

Georg Güntner, Rolf Sint, Rupert Westenthaler

Ein Ansatz zur Unterstützung traditioneller Klassifikation durch Social Tagging

Zusammenfassung

Der vorliegende Beitrag stellt einen Ansatz zur Kombination von traditionellen, geschlossenen Klassifikationsverfahren mit offenen, auf Social Tagging basierenden Klassifikationsverfahren vor. Die Darstellung geht von den grundsätzlichen Anforderungen an die Suche und Navigation in Dokumentenarchiven aus, erörtert die Vor- und Nachteile von geschlossenen und offenen Klassifikationsansätzen und präsentiert schließlich einen kombinierten Lösungsansatz, der im Rahmen eines Prototypen umgesetzt wurde.

Der Lösungsansatz sieht vor, dass Dokumente grundsätzlich mit freien Tags klassifiziert werden können: Die Klassifikation wird jedoch durch ein kontrolliertes Vokabular unterstützt. Freie Tags werden in einem nachgeordneten, moderierten Prozess in das kontrollierte Vokabular übernommen. Das auf diese Weise wachsende und laufend gepflegte Vokabular unterstützt die Suche und Navigation im Dokumentenraum.

1 Motivation und Begriffsdefinition

Die Idee zur Entwicklung und Evaluation des vorgestellten Lösungsansatzes zur Kombination von traditionellen, geschlossenen Klassifikationsverfahren mit offenen, auf Social Tagging basierenden Klassifikationsverfahren entstand im Rahmen eines europäischen Forschungsprojekt, das auf den ersten Blick nicht viel mit Social Tagging zu tun hat: Das integrierten Projekt LIVE¹ („*Live Staging of Media Events*“, FP6-27312,) entwickelt neuartige interaktive Fernsehformate zur Präsentation von Sport-Großereignissen (wie z.B. den Olympischen Spielen 2008 in Peking). Eine wichtige Voraussetzung bei der Produktion derartiger neuer Übertragungskonzepte besteht darin, Live-Datenströme und Archiv-Inhalte mit Hilfe eines im Projekt entwickelten Annotierungs-Werkzeuges zu klassifizieren. Dabei kommt sowohl die geschlossene (d.h. auf ein vordefiniertes Vokabular eingeschränkte) als auch die offene (d.h. freie) Klassifikation zum Einsatz. Im Projekt

¹ Live Staging of Media Events: <http://www.ist-live.org/> - Letzter Zugriff: 15.01.2008

LIVE werden die Klassifizierungen dann unter anderem zur Empfehlung von Video-Datenströmen an das Produktionsteam verwendet.

Dass die Problemstellung an sich nicht auf die Echtzeit-Annotierung von Video-Datenströmen beschränkt ist, war dem Projektteam bereits im Zuge der Anforderungsanalyse bewusst. Aus diesem Grund wurde die Problemstellung in einem etwas breiteren Kontext im Hinblick auf die Anforderungen an Suche und Navigation in digitalen Dokumentenarchiven untersucht. Der vorliegende Beitrag umfasst den aktuellen Stand der Untersuchung und wurde im Rahmen einer an der Universität Salzburg eingereichten Diplomarbeit für die Landesforschungsgesellschaft Salzburg Research erarbeitet.

Wir verstehen in der Folge unter „*Dokument*“ jedwede Art von digitalen Dokumenten, seien sie textueller (z.B. Web-Sites, Einträge in Weblogs oder Wikis, Dokumente im engeren Sinn) oder anderer Natur (z.B. Bilder, Videos, Tonaufnahmen, usw.). Der Begriff „*Archiv*“ wird für eine Sammlung von Dokumenten verwendet, unabhängig davon, ob diese Sammlung im Internet, in geschlossenen Dokumentenmanagement-Systemen oder in einem Dateisystem aufbewahrt und verwaltet wird. Zur Klassifikation der Dokumente wird in unserem Ansatz neben freien Tags ein „*kontrolliertes Vokabular*“ verwendet, das von einer Person oder einem Team erstellt, gepflegt und erweitert wird. Wir haben diesem Team in Kontext dieses Dokuments die Rollenbezeichnung „*Vocabulary-Manager*“ gegeben.

2 Navigation und Suche in Dokumentenarchiven

Für die Suche und Navigation in Dokumentenarchiven stehen grundsätzlich zwei Arten von Informationen zur Verfügung: Zum einen handelt es sich dabei um Informationen, die – meist in unstrukturierter Form – im Dokument selbst gespeichert sind (z.B. der Text in einem PDF-Dokument). Diese Information steht für Suche und Navigation dann zur Verfügung, wenn mit einer geeigneten Dekodierung darauf zugegriffen werden kann. Eine zweite Form der Information sind außerhalb der Dokumente gehaltene beschreibende Daten, sogenannte Metadaten. Diese Metadaten können selbst wieder unstrukturiert (z.B. kann der Dokumentinhalt in Prosaform zusammengefasst werden) oder hoch strukturiert sein. Einfache Formen der Strukturierung von Metadaten bieten Wortlisten, Taxonomien und Thesauri. Komplexere Formen existieren in Form von relationalen Datenmodellen und Ontologien. Die gewählte Komplexität zur Modellierung und Strukturierung von Klassifikationssystemen hängt – unabhängig von der technischen Realisierung und rein auf die praktische Anwendung bezogen – stark von den Anforderungen der Anwendung, dem verfügbaren Aufwand für die Pflege des Klassifikationssystems, der Vertrautheit der AnwenderInnen mit dem Klassifikationssystem sowie von der beabsichtigten Offenheit des Klassifikationssystems ab.

Wir stellen im folgenden drei im Kontext der untersuchten Fragestellung relevante Ansätze für Such- und Navigationsverfahren in Dokumentenarchiven einander gegenüber: 1) die *Volltextsuche* als weit verbreiteten Ansatz in unstrukturierten Dokumentenarchiven, der darüber hinaus auch oft in Verbindung mit anderen Verfahren eingesetzt wird, 2) einen *geschlossenen Klassifikationsansatz* mittels eines kontrollierten Vokabulars und 3) einen offenen, kollaborativen Klassifikationsansatz auf der Basis von *Social Tagging*.

2.1 Volltextsuche

Ein häufig verwendetes Verfahren, bestimmte Dokumente in Archiven aufzufinden, ist die Volltextsuche. Bei einer Volltextsuche werden die vom Benutzer eingegebenen Wörter mit jenen in den gespeicherten Textdokumenten verglichen. Die Dokumente, in denen die Suchbegriffe enthalten sind, werden dann dem Benutzer in einer Ergebnisliste zur Verfügung gestellt (Reimer, 2004).

Ein bekanntes Problem der Volltextsuche ist die Tatsache, dass grundsätzlich nur Dokumente gefunden werden, die den Suchbegriff exakt enthalten. Viele auf Volltext basierte Suchmaschinen erweitern daher die Volltextsuche um zusätzlicher Funktionalitäten, z.B. werden üblicherweise verschiedene Varianten eines Wortes auf einen Wortstamm zurückgeführt und übereinstimmende Wortstämme gesucht („*Stemming*“). Für den Fall, dass sich die Sprache der formulierten Suchanfrage von der Sprache der Dokumente unterscheidet („*Mehrsprachigkeit*“), oder für Wörter mit gleicher Bedeutung („*Synonyme*“) bleibt die Suche erfolglos.

Ein weiterer Nachteil der Volltextsuche besteht in der Problematik der „*Homonymie*“. Darunter versteht man Wörter, die verschiedene Bedeutungen tragen: So steht zum Beispiel das Wort „Jaguar“ einerseits für eine Automarke, andererseits auch für eine Tierart. Da bei einer Volltextsuche der Bedeutungskontext nicht bekannt ist, werden in unserem Beispiel bei einer Suchanfrage sowohl Dokumente selektiert, die mit der Automarke „Jaguar“ zu tun haben, als auch solche, die mit dem Raubtier „Jaguar“ zu tun haben. Die BenutzerIn muss zusätzlichen Arbeitsaufwand in das Selektieren der im Kontext der Suche relevanten Ergebnisse investieren.

Bei Dokumenttypen, deren Inhalt nicht aus natürlichsprachigen Zeichenketten besteht oder in solche umgewandelt werden kann (z.B. Bilder, Videos und Tonaufnahmen), ist die textbasierte Volltextsuche nicht anwendbar.

Trotz der beschriebenen Einschränkungen ist Volltextsuche ein beliebtes Verfahren, um relevante Informationen in Archiven aufzufinden: Viele der führenden Internet-Suchmaschinen und Dokumentenverwaltungs-Systeme basieren auf Volltext-Suchverfahren, die meist mit weiteren Techniken kombiniert und verfeinert

werden. Der Grund für die Verbreitung der Volltextsuche liegt in der Einfachheit dieser Technik. Eine Suchanfrage wird intuitiv formuliert und wenn die Ergebnisse nicht passen, wird die Anfrage einfach mit zusätzlichen, einschränkenden oder anderen Wörtern wiederholt.

2.2 Geschlossene Klassifikationsansätze: Kontrollierte Vokabulare, Thesauri

Eine Möglichkeit, bei der viele der im Zusammenhang mit der Volltextsuche beschriebenen Probleme vermieden werden können, bietet die Klassifikation aller in einem Archiv vorhandenen Dokumente mittels eines kontrollierten Vokabulars. Dabei werden die Dokumente mit Begriffen beschlagwortet, die aus einer eingeschränkten Begriffsmenge stammen. Zusätzlich können die Begriffe (so genannte „Terme“) zueinander in einer Beziehung stehen. Beispielsweise kann eine hierarchische Beziehung zur Modellierung einer Teil-Ganzes-Struktur verwendet werden.

Ein „Term“ eines kontrollierten Vokabulars setzt sich aus einem oder mehreren Wörtern zusammen und darf genau einmal im kontrollierten Vokabular vorhanden sein. Dadurch wird jeder Term eindeutig einem bestimmten Kontext zugeordnet. Damit besteht grundsätzlich die Möglichkeit zur Vermeidung der in der Volltextsuche auftretenden Homonym-Problematik (NIS, 2005). In einem kontrollierten Vokabular könnten im Beispiel des Wortes „Jaguar“ die zwei unterschiedlichen Bedeutungen desselben Wortes durch zwei verschiedenen Terme repräsentiert werden, beispielsweise durch die Terme „Jaguar Tier“ und „Jaguar Auto“. Auf diese Weise wird für jede Bedeutung des Wortes exakt ein Term definiert.

Die Terme des kontrollierten Vokabulars werden zum Annotieren der vorhandenen Dateien verwendet. Auf diese Weise werden zu den Dokumenten eines Archivs „Metainformationen“ zur Verfügung gestellt und die Beschreibung der Dokumente erfolgt unabhängig vom darin enthaltenen Text. Dieses Verfahren kann auch auf Bild-, Video- und Audio-Dokumente angewandt werden, wodurch im Gegensatz zur Volltextsuche die Suche auch in diesen Dokumenttypen angewendet werden kann (NIS, 2005).

Die Bereitstellung von Metainformation mittels eines kontrollierten Vokabulars ermöglicht, dass Dokumente entsprechend der Struktur und den definierten Beziehungen zwischen den einzelnen Termen geordnet bzw. klassifiziert werden.

Ein kontrolliertes Vokabular kann in verschiedenen Formen vorkommen, die sich in Hinblick auf ihre Komplexität unterscheiden. Ein Maß für die Komplexität ist die Anzahl der verschiedenen Beziehungstypen zwischen den Termen: Je mehr

unterschiedliche Beziehungen zwischen Termen ein einem Vokabular vorkommen, desto höher ist seine Komplexität.

Die einfachste Form eines kontrollierten Vokabulars ist eine „Wortliste“, welche eine einfache Sammlung von Termen darstellt (NIS, 2005). Einen höheren Strukturierungsgrad bietet dagegen ein „Thesaurus“: Dieser erlaubt die Definition von hierarchischen Beziehungen, sowie von Gleichheits- und von Verwandtschaftsbeziehungen zwischen den einzelnen Termen:

Die *hierarchische Beziehung* eignet sich zur Modellierung einer Struktur, in der Terme vom Allgemeinen hin zum Speziellen, vom Übergeordneten hin zum Nachgeordneten, vom Ganzen hin zum Teil in Beziehung gebracht werden. Betrachten wir die Terme „Sport“ und „Skifahren“, so kann in einem Thesaurus „Sport“ als Oberbegriff und „Skifahren“ als Unterbegriff bzw. Spezialisierung von „Sport“ definiert werden. Mithilfe der *Gleichheitsbeziehung* werden Synonyme gekennzeichnet. Mit der *Verwandtschaftsbeziehung* können thematische Zusammenhänge zwischen Termen definiert werden (NIS, 2005).

Ein bekannter Vertreter für ein kontrolliertes Vokabular in Form eines Thesaurus sind die „IPTC NewsCodes“. Dabei handelt es sich um eine vom International Press and Telecommunications Council (IPTC) gepflegte Begriffssammlung zur Klassifikation von Nachrichtenbeiträgen (<http://www.iptc.org/NewsCodes/>).

Die Klassifikation von Dokumenten mithilfe eines kontrollierten Vokabulars (z.B. einer Taxonomie oder eines Thesaurus) bietet Such- und Navigationsmöglichkeit, die sich an der Struktur des Vokabulars orientieren können. Sie ist jedoch mit einem zusätzlichen Pflegeaufwand für das Vokabular verbunden: Die Beziehungen zwischen den Termen müssen gesetzt und gegebenenfalls regelmäßig den sich ändernden Bedingungen angepasst werden. Abhängig von der Komplexität der Struktur und der Dynamik der betrachteten Domäne ist der Pflegeaufwand für das Vokabular entsprechend hoch. Mit der Pflege des Vokabulars und dem Aufbau und der Modellierung des Domänenwissens werden in vielen Anwendungsbereichen eigens dafür ausgebildete MitarbeiterInnen betraut, denen wir im Zuge unserer Untersuchungen die Rollenbezeichnung „Vocabulary-Manager“ gegeben haben.

2.3 Offene Klassifikationsansätze: Tagging, kollaboratives Tagging und Folksonomies

Im Gegensatz zu dieser „traditionellen“ Technik Dokumente zu klassifizieren ist gegenwärtig im Internet ein weiterer Trend zu beobachten: Viele unter dem Schlagwort „Web 2.0“ zusammengefassten Web-Anwendungen ermöglichen einer breiten Öffentlichkeit nicht nur die einfache Erstellung von Inhalten im World

Wide Web, sondern sie erlauben auch die einfache Beschlagwortung der veröffentlichten Inhalte (z.B. Weblogs, Fotos, Videos, Bookmarks)

Ein Mechanismus, der es BenutzerInnen gestattet, beliebige Begriffe zur Annotierung von Dokumenten zu verwenden, wird als „Tagging“ bezeichnet. Wenn BenutzerInnen nicht nur ihre selbst veröffentlichten Dokumente annotieren können, sondern auch die der anderer BenutzerInnen wird dies „kollaboratives“ oder als „soziales“ Tagging genannt. Das daraus entstehende Ordnungsschema wird als „Folksonomy“ bezeichnet. Das Wort ist zusammengesetzt aus den Begriffen „folk“ und „taxonomy“ (Weller, 2007).

Typische Repräsentanten, die diesen Klassifikationsansatz unterstützen, sind die Webseiten Del.icio.us² (Social Bookmarking), YouTube³ (Video-Plattform) und Flickr⁴ (Foto-Plattform) – s.a. (Mika, 2005).

Der Vorteil von kollaborativen Tagging-Systemen ist, dass Dokumente unmittelbar mit deren Veröffentlichung von den AutorInnen annotiert und somit klassifiziert werden. Es ist üblicherweise kein zusätzlicher Pflegeaufwand für die Klassifikation von Dokumenten erforderlich. In einigen Fällen wird dieser offene Klassifikationsansatz jedoch missbräuchlich verwendet, um das Ranking in Suchmaschinen zu beeinflussen oder BenutzerInnen durch bewusste Vergabe von falschen Tags auf die angebotenen Inhalte hinzuführen.

Sozialen Klassifikationsansätzen ist gemeinsam, dass jene Personen, die ein bestimmtes Dokument veröffentlichen bzw. annotieren, üblicherweise über das nötige Domänenwissen verfügen, um den Inhalt treffend zu beschreiben. Bei geschlossenen Klassifikations-Ansätzen kann die semantische Lücke zwischen den AutorInnen des kontrollierten Vokabular und dem Domänenwissen der AnwenderInnen zu Problemen führen, die Furnas u.a. in (Furnas, 1987) als das „Vocabulary Problem“ bezeichnet haben: Verschiedene BenutzerInnen klassifizieren Dokumente unterschiedlich abhängig vom Domänenwissen bzw. von der Vertrautheit mit dem kontrollierten Vokabular.

Beim kollaborativen Tagging werden Dokumente von verschiedenen BenutzerInnen annotiert, wodurch eine breite Palette an unterschiedlichen Termen für die Beschreibung eines Dokuments bereitgestellt wird. Auf diese Weise spiegeln Folksonomies das Vokabular der BenutzerInnen und es kommt zu einer Entschärfung des „Vocabulary Problems“, vgl. (Mathes, 2004) und (Marlow, 2006).

Kollaborative Tagging-Systeme bieten eine offene, effiziente Möglichkeit der Klassifikation von Dokumenten.

² Del.icio.us – <http://del.icio.us/> - Letzter Zugriff: 15.01.2008

³ YouTube – <http://www.youtube.com/> - Letzter Zugriff: 15.01.2008

⁴ Flickr – <http://www.flickr.com/> - Letzter Zugriff: 15.01.2008

3 Kombination von offener und geschlossener Klassifikation

Der von uns verfolgte Ansatz zur Kombination von traditionellen, geschlossenen Klassifikationsverfahren mit offenen, auf Social Tagging basierenden Klassifikationsverfahren verbindet die Vorteile der Klassifizierung von Daten mittels eines kontrollierten Vokabulars und eines kollaborativen Tagging-Systems: Der Lösungsansatz sieht vor, dass Dokumente grundsätzlich mit freien Tags klassifiziert werden können: Die Klassifikation wird jedoch durch ein kontrolliertes Vokabular unterstützt. Freie Tags werden in einem nachgeordneten, moderierten Prozess in das kontrollierte Vokabular übernommen. Das auf diese Weise wachsende und laufend gepflegte Vokabular unterstützt die Suche und Navigation im Dokumentenraum.

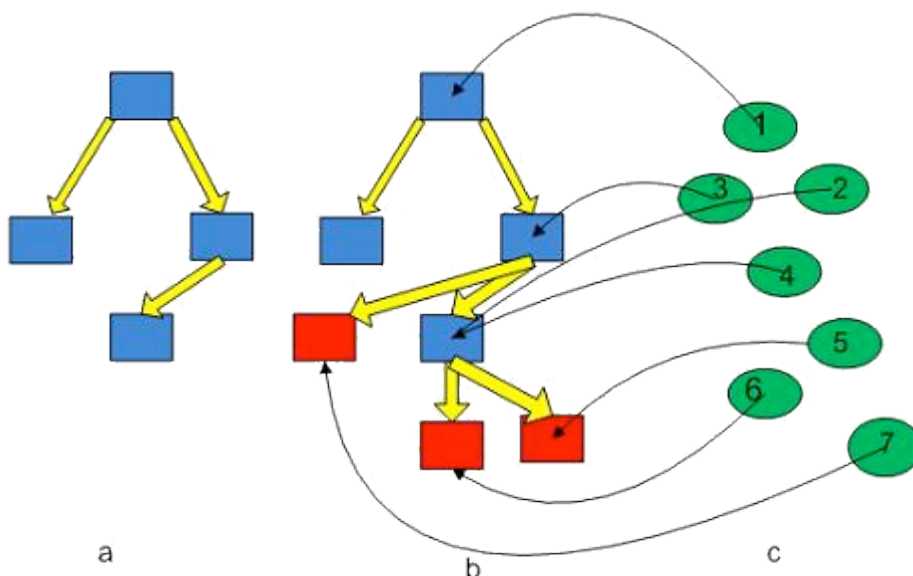


Abb. 1: (a) Kontrolliertes Vokabular; (b) durch Tags angereichertes Vokabular; (c) Tags, die zur Annotierung von Dokumenten verwendet wurden

Aus der Sicht der BenutzerInnen stellt sich der Lösungsansatz wie folgt dar: BenutzerInnen verwenden ihre gewohnte Anwendung (z.B. ein Blogging-Tool), um die verfassten Inhalte zu veröffentlichen und mit frei assoziierten Tags zu annotieren. Der Unterschied zu konventionellen Tagging-Systemen besteht darin, dass ein Vocabulary-Manager die Klassifizierung der Daten in einem nachgeordneten Prozess mittels einer speziellen Software betreut. Diese Software ermöglicht das Einsehen der Dokumente und die Verwaltung der dazugehörigen Tags. Dazu ermöglicht die Anwendung den Zugriff auf ein kontrolliertes Vokabular, welches im Idealfall bereits die gängigen Terme eines bestimmten Wissensgebietes und deren Beziehungen untereinander enthält. Dieses Vokabular könnte beispielsweise in Form eines Thesaurus vorliegen, und ist schematisch in Abb. 1a dargestellt.

Die Menge aller Tags, die zur Annotierung von Dokumenten von den BenutzerInnen verwendet wurde und noch nicht im kontrollierten Vokabular vorhanden sind, werden im folgenden als „neue Tags“ bezeichnet und sind in Abb. 1c dargestellt. Sobald „neue Tags“ zur Annotierung von Dokumenten verwendet wurden, weist die Software den Vocabulary-Manager auf diese Tatsache hin. Dieser kann nun diese „neuen Tags“, unter Angabe eines Beziehungstyps, dem Vokabular beifügen. Wenn es sich bei dem Vokabular um einen Thesaurus handelt, sind das Funktionen zum Eingliedern eines Terms in einer hierarchischen-, Gleichheits- oder Verwandtschaftsbeziehung. Zusätzlich bietet die Anwendung Funktionen an, die es erlauben, die bestehenden Beziehungen zwischen den Termen je nach Bedarf zu verändern. Falls ein Term aus thematischen Gründen nicht in die bestehende Struktur des Vokabulars einzugliedern ist, kann der Vocabulary-Manager selbst Terme zur „Überbrückung“ an bestimmten Stellen einfügen. Dadurch wird ein vorhandenes Vokabular durch Tags, die von BenutzerInnen für die Annotierung verwendet wurden, angereichert (siehe Abb. 1b).

Dieses Verfahren ermöglicht, dass die Klassifizierung der Dokumente auf einem kontrollierten Vokabular basiert, auch wenn diese Tatsache für die BenutzerInnen beim Veröffentlichen und Annotieren der Dokumente nicht notwendigerweise ersichtlich sein muss.

Wir gehen davon aus, dass der Thesaurus in unserem Ansatz nach einer gewissen Zeit und einer entsprechend großen Anzahl an BenutzerInnen und der damit verbundenen Anzahl an vergebenen Tags die gängigsten Wörter innerhalb einer Community enthält. Das bedeutet, dass die meisten Tags, die von BenutzerInnen erstellt werden, bereits im Thesaurus existieren und nicht mehr vom Vocabulary-Manager dem Vokabular zugewiesen werden müssen. Das Vokabular muss ausschließlich mit „neuen Tags“ erweitert werden.

Im Gegensatz zu diesem Verfahren müssen in traditionellen Klassifikationssystemen für jedes Dokument vom Vocabulary-Manager die passenden Wörter für die Annotierung gewählt werden, was mit einem erheblichen Aufwand verbunden ist. Beim vorgestellten Verfahren wird dieser Aufwand auf alle BenutzerInnen verteilt.

Da die Klassifikation auf einem kontrollierten Vokabular basiert, ergeben sich für die BenutzerInnen verbesserte Suchmöglichkeiten. Dabei ergeben sich alle Vorteile die ein Suchverfahren auf Basis eines kontrollierten Vokabulars hat. So könnte eine konkrete Implementierung dieses Ansatzes die Navigation entlang hierarchischer Beziehungen zwischen den Termen gestatten. Dadurch wird ein systematisches Erschließen der Inhalte ermöglicht. Auch können aufgrund der definierten Beziehungen zwischen den Tags die typischen Homonym- und Synonym-Probleme vermieden werden.

4 Prototypische Umsetzung

Unser theoretischer Ansatz zur Kombination von offenen und geschlossenen Klassifikationssystemen wurde im Rahmen der Arbeit prototypisch implementiert. Dadurch wurde eine erste Evaluation des Ansatzes möglich, wenngleich die Evaluationsergebnisse zum gegenwärtigen Zeitpunkt noch nicht vollständig vorliegen:

Der Prototyp besteht aus einer Client Anwendung (einem funktional erweiterten Blogging-Tool zur Erstellung und Annotierung von Inhalten) und dem „Vocabulary Management-Tool“ zur Pflege des kontrollierten Vokabulars. Die Client Anwendung basiert auf dem Open Source Blogging Tool Pebble⁵, welches um zusätzliche Funktionen zur Unterstützung unserer Klassifikationsansätze erweitert wurde.

4.1 Tagging

In einer für Blogging-Tools üblichen Eingabemaske können die BenutzerInnen Beiträge veröffentlichen und mit Tags annotieren. Dabei wird die Benutzerin bei der Auswahl von geeigneten Tags auf folgende Weise unterstützt: Nachdem die BenutzerIn die ersten Buchstaben eines Tags eingegeben hat, wird eine Liste mit allen im kontrollierten Vokabular verfügbaren Termen angezeigt, die mit den von der BenutzerIn eingegebenen Buchstaben beginnen (dies basiert auf AJAX⁶). BenutzerInnen können einen der Terme zum Annotieren auswählen oder einen eigenen „neuen“ Term erstellen, indem sie ein Wort im Eingabefeld formulieren, welches nicht in der Liste enthalten ist.

4.2 Vocabulary Management Tool

Abb. 2 zeigt einen Screenshot der Anwendung, die der Vocabulary-Manager benutzt, um das Archiv und das kontrollierte Vokabular zu verwalten. Diese Software wird „Vocabulary Management Tool“ genannt (VMT):

Der Kalender (1) dient zur zeitlichen Einschränkung und Auswahl der Dokumente und der vergebenen Tags. Im linken Fensterbereich wird das kontrollierte Vokabular (Thesaurus) dargestellt (2), welches im Falle des Prototypen ohne Beschränkung der Allgemeinheit auf den IPTC Newscodes basiert. Der Vocabulary-Manager kann die Schaltflächen (3) benutzen, um die vergebenen Tags dem Vokabular hinzuzufügen bzw. sie mit dem Vokabular in Beziehung zu setzen: Die

⁵ Pebble: <http://pebble.sourceforge.net/> - Letzter Zugriff: 15.01.2008

⁶ AJAX ist die Abkürzung für „Asynchronous JavaScript And XML“

Schaltflächen ermöglichen das Einfügen eines Tags als neue Unterkategorie („<< new Sub-Concept“) oder als neues Synonym („<< new Synonym“) für einen bestehenden Term im Thesaurus. Weiters können mit der Schaltfläche „<< set new related >>“ Terme miteinander verbunden werden, die thematisch in einem Zusammenhang stehen.

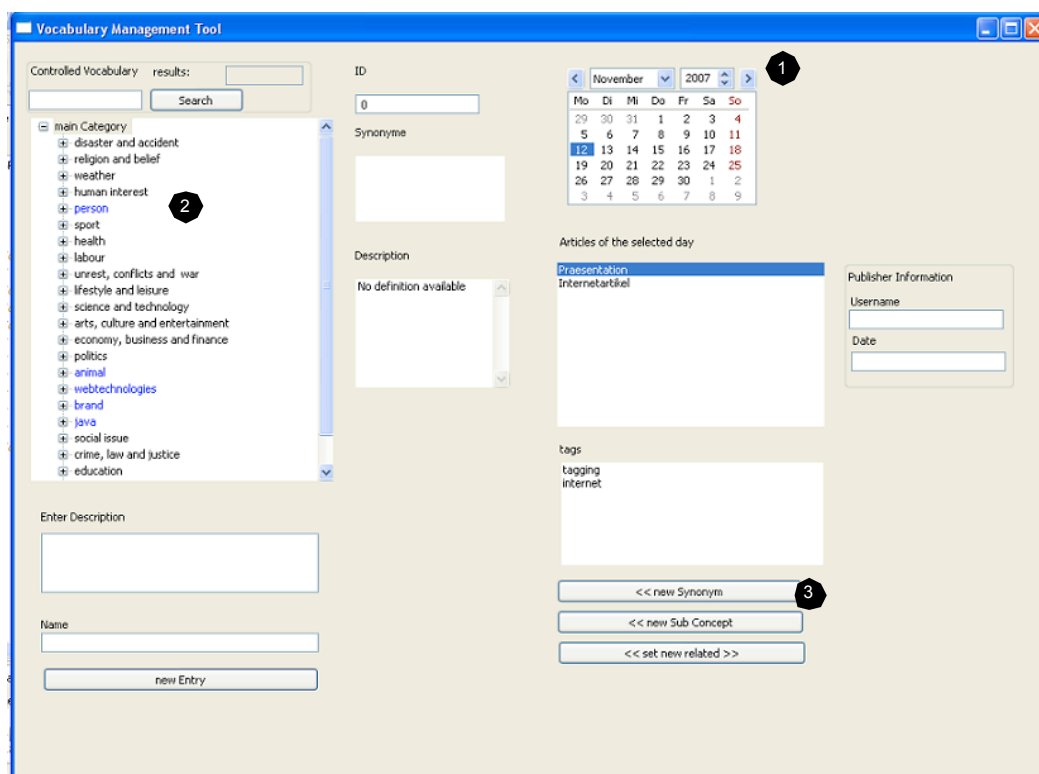


Abb. 2: Vocabulary Management Tool

Die Implementierung des VMTs erlaubt die Erweiterung des kontrollierten Vokabulars mit einem einzelnen Term an mehreren Stellen. Diese Designentscheidung wurde getroffen, da man der BenutzerIn keine Einschränkungen in Bezug auf die zu wählenden Terme zum Annotieren von Dokumenten geben möchte. Die BenutzerIn kann, wie auch in konventionellen Tagging-Systemen, beim Erstellen eines neuen Eintrags beispielsweise den Term „Jaguar“ zur Beschreibung von Dokumenten im Kontext eines Autos oder im Kontext eines Tieres verwenden. Der Vocabulary-Manager fügt den Term dann jeweils im passenden Kontext an der entsprechenden Position im Thesaurus hinzu, falls er dort nicht bereits vorhanden ist.

4.3 Suche und Navigation

Der Prototyp unterstützt die BenutzerInnen in der Wahl der passenden Tags für eine Suchanfrage. Dazu werden den BenutzerInnen zu einem formulierten Term weitere, verwandte Begriffe vorgeschlagen. Diese werden aufgrund der im Thesaurus definierten Beziehungen berechnet. Unter diesen vom System vorgeschla-

genen Termen können die BenutzerInnen diejenigen Begriffe auswählen, die in ihre Suchanfrage mit aufgenommen werden sollen.

So werden der BenutzerIn beispielsweise bei der Formulierung des Wortes „Jaguar“ alle im Thesaurus definierten, synonymen Begriffe vorgeschlagen. Die BenutzerIn kann aus dieser Liste jene Begriffe auswählen, die in ihrer Suchanfrage enthalten sein sollen. Zu Begriffen, die nicht im Thesaurus vorhanden sind, kann die Anwendung keine weiteren Terme empfehlen. Aus diesem Grund werden der BenutzerIn bereits bei der Eingabe der ersten drei Buchstaben eines Tags alle im kontrollierten Vokabular vorhandenen Terme, die mit den eingegebenen Buchstaben beginnen, mittels AJAX-Technologie zur Auswahl angezeigt.

Die Struktur des kontrollierten Vokabulars kann nun von den BenutzerInnen verwendet werden, um die Suche wahlweise auf Synonyme, thematisch verwandte Begriffe oder die auf den Begriffen definierte Strukturbeziehung zu erweitern. Auf ähnliche Weise kann auch die Anzahl der empfohlenen Terme eingeschränkt werden und die Auswahl der passenden Begriffe wird erleichtert.

Im Fall von Homonymen kann durch eine zusätzliche Interaktion der passende Kontext eines eingegebenen Terms identifiziert werden. Damit wird sichergestellt, dass die Software der BenutzerIn die passenden kontextbezogenen Empfehlungen geben kann.

Tagging Assistent

Enter a single tag: jaguar (pref. label)

Choose a Profile: ..

What do you mean? jaguar (pref. label) in context of ..

brand
animal

create Query new Profil

Suggestions:

- jaguar (pref. label)
- jaguarete (synonym)

Abb. 3: Identifikation des Kontextes



Abb. 4: Homonyme im Thesaurus

Abb. 3 zeigt die zur Auflösung der Mehrdeutigkeit erforderliche Interaktion am Beispiel des Begriffs „Jaguar“, welcher an zwei unterschiedlichen Stellen des kontrollierten Vokabulars vorhanden ist (vgl. Abb. 4, in der die doppelten Einträge des Terms „Jaguar“ im Thesaurus dargestellt sind).

Nach der Eingabe des mehrdeutigen Terms wird erkannt, dass der Kontext nicht eindeutig bestimmt werden kann, und die Anwendung fordert die BenutzerIn auf,

einen passenden Überbegriff zu wählen und dadurch den Suchbegriff eindeutig einem bestimmten Kontext zuzuordnen.

5 Evaluierung und Ergebnisse

Auch wenn die Evaluation des Ansatz zur Kombination von traditionellen, geschlossenen Klassifikationsverfahren mit offenen, auf Social Tagging basierenden Klassifikationsverfahren auf der Basis der prototypischen Umsetzung noch nicht abgeschlossen ist, möchten wir nachfolgend erste Ergebnisse vorstellen, die einerseits die angebotene Unterstützung der Suche und Navigation betreffen und andererseits die Unterstützung der Klassifikation der Dokumente untersuchen:

Bei der Unterstützung bei Suche und Navigation zeigte sich, dass vor allem jene Terme zum Entdecken und Auffinden von relevanten Dokumenten beitragen, die den BenutzerInnen aufgrund der im kontrollierten Vokabular definierten Verwandtschaftsbeziehung („related Term“) vorgeschlagen werden.

Die definierten hierarchischen Beziehungen zwischen den Termen können zum systematischen Erschließen eines Themengebietes verwendet werden. Der BenutzerIn bietet sich dadurch die Möglichkeit, einen allgemeinen Begriff zu formulieren und von diesem ausgehend die Suchanfrage mit den von der Anwendung vorgeschlagenen untergeordneten Termen zu spezialisieren.

Werden die Terme im kontrollierten Vokabular in mehrere Sprachen abgelegt, können bei einer Suchanfrage alle Übersetzungen eines Suchbegriffs zur Auswahl angeboten werden bzw. automatisch in die Suche mit einbezogen werden.

Bei der Klassifikation zeigte sich, dass die Unterstützung der BenutzerInnen bei der Eingabe der Tags durch Terme aus dem kontrollierten Vokabular positiv angenommen wurde. Dies gewährleistet eine Konsistenz der Klassifikation einerseits auf terminologischer Ebene, andererseits aber auch auf syntaktischer, lexikalischer Ebene.

Zusammenfassend demonstriert der Prototyp, dass das vorgestellte Verfahren alle Vorteile eines offenen Klassifikationsansatzes aufweist. So kann die BenutzerIn jeden beliebigen Term zur Annotierung der Dokumente verwenden. Weiters werden die Dokumente unmittelbar mit deren Veröffentlichung klassifiziert. Bei der Suche und Navigation können hingegen die Vorteile geschlossener strukturierter Klassifikationssysteme verwendet werden.

Für das Projekt LIVE („Live Staging of Media Events“) sind die Resultate dieser Arbeit aus mehreren Gründen von hoher Relevanz: Erstens hat bei internationalen Sportereignissen die Mehrsprachigkeit eine große Bedeutung. Zweitens erfordert der weit gefächerte Themenkreis von TV Produktionen den Einsatz von offenen

Klassifizierungssystemen. Drittens muss man im Umfeld der Anwendungsdomäne (Live-Produktion) das kontrollierte Vokabular ständig anhand der Liste der vergebenen Tags erweitern, um bei der Archivierung und auch der Suche von Material in Langzeitarchiven ein gut strukturiertes kontrolliertes Vokabular verfügbar zu haben. Während der zeitkritischen Produktionsphase („live“) kommen hingegen Vorschlagssysteme zum Einsatz, die auf den im kontrollierten Vokabular definierten Begriffen beruhen.

Referenzen

- Furnas, G.W., Landauer, T.K., Gomez, L.M., Dumais, S. T. (1987): *The vocabulary problem in human-system communication*. Communications of the Association for Computing Machinery, 30 (11), 964-971. Unter: <http://www.si.umich.edu/~furnas/Papers/vocab.paper.pdf> – Letzter Zugriff: 15.01.2008.
- Marlow, C., Naaman, Mor, boyd, danah., Davis, Marc (2006): *HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, ToRead*. Proceedings of Hypertext 2006, New York: ACM Press. Berkley. Unter: <http://www.danah.org/papers/Hypertext2006.pdf> – Letzter Zugriff: 15.01.2008.
- Mathes, A. (2004): *Folksonomies - Cooperative Classification and Communication Through Shared Metadata*. Academic work. University of Illinois Urbana-Champaign, Computer Mediated Communication - LIS590CMC, Graduate School of Library and Information Science. Unter: <http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.pdf> – Letzter Zugriff: 15.01.2008.
- Mika, P. (2005): *Ontology are us*. Proceedings of the 4th International Semantic Web Conference, ISWC 2005, Galway, Ireland. Vrije Universiteit, Amsterdam.
- NIS - National Information Standards Organization (2005): *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*. Standard: ANSI/NISO Z39.19-2005. Published by NISO Press, July 25 2005.
- Reimer, U. (2004): *Von textbasiertem zu inhaltsorientiertem Wissensmanagement*. Publication, In: R. Hammwöhner (ed): *Wissen in Aktion*, Festschrift für Rainer Kuhlen, Konstanz: UVK Verlagsgesellschaft mbH, 2004. S. 69 – 78. CH-8280 Kreuzlingen. Unter: http://www.informationswissenschaft.org/download/festschrift/cc-festschrift_RK-art5.pdf – Letzter Zugriff: 15.01.2008.
- Weller K. (2007): *Folksonomies and ontologies: two new players in indexing and knowledge representation*. In: *Online Information 2007 Proceedings*; http://wwwalt.phil-fak.uni-duesseldorf.de/infowiss/admin/public_dateien/files/35/1197280560weller009p.pdf – Letzter Zugriff: 15.01.2008.