Edy Portmann, Adrian Kuhn

Extraktion und kartografische Visualisierung von Informationen aus Weblogs

Beim Information Retrieval ist in Anbetracht der Informationsflut entscheidend, relevante Informationen zu finden. Ein vielversprechender Ansatz liegt im semantischen Web, wobei dem System die Bedeutung von Informationen ontologiebasiert beigebracht wird. Sucht der Benutzer nach Stichworten, werden ihm anhand der Ontologie verwandte Begriffe angezeigt, und er kann mittels Mensch-Maschine-Interaktion seine relevanten Informationen extrahieren. Um eine solche Interaktion zu fördern, werden die Ergebnisse visuell aufbereitet. Dabei liegt der Mehrwert darin, dass der Benutzer anstelle von Tausenden von Suchresultaten in einer fast endlosen Liste ein kartografisch visualisiertes Suchresultat geliefert bekommt. Dabei hilft die Visualisierung, unvorhergesehene Beziehungen zu entdecken und zu erforschen.

Inhaltsübersicht

1 Perspektiven im Web

Für ein kongruentes Verständnis werden in diesem Beitrag die Hauptkomponenten definiert, um Informationsextraktion aus Weblogs inklusive kartografischer Darstellung zu realisieren. Um Anwendungen des semantischen Web zu erfassen, wird zuerst auf den Begriff Web 2.0 eingegangen. Dieser verweist auf eine zweite Generation von Anwendungen der Webentwicklung und des Webdesigns, in der Informationsteilung, Interoperabilität, benutzerzentriertes Design und benutzerzentrierte Zusammenarbeit im World Wide Web (WWW) im Vordergrund stehen. Web-2.0-Anwendungen führen zu Netzgemeinschaften und Diensten sowie zu Webapplikationen. Beispiele hierzu reichen von sozialen Netzwerken und Sharing-Seiten über Wikis und Weblogs bis hin zu Mashups und Folksonomies. Viele Konzepte des Web 2.0 dienen heute als Grundlage für Entwicklungen hin zum semantischen Web, das diese kumuliert, mit computerinterpretierbaren Informationen anreichert und zueinander in Beziehung setzt.

1.1 Folksonomy und Weblogs

Das aus den zwei Worten »Folk« (Volk, Leute) und »Taxonomy« (Klassifizierung) gebildete Kunstwort »Folksonomy« beschreibt eine benutzergenerierte Klassifizierung, um Webinhalte wie Webpages, Fotos und Videos zu kategorisieren. Eine Folksonomy ist gemäß Thomas Vander Wal eine von der Praxis des kollaborativen Erschaffens und Bearbeitens von »Tags« (Schlagworte) abgeleitete Klassifikation, um Inhalte zu annotieren und zu kategorisieren [Vander Wal 2007]. Dabei steht das Wort Tag für die Auszeichnung eines Datenbestandes mit zusätzlichen Informationen (wie beispielsweise

einem Internetbookmark, einem digitalen Bild, einem Foto oder einem Video). Diese Metadaten helfen einen Artikel zu beschreiben und durch Browserfunktionen wiederzufinden.

Das Besondere dabei ist, dass dies im Web 2.0 mit vom Benutzer frei wählbaren Schlagwörtern geschieht. So könnte ein Video nicht nur mit Standardeigenschaften, wie dem Namen der Band oder der Auflösung des Videos, sondern mit wertenden Attributen, wie »laut«, »wild« oder »langweilig«, kategorisiert werden. Die Kategorisierung basiert vor allem auf »Social Bookmarking Services« (soziale Lesezeichendienste), wobei computergenerierte Verknüpfungen durch menschliche Assoziationen abgelöst werden. Beispiele solcher Dienste sind Delicious (*http://delicious.com/*), Reddit (*www.reddit.com/*) und Digg (*http://digg.com/*) sowie Film- und Fotoportale wie Flickr (*www.flickr.com/*) oder YouTube (*www.youtube.com/*).

Ein Vorzug der frei wählbaren Schlagwörter ist das Anzapfen der kollektiven Intelligenz. Laut James Surowiecki ist die kollektive Intelligenz ein emergentes Phänomen, wobei zum Beispiel Kommunikation und spezifische Handlungen von einzelnen Individuen gemeinsame, intelligente Verhaltensweisen in sozialen Gemeinschaften hervorrufen können [Surowiecki 2004].

Ein weiteres Element sind die Weblogs. »Weblog« oder kurz »Blog« ist ein Kunstwort, gebildet aus den Wörtern »World Wide Web« und »Log« (Tagebuch). Ein Weblog ist eine spezielle Art eines Content-Management-Systems (Inhaltsverwaltungssystem), das die Erstellung und Bearbeitung von Inhalten ermöglicht.

Im Unterschied zu herkömmlichen Inhaltsverwaltungssystemen werden Weblogs mit chronologisch rückwärts geordneten Einträgen von Kommentaren oder anderen Objekten wie Filmen, Bildern oder Diagrammen normalerweise von einer Person administriert. Zudem besteht für Blogleser die Möglichkeit, einen Kommentar zu hinterlassen, was einen wichtigen Erfolgsfaktor von Weblogs darstellt.

Die wichtigsten Formen von Weblogs sind heutzutage Microblogs wie Twitter (*http://twitter.com/*) oder Plurk (*www.plurk.com/*), die Benutzern erlauben, kurze Textupdates zu vermitteln. Der Inhalt von Microblogs unterscheidet sich üblicherweise nur in der Länge des Eintrags von traditionellen Blogs. So kann ein Eintrag beispielsweise nur aus einem einzigen Satz oder einem Satzfragment bestehen. Vielfach verweisen Einträge in Microblogs auf Webseiten oder Weblogs, wo man zusätzliche Informationen bekommen kann.

Ein gewichtiger Vorteil ist, dass Blogs mit Websuchmaschinen besser gefunden werden und dadurch schneller informieren als andere Medien. Die neusten Informationen zu spezifischen Inhalten werden heute in Weblogs gefunden. Kein anderes Medium vermag es, schneller Neuigkeiten zu verbreiten. »*Guatemala and Iran have both recently felt the Twitter effect, as political protests have been kicked off and coordinated via Twitter*«, wie es Tim O'Reilly und John Battelle in [O'Reilly & Battelle 2009] darlegen. Ein großes Problem ist die Informationsflut (vgl. Abschnitt 1.2). Viele Leser können nicht mehr zwischen für sie relevantem und irrelevantem Inhalt differenzieren. Durch die Möglichkeit der Kommentierung von Einträgen und der starken Verlinkung der Blogs untereinander werden diese von herkömmlichen Websuchmaschinen (vgl. Abschnitt 1.2), wie Google (*www.google.com/*), Yahoo (*www.yahoo.com/*) oder Bing (*www.bing.com/*), vielfach in deren Ergebnisliste auf höherer Stelle platziert.

Für Benutzer hingegen ist es nach wie vor schwierig, zwischen wesentlichen und unwesentlichen Inhalten zu unterscheiden. Manchmal möchte der Benutzer zu einem Suchbegriff verwandte Einträge angezeigt bekommen und sich selbstständig in ein Thema vertiefen können. Dabei kommt im propagierten semantischen Webansatz die Suche mittels Ontologien ins Spiel (vgl. Abschnitt 1.3). Dem Benutzer wird zudem mit der Visualisierung (vgl. Abschnitt 1.4) der Ergebnisse durch Landkarten eine grafi-

sche Orientierung gegeben, wodurch er Ergebnisse und verwandte Konzepte durch Heran- oder Herauszoomen erkennen kann.

1.2 Information Retrieval und Websuchmaschinen

Information Retrieval ist nach Ricardo Baeza-Yates und Berthier Ribeiro-Neto in [Baeza-Yates & Ribeiro-Neto 1999] eine interdisziplinäre Wissenschaft, die sich mit dem (Wieder-)Finden von Informationen aus einer Menge von Dokumenten beschäftigt. Der Begriff Information Retrieval beinhaltet die Suche nach Dokumenten, Informationen und Metadaten.

Üblicherweise werden Information-Retrieval-Systeme genutzt, um den »Information Overload« (Informationsüberflutung) zu reduzieren. Das Vorhandensein »zu vieler« Informationen kann zu einer Blockade in der Entscheidungsfindung führen. Große Mengen an Daten, Widersprüche in vorhandenen Daten sowie ein hohes Rauschen machen es schwierig, Informationen zu filtern, die für eine Entscheidung relevant sein könnten. Unwissen über Methoden des Vergleichens und Aufarbeitens von Informationen verstärken diesen Effekt zudem.

Websuchmaschinen wie beispielsweise die in Abschnitt 1.1 genannten Suchmaschinen Google, Yahoo oder Bing sind Anwendungen aus dem Information Retrieval. Mit Websuchmaschinen sucht der Benutzer im Internet nach bestimmten Informationen. Die Suchresultate werden in einer geordneten Liste präsentiert, wobei die einzelnen Suchresultate »Hits« (Treffer) genannt werden. Die Information kann aus Bildern, Texten, Webseiten und anderen Dokumenttypen bestehen.

Bestimmte Suchmaschinen erlauben, nach Daten in Newsbooks, Datenbanken oder Open Directories zu suchen. Suchmaschinen wie Technorati (*http://technorati.com/*), IceRocket Blog Search (*www.icerocket.com/*) oder Blogdigger (*www.blogdigger.com/*) werden speziell für die Weblog-Suche eingesetzt. Allerdings erlaubt keine der bisherigen Weblog-Suchmaschinen eine Suche nach automatisch erstellten, themenverwandten Suchbegriffen. Dazu muss der Suchmaschine eine Ontologie (vgl. Abschnitt 1.3) hinterlegt werden, wie dies in Abschnitt 2 anhand eines Beispiels demonstriert wird.

1.3 Unscharfe Datensegmentierung und Ontologien

Die Segmentierung von Datenelementen in einzelne Klassen, in denen die einzelnen Elemente einer Klasse sich so ähnlich wie möglich und Elemente anderer Klassen sich so unähnlich wie möglich sind, wird »Data Clustering« (Datensegmentierung) genannt. Die Datensegmentierung ist eine Methode des unbeaufsichtigten Lernens und eine anerkannte Technik der statistischen Datenanalyse und der künstlichen Intelligenz.

Abhängig von der Segmentierungsabsicht und der Beschaffenheit der Daten werden spezielle Zugehörigkeitslevels verwendet, um die Elemente (Schlagworte) in Klassen einzuteilen. Hierbei bestimmt der Zugehörigkeitslevel, wie beispielsweise die Ähnlichkeit, Distanz oder Intensität, wie die Klassen gebildet werden. Ein häufig verwendeter Zugehörigkeitslevel ist dabei der Jaccard-Koeffizient, der durch die Größe der Anzahl gemeinsamer Elemente dividiert durch die Größe der Vereinigungsmenge der Elemente definiert ist:

$$J_{(A,B)} = |A \cap B| / |A \cup B|$$

Des Weiteren unterscheidet man zwischen scharfer und unscharfer Segmentierung, wobei im ersten Fall ein bestimmtes Element nur einer einzigen Klasse, im zweiten Fall auch mehreren Klassen zugewiesen werden darf. In [Bezdek et al. 1999] zeigt James Bezdek, dass durch eine unscharfe Segmentierung Datenelemente zu mehr als einer Klasse gehören können. Dabei beinhaltet jedes Datenelement eine Menge mit einem Zugehörigkeitslevel, der die Zugehörigkeitsstärke zwischen einer Klasse und dem Ele-

ment anzeigt. Insofern ist die unscharfe Segmentierung eine Methode der Bestimmung der Zugehörigkeitslevel und der Zuteilung von Datenelementen $X = \{x_1, \ldots, x_N\} \subset \mathbb{R}^p$ in eine oder mehrere Klassen, abhängig vom Wert des Zugehörigkeitslevels. Der Grad des Zugehörigkeitslevels u_{ik} liegt im Intervall [0..1]. Je größer u_{ik} ist, desto stärker ist die Zugehörigkeit eines Elements x_k zur entsprechenden Klasse i.

Die zu minimierende Zielfunktion lautet:

$$J_{(U,V)} = \sum_{i=1}^{K} \sum_{k=1}^{N} (u_{ik})^m \, d^2(v_i, x_k),$$

wobei $d^2(v_i, x_k)$ die quadrierte euklidische Distanz zwischen den Elementen x_k und den jeweiligen Klassenzentren v_i repräsentiert. K steht für die Anzahl an Klassen, N für die Größe des Datensatzes und $m(> 1)$ beeinflusst die Schärfe der Klassenzugehörigkeit der Elemente. Um eine Ontologie bilden zu können, wird die beschriebene Segmentierungsmethode iterativ wiederholt, wie Edy Portmann und Andreas Meier in [Portmann & Meier 2010] darlegen.

Eine Ontologie – mit begrifflichem Ursprung in der Philosophie – beschreibt ein Modell der Welt, das aus einer Menge von Elementen mit entsprechenden Eigenschaften und Zugehörigkeitslevel gebildet ist. Gemäß Thomas Gruber in [Gruber 1993] ist eine Ontologie eine explizite formale Spezifikation einer Begriffsbildung. Sie enthält Inferenz- und Integritätsregeln, also Regeln für Schlussfolgerungen und zur Prüfung ihrer Gültigkeit.

In der Literatur wird vereinzelt zwischen starken und schwachen Ontologien unterschieden. Eine schwache Ontologie ist nicht so rigoros und erlaubt eine Aufnahme von neuen Gegebenheiten ohne menschliche Interventionen. Gemäß diesem Standard nutzen viele Informationssysteme schwache Ontologien. In der hier präsentierten Arbeit wird auf schwache Ontologien zurückgegriffen. Das Informationssystem sammelt mittels Webagenten Schlagwörter aus Folksonomies, um eine schwache Ontologie zu generieren. Diese Ontologie verändert sich permanent analog den Änderungen verschlagworteter Informationen, wobei Webagenten, eine spezielle Art von Computerprogrammen, die weitgehend autonom sich wiederholenden Aufgaben nachgehen, immer wieder neue Datenelemente und deren Zugehörigkeitslevel der Ontologie hinzufügen.

1.4 Visualisierung und Kartografie

Im WWW wird dem Thema Visualisierung von Informationen in Zukunft wegen der Informationsüberflutung immer größere Beachtung geschenkt werden müssen. Aus diesem Grund sollte ein Schwerpunkt zukünftiger Suchmaschinen auf der Mensch-Maschine-Interaktion liegen, damit große Datenmengen von Benutzern mittels einfach zu bedienender grafischer Benutzerschnittstellen (GUI) und Interaktionsmöglichkeiten besser und angenehmer durchsucht werden können. Dazu wird die Informationsvisualisierung hinzugezogen, die sich nach Colin Ware mit dem Gebrauch von interaktiven, visuellen Repräsentationen von abstrakten Daten befasst, um die Daten mit kognitiven Fähigkeiten zu erschließen. Die menschliche Kognition, aus visuellen Daten Muster auszumachen (»Ein Bild sagt mehr als 1000 Worte«), ist ein entscheidendes Element der Informationsvisualisierung. Wenn abstrakte Daten visuell dargestellt werden, erschließen sie dem menschlichen Betrachter auf einen Blick Strukturen, die bei einer rein tabellarischen Auflistung oder bei einer automatischen Datenaufbereitung (Data Mining) verborgen bleiben [Ware 2000].

Räumliches Denken ist eine ausgeprägte kognitive Fähigkeit. Im Alltag sind wir gewohnt, dass Gegenstände und Dokumente mit den darin enthaltenen abstrakten Informationen in einer räumlichen Beziehung zueinander stehen. Die visuelle Darstellung von Dokumenten in einer Themenlandschaft bedient sich des räumlichen Denkens, indem Dokumente automatisch so auf einer Landkarte platziert werden, dass sie

zu thematischen Inseln zusammenfinden [Wise 1999].

Die Darstellung einer Themenlandschaft ist visuell einer Landkarte nachempfunden, wobei Dokumente als Inseln dargestellt werden und ihre Lage auf der Karte durch das Thema des Dokumentinhaltes gegeben ist. Mittels des mathematischen Verfahrens der multidimensionalen Skalierung (Multidimensional Scaling [Borg & Groenen 2005]) werden die Dokumente so platziert, dass themenverwandte Dokumente zu Inseln zusammenfinden und thematisch unterschiedliche Dokumente weit auseinanderliegen. Dokumente gelten dann als themenverwandt, wenn sie sich desselben Vokabulars bedienen.

Die multidimensionale Skalierung ist ein Verfahren zur Einbettung von Objekten aus einem multidimensionalen metrischen Raum in die zweidimensionale Ebene. Das Verfahren findet nach der Methode der kleinsten Quadrate eine Konfiguration X von Punkten in der Ebene, deren Distanzen $\delta_{ij}(X)$ den Unähnlichkeiten $\delta_{ij}(D)$ der Objekte im metrischen Raum D möglichst ähnlich sind. Im vorliegenden Fall sind die Unähnlichkeiten durch die invertierten Zugehörigkeitslevel der unscharfen Klassifizierung der Webdokumente gegeben.

Die Darstellung von Dokumenten als Themenlandschaften erlaubt, auf einen Blick eine Dokumentmenge und ihre Themen zu erfassen. Ohne Verzögerung kann der Anwender feststellen, welche Dokumente welche Themen abdecken, zudem Anzahl und Umfang der Themen selber abschätzen sowie die Verwandtschaft der Themen zueinander erkennen. Dies erleichtert dem Anwender die Navigation in einem Korpus von Dokumenten mit unbekanntem Inhalt oder unbekannten Themen, was bei einer Internetsuche typischerweise gegeben ist.

Erste Verwendung fanden Themenlandschaften Ende der 90er-Jahre zur visuellen Aufbereitung von Zeitungsartikeln [Wise 1999]. Der damaligen Anwendung blieb jedoch der Durchbruch versagt, da der Einsatz von Themenlandschaften auf dem WWW seiner Zeit voraus war. Einerseits fehlten offene Datenquellen und Folksonomies, wie sie durch die Benutzerpartizipation seit Web 2.0 gegeben sind; andererseits waren die technischen Anforderungen nur bedingt erfüllt. Abseits des WWW fanden Themenlandschaften in vielen Bereichen Verbreitung, wie zum Beispiel in der Politologie [Hermann & Leuthold 2003] oder in der Softwareentwicklung [Kuhn et al. 2009].

Ein Beispiel einer Themenlandschaft ist in Abbildung 1 gegeben. Man identifiziert dabei Hügel, deren Durchmesser und Höhe der Größe der gefundenen Dokumente entsprechen und deren Lage zueinander die Zugehörigkeitslevel der unscharfen Klassifizierung wiedergeben. Je näher sich zwei Hügel liegen, desto verwandter sind die Inhalte der durch die Hügel dargestellten Suchresultate.

In Abbildung 1 wird die Suche nach neuen Technologien in der Bildschirmproduktion verdeutlicht (vgl. Abschnitt 2.2 ff.). Dabei werden stärker verwandte Technologien wie OLED und OEL näher beieinander und Technologien wie LED und LCD, die weniger stark verwandt sind, weiter voneinander entfernt abgebildet. Da die Begriffe OLED und OEL sehr nahe miteinander verwandt sind, werden sie außerdem auf derselben Insel abgebildet, die weniger verwandten Begriffe LED und LCD dagegen auf jeweils einer eigenen Insel.

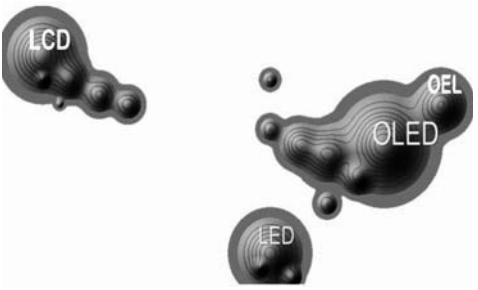

Abb. 1: Beispielhafte Themenlandschaft der unscharfen Suchanfrage

2 Praxisbeispiel

In diesem Abschnitt werden die bisher lose beschriebenen Elemente anhand eines einfachen Beispiels verdeutlicht. Nach der Lektüre sollten der Vorteil und das Zusammenspiel der einzelnen Elemente, der Extraktion und Kartografie von Informationen aus Weblogs, ersichtlich sein.

2.1 Das Zusammenspiel der einzelnen Komponenten

Informationsüberflutung der Benutzer in Weblogs führt zur Frage nach relevanten Informationen und wie diese im semantischen Web besser aufbereitet werden können. Heutzutage wird die Suche für Benutzer häufig erschwert, da Ähnlichkeiten zwischen verschiedenen Begriffen teilweise vage oder gar nicht bekannt sind. Unbekannte Relationen können nicht ohne Weiteres gefunden werden. Ein möglicher Lösungsansatz dieser Probleme ist die präsentierte Methode, in der mithilfe einer unscharfen Segmentierung von Folksonomies eine Ontologie gebildet wird. Diese Ontologie wird als Grundlage für eine verbesserte Suche herangezogen.

In Abbildung 2 wird die Architektur der vorgeschlagenen Weblog-Suchmaschine verdeutlicht. Die Hauptkomponenten sind erstens die grafische Benutzerschnittstelle inklusive entsprechender Webagenten für die Erstellung einer Ontologie, zweitens eine Metasuchmaschine, die nach einmaliger Eingabe einer Suchanfrage mehrere Suchmaschinen mit der Suche betraut, und drittens die Berechnungskomponente für kartografische Suchresultate.

Die Benutzerschnittstelle (inklusive Webagenten)

Die Benutzerschnittstelle dient der Interaktion der Benutzer mit der Suchmaschine. Der Benutzer tippt einen ihm bekannten Suchbegriff in ein leeres Feld und definiert mithilfe eines Schiebereglers den Zugehörigkeitslevel dieses Begriffs (vgl. Abb. 2). Der Zugehörigkeitslevel bestimmt das Intervall, wie weit die Software die Suche nach verwandten Begriffen anhand der zugrunde liegenden, von Webagenten erstellten Ontologie ausdehnen soll (vgl. Abschnitt 1.3). Die Bedienung des Schiebereglers ist äußerst einfach und beinhaltet keine komplexen Berechnungen. Der Benutzer kann eine vage Einstellung seiner Suche vornehmen, die bei einer späteren Interaktion mit der Suchmaschine genauer verfeinert wird.

Die Metasuchmaschine

Eine Metasuchmaschine, wie beispielsweise Dogpile Web Search (*www.dogpile.com/*), ist eine Suchmaschine, deren wesentliches Merkmal darin besteht, dass sie eine Suchanfrage an mehrere andere Suchmaschinen weiterleitet, Ergebnisse sammelt und aufbereitet. Bei der propagierten Methode arbeitet die Suchmaschine die gefundenen Daten auf und eliminiert Dubletten, bewertet die einzelnen Ergebnisse,

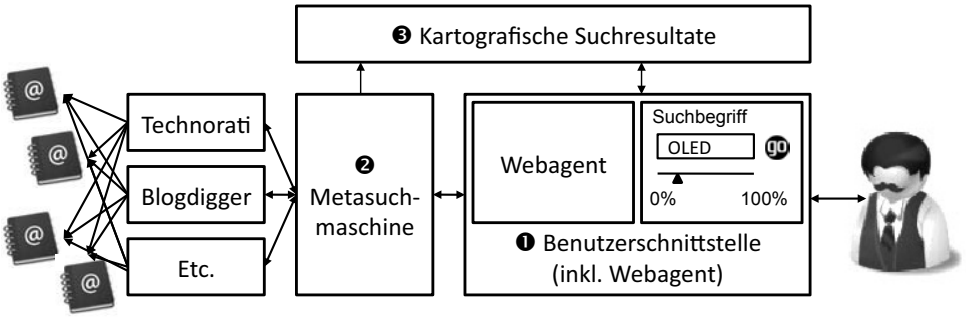

Abb. 2: Das Zusammenspiel der einzelnen Architekturkomponenten

segmentiert diese mithilfe der unscharfen Segmentierung und stellt ein internes Ranking der Ergebnisse auf.

Die kartografierten Suchresultate

Die Suchresultate werden dem Benutzer nicht wie herkömmlich (z.B. bei Google, Bing und Yahoo) in einer einzigen Liste präsentiert, sondern kartografisch aufbereitet, wie zum Beispiel von der Websuchmaschine KartOO (*www.kartoo.com/*) her bekannt. Die verwandten Begriffe werden in Hügeln von Mengen zusammengefasst, wobei die Höhenlinien den Zugehörigkeitslevel widerspiegeln (vgl. Abb. 1). Bei Veränderung des Zugehörigkeitslevels durch Betätigung des Schiebereglers können Mengen von verwandten Suchbegriffen hinzukommen oder wegfallen. In der Karte wird dies durch Auftauchen oder Absenken der Hügel im Meer (der Daten) realisiert. Erst durch einen Mausklick auf eine Höhenlinie werden alle darin enthaltenen Suchresultate als sortierte Liste angezeigt.

2.2 Fallbeispiel aus der Marktforschung

Im Beispiel durchsucht die bildschirmproduzierende Firma Samsung das Internet nach neuen Killerapplikationen potenzieller Konkurrenten, wie organische Leuchtdioden (OLED: »Organic Light Emitting Diode« oder OEL: »Organic Electro Luminescence«).

Eine organische Leuchtdiode ist ein dünnfilmiges, leuchtendes Bauelement aus organischen, halbleitenden Materialien, das sich von den anorganischen Leuchtdioden (LED: »Light-Emitting Diode«) dadurch unterscheidet, dass Strom- und Leuchtdichte geringer sind. Die OLED-Technologie ist vorrangig für Bildschirme und Displays geeignet, weil durch diese hauchdünnen und transparenten Beschichtungen ermöglicht wird, an beliebiger Stelle und in beliebiger Größe einen Bildschirm erscheinen zu lassen. Deshalb ist eine Verwendung der OLEDs als elektronisches Papier ebenfalls denkbar. Im Vergleich zu herkömmlichen Leuchtdioden lassen sich organische Leuchtdioden kostengünstiger herstellen. Der neuartige Herstellungsprozess von OLEDs hat viele Vorteile gegenüber anderen Flachbildschirmen wie etwa Flüssigkristallbildschirmen (LCD: »Liquid Crystal Display«).

2.3 Die Suche

Um nach der bahnbrechenden neuen OLED-Technologie in Weblogs zu suchen, gibt ein Benutzer den Suchbegriff »OLED« ein und definiert die Relevanz der Suche, beispielsweise anhand einer Gewichtung von 0,8 mithilfe eines Schiebereglers (vgl. Abb. 2). Der Benutzer definiert die Gewichtung intuitiv auf einer nicht metrischen, unscharfen Skala, was der menschlichen Natur entspricht. Die Gewichtung (0,8) wird hier nur für das Verständnis des Beispiels erwähnt und ist außerdem arbiträr gewählt.

Im Beispiel in Abbildung 3 ist verdeutlicht, wie die Suche die Begriffe OLED und OEL in Beziehung bringt (mit einem Zugehörigkeitslevel von 0,9). Wegen der Gewichtungsauswahl von 0,8 wird bei dieser Suche im Intervall [0,8..1] der Ausdruck LED mit einer Zugehörigkeit von 0,6 ausgeschlossen. Der Begriff LCD ist in dieser Ontologie zu schwach verwandt, weswegen er in diesem Beispiel nicht gefunden wird.

Bei einer herkömmlichen (booleschen) Suche bekommt der Benutzer der entsprechenden Suchmaschine als Antwortmenge lediglich den Blog mit dem Eintrag zu OLED zurück. Der Blog mit dem Eintrag zu OEL, in dem die gleiche Technologie zugrunde liegt, wird nicht gefunden. Zudem werden die Begriffe LED und LCD nicht gefunden, obwohl sie mit dem Begriff OLED mehr oder weniger verwandt sind.

Bei einer Suche nach OLED und dem Zugehörigkeitsbereich von [0,8..1] werden in diesem Beispiel nur OLED und OEL gefunden. Würde der Zugehörigkeitsbereich auf [0,6..1] ausgeweitet, könnten zudem Einträge zu LED gefunden werden. LCD ist in dieser automatisch generierten Ontologie (vgl. Abschnitt 1.3) momentan nur schwach verwandt. Das kann sich allerdings ändern (beispielsweise nach neuen Forschungen), wenn Informationsressourcen

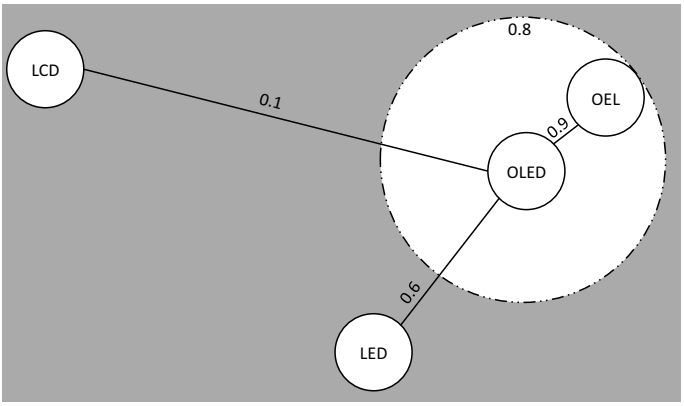

Abb. 3: Beispielhafte Suche mit einem Zugehörigkeitslevel von 0,8

entsprechend verschlagwortet werden. Jedes Schlagwort verändert die der Suche zugrunde liegende Ontologie.

2.4 Auswertung und Aufbereitung der Suchergebnisse

Um eine Ontologie zu kreieren, sucht ein Webagent primär das Internet nach Schlagwörtern ab. Im Fallbeispiel sucht der Webagent nach entsprechenden Folksonomies, um dadurch mithilfe der erwähnten unscharfen Segmentierung (vgl. Abschnitt 1.3) eine Ontologie zu bilden.

Bei der Aufbereitung der Ergebnisse vor der Segmentierung werden die gefundenen Schlagwörter mittels Ähnlichkeitsmaß (vgl. Abschnitt 1.3) im Raum geordnet. Diese Ordnung dient als Grundstruktur, auf die die unscharfe Segmentierung angewandt wird. Mit ihrer Hilfe werden die einzelnen Schlagwörter in mehrere aussagekräftige Klassen, wie beispielsweise in eine OLED-Klasse, geteilt. Wie in Abbildung 3 gezeigt, sind die vom Agenten gefundenen Begriffe OLED, OEL, LED und LCD miteinander verwandt. Außerdem werden in dieser Abbildung die Zugehörigkeitslevel der Begriffe OEL (0,9) und LED (0,6) zu OLED verdeutlicht.

Gleichartige Begriffe wie OLED (Zugehörigkeit von 1) und OEL (0,9) werden hier in Hügeln von Mengen zusammengezogen, wobei die Höhenlinien (in unserem Beispiel von [0,8..1]) den Zugehörigkeitslevel widerspiegeln. Bei wiederholter Betätigung des Schiebereglers können die entsprechenden Höhenlinien verschoben werden. Dementsprechend könnte im Beispiel die Suche ausgeweitet werden, wenn der Zugehörigkeitsbereich beispielsweise auf [0,6..1] ausgedehnt würde. Es käme zusätzlich der Weblog mit dem Eintrag zu LED (0,6) zur Karte hinzu.

Durch einen Mausklick auf eine Höhenlinie, beispielsweise 0,9, würden die Weblogs zu OLED und OEL in einer Liste angezeigt. Durch Klick auf den Link würde der Benutzer zum entsprechenden Weblog geführt.

2.5 Visuelle Interaktion

Die populärsten Suchmaschinen bieten mehrheitlich eine Rückgabe der Suchresultate anhand einer Liste an. Neuere Suchmaschinen wie zum Beispiel KartOO, Clusty (*http://clusty.com/*) oder WolframAlpha (*www.wolframalpha.com/*) versuchen diesen Trend durch innovative Ideen abzufangen. Der Gewinn für den Nutzer ist mit einem besseren Verständnis für die gesuchten Daten gegeben. Der Benutzer kann ähnliche Klassen erkennen, sich besser orientieren und bekommt teilweise eine direkte Antwort auf gestellte Fragen.

Ein gewichtiger Punkt des Ansatzes ist die Ergonomie, unter der handhabbare und kom-

fortabel zu nutzende Produkte verstanden werden. Das bedeutet, dass das Ziel der Weblog-Suchmaschine eine (gegenüber bisherigen Suchmaschinen) verbesserte Interaktion enthalten muss. Der Benutzer muss mit dem System auf einfache und intuitive Weise gemäß seinen Bewegungsabläufen (beispielsweise mittels Mausbewegung) interagieren können.

Um dem Benutzer eine verbesserte Orientierung zu ermöglichen, wird in diesem Ansatz eine Karte der verwandten Begriffe zur Visualisierung der gefundenen Suchresultate angezeigt (vgl. Abschnitt 1.4). Laut Duden bedeutet das Wort Orientierung die »*Ausrichtung, Kenntnis von Weg und Gelände, geistige Einstellung*«. Durch die visuelle Darstellung der Resultate zu Hügeln, die ähnliche Resultate beinhalten, wird dem Benutzer künftig eine intuitivere Orientierung als durch herkömmliche Listen angeboten.

Für den Benutzer besteht die Möglichkeit, direkt im GUI zu interagieren. So kann er die Suche durch Eingabe neuer Begriffe verfeinern, mittels Schieberegler die Zugehörigkeiten erweitern oder verringern oder die geografische Lage durch Drehen der Karte verändern. Dadurch bekommt der Nutzer ein besseres Verständnis für seine Daten und kann unbekannte Verbindungen und Zusammenhänge grafisch erkennen. Dank der Interaktion mit dem Programm erwirbt er Wissen über seine Daten und erkennt Zusammenhänge.

3 Fazit und Ausblick

Obgleich das WWW noch jung ist, kann es schon eine aufregende Historie aufweisen. Mit dem Aufkeimen von sozialer Software wurden Möglichkeiten erschlossen, mit anderen digital in Kontakt zu treten. Auf diese Art organisiert sich der Mensch mittels Weblogs und Folksonomies seine Welt, verbindet sich, tauscht sich aus.

Fest steht, dass die Suche nach unsichtbaren Informationen und Zusammenhängen noch in den Kinderschuhen steckt. Für die methodische Informationssuche in Weblogs gibt es im Web 2.0 spezielle Weblog-Suchmaschinen und Auswertungsservices. Der Schritt hin zum semantischen Web führt jedoch zu einer tief greifenden Veränderung in der Art der Informationsbeschaffung und damit zu ungeahnten Möglichkeiten.

Der hier präsentierte Ansatz ermöglicht nicht nur eine exakte Suche nach Informationen, sondern ein vages Suchen nach themenverwandten und vermutlich relevanteren Informationen. Dazu wird der Ansatz gegenwärtig als Dissertation am Forschungszentrum FM der Universität Fribourg (*www.FMsquare.org*) als Prototyp implementiert. Das Forschungszentrum FM wendet hierbei die Idee der unscharfen Klassifizierung auf verschiedene Anwendungsfelder an.

Um themenverwandte Begriffe zu erkennen, greift die präsentierte Metasuchmaschine auf Folksonomies zurück, wo Internetnutzer beispielsweise durch das Beschreiben von Fotos mittels Schlagwörtern dem Computersystem Assoziationen zwischen den beschriebenen Begriffen auf dem Foto beibringen. Diese Assoziationen können mit dem vorgeschlagenen Ansatz erkannt und (als Hilfe für die Suche) in der Ontologie hinterlegt werden. Die kartografische Aufbereitung hilft dem Benutzer, sich innerhalb der gefundenen Suchbegriffe zurechtzufinden, und erlaubt ihm zugleich, tiefer in eine Thematik einzutauchen oder Beziehungen zwischen Objekten zu erkennen.

Die Zukunft wird zeigen, wie im semantischen Web nach versteckten Informationen in Weblogs gesucht werden kann. Weblog-Suchmaschinen müssen sich kontinuierlich zu funktionalen Programmen weiterentwickeln. So wird es immer wichtiger, dass der Suchende bei seiner Recherche maximal unterstützt wird. Dazu gehören ergonomische und auf den Benutzer abstimmbare Benutzeroberflächen (individuelle Suchmaske, optimierte, an den Bewegungsablauf angepasste Eingabemöglichkeiten usw.), wie auch Hilfen bei einer vertieften

Suche. Wünschenswert wäre, wenn Suchmaschinen automatisch erkennen könnten, in welche Richtung sich die Suche entwickelt. Dadurch könnten Daten zur Verfügung gestellt werden, ohne dass diese zusätzlich manuell gesucht werden müssen. Diese Daten könnten bei Bedarf angeschaut werden.

Als letzter Punkt seien Suchanfragen in deutscher Sprache erwähnt, die englische, spanische oder französische Resultate liefern. Als Schlüssel hierzu könnte die bereits erläuterte Ontologie herangezogen werden. Wenn Tags aller Sprachen zu einer Ontologie aufgearbeitet werden, könnte dieses Ziel erreicht werden.

4 Literatur

[Baeza-Yates & Ribeiro-Neto 1999] *Baeza-Yates, R.; Ribeiro-Neto, B.:* Modern Information Retrieval. Addison-Wesley, Essex, 1999.

[Bezdek et al.1999] *Bezdek, J. C.; Keller, J.; Krisnapuram, R.; Pal, N. R.:* Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. Springer-Verlag, New York, 1999.

[Borg & Groenen 2005] *Borg, I.; Groenen, P. J. F.:* Modern multidimensional scaling: Theory and Applications. Springer-Verlag, New York, 2005.

[Gruber 1993] *Gruber, T.:* A translation approach to portable ontology specifications. In: Knowledge Acquisition, 5 (2), 1993, S. 199-220.

[Hermann & Leuthold 2003] *Hermann, M.; Leuthold, H.:* Atlas der politischen Landschaften der Schweiz. vdf Hochschulverlag AG, ETH Zürich, 2003.

[Kuhn et al. 2009] *Kuhn, A.; Erni, D.; Loretan, P.; Nierstrasz, O.:* Software Cartography: Thematic Software Visualization with Consistent Layout. In: Journal of Software Maintenance and Evolution, John Wiley & Sons, 2009.

[O'Reilly & Battelle 2009] *O'Reilly, T.; Battelle, J.:* Web Squared: Web 2.0 Five Years On, *http://assets.en.oreilly.com/1/event/28web2009_websquared-whitepaper.pdf*; Zugriff am 21.08.2009.

[Portmann & Meier 2010] *Portmann, E.; Meier, A.:* A Fuzzy Grassroots Ontology for improving Weblog Extraction. In: Journal of Digital Information Management, Digital Information Research Foundation, Chennai, India, 2010.

[Surowiecki 2004] *Surowiecki, T.:* The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations. Doubleday, New York, 2004.

[Vander Wal 2007] *Vander Wal, T.:* Folksonomy, Folksonomy Coinage and Definition, *http://vanderwal.net/folksonomy.html;* Zugriff am 20.08.2009.

[Ware 2000] *Ware, C.:* Information Visualization. Morgan Kaufmann, San Francisco, 2000.

[Wise 1999] *Wise, J. A.:* The ecological approach to text visualization, *www.geog.ucsb.edu/~sara/teaching/geo234_02/papers/wise.pdf*; Zugriff am 17.09.09.

Dipl.-Wirtsch.-Inf. (FH)
Edy Portmann MSc
Universität Fribourg
Departement für Informatik
Boulevard de Pérolles 90
CH-1700 Fribourg
edy.portmann@unifr.ch
http://diuf.unifr.ch/is/

Adrian Kuhn MSc
Universität Bern
Institut für Angewandte
Mathematik und Informatik
Neubrückstr. 10
CH-3012 Bern
akuhn@iam.unibe.ch
http://scg.unibe.ch