleme, die Projektleiter, die Mitarbeiter usw. angereichert werden. Die Benennung (Topic Names) der Topic Maps wird laut Bryan et al (vgl. [Bryan1999]) wie folgt eingeteilt: 'Base Names', allgemeine Benennung des Topics; 'Sort Names', Benennung des Sortierschlüssels; 'Display Names', alternative Name für die Anzeige in Anwendungsprogrammen; Bezüglich der Dokumenttypdefinition[7] (DTD) ist der 'Base Name' obligatorisch und die restlichen zwei Namen optional. Eine Topic Map kann durch mehrere Topic Names definiert werden. Das ist in einer Topic Maps mit mehreren unterschiedlichen Sprachen sehr sinnvoll.

Topics in einer Topic Map müssen durch ein eindeutiges ID-Attribut identifiziert werden. Falls zwei Topics dasselbe identisches ID-Attribut haben, müssen sie vor der Verarbeitung durch eine Anwendung zusammengefügt (Merging) werden. Das Merging soll aber nur betrieben werden, wenn es sich bei dem Topic wirklich um dasselbe Objekt handelt. So würden zwei Topics

[7] http://de.wikipedia.org/wiki/Dokumenttypdefinition - Zugegriffen am 20.03.2012

'Geldinstitut' und 'Sitzbank' die mit demselben ID-Attribut 'Bank' gekennzeichnet sind bei einem Merging verschmelzen. Obwohl sie nicht dieselbe Bedeutung haben. Um so ein Problem zu vermeiden, besteht die Möglichkeit Gültigkeitsbereiche (Scope) einzuführen. Der Gültigkeitsbereich kann nicht nur bei Topics sondern auch bei Topic Maps angewendet werden. Dadurch kann zum Beispiel eine ganze Topic Map mit dem Gültigkeitsbereich einer bestimmten Sprache versehen werden.

Durch die Verwendung der Topic Types können Klasse-Intanz-Beziehungen modelliert werden. Die Topic Types stellen wiederum Topics dar. Um eine Aussage 'Ein Maler ist ein Künstler' als eine Topic Map abbilden zu können braucht es zu den Topic Types noch Assoziationen zwischen den Topics. Wie folgt kann die oben genannte Aussage in einer Topic Map mittels XML dargestellt werden:

```
<topic id='Person'>
        <topname><basename>Person</basename></topname>
</topic>
<topic id='Künstler' types='Person'
        <topname><basename>Künstler</basename></topname>
</topic>
<topic id='Maler' types='Künstler'>
        <topname><basename>Maler</basename></topname>
</topic>
<topic id='LeonardoDaVinci' types='Maler'>
        <topname>
        <basename>Leonardo da Vinci</basename>
        <sortname>Da Vinci, Leonardo</sortname>
        <dispname>Leonardo da Vinci</basename>
        </topname>
</topic>
<topic id='tt-klasse'>
        <topname><basename>Topic Klasse</basename></topname>
</topic>
<topic id='tt-superkl'>
        <topname><basename>Superklasse</basename></topname>
</topic>
<topic id='tt-subkl'>
        <topname><basename>Subklasse</basename></topname>
</topic>
<topic id='at-superklasse-von'>
        <topname><basename>Superklasse-von</basename></topname>
</topic>
<assoc type='at-superklasse-von'>
        <assocrl type='tt-supkl'>Maler</assocrl>
```

```
        <assocrl type='tt-superkl'>Künstler</assocrl
</assor>
```

Heckel (vgl. [Heckel2001]) schreibt, dass jedes Topic mit beliebig vielen Informationsressourcen verlinkt werden kann. Diese Ressourcen sind aber nicht näher von dem Standard spezifiziert. Dadurch ist es laut Heckel möglich Dokumente die nicht in elektronischer Form vorliegen mit einem Topic zu referenzieren.

Occurrence Types dienen dazu mehr Kontext bezogenen Informationen in einem Topic anzugeben. So kann zum Beispiel in einer Informationsressource zusätzlich bestimmt werden auf welcher Seite, in welchem Buch, von welchem Autor eine bestimmte Abbildung stammt. Mittels des Rollennamens kann der Zusammenhang zwischen der Ressource und dem ausgewählten Thema ausgedrückt werden.

Topic Types, Occurrence Types und Informationsressourcen dienen bloß zur losen Beschreibungen von Topics. Die Verbindungen zwischen denen werden durch Assoziationen charakterisiert. Sie bestehen ihrerseits wieder aus Rollen, mit denen die Assoziationen näher beschrieben werden können. Folgende Aussage 'Der Fussballspieler 'Sebastian Schweinssteiger' spielt beim F.C. Bayern München' könnte mit Hilfe von Assoziationen wie folgt dargestellt werden:

```
<assoc type='spielt-bei'>
        <assocrl anchrole='Verein' type='Unternehmen'>
                F.C. Bayern München
        </assocrl>
        <assocrl anchrole='Fussballspieler' type='Person'>
                Sebastian Schweinssteiger
        </assocrl>
</assoc>
```

'Addthms'-Elemente dienen dazu Topics aus externen Topic Maps in den Gültigkeitsbereich einzubinden. Dadurch können bereits definierte Topics in anderen Topic Maps wiederverwendet werden. Die Attribute 'cassign' und 'tmdocs' legen den Gültigkeitsbereich fest. Für die Beschreibung von Topic Maps mittels Metadaten werden Facetten eingesetzt. Sie sind Eigenschaft-Wert-Paare die der Topic Map hinzugefügt werden können.

Laut Heckel können Assoziationen in Topics mit semantsichen Netzen verglichen werden. Die Knoten der Netzte werden durch die Topics dargestellt, die Kanten sind ihre Assoziationen. Die Attribute bestimmen den Typ der Kante.

2.4 Tag

Tags sind einfache Schlagwörter die als Metadaten verwendet werden können. Bei der Vergabe des Schlagwortes gibt es keinerlei Struktur und somit auch keine Syntax. Dies erweist sich als größter Unterschied zu Ontologien und Topic Maps. Alle Tags liegen auf derselben Ebene.

Dadurch ist es nicht ersichtlich welche Beziehung zwischen dem Schlagwort und der getaggten Ressource besteht.

Der Begriff Tag stammt aus der englischen Sprache und bedeutet ein Stück bzw. ein Streifen hartes Papier, Plastik, Metall, Leder usw. das als Etikett oder Marke an Etwas angebracht ist.

A piece or strip of strong paper, plastic,
metal, leather, etc.,for attaching by one end
to something as a mark or label

Bei der Definition[8] ist ersichtlich dass einem gewissen Etwas Zusatzinformationen, sogenannte Metadaten, hinzugefügt werden können. Darin wird aber nicht bestimmt wer die Zusatzinformationen hinzufügen darf, ob das jedermann oder nur der Erzeuger des Gegenstandes dürfen. Zudem wird die Anzahl der hinzugefügten Informationen nicht festgelegt, somit können auch mehrere Personen unabhängig voneinander Tags an dieselbe Ressource hinzufügen. Ebenso wird die Art der Informationen nicht weiter bestimmt.

Golder und Hubermann (vgl. [GolHub2005]) definieren 7 verschiedene Arten von Tags:

1. **Idefntifying What (or Who) it is About:**
   Klassifiziert um was (oder wen) sich das Objekt handelt
   Solche Tags beschreiben das Thema oder die Personen und Organisationen über die im Dokument diskutiert wird. Sie werden meist über mittels Nomen ausgedrückt. Es wird dabei nur der formale Aspekt berücksichtigt und nicht auf den Autor oder Dokumententyp eingegangen. Ein Beispiel für die vorliegende Arbeit ist 'Folksnomy', 'Taggingsysteme', 'Ontologien'.

2. **Idefntify What it is**
   Klassifiziert was das Objekt ist
   Bei diesem Tag wird der Bezug auf die Art des Objektes ausgedrückt. Zum Beispiel ob es sich um ein Dokument, Bild, Video, Blogeintrag usw. handelt. Für diese Arbeit wäre die passenden Tags 'Master Thesis' oder 'Diplomarbeit'.

3. **Identify Who Owns it**
   Klassifiziert den Eigentümer des Objektes
   Durch solche Tags können der Autor, Erzeuger bzw. der Eigentümer eines Objektes angegeben werden. Diese können sich auf einzelne Personen, Personengruppen oder auch Unternehmen bzw. Organisationen beziehen. 'Martin Trenkwalder', 'Distribuited System Group' wären passende Tags für diese Arbeit.

4. **Refining Categories**
   Verfeinert die Kategorien
   Manche Tags haben alleinstehend keine große Aussagekraft. Sie sind dazu da andere Kategorien zu verfeinern. Zum Beispiel könnte der Tag 'Master Thesis' mit 'Abschluss eines

[8] http://dictionary.reference.com/browse/tag - Zugegriffen am 10.09.2011

Master-Studienganges' erläutert werden um die Eigenschaften des Dokumentes aufzuzeigen. Aufgrund von fehlenden Konventionen werden Tags, bestehend aus mehreren Wörtern, mit unterschiedlichen Trennzeichen geschrieben. Dies führt zu Problemen da die zusammengesetzten Begriffe von den Taggingsystemen nicht richtig verarbeitet werden können.

5. **Identifying Qualities or Characteristics**
   Klassifiziert die Qualität oder Bewertung
   Diese Art von Tag werden meist mittels Adjektive ausgedrückt, sie beschreiben den Objektinhalt. Sie geben meist die Meinung des Taggers wieder. Zum Beispiel 'sehenswert', 'witzig', 'doof', 'mittelmäßig' usw.

6. **Self Reference**
   Selbst Verweisung
   Mit diesem Tag kann der Tagger seine direkt Beziehung zum Objekt ausdrücken. Meist wird bei solch einem Tag das Wort 'mein'/'meine' vorgesetzt. Für diese Arbeit zum Beispiel 'meine_diplomarbeit', 'meine_master_arbeit'.

7. **Task Organizing**
   Organisation von Aufgaben
   Hierdurch kann der User, der das Objekt tagt (Tagger), ausdrücken welche Aufgaben bzw. Arbeitsschritte mit dem Dokument ausgeführt werden sollen. Für diese Arbeit wär ein solcher Tag zum Beispiel 'lesen', 'zitieren' oder 'drucken'. In der englischen Sprache wird bei solchen Tags meist ein 'to' vor das Verb gesetzt, zum Beispiel 'toread', 'towatch'.

Golder und Hubermann weisen darauf hin, dass die ersten vier Kategorien unabhängig vom User formuliert werden können. Die letzten drei beziehen sich hingegen sehr stark auf den User und können meist nicht als objektiv angesehen werden.

Probleme

Da jedes Wort als Tag verwendet werden kann, entstehen ähnliche Probleme, die auch bei einer natürlichen Sprache auftreten können. Wie zum Beispiel der Gebrauch von Synonymen, Fehlern bei der Rechtschreibung, Begriffe die aus mehreren Wörtern bestehen. Besonders in den Zeiten des World Wide Web bzw. der Vernetzung der Menschen über mehrere Länder entsteht zudem eine Sprachbarriere. Die häufigsten Probleme, die bei Taggingsystemen auftreten können:

1. **Synonyme**
   Zwei Wörter mit gleicher oder ähnlicher Bedeutung werden als Synonyme bezeichnet. Dies stellt beim Tagging-Prozess ein großes Problem dar, da zum Beispiel das Wort 'Bank' mehrere unterschiedliche Bedeutungen hat. Damit kann eine Bank in einem Park gemeint sein, sowohl als auch ein Kreditinstitut. Um solche Verwechslungen zu vermeiden sollte dem Tag ein zusätzliches aussagekräftiges Wort hinzugefügt werden.

2. **Sprachunterschiede**
   Die Wörter 'chips' und 'French fries' haben in der englischen Sprache dieselbe Bedeutung. Jedoch stammt das erste Wort aus dem Britisch English und das zweite Wort aus dem American English. Dabei handelt es sich eigentlich um ein und dieselbe Sprache. Zusätzlich werden Tags in verschiedenen Sprachen verwendet, wie zum Beispiel 'deutsch' und 'tedesco'. Dies sind zwei unterschiedliche Wörter, haben jedoch dieselbe Bedeutung.

3. **Rechtschreibfehler**
   Die Rechtschreibfehler verursachen wohl die meisten Probleme bei dem Tagging-Prozess. Eine Studie von Guy und Tonkin (vgl. [Guy2006]), belegt dass 40% der Flickr[9]-Tags und 28% der del.icio.us Tags Rechtschreibfehler enthielten. Dazu wurden insgesamt 3000 Tags auf Flickr und del.icio.us automatisch überprüft. Bei der Studie wurden nur die Tags in der Sprache berücksichtigt, die von der eingesetzten Rechtschreibprüfung unterstützt wurden. Fehlerhafte bzw. ungenügende Zeichenkodierung stellt ein weiteres Problem dar. Besonders bei Sprachen bei denen das Alphabet nicht ausschließlich aus dem lateinischen Alphabet besteht, kommt es laut Guy und Tonking zu Schwierigkeiten. Darunter fällt auch die deutsche Sprache mit ihren Umlauten. Durch die Nutzung des Unicodes[10] kann man solche Konflikte vermeiden.

4. **Mehrzahl**
   Tags in der Mehrzahl können einem Taging-System so einige Probleme bereiten. Zum Beispiel sind die Tags 'Katze' und 'Katzen' unterschiedlich. Wenn nach einem gesucht wird, wird das andere nicht angezeigt. Durch den Abgleich von Wörtern in Ein- und Mehrzahl wird dem Problem entgegengewirkt.

5. **Genauigkeit**
   Unter Genauigkeit wird die Präzision der Tagbeschreibung verstanden, wie genau der Tagger das Konzept in einem Tag ausdrücken will. Dies wiederspiegelt sich zum Beispiel bei einer Ortsangabe, wo für den 'Karlsplatz' folgende Tags vergeben werden können: 'Wien', 'Wieden' oder 'Karlsplatz'. Die Genauigkeit der Tagvergabe hängt von dem Wissen des jeweiligen Taggers ab, je nachdem wie gut er die Stadt Wien kennt. Zudem hängt es von den Interessen des Taggers ab, ob er nur die Stadt, den Bezirk oder die genaue Straße hinzufügen möchte.

6. **Abkürzungen**
   Für viele Abkürzungen existieren unterschiedliche Bedeutungen aus verschiedenen Gebieten. So kann bei der Abkürzung 'AG' nicht eindeutig zwischen 'Aktiengesellschaft', 'Arbeitsgemeinschaft' oder 'Agendagruppe' bestimmt werden. Deshalb ist es meist besser Wörter auszuschreiben, als Abkürzungen zu verwenden.

7. **Sonderzeichen in Tags**
   Ein weiteres Problem stellen Symbole wie '#' oder '@' dar. Für einige Nutzer haben

[9] http://www.flickr.com/ - Zugegriffen am 16.12.2011
[10] http://de.wikipedia.org/wiki/Unicode - Zugegriffen am 27.02.2012

Identifying what it is	'=excel' oder 'review' Angaben zu dem Typ der Ressource
Identifying who it owns	'in:Quelle' oder 'Quelle' Quelle von der die Ressource stammt
Identifying qualities or characteristics	'*****', ' funny' oder ' cool' Bewertung des Inhaltes, entweder mit Sternen oder mit Adjektiven, eine Tilde wird vorangestellt
Task organizing	'!toread', oder '!towatch' Mit der Ressource assoziierte Aktivität

Tabelle 2.1: Bedeutung der Sonderzeichen in Schlagwörter

sie am Anfang des Wortes eine besondere Bedeutung. Da es dafür aber keinen einheitlichen Standard gibt, werden sie von Nutzern bzw. System unterschiedlich interpretiert und verwendet. In der Tabelle 2.1 ist eine Übersicht, laut Goldner und Hubermann, über die Bedeutung der am meist verwendeten Sonderzeichen dargestellt:

8. **Subjektivität**
   Informationen sind an sich nicht subjektiv, jedoch werden sie von den Taggern aufgrund von unterschiedlichen Motivationen, äußeren Umständen und unterschiedlichen Wissensständen verschieden interpretiert. Diese unterschiedliche Interpretation führt dazu, dass Tags oft als subjektiv angesehen werden.

9. **Zusammengesetzte Begriffe**
   Zusammengesetzte Wörter werden aufgrund von fehlenden Konventionen mit verschiedenen Trennzeichen geschrieben und als fehlerhafte Tags gekennzeichnet. Da sie von den Tag-Systemen nicht richtig erkannt werden. Laut Guy und Tonkin liegt dieser Anteil bei den untersuchten 3000 del.icio.us-Tags bei 10%. Am häufigsten wird für die Zusammensetzung die CamelCase-Technik verwendet, dabei wird der Anfangsbuchstabe der zusammengesetzten Wörter großgeschrieben. Dieses Verfahren wird oft als Konvention bei Namensgebungen von Dateinamen bei Frameworks oder Web Content Management Systeme verwendet.

2.5 Unterschiede

In den nun folgenden Kapiteln werden die Unterschiede der jeweiligen Konzpete aufgezeigt. Ontologien und RDF werden bei den anschließenden Vergleichen oft als Synonyme verwendet, da es sich bei RDF bzw. RDFS um eine einfache Ontologiesprache handelt.

Ontologien und RDF

Die Hauptaufgaben von RDF bestehen in der Strukturierung des Vokabulars, das anschließend in der Ontolgie verwendet wird. RDF bzw. RDF Schema ist eine einfache Ontologiesprache. Eine

Ontologie bzw. OWL bietet mehr Möglichkeiten, Relationen genauer zu definieren. Deshalb ist es möglich komplexere Schlussfolgerungen zu erstellen.

Laut Rasinger (vgl. [Rasinger2005]) können OWL folgende Eigenschaften aufweisen, die ein RDF nicht beinhaltet:

- **Symmetrie:** in einem OWL-Dokument kann eine Relation wie folgt als symmetrisch definiert werden:
  Donau 'istVerbundenMit' March: Die Donau ist verbunden mit der March und die March ist mit der Donau verbunden.

- **Transitivität:** wird in einem OWL-Dokument eine Relation als transitiv definiert, so wird es von OWL wie folgt interpretiert:
  'Breg' -> 'Donau' -> 'Schwarzes Meer' => 'Breg' -> 'Schwarzes Meer'
  'Die Breg fließt in das schwarzes Meer'.

- **Funktionalität:** Funktionalität bedeutet in einem OWL-Dokument folgendes:
  'Rhein' -> 'mündetIn' -> 'Nordsee'
  'Rhein' -> 'mündetIn' -> 'x-3295-01'
  daraus folgt, dass Nordsee und 'x-3295-01' dasselbe sein muss. In der Ontologie kann es aber mit zwei unterschiedlichen Werten belegt ist. Bei RDF muss hingegen Namensgleichheit bestehen, damit es als ident erkannt wird.

- **Inverse:** OWL kann eine Instanz durch bestehenden Ausdrücke durch Inferenzen aufspüren:
  'Lehrveranstalter' -> 'unterrichtet' -> 'Lehrveranstaltung'
  'Lehrversantalter' <- 'wirdUnterrichtetVon' <- 'Lehrveranstaltung'

- **Kardinalität:** Durch das Festlegen von Kardinalitäten kann bestimmt werden, dass eine Ressource mehrere Werte hat:
  'Person' - 'Geburtsort' -> 1 'Ort' besagt, dass eine Person nur an einem Ort geboren werden kann. Wenn im Web zwei verschiedene RDF-Dokumente über die Person 'Martin Trenkwalder' gefunden werden, die folgendes besagen:
  1. Martin Trenkwalder, Geburtsort='Innsbruck'
  2. Martin Trenkwalder, Geburtsort='Tirol'
  So kann OWL durch Inferenzen erkennen dass 'Innsbruck' und 'Tirol' der selbe Ort sein müssen.

- **Aufzählung:** Durch Aufzählung kann in OWL eine Klasse gebildet werden.

- **Equivalenz:** Mit Hilfe von OWL kann bestimmt werden, dass zwei Klassen ident sind.

- **Disjunktion:** Anhand von OWL kann beschrieben werden, dass eine Menge A disjunkt mit der Menge B ist.

RDF/Ontologien und Topic Maps

Bei dem Vergleich von Topic Maps mit RDF bzw. Semantic Web streicht Pepper (vgl. [Pepper2008]) folgende Gemeinsamkeiten heraus:

- Reichern XML mit Semantik an
- Erlauben Definition von Behauptungen für jegliche Dinge der Welt zu erstellen
- Basieren auf eindeutige Identifizierung
- Definieren abstrakte, assoziative Modelle
- Ermöglichen den XML-basierten Austausch der Syntax
- Erlauben Maßnahmen zur Schlussfolgerung und Beweisführung

Pepper weist folgende Unterschiede auf:

- Unterschiedlich Herkunft
  - Topic Maps dienen zur Suchhilfe, Indexierung, Thesauri usw.
  - RDF werden in Metadaten für Dokumente und Prädikatenlogik eingesetzt
- Unterschiedliche Ebenen der Semantik
  - Topic Maps ist in einer höheren Ebene (high level) vertreten
  - RDF ist in einer unteren Ebene (low level) angesiedelt
- Unterschiedlicher Aufbau der Modelle
  - Topic Maps mit Topics, Assoziationen und Occurences
  - RDF mit Subjekt, Prädikat, Objekt
- Unterschiedliche Ziele
  - Topic Maps für die Suchunterstützung von Daten, Visualisierung von Wissen für den Menschen und Navigation durch Themengebiete
  - RDF positioniert sich für den Einsatz für die Integration von großen Daten-Mengen und für den Einsatz in der künstlichen Intelligenz

Aus einer Zusammenfassung von Pepper geht hervor, dass RDF für Maschinen, Topic Maps für den Menschen und OWL für die Nutzung für die künstliche Intelligenz ausgerichtet ist.

Tags und Topic Maps

Topic Maps wurden, unter anderem, für die bessere Navigation und Suche von Internetinhalten und anderen Dokumenten konzipiert. Zusätzlich ermöglicht es Metadaten auszutauschen. Eine Taggingsystem ist eine Topic Map mit je einem Topic pro Tag und jeder getaggte Inhalt wird durch eine Occurence dargestellt. Bei der Abbildung eines Tags als eine Topic Map kommt es jedoch laut Garshol (vgl. [Garshol2006]) zu folgenden Problemen:

1. Eines der großen Probleme der Tags sind Synonyme, sie können nicht gut gehandhabt werden. In Topic Maps können ähnliche Wörter einem Topic hinzugefügt werden.

2. Zwei unterschiedliche Tags können bei Taggingsystemen nicht denselben Namen tragen, z.B. Paris (Frankreich) und Paris (Ontario). Dies wird in den Taggingsystemen meist mit einem Unterstrich gelöst, z.B. paris_france und paris_ontario. In Topic Maps stellt das kein Problem dar, da dort mehrere Topics mit denselben Namen definiert werden können.

3. Tags können in einem System nicht miteinander verbunden werden. Wenn ein Nutzer ein Foto aus 'Meran' taggt, dann fügt er wahrscheinlich auch die Tags 'Südtirol' und 'Italien' hinzu. Bei Topic Maps sind diese zusätzlichen Tags nicht nötig, da es eine Assoziation von 'Meran' zu 'Südtirol' und von 'Südtirol' zu 'Italien' gibt.

RDF/Ontologien und Tags

Für viele Interessensgebiete gibt es noch sehr wenige Ontologien, dabei sind Taggingsysteme bereits weit verbreitet. Allerdings sind Taggingsysteme bei weitem nicht so ausdrucksstark wie Ontologien. Dies liegt unter anderem oft daran, dass die verwendeten Tags Kontext bezogen sind. Dieser Kontext kann aber meist nicht herausgelesen werden. Hinzukommend fließt bei der Taggingvergabe die subjektive Ansicht des Taggers mit ein. Darüber hinaus sind Taggingsysteme oft unvollständig und werden mit redundanten Schlagwörtern versehen. Laut (vlg. [Albrecht2006]) gibt es zwischen Ontologien und Tags folgende Unterschiede:

1. Taggingsysteme besitzen keine Struktur und sind deshalb flach aufgebaut, hingegen ist eine Ontologie hierarchisch strukturiert

2. Bei Ontologien erfolgt die Erstellung von Sachkundigen bzw. Experten, die ein fachkundiges Vorwissen aufweisen müssen. Ein Tag wird hingegen von einem Benutzer, ohne besondere Vorkenntnisse, erstellt und sofort veröffentlicht, ohne dass es von einem Fachkunden auf die Korrektheit überprüft wird

3. Tags werden dezentral abgespeichert und werden von einer großen Gemeinschaft erstellt. Ontologien werden meist zentral verwaltet und zeichnen sich durch eine autoritäre Sicht aus, die bestimmt welche Informationen aufgenommen werden. Dadurch sind Ontologien viel genauer bzw. präziser als Taggingsysteme.

4. Taggingsysteme sind günstig und einfach zu erstellen, da es keine Fachkenntnisse benötigt. Für die Erstellung einer Ontologie jedoch müssen Experten beauftragt werden, was mit hohen Kosten verbunden ist.

5. Ein Taggingsystem ändert sich kontinuierlich, mit dem Hinzufügen neuer Tags. Es ist somit dynamisch bzw. flexibel. Hingegen kann eine Ontologie als starr bzw. inflexibel angesehen werden, da sie neu erstellt werden muss, wenn sich die strukturierten Daten ändern.

6. Synonyme stellen bei Tags ein sehr großes Problem dar, da sie nicht kontrolliert bzw. zusammengefügt werden können. Bei einer Ontologie hingegen wird eine Synonymkontrolle durchgeführt.

Zusammenfassung

In der Tabelle 2.2 werden die Unterschiede zwischen den behandelten Konzepte zusammengefasst.

Daten Modelle	**RDF/Ontologien**	**Topic Maps**	**Tags**
Verwendete Struktur	XML	SGML, XML	Strukturlos
Syntax	RDF / XML, N3	XTM, MyHTM, LTM	Nicht definiert
Vokabular	OWL / RDF Schema	TMCL	Nicht definiert
Aufbau	Subjekt, Prädikat, Objekt	Topics, Assoziationen und Occurences	Tags
Herkunft	Metadaten für Dokumente und Pärdikatenlogik	Suchhilfe, Indexierung und Thesauri	Suchhilfe, Wiederauffinden von Daten
Semantik	Unteren Ebene	Höheren Ebene	Keine
Ziele	Integration von großen Daten-Mengen, Einsatz in der künstlichen Intelligenz, Beschreibung der Ressourcen	Suchunterstützung von Daten, Visualiserung von Wissen für Menschen und Navigation durch Themengebiete, Modell des Wissens	Beschreibung der Ressourcen zur späteren Wiederauffindung
Ausdrucksstärke	Hoch	Hoch	Niedrig
Komplexität	Hoch	Hoch	Sehr gering
Erstellung durch	Experten	Experten	Einfache Benutzer

Tabelle 2.2: Zusammenfassung des Vergleichs zwischen RDF/Ontologien, Topic Maps und Tags

KAPITEL 3

Folksonomies

Folksonomy ist eine von Nutzern erstellte Klassifikation. Thomas Vander (vgl. [Vander2007]) definierte den Begriff als Erster wie folgt:

'Folksonomy is the result of personal free tagging of information and objects (anything with a URL) for oneś own retrieval. The tagging is done in a social environment (usually shared and open to others). Folksonomy is created from the act of tagging by the person consuming the information.
The value in this external tagging is derived from people using their own vocabulary and adding explicit meaning, which may come from inferred understanding of the information/object. People are not so much categorizing, as providing a means to connect items (placing hooks) to provide their meaning in their own understanding.'

Der Begriff Folksonomy besteht aus dem englischen Wort 'folk'[1] (die Leute) und dem altgriechischen Wort 'nomia'[2] (die Verwaltung). Es beschreibt eine von den Leuten erstellte Verwaltung bzw. Ordnung. Im Gegensatz zu der Taxonomie, das aus dem altgriechischen Wort 'taxis' (die Ordnung) und 'nomia' besteht. Somit beschreibt der Mensch bzw. die User selbst die Ordnung. Bei Taxonomie hingegen wird die Ordnung bereits im Vorhinein festgelegt.

In allgemeiner Bedeutung ist es eine Sammlung von ein oder mehreren Schlagwörtern. Die Folksonomy ist sozusagen das Resultat der Tagging-Prozesse, die Gesamtheit aller Tags, die bis zu einem bestimmten Zeitpunkt eingegangen sind. Die Benutzer können zu jedem Zeitpunkt neue Schlagwörter zu einer Ressource hinzufügen und somit die Folksonomy erweitern.

Folksonomy ist ein Mittel für Menschen um Objekte (Webseiten, Fotos, Videos, Podcasts usw.) mit selber gewählten Schlagwörtern zu versehen, um diese Objekte zum Beispiel zu einem späteren Zeitpunkt leichter wiederzufinden. Folksonomy hat meistens auch einen sozialen

[1] http://dictionary.reference.com/browse/folk - Zugegriffen am 20.12.2011
[2] http://www.albertmartin.de/altgriechisch/?q=verwaltung

Aspekt, somit können andere Benutzer die die gleichen Schlagwörter verwenden auch die Objekte finden die von anderen mit diesen versehen wurden. Es ist zu beachten, dass eine Folksonomy am aussagekräftigsten ist, wenn die Nutzer ihre alltäglich verwendeten Wörter für die Beschreibung der Objekte verwenden. Spezielle Fachausdrücke sollten nur äußerst selten verwendet werden.

Ein weiterer Unterschied zu einer Taxonomie liegt darin, dass Nutzer, durch das Hinzufügen eines neuen Schlagwortes, eine Folksonomy erweitern können.

Sobald eine Folksonomy eine gewisse Größe erlangt hat, lässt sich daraus eine gewisse Struktur der Tags ablesen. Die Struktur ist nicht statisch zu sehen, denn sie kann sich im Lauf eines Tages, durch das Hinzufügen von neuen Tags, ändern. Es ist also als ein Ist-Zustand der im System benutzen Tags. Aus der Struktur lässt sich zum Beispiel die am häufigsten genutzten Tags einer Ressource oder die am häufigsten miteinander verwendeten Tags ablesen.

Vander unterscheidet Folksonomies in zwei verschiedenen Formen, die jeweils auf unterschiedlichen Objekten anzuwenden sind. Broad Folksonomy kann jeder Nutzer jedes Objekt taggen. Narrow Folksonomy wird hingegen bei Objekten angewendet, bei denen es schwierig ist aussagekräftige Schlagwörter zu finden und deshalb nur von einer geringen Anzahl von Nutzern getaggt werden können.

3.1 Broad Folksonomy

Bei der Broad Folksonomy sind es viele Nutzer die dasselbe Objekte mit selbst ausgewählten Schlagwörtern versehen. Dabei kann dasselbe Schlagwort auch von verschiedenen Nutzern verwendet werden. Das resultierende Ergebnis tendiert zu einer Power Law-Kurve bzw. Long Tail-Kurve. Bei dem einige wenige Schlagwörter von den meisten Personen zur Beschreibung herangezogen werden. Diese aber von der Anzahl gesehen nur einen sehr geringen Teil ausmachen. Bei der Broad Folksonomy handelt es sich um eine 'Free for all'-Taggingsystem. Dabei kann jeder Nutzer jedes Objekt taggen. Das 'Broad Folksonomy'-Prinzip wird unter anderem von Delicious[3], CiteULike[4] und LibraryThing[5] verwendet.

Betrachtet man die Abbildung 3.1 als Übersicht, so sieht man eine Person (Content Creator) am unterem Rand der ein Objekt erstellt hat und dieses den anderen zugänglich macht. Andere Nutzer fügen dem Objekt (Pfeil von Benutzer zum Tag) selber gewählte Schlagwörter (Nummern von 1 bis 5) hinzu. Personen mit demselben Wortschatz und Bildungsniveau befinden sich in denselben Gruppen. Diese sind mit den Buchstaben A bis F gekennzeichnet. Zudem können von der Abbildung die Schlagwörter entnommen werden, die die Personen (gekennzeichnet durch Pfeile von den Tags zu den Personen) verwenden, um das Objekt zu einem späteren Zeitpunkt wiederzufinden.

Bei genauerer Betrachtung der 3.1, sieht man dass die Personen von der Gruppe 'A', 8 an der Zahl, das Objekt mit dem Tag '1' und '2' versehen haben um es wiederzufinden. Die Gruppe 'B' (2 Personen) hat die Schlagwörter '1' und '2' für das Objekt hinzugefügt. Sie verwenden jedoch auch das Schlagwort '3' um es ausfindig zu machen. Gruppe 'C' (3 Personen) hat das Objekt

[3] http://delicious.com/ - Zugegriffen am 15.12.2011

[4] http://www.citeulike.org/ - Zugegriffen am 01.02.2012

[5] http://www.librarything.com/ - Zugegriffen am 01.02.2012

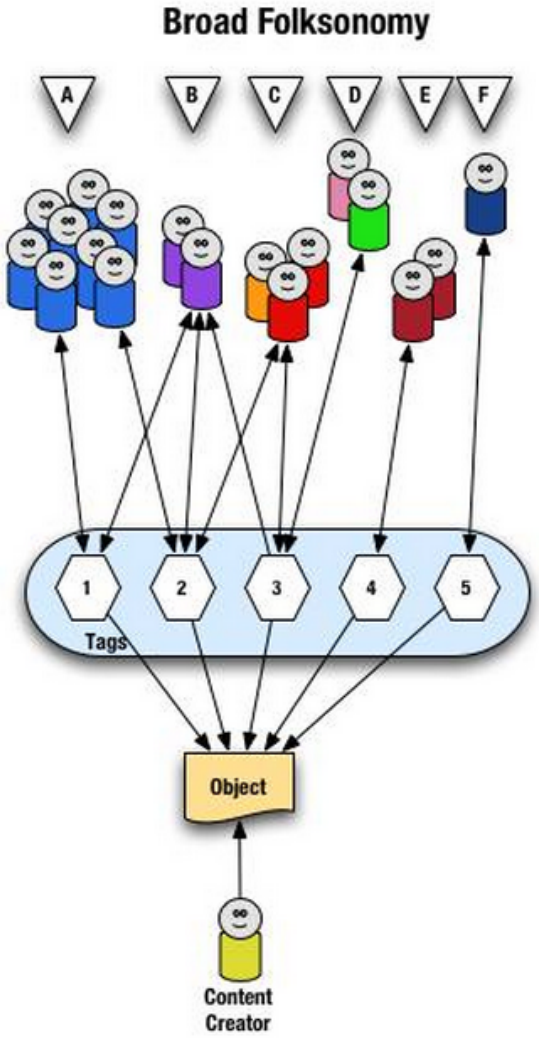

Abbildung 3.1: Broad Folksonomy

mit Tag ’2’ und ’3’ versehen und verwenden auch nur diese um es wiederzufinden. Gruppe ’D’ (2 Personen) verwendet nur Tag ’3’, Gruppe ’E’ (2 Personen) nützen Tag ’4’ und Gruppe ’F’ (1 Person) verwendet Schlagwort ’5’ um das Objekt wieder aufzufinden.

Broad Folksonomy und die Long Tail Kurve

Vander Wal und Shirky (vgl. [Shirky2005]) haben bei der Broad Folksonomy festgestellt, dass die Verteilung der Tags einer Power Law Kurve[6] dem Lotkaś Gesetz gleicht.

$$f(x) = C/x^2$$

[6]http://en.wikipedia.org/wiki/Power_law - Zugegriffen am 02.02.2012

Anzahl wie oft ein Tag verwendet wurde	Anzahl der Tags
1	5.593
2	1.307
3	496
4	296
5	163
6	125
7	107
8	82
9	51
10	47

Tabelle 3.1: Tagverteilung der del.icio.us Feeds

Diese Verteilung folgt der oben genannte Gleichung bei der C eine Konstante, x der Rang des gegebenen Tags und a eine konstanter Wert (meist zwischen 1 und 2) ist.

Sollte diese Annahme richtig sein, wären am linken Rand, also am Anfang der Kurve nur wenige Tags mit einer sehr hohen Häufigkeit. Das rechte Ende der Kurve würde aus sehr vielen Tags bestehen, die alle ungefähr dieselbe geringe Häufigkeit haben. Dieser sehr lange Teil kann als 'Langer Schwanz' gesehen werden und bildet den sogenannte 'Long Tail'.

Kipp et. al (vgl. [Kipp2007]) versuchte anhand einer Studie das Tagging-Verhalten von del.icio.us Usern zu analysieren. Dazu sammelte er Daten von 64 meist getaggten Delicious -Einträgen. Jeder Delicious -Eintrag besteht aus einer URL, einer optionalen erweiterten Beschreibung, einer Reihe von Tags, dem Nutzername und der Zeitpunkt an dem der Tag hinzugefügt wurde. Die Daten wurden von den populärsten Delicious Feeds genommen. Zudem wurden die Artikel die mit 'health', 'productivity' oder 'programming' getaggt wurden dazu genommen, wenn sie von mindestens 500 Usern getagt wurden. Nur alphabetisch identische Tags wurde zusammengefügt. Die Probleme wie Synonyme, Rechtschreibfehler usw. wurden nicht berücksichtigt. Die URL mit den meisten Einträgen hatten 58.728 Tags und die mit den wenigsten 5.172 Tags. Insgesamt wurden 18.904 verschiedene Tags zur Verschlagwortung verwendet.

Von der Studie geht hervor, dass 6% (3462) von den Usern die Webseite markierten ohne ein Tag hinzuzufügen. Die Anzahl der User die mehrere Tags für eine Webseite hinzufügen ist sehr gering. Tatsächlich haben 65% der Teilnehmer nur 1 bis 3 Tags zu einer URL hinzugefügt.

Insgesamt 3.049 (16%) der einzelnen Tags bezogen sich auf aufgaben- oder zeitrelevante Angaben. Sehr interessant ist zudem, dass 5.593 (30%) der Tags nur jeweils einmal verwendet wurden, siehe Tabelle 3.1.

Wie bereits Kipp et. al analysiert hat stellt auch Shirky (vgl. [Shirky2005]) fest, dass die Häufigkeit der Benutzung derselben Tags ziemlich rasch abnimmt. In der Abbildung 3.2 sieht man die Häufigkeitsverteilung der verwendeten Tags um die Webseite 'www.pocketmod.com'[7] zu beschreiben. Die ersten 7 Schlagwörter wurden sehr oft verwendet, hingegen die restlichen nur selten. Zudem ist ersichtlich dass in diesen 7 Wörtern wahrscheinlich die sinnvollste Be-

[7] http://www.pocketmod.com/ - Zugegriffen am 14.02.2012

schreibung enthalten ist. Die repräsentativste Ansicht für diese Webseite wird wohl von den 7 meist genutzten Tags ausgehen.

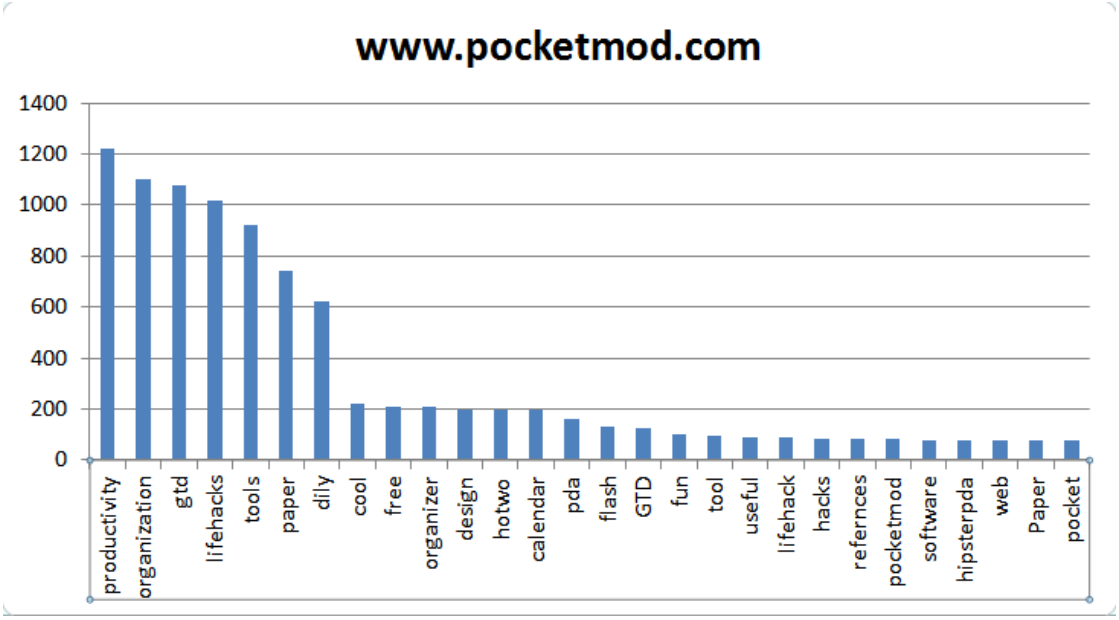

Abbildung 3.2: Meist verwendete Tags der Seite www.pocketmod.com

Nach dem Aufrufen der Webseite kann man erkennen dass die Tags links von der Kurve, den Inhalt der Seite in allgemeiner Form beschreiben. Im Unterschied zu den Long Tail-Tags die ihn sehr speziell und präzise beschreiben. Die Kurve der Webseite www.pocketmod.com und somit die Verteilung der Tags folgt dem Power Law mit einem Exponenten von ungefähr a = 1. Eine idealtypische Power-Law Verteilung wird in der Abbildung 3.3 angezeigt.

In einer Untersuchung von gemeinsam auftretenden Tags können Lux, Granitzer und Kern (vg. [Lux2007]) zeigen, dass rund 80% der Tags derart mit anderen Tags vorkommen, dass diese anderen Terme einem Power Law folgen. Auch Cattuto, Loreta und Pietronero (vgl. [Cattuto2006]) legen dies mit einem sehr interessanten theoretischen Ansatz vor. Dabei analysieren sie die gesammelten Tags auf einer Makroebene.

'Folksonomies ... do exhibit dynamical aspects also observed in human language, such as the emergence of naming conventions, competition between terms, takeovers by neologisms, and more'

Diese Aussage hat im Yule-Simon-Ansatz[8] seinen Ursprung, der die Wahrscheinlichkeit für das Auftreten von Wörtern in Texten thematisiert. Dies besagt, dass ein bestimmtes Wort die Wahrscheinlichkeit p hat, ein neues Wort zu sein, das im bisherigen Text noch nicht vorkommt. Oder die Wahrscheinlichkeit 1 – p, eine Kopie eines bereits vorhandenen Wortes zu sein. Der

[8] http://en.wikipedia.org/wiki/Yule%E2%80%93Simon_distribution – Zugegriffen am 02.02.2012

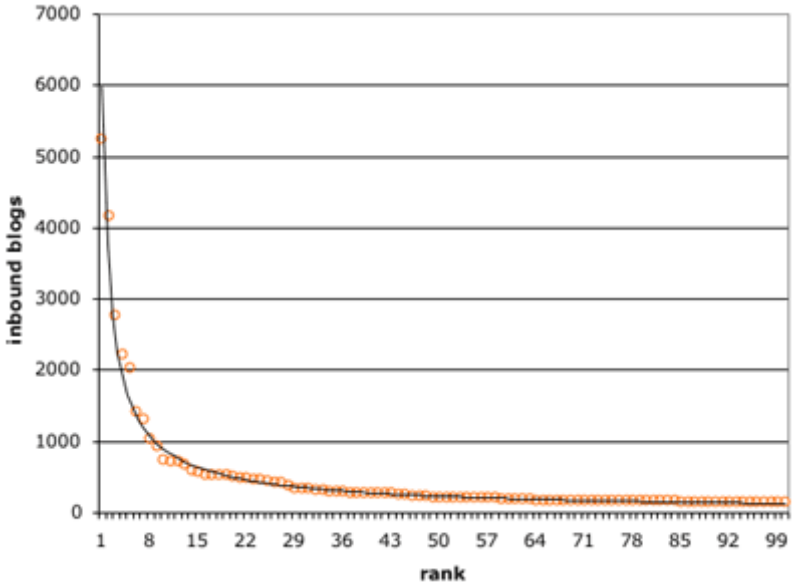

Abbildung 3.3: Idealtypische Power-Law Kurve

Wert von 1 – p hängt davon ab, wie oft das bestimmte Wort schon im bisherigen Text bereits vorkommt. Dabei gilt: Je häufiger das Wort bereits vorkommt, desto wahrscheinlicher ist es, dass es erneut Anwendung findet. Um den Yule-Simon-Ansatz auf Folksonomies anwenden zu können, muss man einfach nur statt von 'Worten in Texten' von 'Tags in Datenbanken' sprechen. Je Häufiger ein Schlagwort in einer Datenbank auftauch, desto wahrscheinlicher wird es sein, dass es erneut zur Indexierung verwendet wird. Cattuto et. al. fügen dem Yule-Simon-Ansatz noch eine zusätzliche Komponente hinzu, die Zeit. Sie besagen, dass neuere bzw. kürzlich verwendete Tags eine höhere Wahrscheinlichkeit haben zu der Verschlagwortung herangezogen zu werden als ältere. In Abbildung 3.4 ist dieses Erklärungsmodell graphisch dargestellt.

In der Abbildung 3.4 drückt die Wahrscheinlichkeit p aus, dass ein völlig neuer Tag zur Indexierung verwendet wird. Hingegen drückt 1 – p aus, dass ein alter bereits verwendeter Tag genutzt wird, unter der Berücksichtigung der Zeit. Die Zeitangabe ist im Gedächtnis Q gespeichert und funktioniert nach dem Motto 'je neuer, desto wahrscheinlicher'.

Es scheint aber nicht immer zuzutreffen dass sich Tags immer nach dem Power Law verteilen. Lux et. al. konnten bei der Tag-Verteilung nur in 80% der Fällen eine Power-Lag-Verteilung vorweisen. In den restlichen 20% gibt es anscheinend eine andere Verteilung. Dazu betrachtet man die Abbildung 3.5, die die Häufigkeit der Tags der Webseite asis.org graphisch darstellt.

Es ist ersichtlich dass keine Power Law Verteilung vorliegt. Es sind vielmehr zwei lang gezogene Kurven, rechts der bekannte 'Long Tail' und links ein 'Long Trunk' (Lange Rüssel). Die Tags im Long Trunk grenzen sich von der Häufigkeit nicht stark genug voneinander ab. Es werden viele Tags sehr oft verwendet. 'Assocations', 'library', 'information', 'ia', 'technology', 'informationscience' und 'professional' können zu den Long Trunk gezählt werden. Danach gibt es einen Wendepunkt, dem der Long Tail folgt.

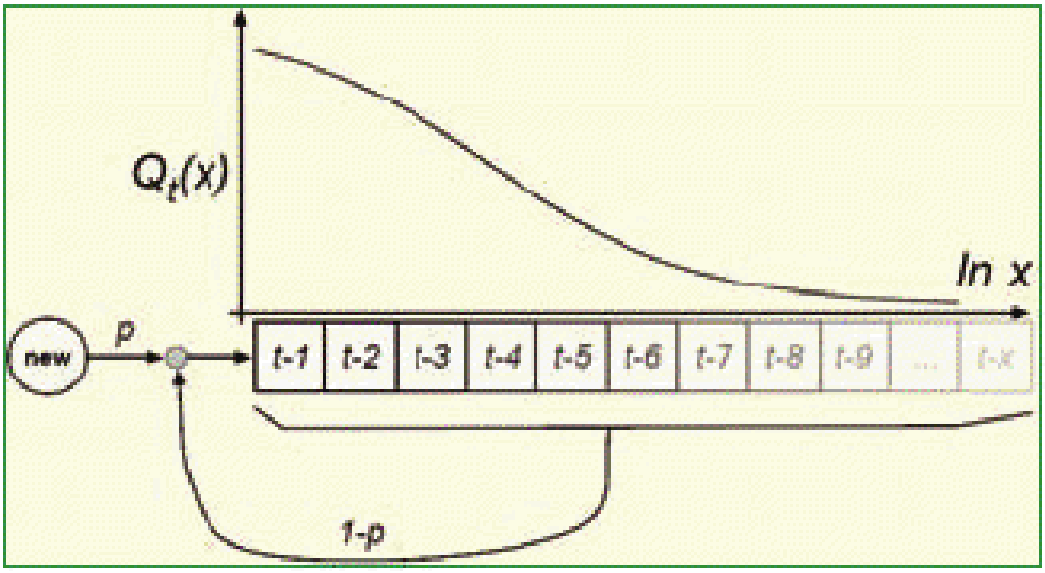

Abbildung 3.4: Ein Yule-Simon-Prozess der Auswahl von Indexierungs-Tags mit der Berücksichtigung des Alters der Tags

Die Tags in dem Long Trunk beschreiben den Inhalt der Seite nur sehr oberflächlich und sehr vage. Anders als die gewöhnlich bei einer Power-Law-Verteilung der Fall ist. Die Tags 'association', 'information' und 'technology' sowie die restlichen Tags sind äußert allgemein gehalten. Anscheinend hat die kollektive Intelligenz Schwierigkeiten hierzu passende und vor allem aussagekräftige Schlagwörter zu finden.

Stock (vgl. [Stock2006]) konnte wenigstens theoretisch beweisen, dass es mindestens zwei verschiedene Verteilungen von Relevanzen gibt. Die bekannte Power-Law-Verteilung und die invers-logische Verteilung. Die invers-logische Verteilung hat sowohl einen Long Trunk als auch einen Long Tail, siehe Abbildung 3.6.

Für die invers-logische Verteilung wird folgende Formel angewendet:

$$f(x) = e^{-C'(x-1)^b}$$

Wobei e die Euler'sche Zahl ist, x der Rang des Tags, C' eine Konstante (in der Abbildung 3.6 mit dem Wert 0,1) und b der Exponent mit dem Wert 3. Bei der Verteilung ist zu bemerken dass in vielen Fällen der Long Trunk kürzer ist als der Long Tail.

Bei der Beobachtung von Tags sollten sowohl die Power-Law Verteilung, als auch die invers-logische Verteilung im Auge behalten werden. Dies ist vor allem bei der Entwicklung und Erstellung von Information Retrieval-Werkzeugen für getagte Dokumente von großer Bedeutung. Sowohl die Tags im Long Trunk als auch die ersten n Tags im Long Tail werden dann für das Relevance Ranking herangezogen.

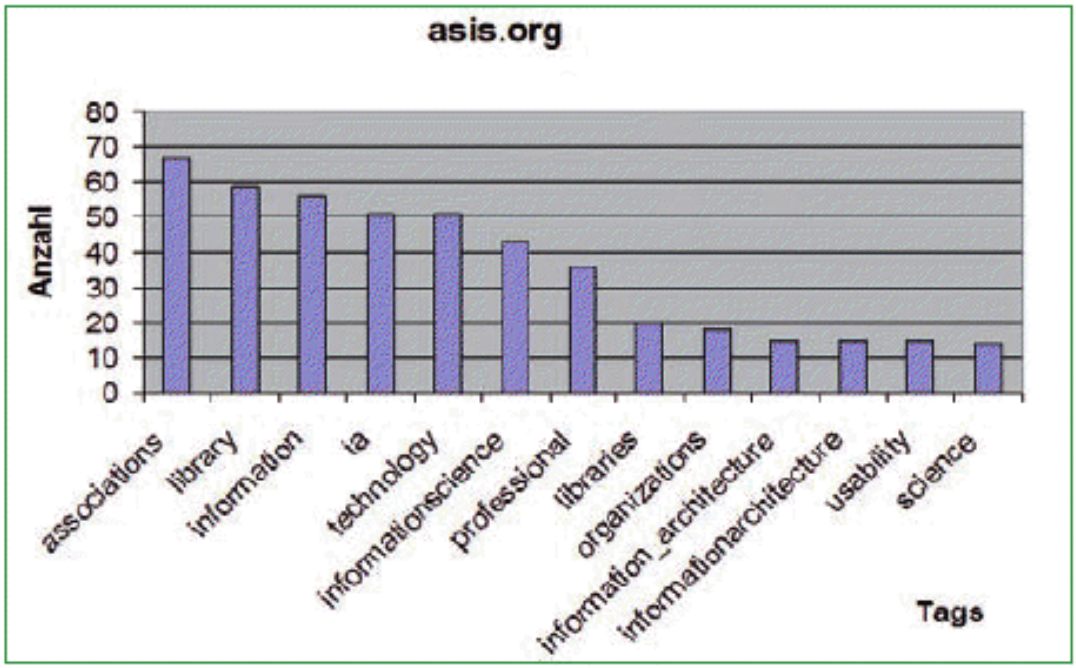

Abbildung 3.5: Tagverteilung zur Webseite www.asis.org

3.2 Narrow Folksonomy

Narrow Folksonomy wird meist bei Inhalten verwendet, die nur sehr schwer auffindbar sind oder wo es sehr schwierig ist aussagekräftige Schlagwörter dafür zu finden. Diese Objekte sind meistens Fotos oder Videos, bei denen eine wörtergebundene Suche oder auf Text basierende Vergleich-Tools nicht funktionieren würden. Dabei schreiben einige wenige Personen vor, meist der Ersteller der Ressource, welche Schlagwörter die Tagger benutzen dürfen. Oft werden Schlagwörter nicht nur zum Beschreiben eines Inhaltes sondern zur Gruppierung verwendet. Dies kommt zum Beispiel bei der Foto-Community Flickr vor. Die Narrow Folksonomy kann als ein Self-tagging System angesehen werden und findet bei YouTube[9], Technorati[10] und Flickr Verwendung.

In der Abbildung 3.7 sieht man, dass der Ersteller des Objektes (Content Creator) den Tag '1' für das Objekt definiert hat. Es gibt im Vergleich zu Broad Folksonomy viel weniger Tags und diese dürfen nur von auserwählten Gruppen ('B' und 'F') und dem Erzeuger hinzugefügt werden.

[9] http://www.youtube.com/ - Zugegriffen am 27.02.2012
[10] http://technorati.com/ - Zugegriffen am 02.02.2012

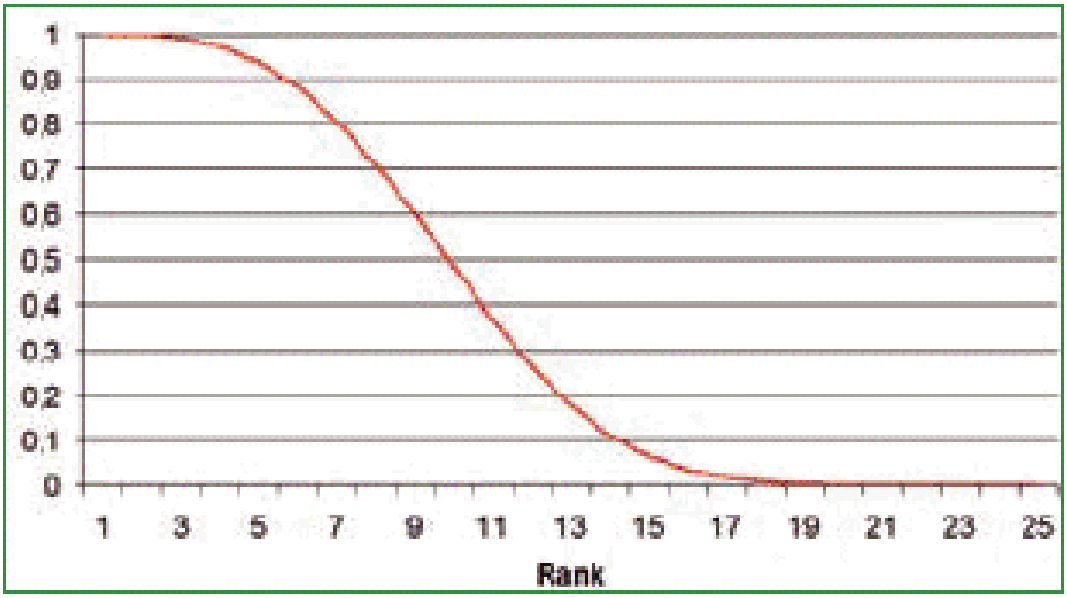

Abbildung 3.6: Invers logische Verteilung

3.3 Folskonomies und assoziierte Daten

Bei einem Tagging-Vorgang werden von dem Nutzer weit mehr Informationen generiert als nur die abgespeicherten Schlagwörter. Betrachtet man ein Webseite mit einer Tag-Funktion, so sind bei dem Tag-Vorgang vier verschiedene Gruppen von Einheiten involviert: die Schlagwörter, das Objekt (wie ein Bild oder ein Video), der Benutzer und die Webseite die die Tagging-Funktion bereitstellt. Bei der Interaktion zwischen den vier Einheiten wird eine große Menge an wertvollen Daten generiert.

Der Benutzer beschreibt eine Ressource mit seinen selbst gewählten Wörtern. Durch die freie Wahl der Schlagwörter fließt oftmals seine subjektive Meinung, sein Wissensstand und sein Interesse zu dem Objekt mit ein. In einem zweiten Schritt müssten diese nicht objektiven Informationen herausgefiltert werden.

Soziale Verbindungen von Nutzern können in einem System sehr aussagekräftige Informationen liefern. Wenn Benutzer gleiche Objekte teilen kann vermutet werden dass diese indirekt miteinander verbunden sind. Man betrachte dazu Abbildung 3.8. Benutzer A und B sind über den Tag 3 verlinkt und Benutzer B und C über das Objekt 5. Im ersten Fall könnte zum Beispiel die soziale Bindung durch die Sprache zustande gekommen sein und im zweiten Fall könnten die Benutzer dieselben Interessen haben.

Werden die Daten weiter analysiert, so kann mit diesen eine Grundstruktur einer Ontologie für diesen bestimmten Interessensbereich erstellt werden. Zum Beispiel müsste es eine signifikante Beziehung zwischen den Objekt 1 und Objekt 5 geben und somit sollten die Schlagwörter dieser Objekte zusammengefasst werden. Desweiteren sollte eine Relation zwischen dem Tag 3

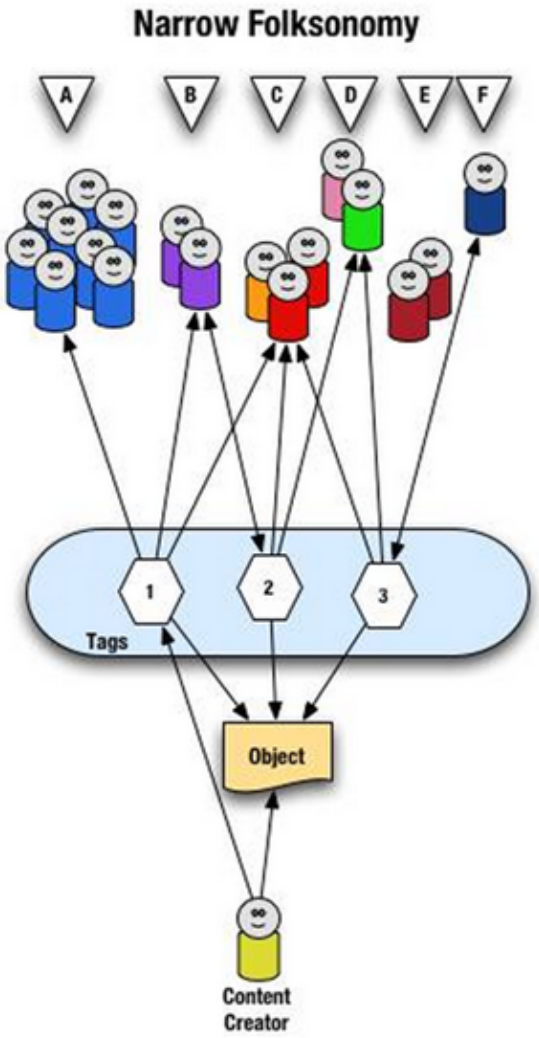

Abbildung 3.7: Narrow Folksonomy

und der Gruppe von Tags (Tag 4, Tag 5 und Tag6) bestehen, da sie alle zur Beschreibung von Objekt 5 verwendet wurden.

Auf manchen Webseiten wie Flickr und Youtube können Benutzer ihre Interessensbereiche und Sachkenntnisse explizit angeben oder Gruppen beitreten. Aus diesen können gemeinsame Interessen abgeleitet werden. In diesen Gruppen können Benutzer mit denselben Interessen Inhalte taggen und teilen. Dabei muss berücksichtigt werden, dass die meisten Community-Portale redundante Gruppen dulden. Zum Beispiel kann auf Flickr jeder Benutzer beliebige Gruppen erstellen, ohne dass kontrolliert wird ob die neu erstellte Gruppe bereits existiert. Dies führt dazu dass es zum Beispiel auf dem Fotoportal Flickr für den Begriff 'Wein' 5204 verschiedene Grup-

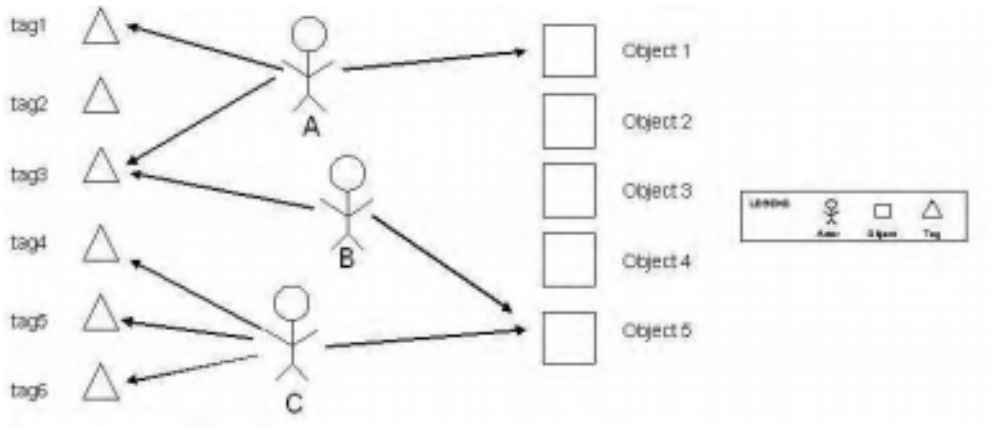

Abbildung 3.8: Sozialer Aspekt in Folksonomies

pen[11] gibt. Das Zusammenlegen dieser redundanten Gruppen offenbart nützliche Daten.

Die Objekte und Ressourcen können somit nicht nur in einem Taggingsystem mehrmals vorhanden sein sondern sich auch auf unterschiedlichen Systemen befinden. So gibt es den Begriff 'Wein' wie vorhin beschrieben des Öfteren auf Flickr sowie auch auf dem Internet Lesezeichen Portal Delicious. Eine solche Verbindung zwischen den Systemen kann als implizite Verbindung bezeichnet werden. Bei solchen Vergleichen zwischen mehreren Tagging-Systemen muss jedoch unterschieden werden, ob es sich bei dem System um eine Narrow- oder einer Broad-Folksonomy handelt.

Systeme können aber auch *explizit* miteinander verbunden sein. Diese Verbindung kommt meist durch die soziale Beziehungen der beteiligten Benutzer zustande. Persönliche Informationen über einen Nutzer können mittels FOAF (Friend of a friend)[12] ausgedrückt werden. Es erlaubt den Nutzer sich selbst (z.B. Vorname, Nachname, Freunde), eigene Online-Benutzerkonten, Gruppen und Dokumente auf eine einfache Art und Weise formell[13] zu beschreiben. Ohmukai et. al (vlg. [Ohmukai2005]) haben ein soziales Bookmarking-System vorgeschlagen, das sich von verschiedenen Metadaten und soziale Netzwerke bedient, um eine Community basierende Ontologie zu erstellen. Das vorgeschlagene System erlaubt es Nutzern die Lesezeichen seiner virtuellen Freunde zu durchforsten. Die Nutzer können anhand von FOAF ihre Beziehung zu den Freunden beschreiben, somit können sie die Tags ihrer Freunde importieren und diese vom System mit ihren eigenen Tags zuordnen lassen. Dies setzt jedoch das Vertrauen zu den eigenen Freunden sowie die ausführliche Angabe der persönlichen Daten voraus.

[11] http://www.flickr.com/search/groups/?q=wines – Zugegriffen am 24.01.2012
[12] http://www.foaf-project.org/ - Zugegriffen am 25.01.2012
[13] http://xmlns.com/foaf/spec/ - Zugegriffen am 24.01.2012

3.4 Vorteile

Folksonomies ist eine neue Methode um Inhalte zu erfassen und diese zu indexieren. Diese bringt einige Vorteile mit sich:

- **Vokabular und Wissensstand der Benutzer**
  Folksonomies spiegeln das aktuelle Vokabular der Benutzer wieder. Auch für Mathes (vgl. [Mathes2004]) liegt der größte Vorteil einer Folksonomy darin, dass es direkt das Vokabular der Nutzer reflektiert. Dies ist auch eines der grundlegendste Unterschied gegenüber einer klassischen Wissenspräsentation. Dort werden die Metadaten für die Dokumenten von dem Ersteller selbst oder von Experten hinzugefügt. Während bei einer Folksonomy die Wortwahl, Terminologie und Präzision der unmittelbaren Nutzer wiedergespiegelt wird. Laut Kipp (vgl. [Kipp2006]) gibt es bei der Indexierung drei verschiedene Akteuere: Autoren, professionelle Indexer und Nutzer. Von den verschiedenen Akteuren wird jeweils eine unterschiedliche Methode zur Indexierung verwendet: Der Autor wendet die textorientierte Methode an, wo die Sprache des Autors im Mittelpunkt steht, z.B. bei der Textwortmethode oder bei der Zitationsindexierung. Bei der Folksonomy wird die Sprache der Nutzer berücksichtigt.

- **Kostengünstige Form der Inhaltserschließung**
  Folksonomies sind eine günstige Form der Inhaltserschließung. Durch die freiwillige und kostenlose Zusammenarbeit der Nutzer, hält sich der Zeit- und Arbeitsaufwand für die Pflege und Indexierung in Grenzen. Jeder Benutzer taggt, was und wie er es für sinnvoll erachtet. Eine Folksonomy wird umso besser, je mehr Leute sich an ihr beteiligen.

  Motho et. al (vgl. [Hotho2006]) sieht den großen Erfolg einer Folksonomy ebenfalls darin, dass es für die Teilnahme keine speziellen Kenntnisse erfordert und es somit keine Einstiegshürden gibt. Der Nutzer muss keine spezielles Vokabular und kein hierarchische Nomenklatur kennen.

- **Konkretes Suchen und Browsing**
  Der Nutzer kann den gewünschten Inhalt durch die Eingabe von Schlagwörtern suchen, die dann mit den Tags verglichen werden. Für den Nutzer besteht ebenfalls die Möglichkeit sich durch die Schlagwörter durchzublättern. Für Mathes (vlg. [Mathes2004]) ist dies eine zentrale Eigenschaft und Stärke für die Recherche mit Folksonomies. Eine Studie von Sinclair et. al (vlg. [Sinclair2008]) hat ergeben, dass einige Nutzer völlig auf die Eingabe von Suchargumenten verzichten und sich durch die Tag Clouds zur gewünschten Information durchklicken. Das ermöglicht das Auffinden von Inhalten, die sonst dem Nutzer verborgen geblieben wären. Dieser Aspekt ist laut Mathes als 'Serendipitous Browsing' (glückliches Blättern) bekannt.

- **Community Bildung**
  Die Entstehung von Communities sind ein Nebeneffekt von Folksonomies. Inhalte werden in Gruppen bzw. Communities mit Gleichgesinnten geteilt. Zudem können sich Nutzer getaggte Inhalte von anderen Personen anzeigen lassen, die dieselben Schlagwörter verwendet haben wie sie.

- **Kollaborative Recommendersysteme**
  Auf Basis der von einem Nutzer verwendeten Schlagwörter, angegebenen Interessensgruppen und indexierten Inhalte wird es möglich, für ihn spezielle Inhalte vorzuschlagen. Diederich et. al (vgl. [Diederich2006]) entwickelten ein implizites kollaboratives Recomendersystem, das für den jeweiligen Nutzer für ihn interessante Inhalte anzeigt.

- **Feedback**
  Sobald ein Nutzer einen Inhalt mit einem Tag versieht, bekommt er weitere Objekte angezeigt, die mit demselben Tag versehen wurden. Er kann gleich entscheiden, ob die Wahl des Tags für den Inhalt passend war. Der Nutzer kriegt somit ein sofortiges Feedback für die vergebenen Tags.

3.5 Nachteile

- **Fehlendes kontrolliertes Vokabular**
  Eines der größten Vorteile ist auch gleichzeitig der größte Nachteil. Ein offenes und unkontrolliertes Vokabular bringt oftmals Uneindeutigkeiten. Wie zum Beispiel Nomen in Plural bzw. Singular, Abkürzungen, verschiedene Wortformen oder mit einem Unterstrich zusammengeschriebene Wörter. Zudem schleichen sich Rechtschreib- oder Tippfehler ein. Schlagwörter werden unterschiedlich geschrieben und Synonyme werden meist nicht zusammengefasst. Laut Guy und Tonkin (vgl. [Guy2006]) sind fast 28 Prozent der Del.icio.us-Tags und fast 40 Prozent der Flickr-Tags mit Rechtschreibfehler behaftet. Wobei falsch kodierte Wörter, zusammengesetzte Wörter oder mehrere Wörter aus verschiedenen Sprachen als Rechtschreibfehler gezählt wurden.

- **Unterschiedliche Motivation**
  Nutzer taggen Inhalten aus verschiedenen Motivationen und Kontexten. Ein Nutzer taggt ein Objekt für seine Freizeitaktivitäten, ein anderer hingegen taggt dasselbe Objekt aus beruflichen Zwecken und noch ein anderer ist ein wahrer Experte.

  Laut Golder et. al (vgl. [GolHub2006]) sind folgende Faktoren für die Erschließung des Inhaltes verantwortlich: die Erfahrung, die sprachliche Ausdrucksstärke, die kognitive Fähigkeit und die Motivationen. Außerdem muss unterschieden werden ob der Nutzer ein Objekt für privaten Nutzen oder aus öffentlichem Interesse taggt.

- **Fehlende Relationen**
  Bei Folksonomies gibt es keinen Gebrauch von Relationen zwischen den Begriffen, sogenannte paradigmatischen Relationen wie z.B. Hierarchierelationen, Assoziationsrelationen oder Äquivalenzrelationen. Solche Relationen sind ein wichtiger Bestandteil von Thesauri und Klassifikationssysteme.

Die Vorteile überwiegen zwar, dennoch sollte weiter versucht werden die Nachteile einer Folksonomy soweit als möglich zu minimieren. Einige Taggingsysteme versuchen bereits Synonyme zu finden und diese dem Benutzer als Alternative vorzuschlagen. Die Verwendung des

Vorschlages darf aber unter keinen Umständen verpflichtend sein, da ein solches Vorgehen gegen die Grundidee einer Folksonomy läuft. Synonyme haben zudem oft nicht die exakt gleiche Bedeutung.

Das Verwenden von Tagging-Werkzeugen und somit das Beitragen zu Folksonomies wird wegen ihrer resultierenden sozialen Aspekte immer gängiger. Dadurch erlauben sie eine persönliche Suche und zeigen dem Nutzer nicht nur genau den Inhalt an den er sucht, sondern auch neue Aspekte seiner Schlagwörter. Wie wir bei Delicious oder Flickr sehen, dient eine Folksonomy nicht nur der Organisation von persönlichen Daten. Viel öfters wird es zum Austausch und Teilen in sozialen Netzwerken verwendet.

3.6 Information Retrieval

Die Wiederauffindung der getaggten Dokumente birgt einige Probleme in sich, da die Inhalte nach dem Tagging-Prozess gänzlich unsortiert sind. Laut Butterfield et. al (vgl. [Butterfield2006]) hat Yahoo! für Flickr bereits 2006 eine Patentanmeldung eingereicht in der sie das Ranking von Suchergebnissen nach sieben Kriterien definieren:

1. Die Anzahl der Tags zu einem Dokument
2. Die Anzahl der Nutzer die ein Dokument taggen
3. Die Anzahl der Nutzer, die das Dokument als Resultat einer Suchanfrage erhalten
4. Die Zeit an dem das Dokument erstellt wurde, desto älter desto geringer die Relevanz
5. Die Relevanz der Tags zu dem Suchbegriff
6. Ortsgebundene Informationen der Dokumente und Aufenthaltsort der Nutzer
7. Vorzüge des Nutzers, aus bereits favorisierten Dokumenten

Der vierte Punkt muss mit Vorsicht behandelt werden, da z.B. ein Bild von Vincent Van Gogh heutzutage nur mehr eine sehr geringe Relevanz hätte. Dies wäre aber besonders Kunsthistorikern ein Dorn im Auge sein.

Peters et. al (vgl. [Peters2008]) sieht für das Relevance Ranking der getaggten Dokumente folgende Faktoren als ausschlaggebend:

- Die (informationslinguistisch 'bereinigten') Tags selbst
- Die Kollaboration in Web 2.0 Diensten
- Nutzerspezifische Rankingkriterien (siehe Abbildung 3.9)

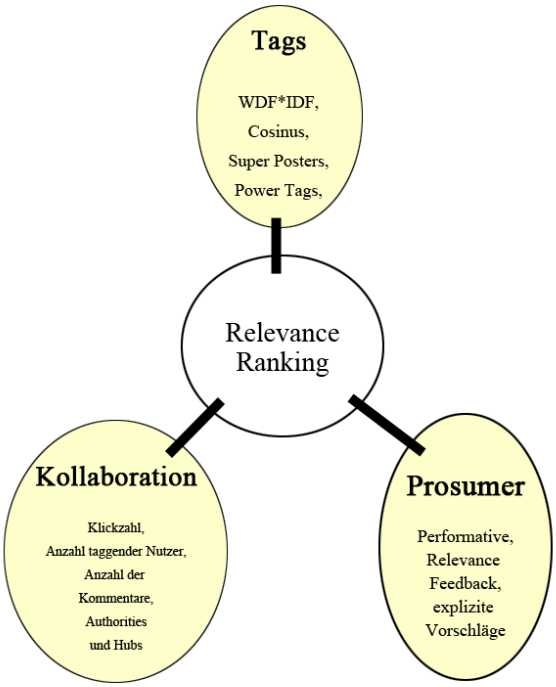

Abbildung 3.9: Kriterien für das Relevance Ranking von getaggten Dokumenten

Faktor Tags

Das Relevance Ranking ändert sich je nach Gewichtung der einzelnen Faktoren. Der erste Faktor 'Tags' (WDF*IDF, Cosinus, Super Poster, Power Tags) beziehen sich auf das Schlagwort selbst. Laut Stock (vlg. [Stock2007]) kann für die Bestimmung der Ähnlichkeit zweier Dokumenten der Vektorenraum verwendet werden. Als Dimension können die verschiedenen Tags in der Datenbank angesehen werden, wobei sich der Wert einer jeweiligen Dimension durch WDF * IDF errechnen lässt:

$$WDF(t,d) = \frac{[ld(freq(d,t)+1]}{ldL}$$

Wo freq die dokumentspezifische Häufigkeit eines bestimmten Tags t ist. Der Wert L die Gesamthäufigkeit aller Tags im jeweiligen Dokument d. Der Wert von IDF lässt sich wie folgt berechnen:

$$IDF(t) = ld(\frac{N}{n})$$

Wobei N die Anzahl aller Datensätze in der Datenbank ist und n die Anzahl der Dokumente die mittels t indexiert worden sind. Der Wert von IDF gibt das Vorkommen eines Tags t in der gesamten Menge aller Datensätze in einer Datenbank wieder.

Die Dokumente werden durch Vektoren abgebildet. Der Cosinus bestimmt die Ähnlichkeit zwischen Dokument und Suchabfrage. In der Broad Folksonomy hängt der WDF Wert von der Anzahl der Index-Tags ab. In der Narrow Folksonomy bestimmt die Anzahl der Such-Tags den WDF-Wert.

Hotho et. al (vgl. [Hotho2006]) fügten der Relevanz Berechnung einen modifizierten PageRank hinzu, der nach dem Motto: 'Ein Inhalt der mit einem wichtigen Schlagwort von einem wichtigen Nutzer getagt wurde, erscheint selbst als wichtig!' arbeitet. Hotho et. al entwickelte zudem den FolkRank-Algorithmus der iterativ die Verbindungsstruktur zwischen Benutzern, Schlagwörter und Quellen analysiert und daraus die Relevanz berechnet. Dabei werden 'Super-Posters' bzw. 'Super-Autors' identifiziert, die eine große Menge an Inhalten veröffentlicht und so als Experten angesehen werden. Daher sollten Inhalte die von einem 'Super-Poster' stammen im Relevance Ranking wichtiger erscheinen als andere.

'Power Tags' sind Schlagwörter, die sehr oft für die Verschlagwortung einer Ressource verwendet werden. Ein solcher Tag hat bei der WDF-IDF-Gewichtsgleichung einen Faktor $f > 1$.

Faktor Kollaboration

Der Faktor 'Kollaboration' hat die Zusammenarbeit der Nutzer als Schwerpunkt. So fließen die Anzahl der Klicks und die Anzahl der unterschiedlichen Tagger in die Berechnung der Relevanz ein. Die Anzahl der Kommentare, zu einem bestimmten Inhalt, können die Relevanz erhöhen. Dabei muss gesagt werden, dass der Inhalt der Kommentare ausgewertet werden müsste. Da sich eine negative Kritik eher nachteilig für das Ranking auswirken würde.

Linktopologische Algorithmen wie der PageRank-Algorithmus[14] oder der Kleinberg- Algorithmus[15] können dazu verwendet werden die Wichtigkeit eines Inhaltes zu definieren. Dabei werden Hub und Authority Werte ermittelt und aus denen wird das Relevance Ranking erstellt. Ein Hub ist ein Inhalt der auf andere Inhalte verlinkt, hingegen eine Authority ist ein Inhalt der von anderen verlinkt wird.

Faktor Nutzer

Der dritte Faktor bezieht sich auf den Nutzer selber. Inhalte die mit 'toRead' bzw. 'toWatch' gekennzeichnet werden haben für den Nutzer einen höheren Stellenwert. Das häufige Vorkommen von negativen Schlagwörtern wie 'boring' oder 'bad' hingegen wirkt sich auf das Ranking negativ aus.

Dem Nutzer kann auch direkt die Gelegenheit gegeben werden sich durch eine Rückmeldung aktiv Einfluss auf das Ranking zu nehmen. Dies kann durch ein Bewertungssystem, z.B. mit der Vergabe von 1 bis 5 Sternen oder einem 'like'- und 'dislike'-Button, geschehen. Es ist ein demokratischer Ansatz der Bestimmung des Relevance Rankings. Diese Bewertung kann nicht nur für das eigene Ranking sinnvoll sein, sondern auch Vorteile für andere Nutzer bringen, zum Beispiel '10 andere Nutzer fanden dieses Dokument sehr nützlich'.

Aus Sicht des Nutzers ist es wichtig, dass die Möglichkeit besteht einzelne Kriterien selbst zu gewichten. Dadurch kann vom Nutzer entschieden werden nach welchen Kriterien die Such-

[14] http://de.wikipedia.org/wiki/PageRank - Zugegriffen am 12.02.2012

[15] http://de.wikipedia.org/wiki/Hubs_und_Authorities - Zugegriffen am 12.02.2012

ergebnisse angezeigt werden sollen. Manche Nutzer möchten zum Beispiel die Ergebnisse lieber nach Anzahl der Aufrufe, andere hingegen nach der Anzahl der verlinkten Seiten und wiederum andere möchten die von ihren Freunden empfohlenen Inhalte ganz oben in der Ergebnisliste sehen.

FolkRank

Eine Folksonomy beschreibt die Nutzer, die Ressourcen, die Tags, sowie die benutzerorientierte Zuteilung von Tags zu Ressourcen. Hotho et. al präsentiert dazu eine formale Definition wie folgt:

Definition: Eine Folksonomy ist ein Tupel $\boldsymbol{F} := (U, T, R, Y)$ wo:

- U, T, und R endliche Mengen sind, deren Elemente User, Tags und Ressourcen heißen
- Y eine tenäre Beziehung zwischen diesen ist, d.h. $Y \subseteq UxTxR$, dessen Elemente 'tag assignments' (TAS) bzw. Tag-Zuordnungen heißen

Die Personomy P_u vom Nutzer u ist die Einschränkung von $\boldsymbol{F}$ auf die Kriterien u.

Ein User wird meist mit einer eindeutigen ID identifiziert und Tags mit einem beliebigen String gespeichert. Die Art der Ressource hängt von dem jeweiligen System ab, zum Beispiel sind es bei Flickr Fotos, bei Delicious Lesezeichen. Auf der Implementierungsebene werden Ressourcen auch als ID's dargestellt.

Damit Motho et. al ihren neuen Algorithmus testen konnten sammelten sie auf Delicious im Zeitraum vom 27ten bis 30ten Juli 2005 folgende Datensätze: 75.242 User $|U|$, 533.191 Tags $|T|$, 3.158.297 Ressourcen $|R|$ und 17.362.212 Tag-Zuordnungen $|Y|$. Diese wurden anschließend in einer MySQL Datenbank abgespeichert.

Der FolkRank Algorithmus orientiert sich an den bekannten PageRank Algorithmus. Jedoch kann er wegen der unterschiedlichen Struktur (ungerichtete dreifach Hyperkanten anstatt gerichtete Zweichfachkanten) nicht direkt auf Folksonomies angewendet werden.

Bei der Anpassung vom PageRank-Algorithmus an die Folksonomie Struktur werden die Hyperkanten zwischen den Usern, Tags und Ressourcen in ungerichtete Kanten transformiert. Motho et. al ging wie folgt vor:

- Dabei wird die Folksonomy $F = (U, T, R, Y)$ in $G_F = (V, E)$ konvertiert. Wo die Knoten $V = U \cup T \cup R$, eine disjunktive Vereinigung von User, Tags und Ressourcen bilden
- Und die Kanten $E = \{\{u, t\}|\exists r \in R : (u, t, r) \in Y\} \cup \{\{t, r\}|\exists u \in U(u, t, r) \in Y\} \cup \{\{u, r\}|\exists t \in T : (u, t, r) \in Y\}$ jeweils die Verbindung zwischen User und Tag, Tag und Ressource, User und Ressource ausdrückt.

Der PageRank Algorithmus verfolgt die Idee, dass eine Seite von großer Bedeutung ist, wenn viele Seiten auf diese verlinken und diese verlinkenden Seiten als wichtig eingestuft werden. Laut Motho et. al wird das auch bei dem FolkRank Algorithmus angewendet, jedoch erweitert auf Tags, Ressourcen und User. So wird eine Ressource als wichtig angesehen wenn sie mit

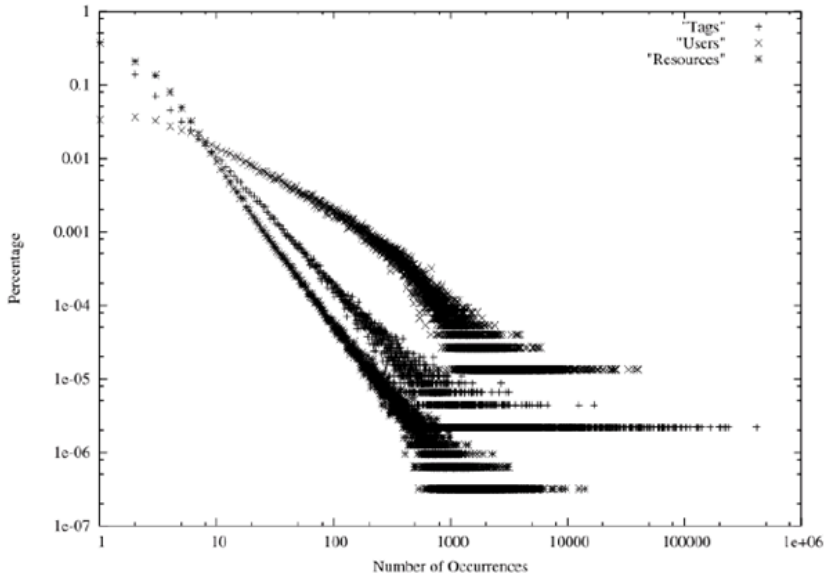

Abbildung 3.10: Verteilung der gesammelten Datensätze

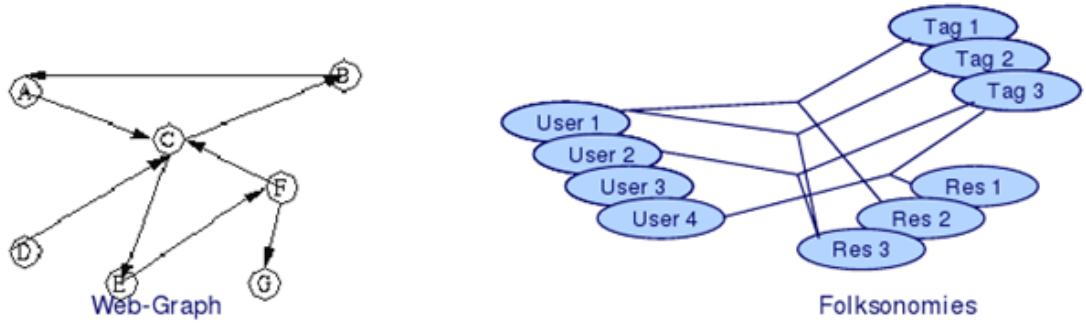

Abbildung 3.11: Strukturunterschied von Web-Graphen und Folksonomies

wichtigen Tags von wichtigen Nutzern versehen worden ist. Dies gilt symmetrisch auch für Tags und Nutzer.

Wie der PageRank verfolgt auch der FolkRank das Modell des Zufallsurfers, das besagt, dass ein typischer Internetbenutzer mittels Hyperlinks durch das Web surft. Jedoch gelegentlich auf eine neue Internetseite kommt ohne einen Hyperlink gefolgt zu sein.

Aus diesen Annahmen ergeben sich für den PageRank folgende Definition:

$$\vec{w} \leftarrow dA\vec{w} + (1-d)\vec{p}$$

Wobei A die Adjazenzmatrix des Graphen ist, $\vec{w}$ der Rankvektor, $\vec{p}$ der Präferenzvektor des jeweiligen Surfers und $d \in [0,1]$ der Gewichtungsfaktor.

Damit dieser Algorithmus für eine Folksonomy, die aus drei verschiedenen Faktoren (Tag, User, Ressource) besteht, angewendet werden kann, muss er wie folgt umgeschrieben werden:

$$\vec{w} \leftarrow \alpha\vec{w} + \beta A\vec{w} + \gamma\vec{p}$$

Wobei A die Adjazenzmatrix der Graphen ist, $\vec{p}$ der Präferenzvektor, $\alpha, \beta, \gamma \in [0, 1]$ sind Konstanten, wobei $\alpha+\beta+\gamma = 1$ gelten muss. Die Konstante α regelt die Stärke der Konvergenz, mit β und γ kann Einfluss auf den Präferenzvektor genommen werden. Durch das höher Bewerten einzelner oder mehrerer Tags, Benutzer und/oder Ressourcen kann ein themenspezifisches Ranking erstellt werden.

Eine direkte Anwendung dieses Ansatzes hat zur Folge, dass das Resultat stark von den global wichtigen Knoten dominiert wird und die Präferenz nur sehr schwach wiederspiegelt. Dies resultiert aus der ungleichen Verteilung verschiedener Elementen in der Folksonomy und aus der Tatsache, dass ungerichtete Graphen verteiltes Gewicht zum Teil unmittelbar wieder entlang derselben Kanten zurückgeben. Um diesen Nachteil zu umgehen berechnet der FolkRank einen Durchlauf mit Präferenzvektoren und einen ohne. Die Differenz der beiden Läufe bildet das eigentliche Ergebnis.

Der FolkRank Algorithmus erstellt laut Motho et. al ein themenspezifisches Ranking in einer Folksonomy wie folgt:

1. Der Präferenzvektor $\vec{p}$ wird zur Bestimmung des Themenbereiches verwendet. Es kann jegliche Gewichtsverteilung besitzen, unter der Voraussetzung dass $\|\vec{w}\|_1 = \|\vec{p}\|_1$ gilt. Spezielle Themen können bestimmt werden indem ein oder mehrere Tags und/oder ein oder mehrere Nutzer und/oder ein oder mehrere Ressourcen stärker gewichtet werden

2. Wo $\vec{w_0}$ ein Fixpunkt der Gleichung (1) mit $\beta = 1$ ist

3. Wo $\vec{w_1}$ ein Fixpunkt der Gleichung (1) mit $\beta < 1$ ist

4. $\vec{w} := \vec{w_1} - \vec{w_0}$ der finale Gewichtungsvektor ist

3.7 NLP eine Problemlösung für Folksonomies

Laut Peters et. al kann ein Großteil der Probleme der Folksonomies auf zwei verschiedenen Verfahren gelöst werden. Bei der ersten Vorgehensweise wird versucht dem Tager bzw. Prosumer zu einem besseren und effektiveren Taggingverhalten zu erziehen. Dazu muss das Taggingverhalten der Nutzer studiert und analysiert werden.

Durch das Vorschlagen, vom System, von geeigneten Tags, wird der Nutzer unterstützt und trainiert. Laut (vgl. [Xu2006]) kann dieses Vorschlagssystem auf zwei Ebenen arbeiten. Durch die syntaktischen Vorschläge (z.B. den Tag 'Auto' durch den Tag 'Autos' zu ergänzen) und den relationalen Hinweisen (z.B. der Nutzer verwendet den Tag 'Fahrzeug' und das System schlägt den Tag 'Auto' vor, da ihn der Benutzer bereits verwendet hat).

Eine ausführlichere Unterstützung durch eine Tag-Empfehlung bringt aber auch Probleme mit sich. Da ein System für ein Dokument meist die bereits am häufigsten vergebenen Tags

vorschlägt. Wenn sich die Nutzer daran orientieren, entsteht automatisch ein von dem System erstellte Tag-Verteilung nach dem Power Law.

Die Tags können zudem als Elemente der natürlichen Sprache angesehen werden und mit der Methode des Natural Language Processing (NLP)[16] verarbeitet werden. Nach einer Studie von Guy et. al (vgl. [Guy2006]) sind ca. 90 Prozent der Tags Nomen. Peters et. al bevorzugt für die NLP-Verarbeitung den Wort-basierten Ansatz mit der Annahme, dass Tags ausschließlich Nomen sind.

Zumal nicht alle Dokumente getaggt sind, müssen zu mindestens die nicht getaggten textuellen Dokumente automatisch mit Tags versehen werden. Bei Blogs stehen dazu der gesamte Text zur Verfügung. Bei Fotos und Videos der Titel und eine eventuelle Beschreibung. Das System bestimmt für jedes ungetaggte Dokument jeweils drei Tags. Der automatische Vorgang sollte nicht zu oft genutzt werden, da laut einer Studie von Kahlifa (vgl. [Khalifa2007]) die automatische Indexierung sehr selten mit der Indexierung des menschlichen Taggers korreliert. Abgesehen davon wiederspricht das automatisierte Tagging dem Grundgedanken des intellektuellen Taggings.

Nach der Sprach- und Wortidentifikation (Abbildung 3.12) werden bei der Identifikation kontextspezfischer Tags, wie z.B. 'me' und 'ich' bearbeitet. Diese werden automatisch durch den Benutzernamen des Taggers ersetzt. Für nicht angemeldete Nutzer werden Suchanfragen mit 'ich' herausgefiltert.

Die weiteren Bearbeitungsschritte Fehlererkennung und –behebung, Wortformzusammenfassung, Identifikation von Eigennamen, Phrasenerkennung und Dekomposition sind typische NLP-Aufgaben dich der Reihe nach abgearbeitet werden. Für diese Aufgaben muss sich das System mit einem Vokabular behelfen. Für die Erkennung von Synonymen und Homonymen kann WordNet[17] eingesetzt werden. Mit mehrsprachigen maschinenlesbaren Wörterbüchern kann der Zugriff auf fremdsprachigen Dokumenten erleichtert werden.

[16] http://de.wikipedia.org/wiki/Computerlinguistik - Zugegriffen am 21.02.2012

[17] http://wordnet.princeton.edu/

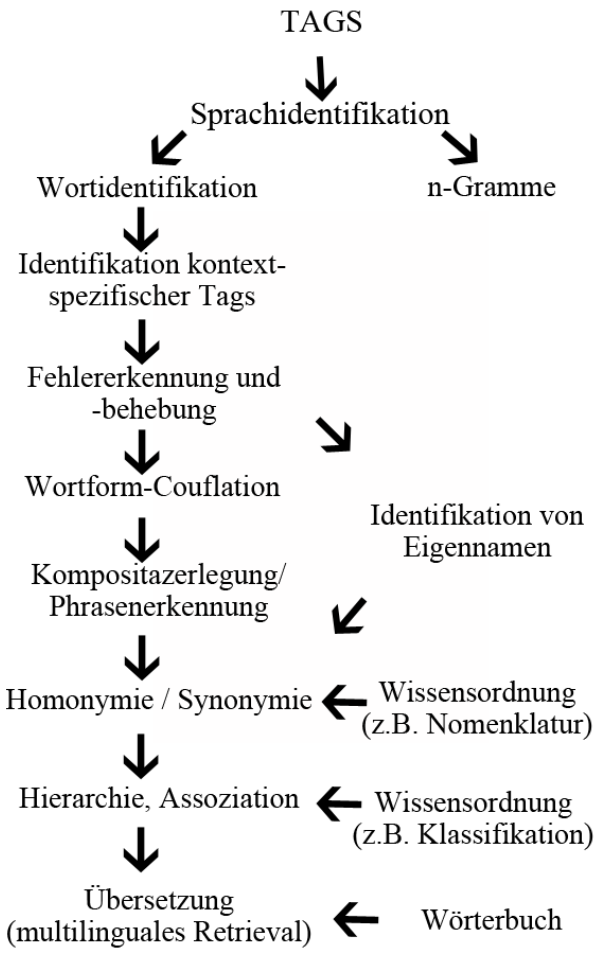

Abbildung 3.12: Der Aufgaben-Algorithmus der Tag-NLP

KAPITEL 4

Tagging Systeme

In diesem Kapitel wird der Tagging-Vorgang vom Anregungen sammeln bis hin zur Vergabe der Tags genauer analysiert und anhand eines Beispiels beschrieben. Die Motivationen, die einen Nutzer dazu bewegen Ressourcen zu taggen werden mittels einer Studie aufgezeigt. Dabei treffen für die meisten Nutzer mehrere Motivationen gleichzeitig zu. Anschließend werden die verschiedenen Merkmale eines Taggingsystems beschrieben und bei dem Vergleich der bereits verfügbaren Taggingsysteme gegenübergestellt. Die graphische Darstellung der gesammelten Tags in einem Taggingsystem ist ein weiterer Punkt dieses Kapitels. Abschließend werden vier der vorhandenen Information retrieval Tools analysiert und miteinander verglichen.

4.1 Tagging-Prozess

Eine Ressource zu tagen ist an sich sehr einfach, der Gedankenprozess der dabei im Gehirn abläuft ist jedoch viel komplexer. Rashmi teilt in 'Theorie of tagging – My hypothesis' (vgl. [Rashmi2005]) den Tagging-Prozess in zwei Schritte ein:

1. **Related Category activation:**
   Anregung ähnlicher Kategorien:
   Der erste Schritt besteht darin die Ähnlichkeit zwischen den zu taggenden Gegenstand und dem möglichen Konzept einzuschätzen. Zum Beispiel, in der Bibliothek bei der Betrachtung des Buches 'Der Herr der Ringe'. Sofort werden im Gehirn ähnliche semantische Konzepte aktiviert, wie 'Buch', 'Fantasy', 'Frodo', 'John Tolkien'. Andere Konzepte sind persönlicher, wie 'Lieblingsbuch', 'spannend'. Wiederum andere Konzepte die aktiviert werden betreffen den physischen Zustand des Buches: 'neu', 'gebunden'. In diesem Abschnitt werden noch keine Konzepte aussortiert, sonder vorerst nur gesammelt.

2. **The decision:**
   Die Entscheidung:

Sobald einige mögliche Konzepte gesammelt wurden, muss bestimmt werden welche Tagging Wörter für eine Ressource wohl die passendsten sind. Das menschliche Gehirn ist für solche Entscheidungen sehr gut ausgerüstet. Aber meist fällt es uns trotzdem schwer, da es bei virtuellen Ressourcen meist einen geringeren kulturellen Konsens gibt. In der digitalen Welt versuchen wir nicht nur Objekte zu kategorisieren, sondern sie auch für eine zukünftige Wiederauffindung zu optimieren. Dazu sollte sich ein gewisser persönlicher Standard verfolgt werden. Damit sich nach einigen getaggten Objekten nicht zu viele Einträge in derselben Kategorie befinden. Außerdem wird ein Objekt von einer Person meist nur einmalig getaggt, da eine Änderung der Schlagwörter meist zu aufwendig ist. Die wirkt sich nicht positiv auf die Entscheidung der passenden Schlagwörter aus.

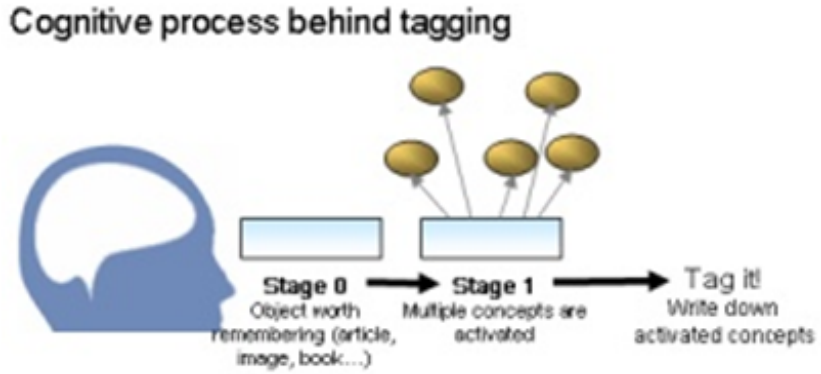

Abbildung 4.1: Taggingprozess

4.2 Studie zur Motivation von User Tagging

Ansporn und Motivation spielen bei dem Tagging-Vorgang eine signifikante Rolle. Benutzer tagen sowohl aus persönlichem Nutzen als auch aus sozialem Interesse. Tags sind meist für die Allgemeinheit nützlich, jedoch werden die meisten Tags nur für den Eigenbedarf genutzt. Das Tagverhalten bzw. den Anreiz Objekte zu Tagen hängen sehr stark von dem zur Verfügung gestelltem System und von der zu taggenden Ressource ab.

Marlow et. al (vgl. [Marlow2006]) teilt die Motivation in zwei Oberkategorien ein: organisatorisch und sozial. Der organisatorische Anreiz kann als Alternative zur strukturierten Ablage angesehen werden. Die Benutzer versuchen für dies meist einen persönlichen Standard zur Vergabe von Schlagwörtern zu entwickeln, mit dem sie ein Wiederfinden zu einem späteren Zeitpunkt erleichtern. Hingegen bei der sozialen Motivation steht der kommunikative Ansatz des Taggens im Vordergrund. Dabei versuchen die Benutzer ihre eigene Meinung auszudrücken oder der Ressource Aufgaben, wie 'lesen' oder 'drucken', anzufügen.

Marlow et. al beschreibt sechs verschiedene Motivationen, dabei ist zu beachten, dass sie sich nicht gegenseitig ausschließen. Für die meisten Nutzer treffen größtenteils mehrere Motivationen gleichzeitig zu.

1. **Future retrieval**
   Späteres Wiederauffinden
   Objekte werden für den Eigennutzen getaggt, damit ein Wiederauffinden in der Zukunft erleichtert wird. Diese Schlagwörter können gelegentlich zu einer Aktion anregen, wie zum Beispiel das Tag 'to read', oder aber auch als Erinnerung dienen.

2. **Contribution and sharing**
   Beitragen und Teilen
   Dazu zählen Ressourcen die hinzugefügt wurden um andere Nutzer auf bestimmte Ressourcen aufmerksam zu machen, wie zum Beispiel das Taggen von Veranstaltungs- Webseiten bzw. Aktionen. Dabei kann das Schlagwort für einen Freund als auch für einen Unbekannten hinzugefügt werden.

3. **Attract Attention**
   Aufmerksamkeit erwecken
   Um andere Benutzer dazu zu bewegen auf selber getaggte Ressourcen zu bringen, sollte bei der Wahl der Tags auf gängige Wörter zurückgegriffen werden. Einige Nutzer versuchen bestimmte Wörter zu pushen um diese in 'Tag clouds'(Querverweis auf Artikel) wichtig erscheinen zu lassen. Dieses Ziel wird laut Marlow et. al vor allem von Spammern verfolgt.

4. **Play and Competition**
   Spiel und Wettbewerb
   Die Motivation wird durch ein Computerspiel, einem 'Game with a purpose' (GWAP)[1], erweckt. Dabei wird spielend die Fähigkeit des Menschen genutzt, um ein gewisses Ziel zu erreichen. Die Mitspieler müssen Tags zu einer gezeigten Ressource hinzufügen. Dies wird vor allem bei nicht textuellen Ressourcen, wie Bildern und Videos, angewendet. Bei dem Online Game 'ESP Game'[2] wird dasselbe Bild zwei Spielern gezeigt, sie müssen anschließend versuchen die gleichen Tags für das Bild zu vergeben.

5. **Self Presentation**
   Selbst Präsentation
   Der Nutzer versucht sich selber, durch das Taggen von Ressourcen und das anschließende Teilen von denen in sozialen Netzwerken, zu präsentieren. Ein Beispiel dazu ist 'seen live' Tag bei Last.FM[3], dabei kann man anderen Nutzern mitteilen, dass man die Band bereits live gesehen hat.

6. **Opinion Expression**
   Meinungsäußerung
   Der Nutzer möchte seine eigene Meinung einer Ressource hinzufügen und seine Ansicht mit andern Nutzern teilen. Der Nutzer fügt ein subjektives Tag, zum Beispiel 'sehenswert', 'langweilig' oder 'lustig', zu einem Video hinzu.

[1] http://www.lrec-conf.org/proceedings/lrec2010/pdf/476_Paper.pdf - Zugegriffen am 14.03.2012
[2] http://www.gwap.com/gwap/gamesPreview/espgame/ - Zugegriffen am 09.12.2011
[3] http://www.lastfm.de/ - Zugegriffen am 09.12.2011

	Function		
		Organisation	Communication
Sociality	Self	Retrieval, Directory Search	Context for self Memory
	Social	Contribution, attention, Ad hoc photo pooling	Content descriptors, Social Signaling

Tabelle 4.1: Taxonomie der Motivation von Taggern

Bei der von Marlow et. al erstellte Motivation-Liste ist ersichtlich, dass nur der Punkt 'Future retrieval' ausschließlich der Motivations-Kategorie 'organisatorisch' zugeschrieben werden kann. Bei den Punkten 'Contribution and sharing' und 'Attraction Attention' hingegen können der Kategorie 'sozial' eingegliedert werden. Bei den restlichen steht eine organisatorische als auch eine soziale Motivation dahinter.

Morgan Ames et. al (vgl [Morgan2007]) versuchten anhand einer Studie die Gründe zu finden, wieso Flickr-User und ZoneTag[4]-User ihre Bilder taggen. Flickr ist eine kommerzielle Foto-Sharing-Webseite die es allen Mitgliedern ermöglicht Fotos Online zu speichern und diese in verschiedene Alben zu sortieren. Zudem können die Fotos betitelt, getaggt und veröffentlicht werden. Bei ZoneTag handelt es sich um eine Applikation für Smartphones um Fotos auf einfachem Wege, per zwei Klicks, auf Flickr hinauf zuladen.

Bei der Studie haben 13 Probanden (4 Frauen und 9 Männer) teilgenommen, die regelmäßig Flickr bzw. ZoneTag verwenden. Das Alter der Teilnehmer reichte von 25 bis 45 Jahren, wobei die Meisten im Alter von 25 bis 35 Jahren waren. Bei 9 der 13 Teilnehmer handelte es sich um technisch sehr versierte Personen. Am Anfang des Interviews wurde eine Diskussion über den Gebrauch von Flickr bzw. ZoneTag aufgebaut, anschließend genauer Fragen über die Tag-Verwendung und die Motivation gestellt und am Ende wurde auf die Verwendung spezieller Tag-Wörter eingegangen.

Die Studie ergab eine Vielzahl von Motivationen, dabei war erstaunlich, dass nur einige Wenige ausschließlich für den eigenen Nutzen Tags vergeben. Meistens waren es zwei oder drei Motivationen, die dazu führten. Ein Proband hat zum Beispiel Fotos aus Hawaii deswegen getaggt, damit sie von anderen Usern leicht gefunden werden und um seine eigenen Eindruck über das Bild auszudrücken, wie z.B. 'Aloha Air sucks' oder 'good restaurant'. Der Ansporn eines anderen Users bestand darin, dass er durch das Tagen die Bilder zu einem späteren Zeitpunkt leichter findet und darin, dass Freunde und Bekannte nähere Informationen über das Foto erlangen.

Morgan Ames et. al entwickelten anhand der oben genannten Studie zwei Taxonomien (siehe Tabelle 4.1), um die Motivation grafisch zu veranschaulichen. Die Tabelle ist in zwei Achsen aufgeteilt. Auf der y-Achse werden die sozialen Aspekte berücksichtigt, ob die Fotos für den eigenen Nutzen oder für den Nutzen für andere getaggt wurden. Auf der x-Achse wird der funktionale Aspekt berücksichtigt, dabei kann das Taggen für die Verwaltung bzw. Organisation oder für die Verbreitung bzw. Teilen von Fotos genutzt werden.

- **Self/Organisation: Search and Retrieval**

[4]http://zonetag.research.yahoo.com/ - Zugegriffen am 26.02.2012

Self/Organisation: Suche und Wiederauffindung
Unter diesem Punkt fällt die traditionelle Beschreibung eines persönlichen Fotoalbums. Der Fotograf taggt Bilder um sie ordnen und sie zu einem späteren Zeitpunkt leichter auffinden zu können.

- **Self/Communication: Memory and Context**
  Self/Communication: Erinnerungen und Zusammenhänge
  Bei diesem Punkt fallen solche Motivationen hinein, die zum Tagen von Personen und Orte führen. Auf einem Bild erscheinende Personen können bei Flickr sehr einfach mit dem jeweiligen Benutzername versehen werden. Mit diesen Informationen können zu einem späteren Zeitpunkt Erinnerungen an Ereignisse und Orte geweckt werden. Diese Art Information hinzuzufügen kennt man von dem traditionellen Fotoalbum.

- **Social/Organisation: Public Search and Photo Pools**
  Social/Organisation: Öffentliche Suche und Foto Sammlungen
  Unter diesem Punkt falle jene Motivationen hinein, die das Auffinden von Fotos für andere User erleichtern. Fotos werden oft dazu verwendet um gemeinsame Erlebnisse festzuhalten und sie mit Freunden und Bekannte zu teilen. Bei dem Fotoportal Flickr können Fotos und ganze Fotoalben sehr einfach mit der Öffentlichkeit geteilt werden.

  Flickr umfasst ein Feedback System, das den User, anhand von Statistiken über die Zugriffe, Anzahl der Kommentare, Bewertungssysteme, dahingehend motiviert, seine Bilder mit noch geeigneteren Tags zu versehen. Damit möglichst viele Besucher auf seine Bilder stoßen. Das kann laut Morgan Ames et. al den Ruf eines Users gesteigert werden.

  Einige Teilnehmer der Studie fügten Tags zu ihren Bildern hinzu, die dann in der Liste der vorgeschlagenen Wörter für andere Benutzer aufschienen. Dabei handelt es sich um eine implizite Zusammenarbeit. Wenn diese Art von Koordination erfolgreich ausgeführt wird, ermöglicht es Benutzern nach Fotos zu suchen, die zum Beispiel auf dem gleichen Event oder an demselben Ort geschossen wurden.

- **Social/Communication: Context and Signaling**
  Social/Communication: Kontext und Andeutung
  Die letzte Kategorie der erstellten Taxonomie entspricht der Motivation, einem Bild Kontext bezogene Informationen hinzuzufügen. Um diese mit anderen Benutzern zu teilen. Die er dann mit anderen Benutzern teilen kann. Solche Informationen beinhalten Namen des Fotografs, sowie Erfahrungsberichte über Restaurants oder Hotelaufenthalte. In einigen Fällen geben solche Zusatzinformationen nur für Freunde und Bekannte einen Sinn und werden von Außenstehenden nicht verstanden.

Zusammenfassend kann gesagt werden, dass spezielle Tags in mehrere Kategorien der erstellten Taxonomie fallen können. Zum Beispiel kann ein Ortsname für die Suche für den Fotograf als auch für andere Benutzer hilfreich sein.

Wie in der Tabelle 4.2 dargestellt, fiel die Motivation für die meisten Teilnehmer in die Kategorie 'Social/Organisation'. Gefolgt von 'Self/Organisation' und 'Social/Communication', die fast gleichauf lagen.

	Function	
	Organization	Communication
Self	4, 3	0, 3
Social (friends/family)	1, 1	4, 2
Social (public)	3, 7	0, 1

Tabelle 4.2: Primäre (fett) und sekundäre (kursiv) Motivationen zum Tagen von den 13 Interview Teilnehmern

Nur eine geringe Anzahl von Teilnehmern fügt Tags für Freunde und Bekannte hinzu, damit diese die Fotos leichter auffinden können. Jene Teilnehmer bevorzugen die Fotos gleich per Email zu versenden. Organisation von Ressourcen für die Allgemeinheit ist für Viele häufig wichtiger als die Kommunikation von denen.

4.3 Merkmale der Tagging Systeme

Unterschiede in den Taggingsysteme haben eine große Auswirkung auf die resultierenden Tags und den Informationsgehalt. Zudem beeinflusst der persönliche und soziale Nutzen maßgeblich die Qualität und die Häufigkeit der Tags. Das Design und die Benutzerfreundlichkeit der Systeme sind ebenso ein ausschlaggebender Faktor.

Marlow et al. (vgl. [Marlow2006]) beschreibt einige grundlegende Merkmale von Taggingsysteme die beträchtliche Auswirkungen auf den Inhalt und die Nützlichkeit der hinzugefügten Tags haben:

1. **Tagging Right**
   Tagging Rechte
   Die wohl wichtigste Kennzeichnung eines Taggingsystems ist die Rechtevergabe, welchen Benutzer es erlaubt ist einen oder mehrere Tags hinzuzufügen. Ein Taggingsystem kann auf *self-tagging* eingeschränkt sein, wodurch nur der Erzeuger der Ressource das Objekt taggen darf (z.B. Technorati) oder *free-for-all-tagging*, wo jeder Nutzer die Ressource taggen kann (z.B. Flickr). Die Tag-Rechte können überdies verfeinert werden, sodass nur eine eingeschränkte Gruppe, wie zum Beispiel Freund und Bekannte die Ressource tagen können. Diese Art der Restriktion kann bei Facebook[5] angewendet werden.

   In einem Taggingsystem besteht oftmals auch die Möglichkeit hinzugefügte Tags wieder zu löschen. Dabei wird zwischen Niemand (z. B. Yahoo! Podcast), jeder (z.B. Odeo), dem Tag-Ersteller (z.B. LastFm) oder dem Ersteller der Ressource (z.B. Flickr) unterschieden. Bei einem free-for-all Sytem werden klarerweise mehrere Tags hinzugefügt. Dabei ist zu beachten, dass sich die Systeme meist sehr voneinander unterscheiden. So fügen laug Stoyanovich et. al (vgl. [Stoyanovich2008]) User mit unterschiedlichen Interessen sehr verschiedenartige Tags hinzu.

[5] https://www.facebook.com/ - Zugegriffen am 27.02.2012

2. **Tagging Support**
   Tagging Unterstützung
   Die Hilfestellung beim dem Tagvorgang wirkt sich wesentlich auf die Qualität und die Divergenz der Schlagwörter aus. Sie können in drei verschiedene Kategorien eingeteilt werden: *blind tagging*, wo der Nutzer während des Tagvorganges die bereits vergebenen Tags von den anderen Nutzern nicht sehen kann (z.B. Del.icio.us); *viewable tagging*, wo dem Tagger die bereits assoziierten Tag sichtbar gemacht werden (z.B. Yahoo! Podcast); *suggestive tagging*, wo dem Nutzer eine Liste möglicher Tags vorschlägt (z.B. Yahoo! MyWeb2.0). Die Liste der vorgeschlagenen Tags besteht meist aus bereits assoziierten Tags oder wird automatisch durch das Analysieren der Ressource vom System erstellt. Ein 'suggestive tagging'-System verhilft dazu, dass der Taggebrauch für eine Ressource viel schneller konsolidiert bzw. zusammengeführt wird. Bei einem 'blind tagging'-System ist dies nicht der Fall.

3. **Aggregation**
   Zusammenlagerung
   Ein weiteres Merkmal der Taggingsysteme ist die physische Speicherung der einzelnen Tags. Dies kann zentral oder lokal erfolgen. Bei der zentralen Speicherung wird ein Tag nicht mehrmals mit einer Ressource assoziiert abgespeichert, dies erfolgt zum Beispiel bei Youtbe und Flickr. Marlow et. al nennt dieses Carakteristik *set-model*. Bei der lokalen Speicherung hingegen wird ein identisches Tag mehrmals abgespeichert. Dadurch kommt es zu Duplikaten, wie dies zum Beispiel bei Del.icio.us geschieht. Marlow et. al nennt diese Art der Speicherung *bag-model*.

4. **Type of Object**
   Art des Objektes
   Die Taggingsysteme unterscheiden sich darin, welche Art von Objekten getaggt werden können. Grundlegend können die Objekte in 'textuelle' und 'nicht textuelle' Ressource unterteilt werden. Einige Beispiele für Objekte die heutzutage sehr oft getaggt werden sind: Webseiten (z.B. Del.icio.us, Yahoo! MyWeb2.0), bibliographisches Material (z.B. CiteULike), Blog Einträge (z.B. Technorati, LiveJournal), Bilder (z.B. Flickr, Picasa[6]), Videos (z.B. Youtube), Podcasts (z.B. Yahoo! Podcasts) und Lieder (z.B. LastFm). In Wirklichkeit kann jede Ressource getagt werden die virtuell repräsentiert werden kann.

5. **Source of material**
   Herkunft der Ressource
   Ein weiteres Merkmal ist die Herkunft der zu taggenden Ressource. Sie können von den Teilnehmern selber (z.B. Youtube, Picasa, Flickr) oder von dem System (z.B. ESP Game, Last.fm) bereitgestellt werden.

6. **Ressource connectivity**
   Zusammenhang der Ressourcen
   Die Ressourcen können, unabhängig von den assoziierten Tags, im System auf unterschiedlichste Weise miteinander verlinkt sein. Verbindungen können laut Marlow et. al in

[6] http://picasaweb.google.com – Zugegriffen am 27.02.2012

Dimension	Main categories	Auswirkungen
Tagging rights	Self-tagging, permission-base, free-for-all	Rolle im Tag System, Art und Typ der resultierenden Tags
Tagging support	Blind, suggested, viewable	Konvergenz zu einer Folksonomy, Übergewichtige Tags
Aggregation model	Bag, Set	Verfügbarkeit von Aggregations-Statistiken
Object Type	Textual, non-textual	Art und Type der resultierenden Tags
Source of material	User-contributed, system, global	Unterschiedliche Anreize, Art und Qualität der resultierenden Tags
Ressource connectivity	Links, groups, none	Zusammenführung von ähnlichen Tags für die selben Ressourcen
Socil connectivity	Links, groups, none	Lokale Folksonomy die auf einer sozialen Struktur basiert

Tabelle 4.3: Merkmale der Taggingsysteme

folgende Kategorien eingeteilt werden: '*linked*' (z.B. Webseiten sind per direkten Links verbunden), '*grouped*' (z.B. Picasa Fotos können in Alben eingeteilt werden) und '*none*'. Bei verbundenen Ressourcen ist es sehr hilfreich wenn ähnliche Tags miteinander vereint werden. Dies gilt besonders bei 'suggested-' und 'viewable-Tagging'.

7. **Social connectivity**
   Soziale Verbindung
   Soziale Netzwerke laden gerade dazu ein Nutzer mit Ressourcen, wie Fotos und Videos, zu verbinden. Wie bei der 'Ressource connectivity' können laut Marlow et. al auch hier die Verbindungen in '*linked*', '*grouped*' und '*none*' eingeteilt werden.

In der Tabelle 4.3 sind die Merkmale der Taggingsysteme zusammengefasst. Zusätzlich zu den Kategorien werden die jeweiligen Auswirkungen gelistet.

4.4 Übersicht vorhandener Tagging Platformen

Ein Taggingsystem ist Carlin (vgl. [Carlin2006]) zufolge ein Informationssystem, das den Nutzern die Möglichkeit gibt, eigene oder fremde Dokumente im System bekannt zu machen und mit einem oder mehreren Tags zu versehen. Abhängig vom System müssen die Nutzer keine oder nur wenige formale Daten erschließen. Dem Nutzer soll die Eingabe der Schlagwörter so einfach wie möglich gemacht werden. Infolge werden die bekanntesten Taggingsysteme, wie Evernote, Delicious, Flickr, Picasa, Zootool und Mister Wong vorgestellt.

Evernote

Mit der Software und Webanwendung Evernote[7] steht ein Dienst bereit, die das Finden, Sammeln und Ordnen von Dokumenten, Notizen (Text, Audio und Video) und Fotos ermöglicht. Der Dienst steht seit Juni 2008 für alle gängigen Betriebssysteme und auch als webbasierte Applikation zur Verfügung. Unterdessen gibt es einige Zusatzanwendungen wie zum Beispiel Evernote Hello[8] (Organisieren und Sammeln von Kontakten), Evernote Food[9] (Verwalten von Mahlzeiten, ob Selbstgemacht oder von Restaurantbesuchen) und Evernote Clearly[10] (Benutzerfreundliches Lesen von Webseiten zu einem späteren Zeitpunk). Bei all diesen Diensten kann der Nutzer neue Tags frei hinzufügen und diese chronologisch, nach Tags oder Titeln sortieren lassen.

Der Nutzer kann die Schlagwörter bei der Tagvergabe frei wählen (blind-tagging). Die hinzugefügten Objekte können nur von dem Nutzer getaggt werden, der sie auf Evernote hinzufügt. Die Dokumente und Notizen können mittels Facebook, Twitter und Email mit den Freunden geteilt werden.

Delicious

Bei Delicious handelt es sich wohl um das bekannteste Social-Bookmark-System das dem Nutzer als Webandwendung zur Verfügung steht. Der Nutzer kann damit persönliche Lesezeichen anlegen und diese mit Tags versehen. Die Lesezeichen werden nicht lokal sondern auf einen zentralen Server abgespeichert, somit kann von jedem internetfähigem Gerät mit einem Webbrowser darauf zugegriffen werden. Für eine benutzerfreundliche Bedienung stehen Erweiterungen[12] für den Internet Explorer, Mozilla Firefox und Google Chrome zur Verfügung.

Bei dem Tagvorgang werden dem Nutzer bereits passende Schlagwörter vorgeschlagen (suggestive- tagging). Ein Lesezeichen kann auf ein beliebiges Bild, Text oder Video verweisen (free-for-all). Die Lesezeichen können zudem in Ordner gruppiert werden. Delicious bietet die Möglichkeit mit anderen Nutzern der Plattform zu kommunizieren und den von ihnen getagten Lesezeichen zu sehen.

Flickr

Flickr ist ein Online-Dienst zur Bildbearbeitung aus dem Hause Yahoo!. Der Dienst hat eine sehr starke soziale Netzwerk-Komponente integriert. Sie erlaubt es Bilder und Videos zu kommentieren, Personen zu verlinken, neue Tags hinzuzufügen und Fotos auch nur für gewisse Nutzer sichtbar zu machen. Flickr ist eine Tochtergesellschaft von Yahoo! und ist im Februar 2004 zum ersten Mal ins Netz gegangen. Mittlerweile gibt es auch eine Android Version[13] und eine iOS Version[14].

[7] http://www.evernote.com – Zugegriffen am 15.12.2011

[8] http://www.evernote.com/hello/ - Zugegriffen am 15.12.2011

[9] http://www.evernote.com/food/ - Zugegriffen am 15.12.2011

[10][11]

[12] http://delicious.com/help/tools - Zugegriffen am 15.12.2011

[13] https://market.android.com/details?id=com.yahoo.mobile.client.android.flickr – Zugegriffen am 16.12.2011

[14] http://itunes.apple.com/at/app/flickr/id328407587?mt=8 – Zugegriffen am 16.12.2011

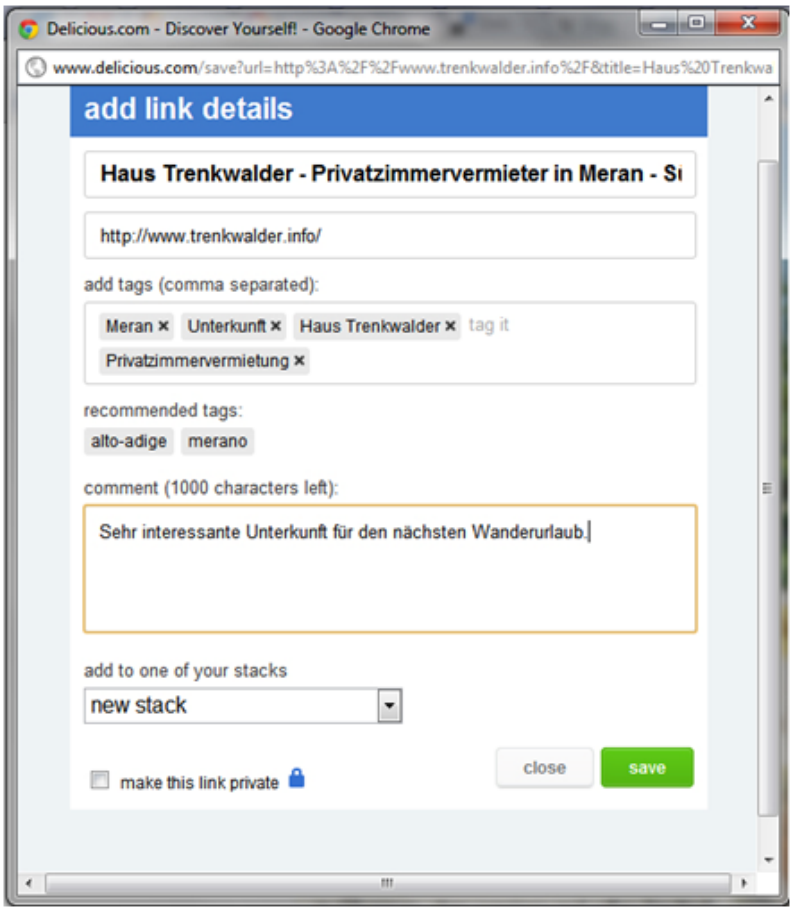

Abbildung 4.2: Delicious Eingabemaske für eine neue Ressource

Dank eines sehr ausgeklügelten Statistik-Eweiterung wird der Nutzer dazu motiviert immer wieder neue Bilder hinauf zuladen, diesen sehr passende Tags zu geben und Zusatzinformationen wie 'Ort der Aufnahme' hinzuzufügen. Jedoch ist diese Erweiterung kostenpflichtig. In dem Online-Dienst ist auch ein Web-Shop integriert, mit dem Bilder als Abzüge, Bücher, Poster usw. bestellt werden können.

Picasa

Bei Picasa[15] handelt es sich ebenfalls um einen Bildbearbeitungsdienst der von Google entwickelt wurde. Die Anwendung ermöglicht es anhand eines Client-Programmes und auch Online, Bilder zu bearbeiten, Personen darauf zu taggen, Fotos in Fotoalben einzuteilen, Ort des aufgenommenen Fotos und Tags hinzuzufügen. Dem Nutzer werden keine Tags von anderen Nutzern vorgeschlagen sondern nur die bereits von ihm hinzugefügten Tags bereit gestellt. Zudem kann

[15] http://picasaweb.google.com/ - Zugegriffen am 16.12.2011

nur derjenige Tags hinzufügen der das Foto hinauf geladen hat. Die Fotos können auf einem einfachen Weg mit dem Web synchronisiert werden, um diese dann dort mit anderen Leuten zu teilen oder diese für jeden Picasa-Nutzer öffentlich zu machen. Seit dem November 2011 gibt es eine Integration von Google+[16], die dem Online-Dienst eine sehr starke soziale Komponente hinzufügt. Dieses Bildverarbeitungssystem ist mehr für den eigenen Gebrauch und für das Teilen mit Freunden und Bekannten gedacht. Flickr, im Gegensatz, ist primär für das Veröffentlichen von Fotos konzipiert.

Zootool

Zootool[17] ist ein sozialer Bookmarking-Dienst mit dem man Links (Fotos, Videos, Dokumente und Lesezeichen) sammeln, sie mit Tags organisieren und mit Kommentaren versehen kann. Der Dienst ist webbasiert, jedoch gibt es inzwischen auch einen Android- und einen iOS-Client. Für die wichtigsten Web-Browser steht eine Extension zur Verfügung, die das Abspeichern von Lesezeichen erleichtert. Wenn keine Erweiterung installiert ist, können die Seiten über einen abgespeicherten Javascript-Code in einem Lesezeichen gesichert werden. Die Organisation der Links kann über Tags sowie über Ordner erfolgen. Zootool zeichnet sich für seine sehr gute graphische Oberfläche aus, die das Bedienen der Webseite vereinfacht. Die abgespeicherten Links können als öffentlich oder privat markiert werden.

Mister Wong

Mister Wong[18] ist ein 'social-bookmarking'-System, das im Frühling 2006 zum ersten Mal Online ging. Das System ermöglicht den Nutzern sich gegenseitig Nachrichten zu schicken, Netzwerke bzw. Gruppen zu bilden und in diesen Lesezeichensammlungen von anderen Gruppenmitgliedern zu verfolgen. Außerdem werden User angezeigt, die sich für seine geteilten Webseiten interessieren. Folglich kann mit dsiesen Nutzern Kontakt aufgenommen werden. Für die gängigsten Web-Browser wird eine Extension zur Verfügung gestellt. Als zusätzliche Funktion können PDF-Dokumente und Präsentationen hochgeladen werden, um diese dann von anderen Nutzern bewerten und kommentieren zu lassen. Dies eignet sich besonders für Schüler und Studenten. Mister Wong ist im Gegensatz zu Delicious Regional und Sprachen abhängig. Das heißt, dass unter der deutschen Domäne (www.mister-wong.de) hauptsächlich deutschsprachige Webseiten getaggt und angezeigt werden. Mister Wong gibt es in den folgenden Sprachen: Englisch, Französisch, Spanisch, Deutsch, Chinesisch und Russisch.

Vergleich

In der Tabelle 4.4 werden die analysierten Taggingsysteme gegenübergestellt und miteinander verglichen. Dabei fällt auf, dass der Nutzer bei der Verschlagwortung nur von einem Taggingsystem mit vorgeschlagenen Schlagwörter unterstützt wird. Bei den restlichen kann der

[16] https://plus.google.com/ - Zugegriffen am 16.12.2011
[17] http://zootool.com/ – Zugegriffen am 16.12.2011
[18] http://www.mister-wong.de – Zugegriffen am 20.12.2011

Nutzer die bereits vergebenen Tags von anderen Nutzern nicht sehen. Bei den beiden Foto-Taggingsystemen, Flickr und Picasa, stammen die Ressourcen von den Nutzern selber bzw. wurden zumindest von denen hochgeladen. Bei der restlichen Taggingsystemen (ausgenommen Evernote) kommen die Ressourcen aus dem Internet. Alle Taggingsysteme bieten die Funktion, Ressourcen untereinander in Gruppen zu verlinken.

	Taggin Rights	**Tagging Support**	**Aggregation**	**Type of Object**	**Source of material**	**Ressource connectivity**	**Social connectivity**
Evernote	Self tagging	Blind tagging	Bag model	Notizen, Webseite, Bilder	Web, Self	Grouped (Notizbücher)	Twitter, Facebook posten
Delicious	Free for all	Suggestive tagging	Set model	Lesezeichen (Url)	Web	Grouped (Ordner)	none
Flickr	Self Tagging, Free for all, Restricetd	Blind tagging	Set model	Foto	Self	Grouped (Album)	Personen taggen
Picasa	Self Tagging	Blind Tagging	Bag model	Foto	Self	Grouped (Album)	Personen taggen, Google+
Zootool	Self Tagging	Blind Tagging	Bag model	Webseiten, Bilder, Videos, Dokumente	Web	Grouped (Ordner)	Twitter, Facebook, Delicious posten
Mister Wong	Self Tagging	Blind Tagging	Bag model	Dokumente, Webseiten	Web	Grouped (Gruppe)	Twitter, Facebook

Tabelle 4.4: Vergleich der analysierten Tagging-Tools

4.5 Graphische Darstellung der Tagging-Struktur

Es gibt mehrere verschiedene Applikationen mit der Folksonomies graphisch dargestellt werden können. Diese stellen die Beziehung der einzelnen Tags visuell dar. Dabei kann die Häufigkeit der Schlagwörter durch unterschiedliche Schriftgröße ausgedrückt werden. Mehrfach gemeinsam verwendete Tags werden oftmals durch eine Linie miteinander Verbunden. Dabei gilt, desto näher die beiden Schlagwörter graphisch dargestellt werden, desto öfters wurden sie zusammen verwendet.

Im Folgenden werden 6 sehr unterschiedliche Tools zur graphischen Darstellung von Folksonomies beschrieben.

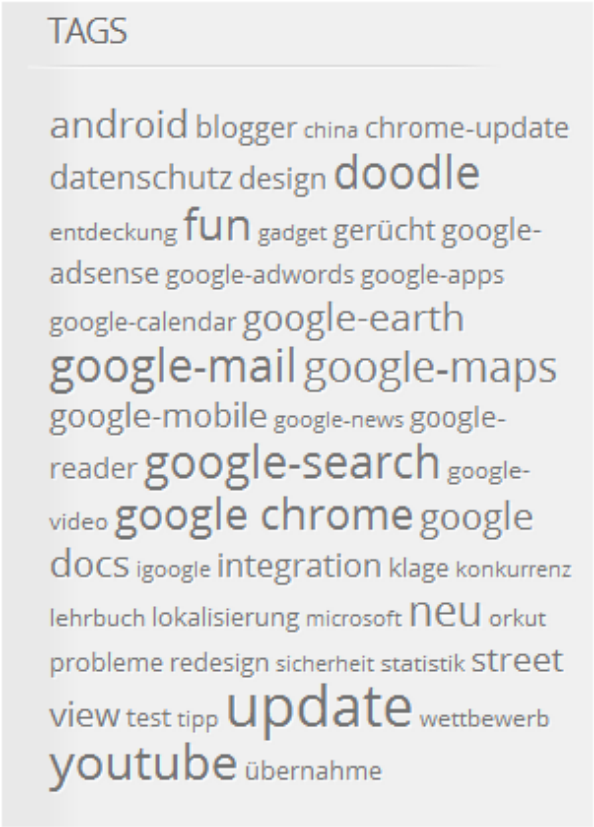

Abbildung 4.3: Tag Cloud von googlewatchblog.de

Tag Clouds

Die bekannteste und am meisten verwendetet Art zur Visualisierung von Folksonomie ist die Anordnung von Tags in Tag Clouds (Tagwolken). In einer Tagwolke, siehe Abbildung 4.3, werden die beliebtesten Schlagwörter der Nutzer meist alphabetisch angeordnet. Es ist eine Submenge von den am häufigsten genutzten Schlagwörtern. Die Größe der Schriftart der Tags wiederspie-

gelt ihre Beliebtheit. Desto größer umso öfters wurde das Schlagwort verwendet. Zusätzlich kann der Effekt mit der Verwendung von unterschiedlichen Farben verstärkt werden. Delicious verwendet unterschiedliche Farben, um die Schlagwörter zu kennzeichnen die bereits selbst verwendet wurde.

Helic et. al (vgl. [Helic2010]) wiederlegt anhand von Experimenten auf den Seiten Austria-Forum[19], BibSonomy[20] und CiteULike die Annahme, dass Tag Clouds für die Navigation von Internetseiten sehr hilfreich ist. Oft kommen redundante Schlagwörter in den Tag Clouds vor, wie z.B. die Schlagwörter 'advertisers' und 'advertising', 'List' und 'Lists', 'social networking' und 'social networks' in der Abbildung 4.4.

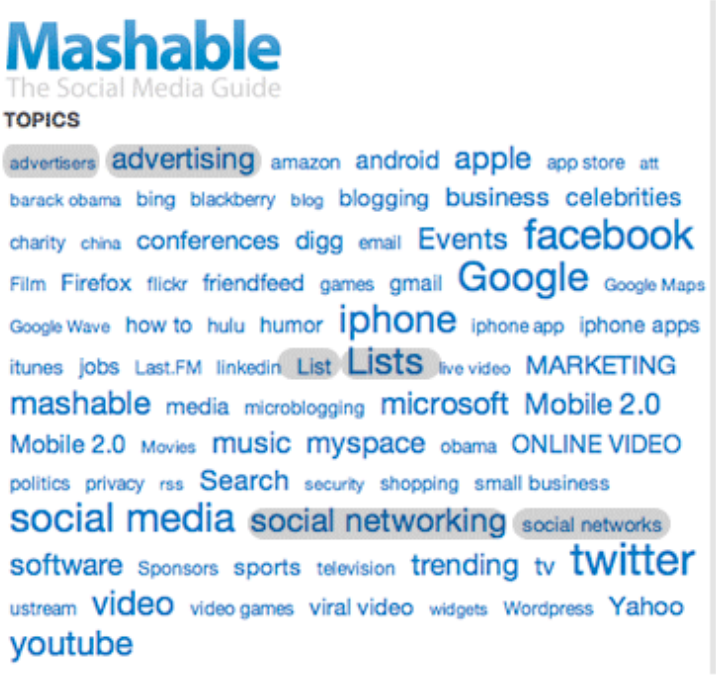

Abbildung 4.4: Tag Cloud von mashable.com

Außerdem überfliegt ein Besucher die Tag Clouds meist nur und liest nicht jedes einzelne Wort. Das bewirkt, dass von den Nutzern hauptsächlich nur Wörter mit einer größeren Schriftgröße wahr genommen werden. Das hat zur Folge, dass nicht alle Inhalte der Seite den gleichen Wert und dieselbe Relevanz haben. Für manche bleibt somit für sie interessanter Inhalt verborgen.

[19] http://www.austria-lexikon.at – Zugegriffen am 06.02.2012
[20] http://www.bibsonomy.org/ - Zugegriffen am 06.02.2012

TagXedo

TagXedo[21] ist eine Tag Cloud, bei der die Wörter nicht in einer Linie untereinander dargestellt werden, sondern in Form eines davor festgelegten graphischen Objektes. Dieses Tool basiert auf der Programmiersprache Java und lässt sich in jedem Webbrowser, der Java-Applets unterstützt, öffnen. Es können dabei zwischen verschiedenen Farben, Schriftarten, Themes und Layouts gewählt werden. Die Tag Cloud kann durch die Eingabe einer URL, den Namen eines Twitter-Accounts, Delicious-Account, News- oder Suchanfrage erstellt werden.

Abbildung 4.5: Word Cloud vom Wort 'Folksonomy'

Delicious Soup

Bei Delicious Soup[22] handelt es sich um eine Applikation, die die Beziehung zwischen den Tags veranschaulicht. Somit kann auf Anhieb gesehen werden welche Tags oftmals miteinander verwendet werden. Die Daten stammen aus der Datenbank von Delicious. Der Anwender muss sich vorher mit seinem Delicious Benutzerkonto anmelden. Durch das Klicken auf die Kugeln, die mit den Tags beschriftet sind, werden die Beziehungen neu berechnet und sie richten sich zueinander aus. Je näher die Kugeln nebeneinander liegen, desto öfters wurden die beiden Schlagwörter gemeinsam verwendet. Je größer die Kugel dargestellt wird, desto beliebter ist sie.

[21] http://www.tagxedo.com/ - Zugegriffen am 06.02.2012
[22] http://www.zitvogel.com/delicioussoup/ - Zugegriffen am 06.02.2012

Bei der graphischen Darstellung werden die Verbindungen im rechten unteren Rand in Zahlen ausgedrückt.

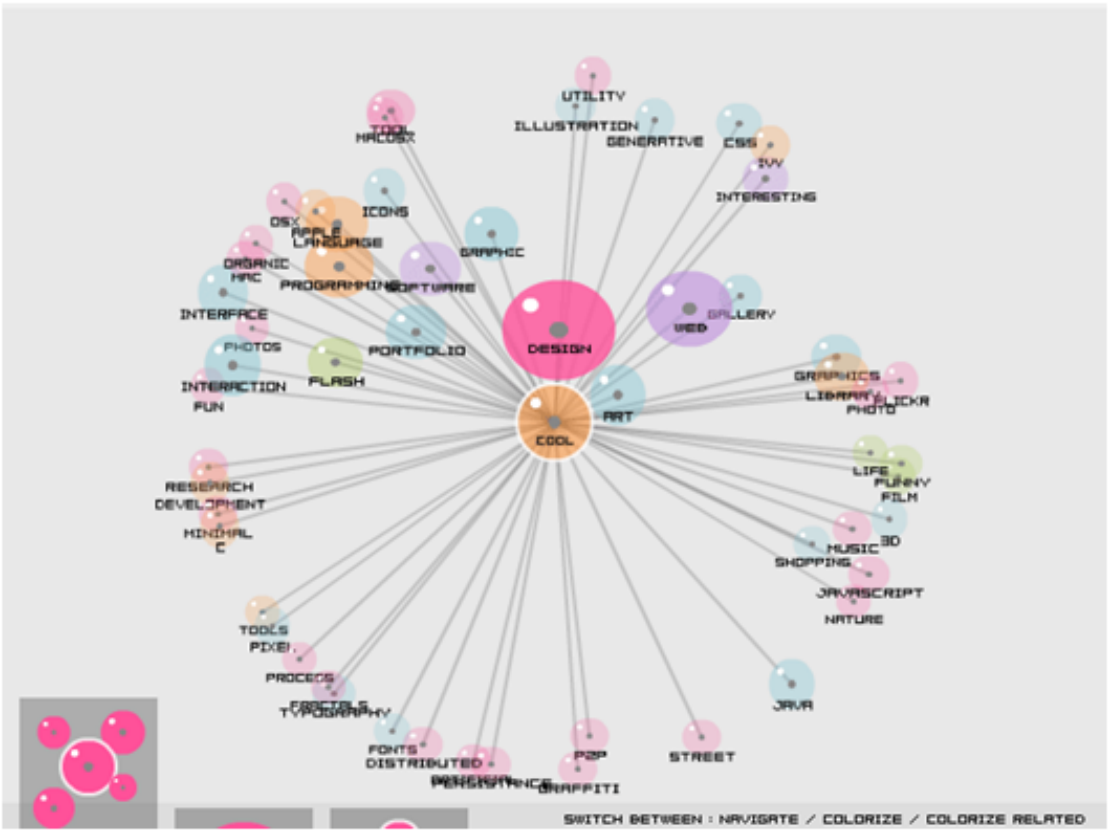

Abbildung 4.6: Darstellung der eigenen Tags durch Delicious Soup

Die Applikation kann als graphisches Hilfsmittel gesehen werden. Mit dem der Nutzer, durch das Klicken auf die Seifenblasen, sich ein Bild über seine verwendeten Schlagwörter machen kann. Aus der graphischen Darstellung, siehe Abbildung 4.6, können die Zusammenhänge und die Beziehungen von den einzelnen Tags ablesen werden. Dies ist bei den Tag Clouds nicht möglich.

Wiki Mind Map

Wiki Mind Map[23] dient zur Erschließung und visuellen Darstellung von eingegebenen Themengebieten. Das gewählte Themengebiet wird in der Mitte des Blattes dargestellt. Nach außen hin verlaufen die Hauptäste, die ihrerseits mit weiteren Unterästen ausgestattet sind. Auf einem Zweig befindet sich immer nur ein Schlüsselwort. Durch das Klicken auf das '+'-Zeichen werden

[23] http://www.wikimindmap.org/ - Zugegriffen am 06.02.2012

zusätzliche Tags angezeigt. Somit kann die Wiki Mind Map zur Strukturierung der Beziehungen der Tags untereinander verwendet werden.

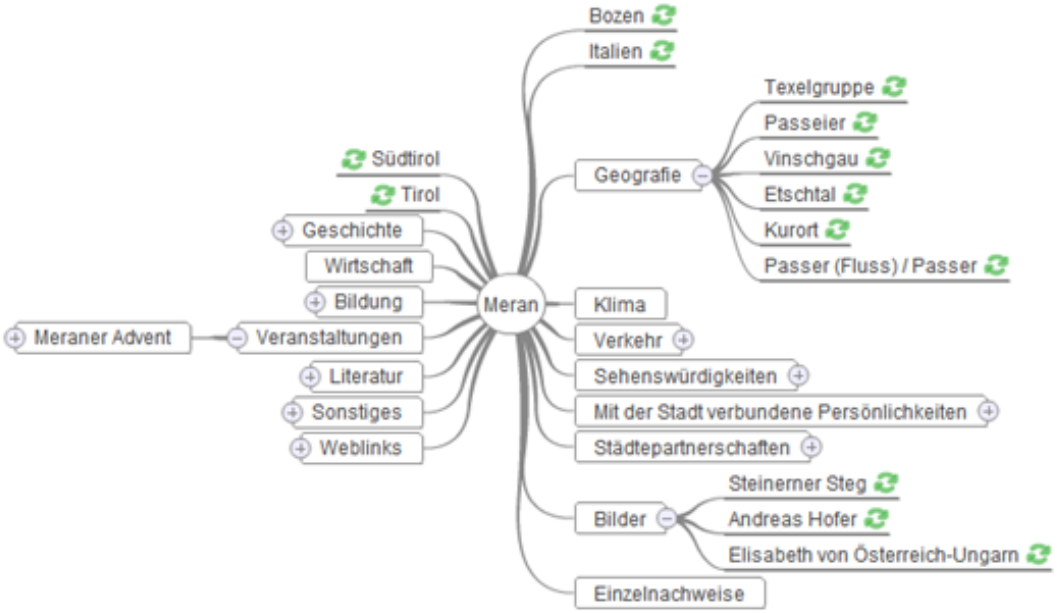

Abbildung 4.7: Wikimindmap.org Mind Map des Tag 'Meran'

TouchGraph

TouchGraph hat mit den TouchGraph Navigator 2[24] ein in der Programmiersprache Java geschriebenes Visualisierungs- und Analysetool entwickelt. Mit dem sich Netzwerke jeglicher Art darstellen lassen. Das Programm lässt sich in jedem Java Applets kompatiblen Webbrowser öffnen. TouchGraph GoogleBrowser[25] wurde anhand des TouchGraph Frameworks entwickelt und zeichnet mithilfe von Google's[26] Suchergebnissen Netze aus Objekte (URLs) und deren Verbindung zueinander.

In der Abbildung 4.8 wird die Verbindung von dem Suchergebnis von 'Folksonomy' dargestellt. Zunächst wurde die Relation zu 'Wired.com' aufgezeichnet, welches über das Wort 'Wired' mit dem Wort 'Folksonomy' verbunden ist. Anschließend wurde durch ein Doppelklick auf 'Folksonomy.co' der Zusammenhang zu 'Folksonomy' visualisiert. Dadurch wird sichtbar, dass 'Folksonomy.co' zusätzlich über 'davidsturtz', 'nytimes' und 'adammathes' mit dem Suchergebnis 'Folksonomy' verbunden ist.

[24] http://www.touchgraph.com/navigator - Zugegriffen am 06.02.2012
[25] http://www.touchgraph.com/TGGoogleBrowser.php - Zugegriffen am 07.02.2012
[26] http://www.google.at/ - Zugegriffen am 07.02.2012

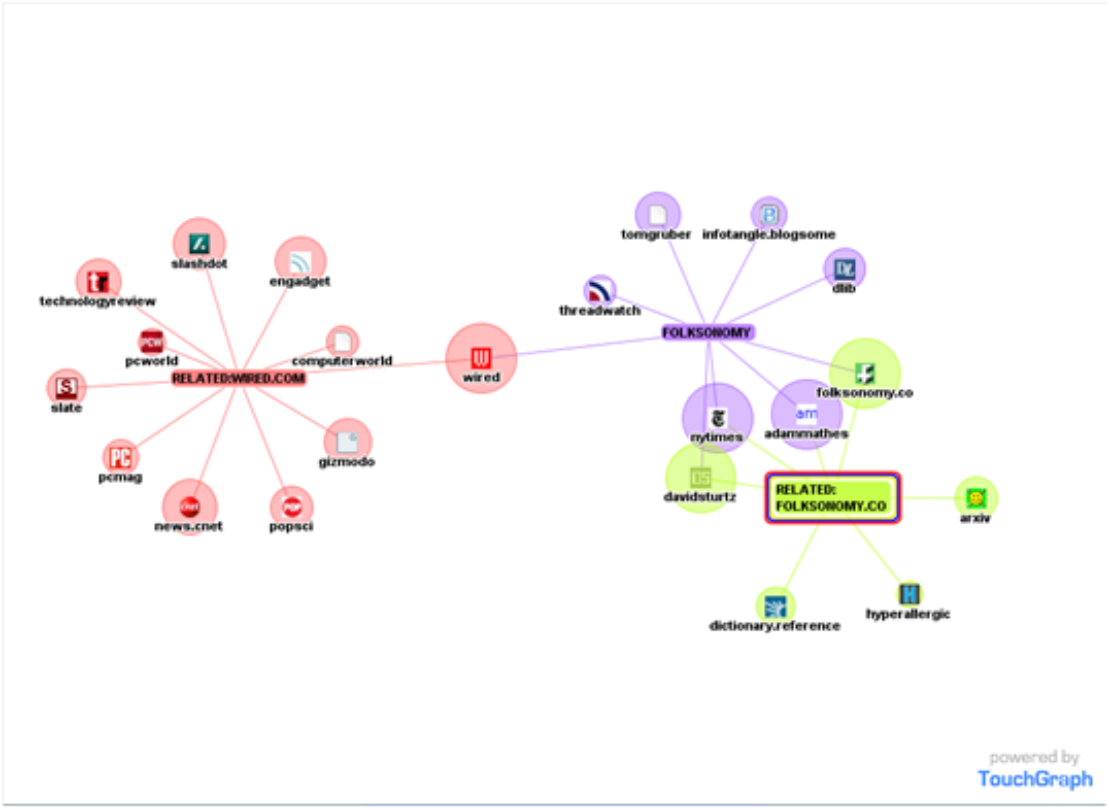

Abbildung 4.8: TouchGraph Google Browser

4.6 Übersicht vorhandener Information retrieval Tools

Es folgt eine Übersicht der bekanntesten Information retrieval Tools, die zu einem bestimmten Schlagwort eine personalisierte Zeitung erstellt. Die Informationen werden aus den sozialen Netzwerken wie Facebook, Twitter und Google+ entnommen. Informationen zu den serverseitig verwendete Relevance Ranking Algorithmen konnte bei keinem der analysierten Tools gefunden werden. Dies ist wohl bei den meisten Suchdiensten ein sehr gut gehütetes Geheimnis.

Paper.li

Der kostenlose Online Dienst Paper.li[27] aus der Schweiz bringt Twitter-Nachrichten, Facebook-Nachrichten und Google+ Einträge in ein Tageszeitungslayout. Die Nachrichten werden in einem Webbrowser dargestellt und lassen sich individuell zusammenstellen. Die sich entweder auf Basis eines bestimmten Thema, aus den letzten Tweets eines Twitter Users, aus den von ihm gefolgten Nutzer, aus den letzten Beiträgen eines Google+ Nutzers oder auf Basis eines Stichwortes in Facebook zusammensetzt. Bei der Erstellung kann außerdem eine Sprache ausgewählt

[27] http://paper.li/ - Zugegriffen am 22.02.2012

werden. Dadurch werden nur Artikel der ausgewählten Sprache angezeigt. Zusätzlich können zwischen Rubriken wie Erziehung, Gesundheit, Technologie, Wirtschaft usw. ausselektiert werden. Das Resultat wird als eine individualisierte Nachrichtenübersicht dargestellt.

Die Informationen eines Nutzers oder eines #Hashtags bzw. Schlagwortes werden graphisch sehr ansprechend dargestellt. Links in Tweets werden automatisch in einen Teaser-Artikel umgewandelt, Bilder und Videos werden stilvoll eingebunden. Die erstellte Seite kann öffentlich gemacht werden, damit auch andere Nutzer die erstellte Online-Zeitung lesen können. Sie wird anschließend alle 24 Stunden aktualisiert und kann über die in Paper.li integrierte Suchfunktion gefunden werden. Die täglich neu erstellte Zeitung kann durch die Eingabe einer Emailadresse automatisch nach Erscheinen direkt zugesendet werden. Selbst erstellte Online-Zeitungen können mittel Widgets in andere Webseiten eingebunden werden.

Der Nutzer hat die Möglichkeit sich über einen bereits registrierten Twitter oder Facebook Account einzuloggen. Die getwitterten Artikel, Google+ und Facebook Einträge werden vom System nach Keywords durchsucht um sie anschließend in Kategorien einzuordnen. Die Aktualität dieser Einträge wird durch die Relevanz des Tags bestimmt. Viel diskutierte Nachrichten werden dadurch öfters angezeigt.

Abbildung 4.9: Paper.li Zeitung vom Twitter User Ikangai

Flipboard

Flipboard[28] ist eine kostenlose App für die mobilen Betriebssysteme iOS und Android. Die App vereint das soziale Netzwerk eines Nutzers mit einem virtuellen Magazin. Flipboard stellt jene Beiträge, Fotos und Videos dar, die von den Freunden eines Facebook, Twitter, LinkedIn[29], Tumblr[30] und Instagram[31] Accounts geteilt wurden. Zusätzlich können RSS-Feed eingebunden werden. Dies funktioniert aber ausschließlich über den Goolge Reader[32]. Für jeden Inhalt wird ein Titel und einen kurzen Teaser mit einem hinterlegten Bild im Stil eines modernen Magazins angezeigt.

Der Nutzer kann sich durch die Seiten der Schlagezeilen durchwischen und bei Interesse auf einen Artikel klicken um sie vollständig lesen zu können. In dieser Vollansicht kann der Artikel kommentiert, mit einem 'gefällt mir' versehen oder dem ursprünglichen Twitter Poster geschrieben werden. Sehr aktive bzw. nervige Freunde können durch einen Button geblockt werden und erscheinen so nicht mehr im Flipboard. Je länger man Flipboard nutzt, desto besser sollen die Interessen der Nutzer getroffen werden. Die Darstellung der einzelnen Inhalte wird sehr schnell geladen und das Durchblättern erscheint sehr flüssig.

Der Nutzer kann sich aus verschiedenen Kategorien wie z.B. Technologie, Design, Fotografie, Sport, Reisen, Politik, Essen, Wissenschaft oder Filme, eigene Flipboards zusammenstellen lassen. Zusätzlich können auch neue Themengebiete erstellt werden.

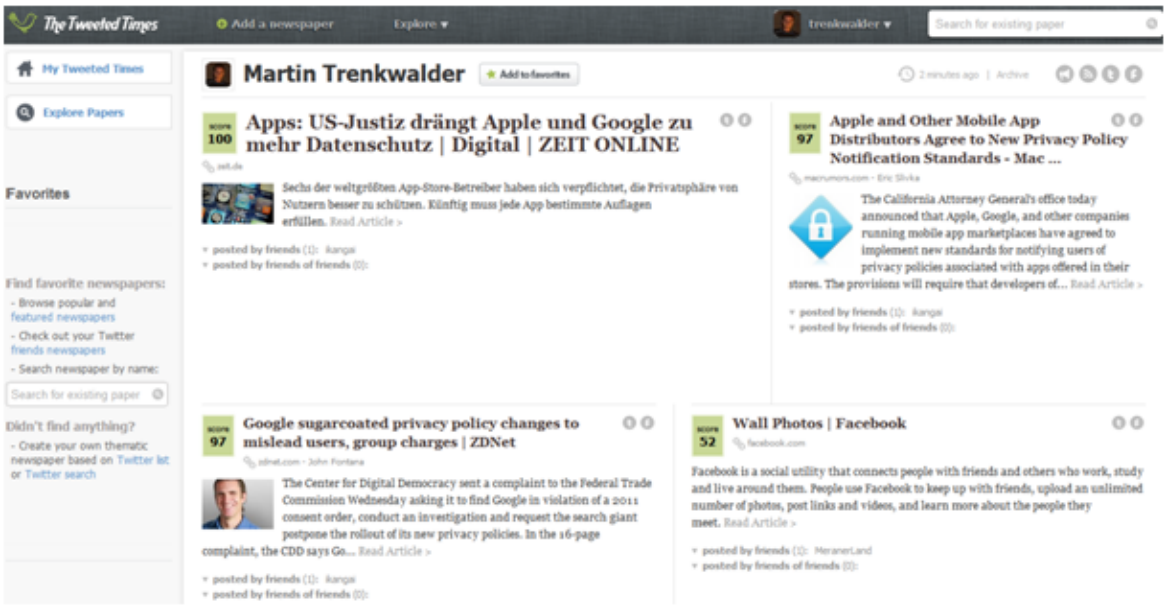

Abbildung 4.10: The Tweeted Times vom Twitter User trenkwalder

[28] http://flipboard.com/ - Zugegriffen am 22.02.2012
[29] http://at.linkedin.com/ - Zugegriffen am 22.02.2012
[30] https://www.tumblr.com/ - Zugegriffen am 22.02.2012
[31] http://instagr.am/ - Zugegriffen am 22.02.2012
[32] http://www.google.at/reader/view/ - Zugegriffen am 29.02.2012

Tweeted Times

Tweeted Times[33] wurde von Semantic Dimension Inc. entwickelt und erstellt Artikel anhand der eigenen individuellen Twitter-Timeline. Dabei werden die News, abhängig von der Popularität einer Nachricht, im eigenen Twitter-Freundeskreis ausgewählt und angezeigt. Neben jeden Post steht die Anzahl der Freunde, die ihn geteilt haben. Eine individuelle Newsseite kann mit der Eingabe eines spezifischen Suchwortes erstellt werden. Die selbst spezifizierte online Twitter-Zeitung wird stündlich aktualisiert.

Tweeted Times steht als Webbrowser- und als iPad-Variante für das iOS zur Verfügung. Durch ein Score Punktesystem (0 – 100) berechnet das System wie relevant ein Tweet für den jeweiligen Nutzer ist. Diese Punkteanzahl wird neben jeden Artikel angegeben, somit kann der Benutzer sofort sehen wie relevant das Dokument für ihn ist. Die geposteten Twitter Links werden, wie bei den davor erwähnten Diensten, geladen und ein Artikel-Teaser sowie ein Bild, falls vorhanden ist, angezeigt.

Journal+

Journal+[34] ist ein News-Aggregator, der sich aus öffentlichen Google+[35] Posts generieren lässt. Die vom System ausgewählten Posts können aus Text, Video oder Bild bestehen. Das Web-Magazin organisiert die beliebtesten Inhalte in Kategorien und bietet einen Contentfilter an, der es ermöglicht Inhalte nur aus einer bestimmten Region herauszufiltern.

Wer ein eigenes Google+ Profil besitzt bekommt eine Übersicht über die erfolgreichsten eigenen Inhalte. Zudem gibt es eine sehr ausgereifte Suchfunktion die es erlaubt gezielt nach bestimmten Inhalten in den Kategorien Überschriften, News, Bilder oder Videos zu suchen. In der Kategorie 'User' können neue interessante Inhalte entdeckt werden, dieser Funktion geht weit über die eigenen Google+ Empfehlungen hinaus.

Journal+ geht bei der Aggregation der News einen anderen Weg als die drei vorherigen online Zeitschriften, es bedient sich den öffentlichen Posts und berücksichtigt nicht den sozialen Faktor bzw. die Bindung zu den eigenen Freunden.

Vergleich

In der Tabelle 4.5 werden die untersuchten Information retrieval Tools gegenübergestellt und verglichen. Dabei ist ersichtlich, dass sich die analysierten Tools hauptsächlich nur bei der Aktualisierungsperiode unterscheiden. Einige befüllen die Online-Zeitschrift bei jedem Aufruf neu, andere hingegen machen dies nur stündlich oder täglich. Zum Beispiel verfolgt Paper.li diese Strategie und versendet einmal im Tag die Online-Zeitschrift per Email. Dadurch muss der User nicht eigens die Webseite ansurfen, sondern kann die erstellte Zeitschrift bei dem morgendlichen Emailcheck gleich lesen. Ein weiterer Unterschied ist bei der Herkunft der Daten ersichtlich. Dabei bindet Flipboard die meisten Taggingsysteme ein und Journal+ mit Google+ die wenigsten.

[33] http://tweetedtimes.com – Zugegriffen am 23.02.2012
[34] http://www.journalplus.net/ - Zugegriffen am 23.02.2012
[35] https://plus.google.com/ - Zugegriffen am 23.02.2012

	Herkunft der Daten	**Aktualisierung**	**Ressourcen**	**Plattformen**
Paper.li	Twitter, Facebook, Google+	Morgens und Abends, täglich oder wöchentlich	Bilder, Videos, Links, Webseiten	Webbrowser, Apple iOS
Flipboard	Facebook, Twitter, LinkedIn, Tumblr, Instagram, Soundcloud	Sofort	Webseiten, Fotos, Videos, Musik	Apple IOS, Android (Beta)
Tweeted Times	Twitter	Stündlich	Tweets, Links	Webseite, Apple iOS
Journal+	Google+	Sofort	Text, Video, Bilder	Webseite

Tabelle 4.5: Vergleich der analysierten Information retrieval Tools

KAPITEL 5

Web Content Maps

5.1 Ansatz

Bei der Erstellung des neuen Taggingsystems wurden die bereits bestehenden genau analysiert. Dabei wurden Funktionen heraus gesucht, die sehr nützlich sind, aber nicht implementiert wurden. Alle Systeme beinhalten ausschließlich die Möglichkeit ganze Webseiten, Videos, Fotos und Links zu taggen. Für einige Nutzer erscheint es aber sehr nützlich nur einzelne Absätze bzw. einen Satz zu taggen. Zu der zusätzlichen Funktion sollten aber auch alle bereits genannten Ressourcen wie Bilder, Videos, Links und Webseiten getaggt werden können.

Der getaggte Text bzw. Absatz sollte bei dem Tagvorgang modifizierbar sein. Da es bei dem Markiervorgang des Textes passieren kann, dass zu viele Wörter ausgewählt wurden. Bei einer modifizierten Ressource wäre es sinnvoll, dies für den Nutzer sichtbar zu machen. Diese Funktion stellt keines der analysierten Taggingsysteme bereit.

Um den Taggingprozess für den Nutzer so einfach wie möglich zu halten und ihm dabei zu unterstützen, wäre es sinnvoll die zu taggende Ressource zu analysieren und ihm bereits passende Schlagwörter zur Auswahl zur Verfügung zu stellen. Ein wichtiger Faktor bei der Analyse der Ressource ist die Sprache, in der der Text verfasst wurde. Es sollten möglichst viele Sprachen unterstützt werden. Nur ein Taggingsystem schlägt für die ausgewählte Ressource passende Schlagwörter vor (suggestive tagging). Von Delicio.us vorgeschlagene Tags halten sich in der Anzahl aber sehr in Grenzen, dabei werden durchschnittlich circa 10 Schlagwörter vorgeschlagen. Die vorgeschlagenen Tags sind meist sehr generell gehalten und sind nicht in Kategorien eingeteilt. Das neu entwickelte Taggingsystem sollte dem Nutzer, wenn möglich, mehrere Tags zur Auswahl vorschlagen. Durch die größere Anzahl von Tags wird es nützlich sein, diese zu strukturieren.

Der praktische Teil beinhaltet nicht nur das Taggen von Ressourcen, sondern auch das Wiedergeben der getaggten Ressourcen mit verschiedenen Filterfunktionen. Die Ergebnisse der Suchfunktion sollte in einer angemessenen Form dargestellt werden. Für den Nutzer erscheint es als hilfreich Zusatzinformationen zu den dargestellten Ressourcen anzuzeigen. Diese Infor-

mationen können aus dem Internet von Onlinelexikon, wie zum Beispiel Wikipedia oder IMDB abgefragt werden.

Die meisten Ressourcen in einer Folksonomy können oft nur sehr schwer in ein Themenbereich eingegliedert werden. Wie die Analyse der Information Retrieval Tools von Folksonomies ergeben hat, ist das Filtern nach Rubriken die wichtigste Filter-Funktion. Jedoch werden bei den analysierten Tagging-Tools jeweils nur die Tags aber nicht der Themenbereich abgespeichert. Er kann weder angegeben werden, noch ist es ersichtlich dass er intern abgespeichert wird. Das neu entwickelte Tagging-Tool sollte selbst in der Lage sein, durch die ausgeführte Analyse, den Themenbereich zu bestimmen oder zumindest den Nutzer die Möglichkeit zu geben diesen zu definieren.

Der wichtige soziale Aspekt von Folksonomy soll durch ein soziales Netzwerk abgedeckt werden. Dabei kann sich jeder User einen Freundeskreis aufbauen und mit denen die getaggte Objekte austauschen. Dies wird von all den analysierten Tools angeboten und gehört eigentlich zum Funktionsstandard. Die meisten Tools verfügen über ein eigenes soziales Netzwerk, sowie die Möglichkeit über die weit verbreiteten sozialen Netzwerken, wie Facebook, Twitter und Google+, anzumelden. Infolge werden die zusätzlichen Anforderungen zu den bereits bestehenden Tools aufgelistet:

- Bereiche einer Seite, wie Absätze oder Sätze, als Ressource zu taggen
- Die Möglichkeit ausgewählten textuellen Bereich bearbeiten zu können
- Dem Nutzer Tags, inklusive ihrer Relevanz, vorzuschlagen und die Analyse möglichst in vielen Sprachen zu unterstützen
- Zusätzlich ein Themenbereich bzw. Rubrik der Ressource abzuspeichern
- Zusatzinformationen aus Onlinelexika abfragen und bei der Darstellung der Ressource darzustellen

5.2 Das System

Das hier beschriebene System kann in zwei Teile untergliedert werden. Es dient einerseits zur Erfassung und Abspeicherung von Tags und Verweise der Ressourcen. Andererseits beinhaltet es ein Information Retrieval Funktion mit der in den gesammelten Ressourcen gesucht werden kann. Diese Informationen können in einer anspruchsvollen Sicht dargestellt werden.

Als Ressourcen können Texte, Textabschnitte, Bilder und Videos gesammelt werden. Wobei für die Bilder und die Videos nur die URL abgespeichert wird. Die Inhalte müssen für die spätere Verwendung online zur Verfügung stehen. Bei den Videos begrenzt sich die Auswahl auf die Video-Plattform Youtube. Markierte Textbereiche können während dem Tagvorgang abgekürzt bzw. verändert werden.

Das System enthält eine User Komponente, wo sich neue User anmelden können. Um eine Ressource zu taggen muss sich der Nutzer einmalig registrieren und bei dem System angemeldet sein. Der eingeloggte User kann seine gesammelten Inhalte mit Tags versehen. Diese bekommt

er vom System vorgeschlagen. Durch eine Suchfunktion können bereits gesicherte Inhalte angesehen, verwaltet und gelöscht werden. Das Tool beinhaltet ein soziales Netzwerk, bei dem ein Nutzer andere Nutzer als Freunde hinzufügen kann. Nach der Bestätigung der Freundschaftsanfrage können die gespeicherten Inhalte des jeweils anderen angesehen werden. Außerdem sieht man auf dem Profil des Freundes mit welchen Nutzern er befreundet ist. Anhand der gesammelten Ressourcen kann ein Online Magazin erstellt werden. Dazu muss der Nutzer ein Template aussuchen und eine Kategorie oder ein Schlagwort dazu wählen. Zur Verfeinerung des Online-Magazins kann der Nutzer die Zeitperiode eingrenzen, sowie nur Artikel anzeigen lassen die positiv bewertet wurden. Zur Befüllung des Templates können zwischen selbst getaggten und von Freunden getaggten Inhalte entschieden werden. Das Online-Magazin wird bei jedem Aufruf neu erstellt und mit zusätzlichen Informationen aus Online-Lexika angereichert.

Die Taggingrechte des Systems 'Web Content Maps' beschränken sich auf das 'self tagging', da eine Ressource immer nur von dem getaggt wird, der den Inhalt über die Chrome Extension hinzufügt. Bei der Auswahl der Schlagwörter wird dem Nutzer eine Reihe von Tags vorgeschlagen. Der Inhalt der Internetadresse wird davor von dem Web-Service OpenCalais[1] gescannt und die geeignetsten Schlagwörter werden mit einer Prozentzahl angezeigt. Mit dem Tool können nur englische, spanische und französische Webseiten getaggt werden. Da zur Zeit nur diese Sprachen von OpenCalais unterstützt werden. Die getaggten Inhalte werden jeweils getrennt gespeichert, es handelt sich also hierbei um ein 'Bag model'. Der Nutzer kann Texte, Links, Bilder, Videos und komplette Webseiten taggen bzw. zu seiner Sammlung hinzufügen. Die Inhalte müssen ausschließlich über das Internet abrufbar sein, einige Inhalte wie Videos und Bilder werden nicht lokal gespeichert und wären somit bei dem Online-Magazin nicht verfügbar. Die Inhalte werden automatisch vom System in Rubriken eingeteilt und gruppiert. Web Content Maps bietet keine Möglichkeit sich mit anderen sozialen Netzwerken, wie Facebook, Twitter oder Google+, zu verbinden. Deshalb muss selbst ein Freundeskreis aufgebaut werden.

	Taggin Rights	**Tagging Support**	**Aggregation**	**Type of Object**	**Source of material**	**Ressource connectivity**	**Social connectivity**
Web Content Maps	Self tagging	Suggested tagging	Bag model	Texte, Webseiten, Bilder, Videos	Web	Grouped (Rubriken)	None

Tabelle 5.1: Merkmale des Taggingsystems 'Web Content Maps'

[1] http://www.opencalais.com/ - Zugegriffen am 25.02.2012

5.3 Eingesetzte Technologien

Zur Speicherung der Daten wird eine frei und quelloffene verfügbare MySql-Datenbank[2] zum Einsatz gebracht. Das Datenmodell wird in Kaptiel 5.6 genauer beschrieben. Als Serverseitige Skriptsprache wird PHP[3] mit dem Framework CakePHP[4] eingesetzt. CakePHP ist ein Web-Framework, das dem Schema des Model-View-Controller (MVC) folgt. In dem Projekt kam die Version 1.3 zum Einsatz.

Als Clientseitige Skriptsprache wird JavaScript[5] mit der Java-Script-Klassenbibliothek jQuery[6] und jQueryUI[7] eingesetzt. jQueryUI stellt die Goolge Chrome Extension in einfacher Weise schöner dar. jQuery wurde hauptsächlich zur Selektierung von Elementen verwendet.

Die Formatierung der Seitenelemente wurde in CSS3[8] umgesetzt. Zur Strukturierung der Inhalte auf der Internetseite wurde als Standard HTML verwendet.

Für das Taggingsystem wurde eine Google Chrome Extension entwickelt, die auf HTML, Javascript und CSS basiert.

5.4 Systemarchitektur

Google Chrome Extension sind einfache Webseiten, die mithilfe von HTML5, CSS3 und Javascript erstellt werden. Jedoch unterscheidet sich der Zugriffsumfang von einer normalen Webseite aus dem Internet. So kann durch die Extension auf alle geöffneten Tabs zugegriffen werden, die Browser-History einsehen werden, geöffnete Webseiten bearbeiten und Ajax Request auf jegliche Seiten abgeschickt werden. All diese Zugriffsrechte müssen in einem JSON-File ('manifest.json') angegeben werden.

Die Architektur einer Google Chrome Extension ist in vier Komponenten aufgeteilt:

- **Content Script**
  Diese Komponente ermöglicht den Zugriff auf das Document Object Model (DOM) der geöffneten Webseiten. Der Zugriff beschränkt sich nicht nur auf das Lesen der Seite sondern erlaubt es die Seite zu bearbeiten bzw. zu verändern.

- **Background Page**
  Die Extension enthält eine Webseite mit einem long-running Script das immer geöffnet bleibt. Während dem kompletten Lifecycle ist sie immer aktiv und es kann davon nur eine aktive Instanz geben.

- **UI Pages**
  Anhand der UI Pages werden dem Nutzer Informationen angezeigt, wie zum Beispiel ein

[2] http://mysql.de/ - Zugegriffen am 24.02.2012
[3] http://www.php.net/ - Zugegriffen am 24.02.2012
[4] http://cakephp.org/ - Zugegriffen am 24.02.2012
[5] http://de.wikipedia.org/wiki/JavaScript - Zugegriffen am 24.02.2012
[6] http://jquery.com/ - Zugegriffen am 24.02.2012
[7] http://jqueryui.com/ - Zugegriffen am 24.02.2012
[8] http://de.wikipedia.org/wiki/Cascading_Style_Sheets - Zugegriffen am 24.02.2012

Popup oder eine Optionenseite. Durch diesen Seiten kann der Nutzer mit der Extension interagieren.

- **Message Passing**
  Die Kommunikation zwischen den Content Script und der Background Page erfolgt über die 'Message Passing'-Komponente. Sie ist bidirektional und ermöglichen es JSON- Objekte auszutauschen. Die dafür zu verwendende API beinhaltet einen einfachen 'one-time request' und eine komplexe 'long-lived connection'.

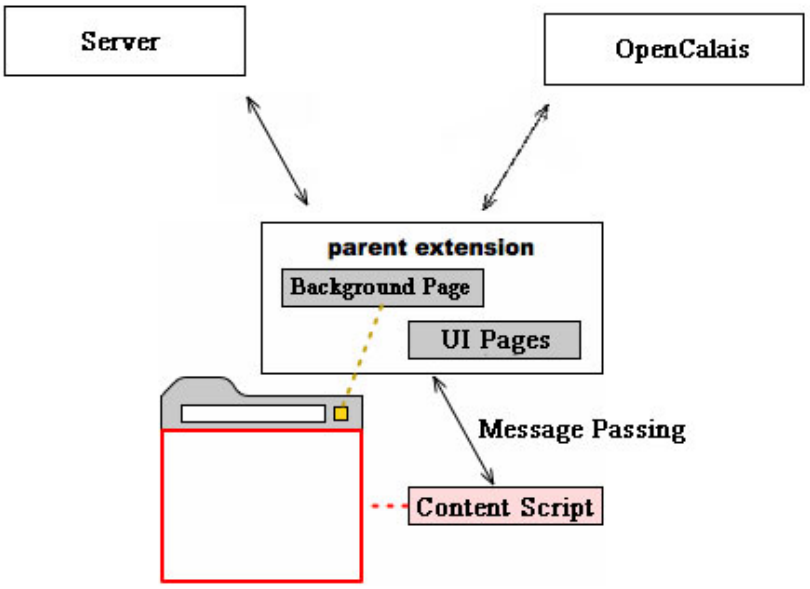

Abbildung 5.1: Systemarchitektur einer Google Chrome Extension

Die Systemarchitektur wird nun anhand des Taggingvorganges genauer erläutert. Der Vorgang startet mit der Markierung der gewählten Ressource, durch den Nutzer. Daraufhin öffnet der Nutzer durch ein Klick auf das Extension Icon die Background Page. In dem realisierten Tool ist es die einzige HTML Seite. Aufgrund der geringen Anzahl an Anzeige-Seiten wurden alle Ausgaben an den Nutzer in die Background Page gestellt. Für eine interaktiveres Projekt mit Optionen-Seite würde es sich anbieten jede Seite in ein UI Page zu packen.

Die Background Page öffnet den 'Message Passing'-Kanal und sendet anhand eines JSON-Objektes eine Anfrage an die 'Content Script'-Komponente, die dadurch aktiviert wird. Sie analysiert den markierten Bereich und gruppiert sie in verschiedene Ressourcearten (Bilder, Videos, Text, URL, Webseiten) ein. Diese Informationen werden in ein JSON-Objekt verpackt und über das 'Message Passing' an die Background Page zurück geschickt. Nach dieser Kommunikation wird der 'Message Passing'-Kanal geschlossen. Da es keine weitere Interaktion zwischen den beiden Komponenten gibt.

Die erhaltenen Ressourcen werden graphisch aufbereitet und anhand der Background Page dem Nutzer sichtbar gemacht. Dieser muss nun die gewünschte Ressourceart auswählen. Hat sich der Nutzer für eine Art entschieden, so gelangt er auf eine weiteren Seite. Dabei handelt es sich immer noch um die Background Page. Die verschiedenen Seiten sind in einer HTML-Datei enthalten und werden anhand CSS auf sichtbar oder unsichtbar gesetzt.

Zu diesem Zeitpunkt wird über AJAX ein Anfrage zu dem OpenCalais Webservice abgeschickt. Dieser erfolgt in der Background Page, da nur von dort aus externe Aufrufe erlaubt sind. Die Anfrage beinhaltet ein HTML-Dokument oder eine URL, je nachdem welche Ressourceart ausgewählt wurde. Als Antwort wird ein JSON-Objekt zurückgeschickt. Die erhaltenen Tags werden nach Kategorien eingeteilt und wiederum in der Background Page dargestellt.

Hat der Nutzer all seine gewünschten Tags ausgewählt und bestimmt welche Sichtbarkeit sein Eintrag besitzen soll. So wird durch den Klick auf den 'Upload'-Button die in den HTML Input-Elementen gespeicherten Informationen serialisiert und als ein JSON-Objekt an den Server gesendet. Diese Kommunikation erfolgt ein weiteres mal über die Background Page. Auf dem Server werden die Informationen gespeichert und bei einem erfolgreichen Abschluss ein 'success', als Antwort, zurück an die Extension gesendet. Erhält die Extension vom Server ein 'success' so informiert sie den Nutzer über den erfolgreichen Ausgang des Vorgangs.

5.5 Aktivitäts Diagramm

In der Abbildung 5.2 wird der Taggingvorgang mit 'Web Content Maps' als ein Aktivitäts Diagramm dargestellt. Dies ist der interessanteste und gleichzeitig komplexeste Ablauf des Tools. Der Start-Knoten beginnt bei dem Öffnen der Google Chrome Extension und der End-Knoten endet bei dem Versand von den Tagging-Informationen zu dem Server.

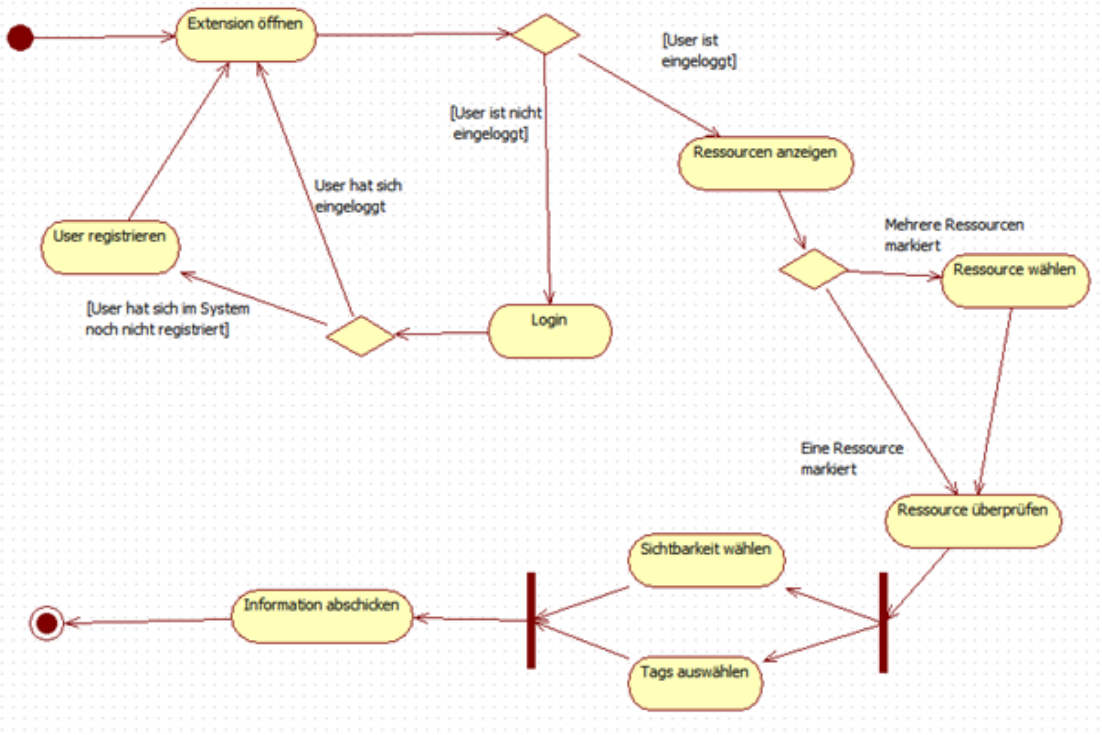

Abbildung 5.2: Aktivitäts Diagramm, Ressource taggen

5.6 Datenbankmodell

Das Datenbankmodell besteht aus 49 verschiedenen englisch lautende Tabellen. Das eingesetzt Framework CakePHP setzt als Konvention englischsprachige Tabllennamen voraus. Eine große Anzahl von Tabelle, 38 an der Zahl, beinhalten Informationen über die verschiedenen Ressourcen. OpenCalais kann den analysierten Text in 39 verschiedenen Kategorien einteilen. Genau so viele verschiedene Ressource-Tabellen sind im Datenbankmodell vorhanden. Sie weisen unterschiedlichste Attribute auf, wie zum Beispiel 'latitude', 'longitude', 'nationality', 'socre', 'shortname' usw. All jene Tabellen sind mit der Tabelle 'Ressource' verbunden.

Die restlichen Tabellen und Verbindungen zueinander lassen sich aus der Abbildung 5.3 ablesen.

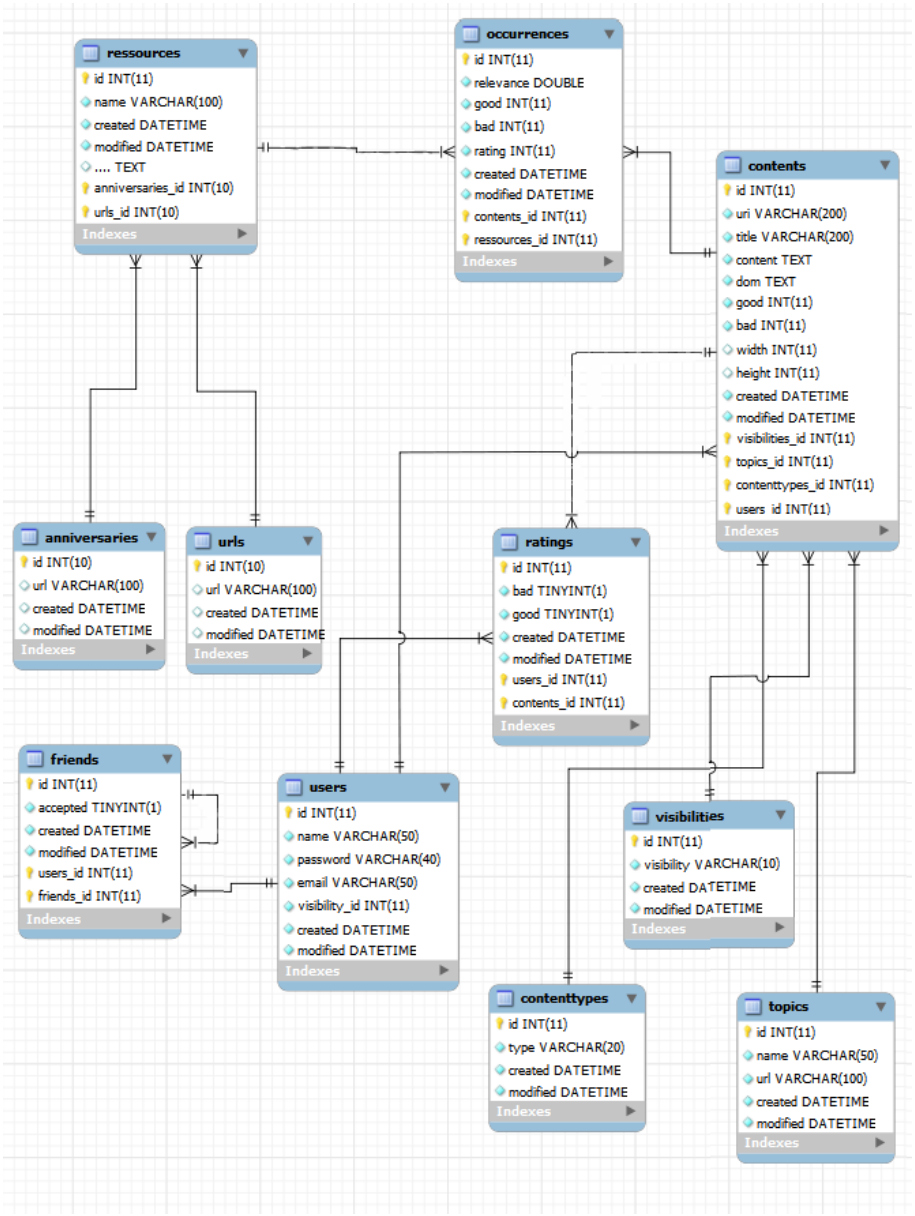

Abbildung 5.3: Das Datenbankmodell

KAPITEL 6

Ausblick

Das in dieser Arbeit entwickelte Tagging-Tool versucht einige Nachteile der Folksonomies zu beheben. So wird ein fehlendes kontrolliertes Vokabular durch die vorgeschlagenen Schlagwörter zwar behoben. Es funktioniert aber nur für die unterstützen Sprachen (Englisch, Französisch und Spanisch) und bei textuellen Ressourcen. Möchte der User ein Bild oder ein Video taggen, das keinen Text auf der Webseite beinhaltet, so kann das Tool keine Schlagwörter vorschlagen. Zusätzlich widerspricht die Einschränkung von den zu vergebenen Schlagwörtern dem Konzept einer Folksonomy.

Die Abgeschlossenheit von vielen Tagging-Systeme, unter anderem auch das 'Web Content Maps'-Tool, birgt ein weiteres Problem in sich. Die meisten Tagging-Systeme besitzen eine eigene Datenbasis, die von den anderen Tagging-Systemen nicht verwendet werden kann. Hierfür müsste ein einheitlicher Standard bzw. eine Schnittstelle definiert werden, die einen Systemübergreifenden Austausch von Folksonomies ermöglicht. Die semantischen Informationen in den Folksonomies müssten plattformübergreifend miteinander verbunden werden.

Die Motivation zum Tagging besteht meistens aus funktionalen und sozialen Aspekten. Leider lässt sich aus einer Folksonomy nicht herauslesen, mit welchem Ansporn ein Nutzer ein Tag einer Folksonomy hinzugefügt hat. Für den Nutzer steht meistens die Wissensorganisation und/oder die Wissenskommunikation im Vordergrund. So gibt es Nutzer die Tags hinzufügen um sich im Netz besser darzustellen und aufzufallen. Andere Nutzer versuchen damit ein späteres Wiederauffinden zu erleichtern. Einige Nutzer fügen neue Tags hinzu um es mit der Netzöffentlichkeit oder nur mit Freunden zu teilen. Es wird nur schwer möglich sein, diesen Nachteil von Folksonomies zu beheben.

Eines der größten Unterschiede zu Ontologien und Topic Maps, und gleichzeitig der größte Nachteil von Folksonomy ist ihre flache Struktur. Heymann et. al (vgl. [Heymann2006]) haben eine Methode vorgeschlagen, die es ermöglicht eine flache Folksonomy in eine Folksonomy mit mehreren Ebenen umzuwandeln. Sie starten dabei von einem triparaten Graphen und zählen die Häufigkeit wie oft dieser vorkommt. Anschließend wird die Häufigkeit der Tags gezählt, die in Verbindung mit dem ersteren auftreten. Daraus erstellen sie einen ungewichteten Graphen und löschen die Verbindungen zwischen den Tags, die selten zusammen auftreten. Nun werden die

Tags untereinander nach Ähnlichkeiten geordnet. Heyman et. al bemerken, dass ihr Verfahren umso besser funktioniert, desto größer die Folksonomy ist.

Ein sehr interessanter Ansatz liegt im Simple Knowledge Organisation System[1] (SKOS). Dabei handelt es sich um einen aktuellen W3C-Standard von Wissensorganisationen. Er besteht aus einem RDF-Vokabular zur Kodierung von kontrollierten Vokabularen wie Taxonomien, Schlagwörter, Klassifikationen etc. Eine Ressource wird anhand einer RDF-Klasse (skos:Concept) definiert und mittels einer Property (skos:subject) kann einer Ressource mit einem Konzept verbunden werden. Die Konzepte können wiederum untereinander mit skos:narrow und skos:boarder hierarchisch verknüpft werden. Der Vorteil in der Trennung von Begriffen und deren Bennenung mit SKOS besteht darin, dass die Begriffe zur Verschlagwortung eine eindeutige URI besitzen. Somit besteht die Möglichkeit sie in das Semantic Web einzubinden. Diese Trennung erfolgt auch bei dem entwickelten 'Web Content Maps'-Tool, wo jedes Schlagwort durch eine eindeutige URI identifiziert wird.

[1] http://www.w3.org/2001/sw/wiki/SKOS/Datasets

ANHANG A

Benutzeranleitung für Web Content Maps

Die Anleitung baut sich auf die chronologische Abfolge eines neuen Nutzers auf. Zuerst muss sich der neue User mit der Eingabe von den geforderten Daten registrieren. Anschließend kann er sich durch das hinzufügen von Freunden ein soziales Netzwerk aufbauen.

Der nächste Schritt besteht in dem Hinzufügen der Google Chrome Extension 'Web Content Map'. Diese erlaubt es künftig Inhalte zu taggen und diese hinzuzufügen. Sobald einige Daten gesammelt wurden kann nach der Auswahl des Templates und des gewünschten Users bzw. eines Schlagwortes ein Online Magazin erstellt werden. Dieses wird nach der Speicherung bei jedem Aufruf mit neuen Daten befüllt.

A.1 User erstellen

Die Erstellung eines Nutzers ist der erste Schritt zur Nutzung der Web Content Map. Bei der Login-Maske wird der Nutzer nach seinem Vor- und Nachname, Emailadresse und Passwort gefragt. Diese Daten werden validiert und darauf kontrolliert ob ein User mit der Emailadresse bereits angemeldet wurde. Ist dies erledigt, wird eine Profilseite des neu angelegten Nutzers erstellt.

A.2 Soziales Netzwerk aufbauen

Ein wichtiger Schritt um Inhalte mit anderen Nutzern zu teilen ist der Aufbau eines sozialen Netzwerkes. Der Nutzer kann über den 'Freund hinzufügen'-Button eine Freundschaftsanfrage versenden und nach der Bestätigung sein Profil mit den kompletten Inhalten anschauen. Eine Freundschaft kann zu jedem Zeitpunkt von beiden Seiten wieder gelöscht werden. Bei befreundeten Nutzern können die getaggten Inhalte angesehen werden. Eine Tag-Cloud stellt dabei die von ihm am öftesten verwendeten Schlagwörter dar.

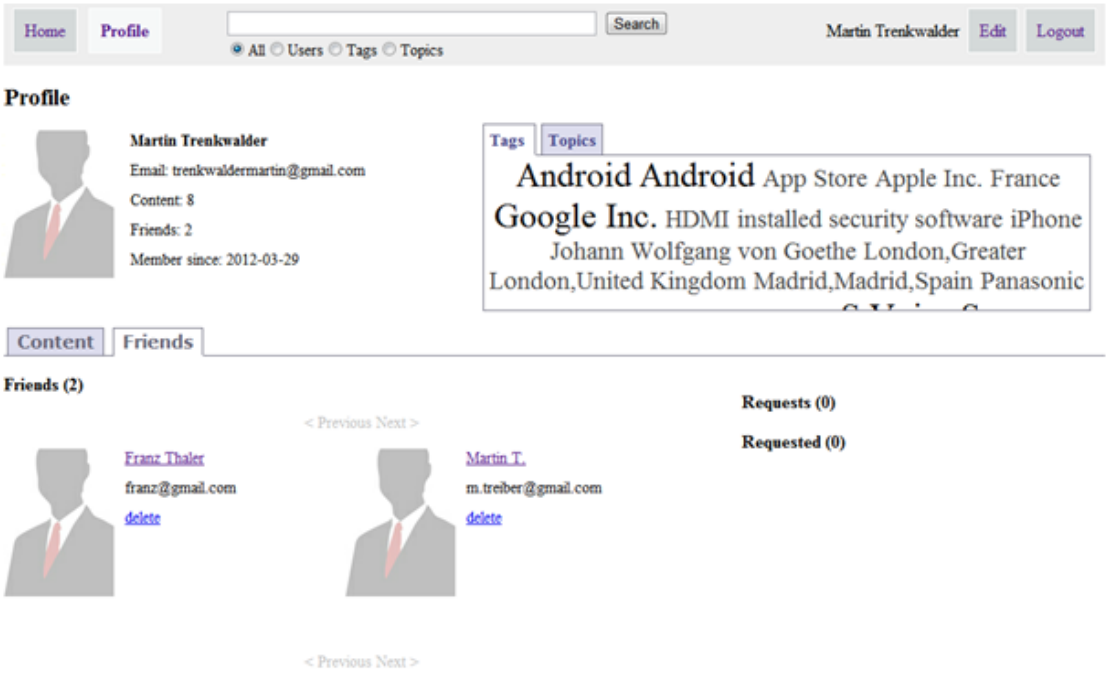

Abbildung A.1: Profil eines Nutzers

A.3 Inhalte taggen und verwalten

Um Inhalte zu taggen und sie zum eigenen Profil hinzufügen zu können, muss der Nutzer eingeloggt sein und den Webbrowser Google-Chrome mit der Extension 'Web Content Map' verwenden. Sobald der Nutzer auf eine interessante Seite im Internet trifft, die er gerne zu einem späteren Zeitpunkt leicht wiederfinden möchte oder die er einem Freund zeigen möchte, geht er wie folgt vor:

1. Gewünschten Bereich markieren
2. Auf das 'Web Content Map'-Icon klicken
3. Zwischen den angezeigten Links, Videos, Bilder, Text oder Seite wählen
   - Bei Link, wird nur die URL gespeichert
   - Bei Youtube-Videos, wird der Link zum Video gesichert
   - Bei Text, wird der ausgewählte Inhalt in die Datenbank geschrieben
   - Bei Seite, der komplette Inhalt der Seite gespeichert
4. Wenn die Ressource 'Text' ausgesucht wurde, kann dieser durch den Klick darauf geändert werden

5. Zwischen den Privatsphären ’public’, ’friends’ und ’private’ wählen
6. Schlagwörter aus der vorgeschlagenen Liste hinzufügen
7. Das System fügt der Ressource selbst eine Rubrik hinzu.
8. Der ausgewählte Text, die Url der getaggten Seite, der Titel der im <title>-Tag der Seite steht, sowie die ausgesuchten Schlagwörter anhand des ’Submit’-Buttons an den Server schicken

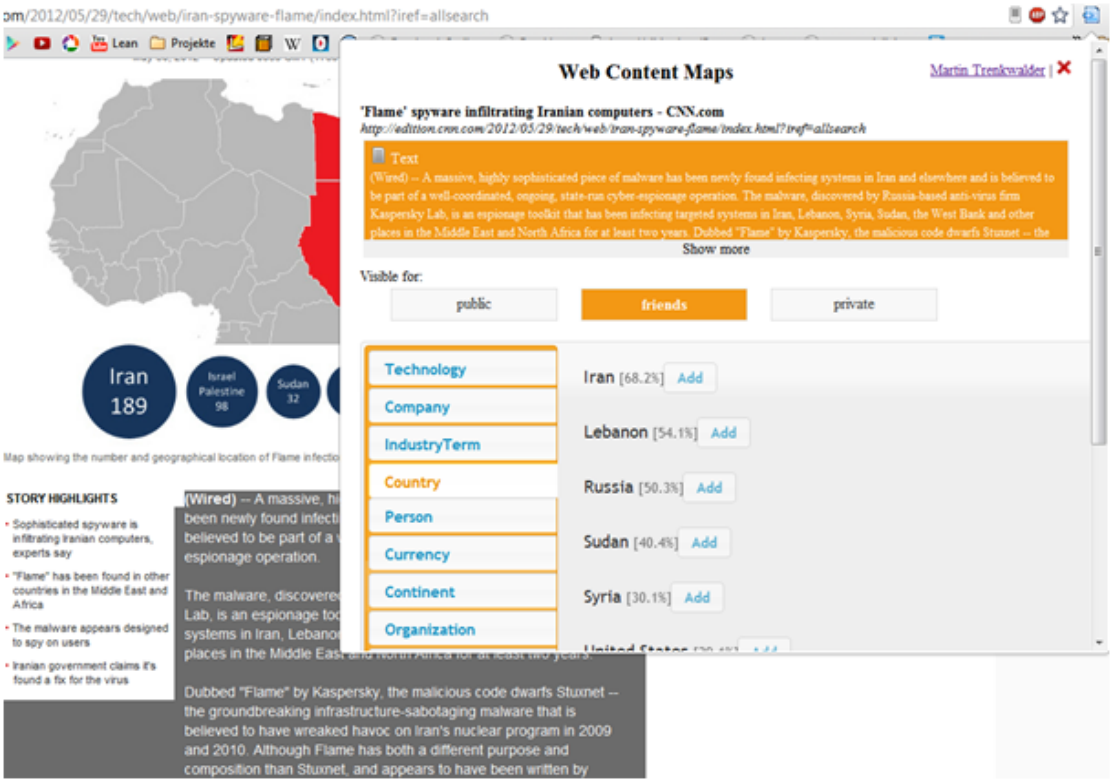

Abbildung A.2: Google Chrome Extension mit der eine Ressource hinzugefügt werden kann

A.4 Ressourcen anzeigen

Auf der ’Home’-Seite werden die neuesten getaggten Ressourcen angezeigt. Diese können nach Inhaltstyp (Text, Bilder, Links und Videos) und nach Sichtbarkeit (Freunde, Privat und fremd) gefiltert werden. Jede getaggte Ressource kann durch einen ’Like’ oder ’Dislike’-Button bewertet werden. Zu den Inhalt und den Tags ist auch die automatisch hinzugefügte Rubrik sichtbar. Durch einen Klick auf den Tag werden Zusatzinformationen von den Online-Lexika Wikipedia oder IMDB angezeigt.

Handelt es sich bei dem Tag um einen Ort so wird dieser mittels Google Maps rechts von den Ressourcen eingebunden. Unter den Ressourcen werden, falls vorhanden, mit dem Tag verwandte Begriffe dargestellt.

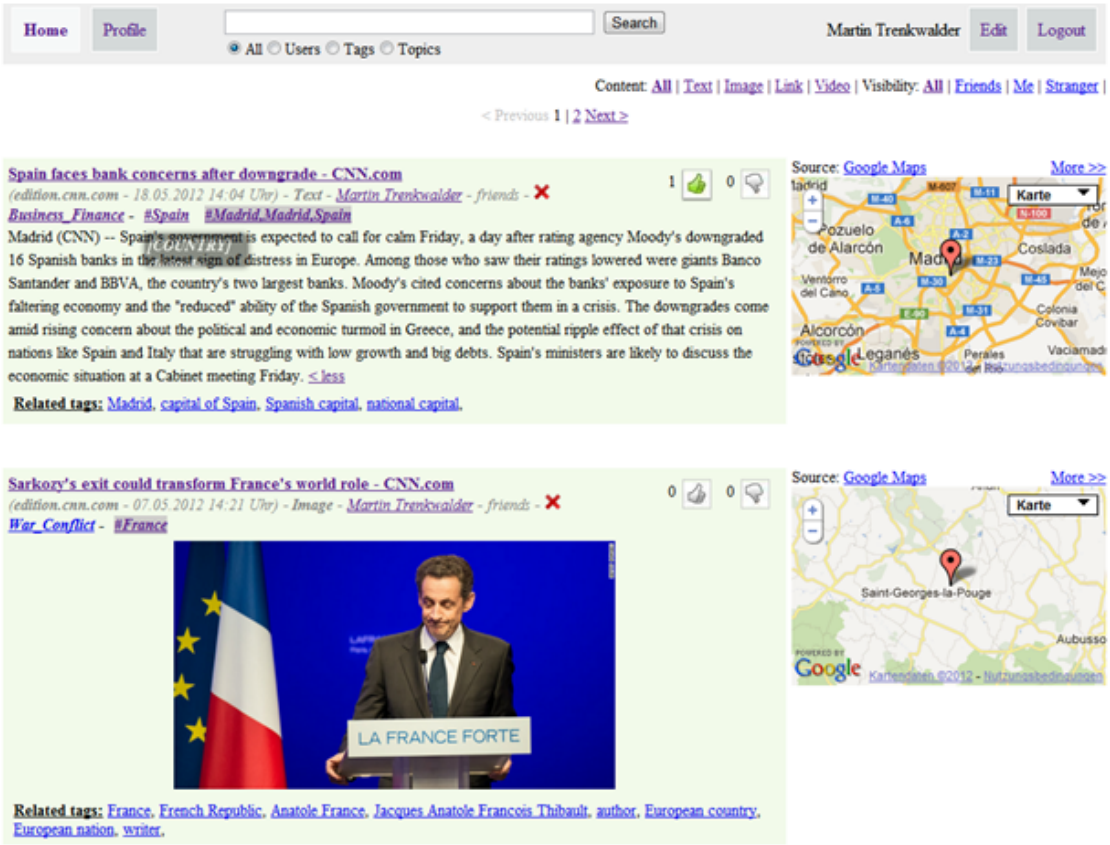

Abbildung A.3: Startseite eines angemeldeten Nutzers

A.5 Online Magazin erstellen

Für das Erstellen des Online-Magazins kann im Profil auf den 'Magazin erstellen'-Button geklickt werden. Auf der folgenden Seite kann der Nutzer den Titel für das Magazin bestimmen. Als nächstes muss ein Suchbegriff eingegeben werden. Dieser Begriff bestimmt das Thema des Online-Magazins. Als weitere Option kann zwischen ein und mehreren Rubriken, wie z.B. 'Politik', 'Wirtschaft', gewählt werden. Diese Unterteilung in Rubriken ist hilfreich, wenn es einen Begriff in mehreren verschiedenen Bereichen gibt. Als letzte Einstellung kann bei der Erstellung des Online-Magazins die Quelle des Artikels gewählt werden. Dabei stehen folgende Optionen zur Verfügung: selbst getaggte Ressourcen, von Freunden getaggte oder von allen Nutzern des Systems. Sobald alle Kriterien für die Suche ausgefüllt wurden, kann der Nutzer noch das gewünschte Template wählen, wie das Online-Magazin aussehen soll.

Literaturverzeichnis

[Albrecht2006] Christine Albrecht; Folksonomy; http://www.cheesy.at/download/Folksonomy .pdf - Zugegriffen am 29.05.2012

[Bryan1999] Martin Bryan, Steven R. Newcomb, Michael Biezunski; Topic Maps; http://www1.y12.doe.gov/capabilities/sgml/sc34/document/0129.pdf - Zugegriffen am 16.03.2012

[Butterfield2006] Daniel Stewart Butterfield, Eric Costello, Caterina Fake, Callum James Henderson-Begg, Serguei Mourachov, Joshua Eli Schachter; Media object metadata association and ranking; http://www.patentlens.net/patentlens/patents.html?patnums=US_2006/0242178_A1 &returnTo=quick.html – Zugegriffen am 07.02.2012

[Carlin2006] Sascha Carlin, Schlagwortvergabe durch Nutzende (Tagging) als Hilfsmittel zur Suche im Web. http://itst.net/wp-content/uploads/2007/02/diplomarbeit-tagging-sascha-a-carlin-volltext.pdf - Zugegriffen am 27.02.2012

[Cattuto2006] Ciro Cattuto, Vittorio Loreto, Luciano Pietronero; Semiotic dynamics and collaborative tagging; http://www.pnas.org/content/104/5/1461.full - Zugegriffen am 02.02.2012

[Diederich2006] Jörg Diederich, Tereza Iofciu; Finding Communities of Practice from User Profiles Based On Folksonomies; http://www.l3s.de/ diederich/Papers/TBProfile-telcops.pdf - Zugegriffen am 06.02.2012

[Fensel2004] Dieter Fensel; Ontologies: A Silver Bullett for Knowledge Management and Electronic Commerce, ISBN-13: 978-3540416029

[Garshol2006] Lars Marius Garshol; Tags/folknomies and Topic Maps; http://www.garshol.priv.no/blog/33.html - Zugegriffen am 07.02.2012

[GolHub2005] Scott A. Golder, Bernardo A. Huberman, The Structure of collaborative Tagging Systems http://arxiv.org/ftp/cs/papers/0508/0508082.pdf - Zugegriffen am 29.11.2011

[GolHub2006] Scott A. Golder, Bernardo A. Huberman; Usage patterns of collaborative tagging systems; http://www.jasonmorrison.net/iakm/cited/Golder_usage_patterns_collaborative_ tagging.pdf - Zugegriffen am 06.02.2012

[Guy2006] Marieke Guy, Emma Tonkin; Folksonomies: Tidying up Tags? http://www.dlib.org/dlib/january06/guy/01guy.html - Zugegriffen am 06.02.2012

[Heckel2001] Ronald Heckel; Einsatzmöglichkeiten von Topic Maps zur flexiblen Navigation in elektronischen Dokumenten; http://www.topicmap-design.com/download/ger/Topic%20Maps.pdf – Zugegriffen am 19.03.2012

[Helic2010] Denis Helic, Christoph Trattner, Markus Strohmaier, Keith Andrews; Are Tag Clouds Useful for Navigation? A Network-Theoretic Analysis; http://kmi.tugraz.at/staff/markus/documents/2011_JoSCCPS-socialcom2010_extended.pdf - Zugegriffen am 06.02.2012

[Heymann2006] Paul Heyman, Hector Garcia-Molina, Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems; http://ilpubs.stanford.edu:8090/775/1/2006-10.pdf - Zugegriffen am 22.07.2012

[Hotho2006] Andreas Hotho, Robert Jäschke, Christoph Schmitz, Gerd Stumme; Information Retrieval in Folksonomies: Search and Ranking; http://www.kde.cs.uni-kassel.de/stumme/papers/2006/hotho2006information.pdf - Zugegriffen am 06.02.2012

[Khalifa2007] Hend S. Al Khalifa, Hugh C. Davis; Exploring the value of Folkonsomies for creating semantic metadata; http://eprints.ecs.soton.ac.uk/13555/1/IJSWIS_2007.pdf - Zugegriffen am 22.02.2012

[Kipp2006] Margarete E. I. Kipp; Complementary or Discrete Contexts in Online Indexing: A Comparison of User, Creator and Intermediary Keywords; http://eprints.rclis.org/bitstream/10760/8771/1/mkipp-caispaper.pdf - Zugegriffen am 03.02.2012

[Kipp2007] Margaret E.I. Kipp, D. Grant Campbell; Patterns and Inconsistencies in Collaborative Tagging Systems: An Examination of Tagging Practices; University of Western Ontario; http://eprints.rclis.org/bitstream/10760/8720/1/KippCampbellASIST.pdf - Zugegriffen am 13.12.2011

[Lauer2003] Markus Lauer; Techniken des Semantic Web; http://www.informatik.uni-ulm.de/ki/Edu/Seminare/Semantic.Web/WS0203/3-Lauer-Ontologien.pdf - Zugegriffen am 15.03.2012

[Lux2007] Mathias Lux, Michael Granitzer, Roman Kern; Aspects of Broad Folksonomies; http://www.uni-weimar.de/medien/webis/research/events/tir-07/tir07-papers-final/lux07-aspects-of-broad-folksonomies.pdf - Zugegriffen am 02.02.2012

[Maicher2008] Lutz Maicher; Was ist Topic Maps?; http://www.topicmapslab.de/introduction - Zugegriffen am 07.02.2012

[Marlow2006] Cameron Marlow, Nor Naaman, Danah Boyd, Marc Davis; Position Paper, Tagging, Taxonomy, Flickr, Article, ToRead; http://www.danah.org/papers/WWW2006.pdf - Zugegriffen am 10.12.2011

[Mathes2004] Adam Mathes; Folksonomies – Cooperative Classification and Communication Through Shared Metadata; http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html - Zugegriffen am 03.02.2012

[Morgan2007] Morgan Ames, Nor Naaman; Why we Tag: Motivation for Annotation in Mobile and Online Media; http://www.stanford.edu/ morganya/research/chi2007-tagging.pdf - Zugegriffen am 16.02.2012

[Ohmukai2005] Ikki Ohmukai, Masahiro Hamasaki, Hideaki Takeda, A Proposal of Community-based Folksonomy with RDF Metadata; www-kasm.nii.ac.jp/papers/takeda/05/ohmukai05iswc.pdf – Zugegriffen am 25.01.2012

[Pepper2008] Steve Pepper; Topic Maps and the Semantic Web; http://topicmaps.wordpress.com/2008/05/11/topic-maps-and-the-semantic-web/ - Zugegriffen am 19.03.2012

[Peters2008] Isabella Peters, Wolfgang G. Stock; Folksonomies in Wissensrepresentation und Information Retrieval; http://www.phil-fak.uni-duesseldorf.de/fileadmin/Redaktion/Institute/Informationswissenschaft/1204545101 folksonomi.pdf&pli=1 – Zugegriffen am 07.02.2012

[Rashmi2005] Rashmi Sinha, A cognitive analysis of tagging, http://rashmisinha.com/2005/09/27/a-cognitive-analysis-of-tagging/ - Zugegriffen am 13.12.2011

[Rasinger2005] Marcel Rasinger, Vergleich von RDF/RDFS Schema und OWL; http://www.dke.univie.ac.at/semanticweb/history/ws0405/km_resources/studenten_ hoefferer/Vergleich%20von%20RDF%20RDFS%20und%20OWL.pdf – Zugegriffen am 29.05.2012

[Schön2005] Eckhardt Schön; Das Resource Description Framework (RDF) – ein neuer Weg zur Verwaltung von Metadaten im Netz; http://www.eckhardt-schoen.de/res/Beruf/rdfeinfuehrung.pdf - Zugegriffen am 20.02.2012

[Shirky2005] Clay Shirky; Ontology is Overrated: Categories, Links, and Tags; http://www.shirky.com/writings/ontology_overrated.html - Zugegriffen am 15.12.2011

[Sinclair2008] James Sinclair, Michael Cardwell-Hall; The Folksonomy Tag Cloud: When is it Useful?; http://www.jrsinclair.com/academic/JIS-0449%20-v3%20-%20The%20Folksonomy%20Tag%20Cloud%20-%20When%20is%20it%20useful.pdf – Zugegriffen am 06.02.2012

[Stock2006] Wolfgang Stock; On relevance distributions; http://onlinelibrary.wiley.com/doi/10.1002/asi.20359/full - Zugegriffen am 02.02.2012

[Stock2007] Wolfgang G. Stock; Information Retrieval. Informationen suchen und finden; ISBN-13: 978-3486581720

[Stoyanovich2008] Julia Stoyanovich, Sihem Amer Yahia, Cameron Marlow, Cong Yu; A Study of the Benefit of Leveraging Tagging Behavior to Model Users'Interests in del.icio.us; http://cameronmarlow.com/media/stoyanovich-2008-study.pdf - Zugegriffen am 14.03.2012

[Vander2007] Thomas Vander Wal; Folksonomy Coinage and Definition; http://vanderwal.net/folksonomy.html - Zugegriffen am 20.12.2011

[Xu2006] Zhichen Xu, Yun Fu, Jianchang Mao, Difu Su; Towards the Semantic Web: Collaborative Tag Suggesiont; http://semanticmetadata.net/hosted/taggingws-www2006-files/13.pdf - Zugegriffen am 21.02.2012