(Semi-)Automatische Annotation von Textdokumenten*

Work in Progress

Heiko Betz, Daniel Klan, Kai-Uwe Sattler
Department of Computer Science & Automation
TU Ilmenau, Germany
{*first.last*}@tu-ilmenau.de

Abstract

Das Erfassen der Bedeutung von geschriebener oder gesprochener Sprache ist bis heute eine der größten Herausforderung in der Informatik. Für eine effiziente computergestützte Analyse und Suche ist dies aber unumgänglich. Gegenwärtig ist es nicht möglich Informationen in ausreichender Qualität vollständig automatisch auf ihre Bedeutung hin zu analysieren und entsprechend zu annotieren. Häufig wird daher auf teilautomatische Systeme zurückgegriffen, welche eine Nutzerinteraktion erfordern. In der folgenden Arbeit wollen wir zwei neue Ansätze zur automatischen Annotation von Dokumenten präsentieren.

1 Einleitung

Mit der stetig steigenden Menge an digitalen Dokumenten wird die automatische Computergestützte Auswertung und Verknüpfung von Informationen immer wichtiger. Das Finden relevanter Informationen mit Internet-Suchmaschinen, wie zum Beispiel Google, würde ohne eine entsprechende Vorverarbeitung und Bewertung durch diese häufig zu nicht relevanten Informationen führen. So leistungsfähig aktuelle Suchmaschinen auch sind, im Allgemeinen weisen sie allerdings alle das gleiche Problem auf. Sie können ausschließlich über dem Inhalt von Dokumenten suchen. Deren Bedeutung bleibt ihnen meist verschlossen. Sind die Suchbegriffe nicht Teil der untersuchten Dokumente, so werden diese nicht gefunden.

Semantische Suchmaschinen, wie zum Beispiel Wolfram Alpha[1], wollen dieses Problem lösen, indem sie versuchen sowohl die Bedeutung der durchsuchten Dokumentenbasis als auch der Nutzeranfragen zu erfassen und entsprechende Ergebnisse zu präsentieren. Größtes Problem ist dabei die computergestützte Erfassung der Semantik von Dokumenten. Wie komplex sich dies gestaltet, zeigt sich sowohl am geringen Verbreitungsgrad entsprechender Systeme als auch an der häufig dramatisch kleineren Datenbasis als bei herkömmlichen Suchmaschinen.

Das Erfassen der Bedeutung nativer gesprochener oder geschriebener Sprache zählt bis heute zu einer der größten Herausforderungen in der Computerlinguistik. Gegenwärtig ist kein System bekannt, welches Dokumente in hinreichender Qualität automatisch annotieren kann, so dass sinnvolle komplexe Suchanfragen möglich werden. Die Autoren in [12] haben gezeigt, dass selbst eine vollständig manuelle Annotation von Dokumenten durch Mitglieder einer Community keine Garantie für eine hohe Qualität an Schlagworten darstellt. Häufig werden weniger als 50 % der verwendeten Taggs als nützlich erachtet [12].

Es hat sich gezeigt, dass eine vollständig automatisierte Annotation durch den Computer nur in beschränktem Umfang zu den gewünschten Ergebnissen führt. Eine ausschließlich manuelle Verschlagwortung ist sowohl aufgrund der enormen Datenmenge als auch dem häufig unterschiedlichen Verständnis für die Bedeutung von Information nur bedingt sinnvoll. Im Weiteren wollen wir zwei Ansätze präsentieren, welche den Anwender bei der Annotation von Dokumenten unterstützen sollen. Der Nutzer bekommt bei diesen eine Menge möglicher passender Taggs präsentiert, aus denen er die seiner Meinung nach zutreffendsten Schlagwörter auswählt.

2 Verwandte Arbeiten

Für die Automatisierung des Verschlagwortungsprozesses existieren bereits einige etablierte Verfahren. Prinzipiell lassen sich diese in zwei unterschiedlichen Klassen einordnen: graphbasierte und inhaltsbasierte Verfahren.

Graphbasierete Verfahren [11] kommen häufig beim kollaborativen Tagging, welches zum Beispiel typisch für soziale Netzwerke ist, zum Einsatz. Die Verfahren arbeiten auf einem vorab selbst definierten Graphen. Die finden häufig dann Anwendung, wenn Schlüsselworte zu Objekten zugewiesen werden, die keinen oder einen geringen Anteil an maschinell lesbaren Inhalt aufweisen (zum Beispiel Bilder, Musik, Videos, etc). Knoten in den Graphen entsprechen den zu klassifizierenden Objekten (Schlüsselworte und Benutzer). Die Kanten beschreiben Zusammenhänge zwischen den verschiedenen Objekten. Hierüber wird definiert, welcher Benutzer welches Dokument mit welchen Stichworten versehen hat. Um Stichworte aus den Graphen zu extrahieren, kommen häufig Min-Cut-Algorithmen oder Modifikationen von diesen zum Einsatz.

In [10] wird die Anwendung eines graphbasierten Verfahrens auf ca. 52 Millionen Bilder aus Flickr gezeigt. Die Bilder werden maschinell analysiert und Empfehlungen anhand von bereits getätigten Zuweisungen ausgesprochen. Die Autoren in [5] präsentieren ein graphbasiertes Verfahren, welches eine Erweiterung des PageRank-Algorithmus [6] darstellt und im wesentlichen auf dem in [4] präsentierten Ranking-Verfahren für Folksonomies basiert. Das präsentierte Verfahren berechnet die Wahrscheinlichkeit dafür, dass ein Tag für einen bestimmten Benutzer zu einem bestimmten Objekt vorgeschlagen wird.

Im Gegensatz zu den graphbasierten Verfahren verwenden inhaltsbasierte Verfahren [7] den Inhalt von Doku-

*Funded by the TAB under grant 2010FE9007

[1] http://www.wolframalpha.com/

Input: Textdokument
Output: Menge von Schlüsselworten
1 Erkennung der Sprache
2 Absätze erkennen
3 Zerlegung von Sätzen
4 Splitten von Worten
5 Überführung in Kleinbuchstaben
6 Worttrennungen entfernen
7 Grundwortreduktion
8 Filtern von Stoppworten
9 Filtern von unwichtigen Textstellen
10 Filtern von Substantiven
11 Synonymerkennung
12 Homonymerkennung
13 Abkürzungen ersetzen
14 Relativpronomen erkennen und ersetzten
15 Kompositazerlegung
16 Fachterme erkennen

Algorithm 1: NLP-Vorverarbeitungsschritte

menten um Stichworte aus diesen zu extrahieren. Somit können sie nur auf Objekte angewandt werden die selbst ausreichend maschinell lesbaren Inhalt aufweisen. Hierunter fallen Dokumente wie zum Beispiel Webseiten oder E-Mails. In [7] werden u.a. Empfehlungssysteme vorgestellt, die Objekte anhand des Benutzerprofils und des Inhaltes vorschlagen. In [1] wird gezeigt, wie Taggs anhand von ähnlichen und bereits mit Stichworten versehenen Webseiten vorgeschlagen werden können. [3] enthält eine detaillierte Übersicht über die wichtigsten Textmining-Verfahren um wichtige Worte aus Texten zu extrahieren.

Neben diesen Ansätzen werden zusehends auch Verfahren für das Semantic Web entwickelt. Diese zielen darauf ab Webseiten für Maschinen verständlich zu gestalten, indem der extrahierbare Textinhalt mit zusätzlichen Elementen versehen wird. Es existieren bereits einige Ansätze, die diese Aufgabe in hinreichendem Maße erledigen [8]. Durch die Feingranularität des Prozesses sind den Systemen jedoch Grenzen gesetzt. So existiert bisher noch kein vollautomatisches Programm, dass diese Aufgabe in zufriedenstellendem Maße erledigt. Häufig wird aus diesem Grund auf Systeme gesetzt, welche zunächst manuell angelernt werden müssen und anschließend das erlernte kontinuierlich verbessern [13]. Durch den hohen manuellen Aufwand sind die entwickelten Lösungen jedoch zumeist domänenspezifisch und nicht auf allgemeine Probleme anwendbar.

3 Tag-Empfehlunssystem

Im folgenden sollen zwei Ansätze zur automatischen Annotation von Dokumenten beschrieben werden. Die beiden Ansätze unterscheiden sich dabei in der verwendeten Wissensbasis. Der erste Ansatz beschränkt sich im Wesentlichen auf den Inhalt des zu prüfenden Dokumentes und versucht relevante und beschreibende Begriffe aus diesen zu identifizieren und vorzuschlagen. Der zweite Ansatz vergleicht das zu annotierende Dokument mit bereits korrekt getaggten Dokumenten und versucht auf Grundlage dieser Empfehlungen für das neue Dokument zu geben.

3.1 Inhaltsbasierte Annotation

Die computergestützte Extraktion relevanter Informationen aus sprachlichen Dokumenten ist ein Teilgebiet der Computerlinguistik (NLP - *natural language processing*). Typischerweise durchläuft ein Text dabei eine Prozesskette, deren Ziel das Entfernen nicht relevanter Begriffe, die Zurückführung auf Grundformen, sowie die syntaktische und semantische Analyse ist. Algorithmus 1 zeigt exemplarisch die wesentlichen NLP-Schritte.

Input: Textdokument
Output: Menge von Taggs
1 NLP-Vorverarbeitungsschritte nach Algorithmus 1
2 Korpusbasierte Extraktion
3 Bestimmung von Wortrelationen
4 Part-of-speech Tagging

Algorithm 2: Inhaltsbasierte Annotation

Im folgenden sollen diese NLP-Vorverarbeitunsschritte derart erweitert werden das signifikante Worte extrahiert werden, welche anschließend als Taggs vorgeschlagen werden können (siehe Algorithmus 2).

Ein erster zusätzlicher NLP-Verarbeitungschritt ist die Verwendung eines Korpus. Ein Korpus ist eine Ansammlung von vielen natürlich sprachlichen Texten aus verschiedenen Quellen mit unterschiedlichen Themen und Textarten (z.B. Berichte, Fachliteratur, Lyrik, etc.). Das Ziel eines Korpus ist es, eine relative Auftrittshäufigkeit von Worten in einer Sprache abzuschätzen. Basierend auf dieser kann eine Differenzanalyse durchgeführt werden, d.h. nach der Extraktion eines jeden Wortes aus einem Dokument kann verglichen werden wie häufig dieses Wort im aktuellen Dokument im Vergleich zur Allgemeinsprache auftritt. Ein signifikanter Unterschied kann auf einen Fachterm oder ein wichtiges Wort hindeuten. Weiterhin lassen sich mittels eines Korpus signifikante Nachbarschaftskookkurenzen bestimmen, d.h. es werden Worte identifiziert, die im Dokument häufig in Kombination zusammen auftreten.

Ein weiterer wichtiger zusätzlicher NLP-Schritt ist die Verwendung eines Thesaurus. In diesem kontrollierten Vokabular werden Relationen zwischen Worten definiert, die ein bestimmtes Verhältnis innehaben. Hierunter fällt z.B. die Synonymrelation, Über- und Untergeordnete Begriffe usw.

Weiterhin ist der Einsatz eines Part-of-speech-Tagger (POS-Tagger) sinnvoll. Dieser ermittelt durch Wahrscheinlichkeitsabschätzung welches Wort welcher Wortform angehört[2]. Intern arbeiten solche Programme häufig mit Hidden-Markov-Modellen, die zuvor durch Trainingsdaten angelernt werden müssen.

Reguläre Ausdrücke sollen hier nur der Vollständigkeit halber erwähnt werden. Diese sind auch über die NLP-Disziplin hinaus ein häufig anzutreffendes Vorverarbeitungswerkzeug. Mit regulären Ausdrücken kann auf einem Text, eine einfache Suche auf Eigennamen und normale Nomina, sowie Kombinationen hiervon, angewendet werden um wichtige Worte sowie Fachterme zu filtern.

Durch eine geschickte Kombination all dieser NLP-Techniken lassen sich signifikante Worte aus einem Dokument ermitteln. Diese beschreiben bereits ein Dokument recht gut [3]. Hierbei tritt jedoch ein Problem auf. Durch diesen Ansatz existiert eine sehr eingeschränkte Sicht auf ein einzelnes Dokument. Dies bedeutet in ihrer Konsequenz, dass nur Worte aus dem aktuellen Dokument vorgeschlagen werden können. Dies schränkt den Vorschlag ungemein ein, da ein Tag eine Abbildung von konkreten Worten und Inhalt (dem Urbild) auf häufig ein Wort (dem Bild) ist. Tritt dieses Bild jedoch nicht im Dokument auf,

[2] Für die deutsche Sprache hat sich das Stuttgart-Tübingen Tagset (STTS) etabliert.

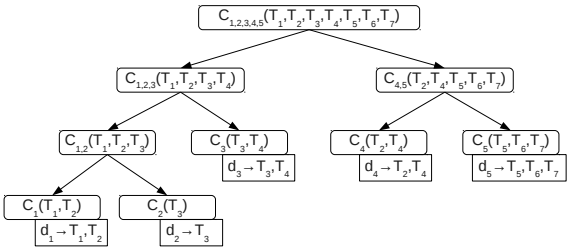

Abbildung 1: COBWEB-Datenstruktur und Taggingklassen

so kann aus dem gegebenen Urbild keine eindeutig treffende und über mehreren Dokumenten hinweg deterministische Zuordnung durchgeführt werden. Letztendlich werden für das gleiche Urbild verschiedene Bilder ermittelt. Dies schränkt schließlich die Eindeutigkeit von Stichworten auf Dokumente ein und verschlechtert die Suche. Der Benutzer muss für das Auffinden gleicher Dokumente alle Bilder kennen. Apriori kann von dieser Annahme jedoch nicht ausgegangen werden. Bis zu diesem Zeitpunkt wurde immer von einem eindeutigen Urbild ausgegangen. Dies ist jedoch nicht gegeben. Somit muss eine probabilistische Abbildung von Urbilder auf Bilder existieren. Ein Grund hierfür ist die Komplexität der natürlichen Sprache. Letztendlich können die automatisch generierten Vorschläge auf dieser Basis eine geringere Qualität aufweisen als Taggs, die durch den im nächsten Kapitel beschriebenen Ansatz generiert werden.

3.2 Kollaborative Annotation

Um dem Problem der rein inhaltsbezogenen Extraktion zu begegnen, soll im weiteren ein kollaborativer Ansatz beschrieben werden. Kollaborativ bedeutet in diesem Zusammenhang, dass andere Dokumente, die bereits mit Schlagworten versehen sind, in die Berechnung mit einbezogen werden. Kollaborative Maßnahmen, welche zur Zeit noch in der Forschung sind, sind nach [5] und [9] ein häufig besserer Ansatz als die inhaltsbezogene Extraktion. Im folgenden wollen wir einen neuen Data Mining basierten Ansatz zur kollaborativen Annotation präsentieren.

Im weiteren soll davon ausgegangen werden, dass Dokumente ähnlichen sprachlichen Inhalts auch über eine ähnliche Bedeutung verfügen und somit analog annotiert werden können. Für einen entsprechenden algorithmischen Ansatz müssen somit für ein Dokument die folgenden drei Fragestellungen geklärt werden:

- Wie lässt sich der wesentlichen sprachliche Inhalt eines Dokumentes erfassen?
- Wie lassen sich Dokumente ähnlichen Inhalts bestimmen?
- Welche der verwendeten Stichworte sollen zur Annotation verwendet werden?

Im weiteren sollen für die beschriebenen Probleme Lösungen aufgezeigt werden.

Die Inhaltsextraktion ist Teil des bereits im letzten Abschnitt beschriebenen NLP-Prozesses (siehe Algorithmus 1). Wie beschrieben, wird bei diesem Ansatz ein Dokument auf seine dominierenden Worte reduziert.

Das Problem des Findens (und Gruppierens) möglichst ähnlicher Objekte (Cluster) gehört zu den klassischen Aufgaben im Bereich des Data Mining. Üblicherweise basieren entsprechende Verfahren auf der Definition geeigneter Distanzfunktionen, so dass eine qualitative Aussage über aktuelle Gruppierungen möglich ist. Die Wahl einer geeigneten Distanzfunktion ist dabei essentiell für die Güte des Clusterergebnisses. Zwar existieren für das Textclustering entsprechende Maße, aufgrund der notwendigen Überführung in numerische Werte leidet die Clusterqualität allerdings häufig. Einen anderen Ansatz verfolgt COBWEB [2]. COBWEB erzeugt eine Konzepthierarchie (Entscheidungsbaum), deren Blätter jeweils ein einzelnen Objekt enthält. Jeder Knoten enthält die Wahrscheinlichkeit dafür, dass ein Objekt mit den Knoteneigenschaften in diese Klasse fällt. Das Erstellen der Datenstruktur erfolgt dabei auf Basis eines Reinheitsmaßes (*Category Utility*), welches für jede Klasse (Cluster) berechnet wird. Wird der Datenstruktur ein neues Dokument hinzugefügt, dann wird dieses testweise in jeden Cluster eingefügt und überprüft, ob sich das Reinheitsmaß verbessert. Abschließend wird das Objekt dem Cluster mit dem größten Reinheitsgewinn hinzugefügt.

Mit Hilfe dieses Clusterverfahrens lassen sich nun Gruppen von Dokumenten identifiziert, deren Inhalt ähnlich zueinander ist. Abbildung 1 zeigt ein einfaches Beispiel. Es wurden 5 Dokumente $(d_1, ..., d_5)$ mittels COBWEB geclustert. Anschließend wurden die Taggs T_i der jeweiligen Dokumente in die Knoten C_i der Konzepthierarchie übernommen. Wird ein neues Dokument eingefügt, so wird unterhalb des am besten passenden Knotens ein neues Blatt erzeugt. Die Taggs des entsprechenden Elternknotens können anschließend als Empfehlungen das neu eingefügte Dokument übernommen werden.

Neben der Verwendung nominaler Attribute bietet der COBWEB-Algorithmus noch weitere Vorteile. So erlaubt die inkrementelle Arbeitsweise prinzipiell eine schnelle Klassifizierung neuer Dokumente, was bei einem Empfehlungssystem für Nutzer essentiell ist. Des weiteren ist es denkbar, die Konzepthierarchie insoweit auszunutzen, dass dem Nutzer nicht nur die Schlagworte des Elternknotens eines Dokumentes präsentiert werden, sondern auch solche die sich in einer höheren Ebene der Hierarchie befinden.

Zu den wesentlichen Vorteilen des kollaborativen Ansatzes gehört, dass ausschließlich durch den Nutzer verifizierte Tags zum Einsatz kommen. Einer der Nachteile des Verfahrens ist die abhängig von einer ausreichend großen bereits annotierten Dokumentenmenge. Für das Verfahren ist es zwingend erforderlich, dass bereits inhaltlich ähnliche Dokumente existieren. Ist dies nicht der Fall werden zwar ebenfalls Taggs vorgeschlagen (im schlimmsten Fall von der Wurzel der Konzepthierarchie), diese haben aber unter Umständen keinen wirklichen Bezug zum eigentlichen Dokument.

3.3 Kombination und Ranking

Beide oben präsentierte Verfahren weisen verschiedene Vor-und Nachteile auf. Der inhaltsbezogene Ansatz kann lediglich Taggs vorschlagen, welche tatsächlich in den Dokumenten enthalten sind (in ihrer Grundform bzw. Bedeutung). Der kollaborative Ansatz ist auf eine aktive Community, korrekte Annotationen durch diese, sowie eine ausreichend große Datenbasis angewiesen. Eine Kombination beider Verfahren kann viele dieser Probleme lösen. Allerdings erschwert die durch eine Kombination beider Verfahren entstehende Menge an möglichen Taggs die Auswahl für den Nutzer. Eine mögliche Unterstützung bei der Entscheidungsfindung ist die Gewichtung (Ranking) der beiden Ergebnislisten. Für beide präsentierten Ansätze lassen

sich entsprechende Lösungen finden.

Die einfachste Möglichkeit des Ranking für die inhaltsbasierte Lösung ist ein punkbasierter Ansatz. Dieser bestimmt die Vorkommenshäufigkeit der Schlagworte in den jeweiligen Dokumenten und gibt die Sortierung als Ranking aus. Im Gegensatz dazu bestimmen probabilistischen Ansätze mit welcher Wahrscheinlichkeit ein Tag ein Dokument beschreibt, unter der Bedingung, dass dieses Tag von anderen (ähnlichen) Dokumenten verwendet wurde. Letztendlich beschreibt dieses Vorgehen den Bayes-Klassifikator. Ähnliche Dokumente lassen sich wiederum mittels Clustering bestimmen.

Für den Fall des oben beschrieben kollaborativen Ansatzes kann auch der hierarchische Gedanke und das berechnete Reinheitsmaß von COBWEB mit einbezogen werden. Dies bedeutet, dass ein Stichwort dann signifikant ist, wenn es in einem Subcluster häufig auftaucht, jedoch in anderen signifikant weniger häufig vorhanden ist (auch übergeordnete Cluster). Dieser Ansatz kann rekursiv über allen Ebenen von COBWEB ausgenutzt werden. Hiermit kann eine Hierarchie von übergeordneten Kategorisierungsbegriffen ermittelt werden, die den Suchprozess bzw. das Tagging unterstützen können.

4 Integration in TOPIEK

Beide oben vorgestellte Ansätze werden gegenwärtig in das von uns entwickelte Dokumentenmanagementsystem (DMS) TOPIEK integriert. Ziel des Projektes ist es die Funktionalitäten herkömmliche DMS um Techniken aus dem Bereich des Semantic Web zu erweitern. So wurden bereits unter anderem Funktionen zur semantischen Suche in das System integriert. Neben der Verwendung Domainspezifischer Ontologien zählt das (semi-)automatische Annotieren von Dokumenten zu einer der Forschungsschwerpunkte in diesem Projekt.

Der oben präsentierte kollaborative Ansatz wurde bereits weitestgehend in TOPIEK integriert. Erste praktische Erfahrungen haben gezeigt, dass die in Algorithmus 1 präsentierte NLP-Vorverarbeitung und das anschließende integrieren in COBWEB ausreichend schnell Taggs für den Nutzer vorschlägt. D.h. zu einem neuen Dokument werden passende Schlagworte nahezu umgehend nach dessen hinzufügen zum Datenbestand präsentiert. Für die Vorverarbeitung kommen sowohl Rapidminer[3] als auch GATE[4] zum Einsatz. Die COBWEB Implementierung entstammt der Weka-Bibliothek[5].

Der inhaltsbezogene Ansatz, sowie die Kombination der beiden Verfahren befinden sich aktuell in der Implementierungsphase. Für die Umsetzung wurde ein eigenes Workflowkonzept entwickelt, welches die bestehenden Nachteile von GATE und Rapidminer beseitigen soll. In diesem ist es möglich Operatoren in beliebiger Art und Weise zu verschachteln, sowie beliebige Datenobjekte zu verwenden.

5 Zusammenfassung und Ausblick

Die automatische Erfassung der Semantik von Informationen zählt bis heute zu den großen Problemen in der Computerlinguistik. In der vorliegenden Arbeit wurden zwei Ansätze zur automatischen Annotation von Dokumenten präsentiert. Beide Lösungen versuchen aktuelle Techniken aus den Bereichen NLP und Data Mining so zu kombinieren, das qualitativ hochwertige Empfehlungen für die Annotation von Dokumenten entstehen. Erste Ergebnisse zum vorgestellten kollaborativen Ansatz haben gezeigt, dass, insofern eine ausreichend große Dokumentenbasis existiert, das Übertragen der Annotation von Dokumenten ähnlichen sprachlichen Inhalts zu guten Tagg-Vorschlägen führt. Weiterhin hat sich gezeigt, dass dieses Verfahren ausreichend schnell ist, was insbesondere bei Empfehlungssystemen für Nutzer von Interesse ist.

Im Weiteren Projektverlauf wollen wir die Umsetzung des inhaltsbasierten Ansatzes vorantreiben, sowie die Kombination der beiden präsentierten Lösungen untersuchen. Zu den wesentlichen Herausforderungen zählen hierbei eine effiziente Umsetzung der inhaltsbasieren Lösung, die quantitative Beurteilung der erreichten Ergebnisgüte, sowie eine geeignete Kombination der verschiedenen Techniken.

[3] http://rapid-i.com/

[4] http://gate.ac.uk/

[5] http://www.cs.waikato.ac.nz/ml/weka/

Literatur

[1] A. Byde, H. Wan, and St. Cayzer. Personalized tag recommendations via tagging and content-based similarity metrics. In *ICWSM*, March 2007.

[2] D. H. Fisher. Knowledge acquisition via incremental conceptual clustering. In *Machine Learning*, pages 139–172, 1987.

[3] U. Quasthoff Th. Wittig G. Heyer. *Text Mining: Wissensrohstoff Text - Konzepte, Algorithmen, Ergebnisse*. W3L, 2006.

[4] R. Jäschke, L. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme. Tag recommendations in folksonomies. In *PKDD 2007*, volume 4702 of *Lecture Notes in Computer Science*, pages 506–514. Springer, 2007.

[5] Zh. Liu, Ch. Shi, and M. Sun. Folkdiffusion: A graph-based tag suggestion method for folksonomies. In *Information Retrieval Technology*, volume 6458 of *Lecture Notes in Computer Science*, pages 231–240. Springer, 2010.

[6] L. Page, S. Brian, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Standford Universität, 1998.

[7] M. Pazzani and D. Billsus. Content-Based Recommendation Systems. In *The Adaptive Web*, volume 4321 of *Lecture Notes in Computer Science*, chapter 10, pages 325–341. Springer, 2007.

[8] L. Reeve. Survey of semantic annotation platforms. In *SAC '05*, pages 1634–1638. ACM Press, 2005.

[9] J. B. Schafer, D. Frankowski, J. Herlocker, and Sh. Sen. The adaptive web. chapter Collaborative filtering recommender systems, pages 291–324. Springer-Verlag, 2007.

[10] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *WWW '08*, pages 327–336, New York, NY, USA, 2008. ACM.

[11] Y. Song, Z. Zhuang, H. Li, Q. Zhao, J. Li, W.-Ch. Lee, and C. L. Giles. Real-time automatic tag recommendation. In *SIGIR '08*, pages 515–522, New York, NY, USA, 2008. ACM.

[12] S. C. Sood, S. H. Owsley, K. J. Hammond, and L. Birnbaum. Tagassist: Automatic tag suggestion for blog posts. 2007.

[13] M. Yasrebi and M. Mohsenzadeh. Semi-automatic approach for semantic annotation. 2009.