Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Information Retrieval

Grundlage für Journalismus im Web 2.0

Nina Hälker

Hausarbeit zur Ringvorlesung
Content & Technology

Wintersemester 2013/14

Fakultät Technik und Informatik
Department Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Inhalt

Einleitung

Die vorliegende Arbeit befasst sich mit dem Thema Information Retrieval (IR). Dargestellt werden verschiedene Methoden und Möglichkeiten des IR vor dem Hintergrund der sich stark transformierenden Medienlandschaft.
Einem kurzen Einblick in die sich wandelnde Medienlandschaft und den neuen Aufgaben, denen Journalisten gegenüber stehen, folgt ein Abschnitt mit Definitionen, um von dort aus überzuleiten zu den Anwendungsbereichen und Verfahren des Information Retrieval.
Die Arbeit schließt ab mit einem Überblick über die Konferenzen, auf denen das Thema weiterhin diskutiert werden wird und stellt zu guter Letzt Fragen hinsichtlich der noch zu bewältigenden Aufgaben von Information Retrieval für eine Unterstützung der journalistischen Arbeit in Zeiten des Web 2.0.

1. Die Medien und der Journalismus

Verfahren, die im Information Retrieval derzeit Anwendung finden, werden in dieser Arbeit in Bezug auf die Erschließung von Datenbanken und dem World Wide Web thematisiert. Aus diesem Grund widmet sich der erste Teil der Arbeit dem Thema Journalismus und den Erosionen, die seit einigen Jahren für Journalisten zu spüren sind.

1.1 Informationen und ihre Vermehrung

Wir leben in einer Zeit explodierender Datenmengen. Dem amerikanischen Suchspezialisten Qmee zufolge werden innerhalb von 60 Sekunden zwei Millionen Suchanfragen bei Google gestellt, 1,8 Millionen Facebook-Likes geklickt, 287.000 Kurznachrichten über Twitter verschickt und 72 Stunden Videomaterial bei Youtube hochgeladen – um nur einen Teil des Wachstums zu benennen.[1] Im Jahr 2010 sagte der damalige CEO von Google, dass sich das weltweite Datenvolumen alle zwei Jahre verdoppele.[2] Die Konsequenz daraus ist, dass es minütlich unmöglicher wird, einen Überblick über die im World Wide Web vorhandenen Datenmengen zu behalten.

[1] Nina Maaßen: 60 Sekunden Internet. 600 neue Websites, 200 Millionen Emails. Bayerischer Rundfunk, 30.7.2013. Online unter: http://www.br.de/puls/themen/welt/morningstory-60-sekunden-internet100.html (Stand: 20.2.2014).
[2] Michael Seemann: Das neue Spiel – Prism vs. Kontrollverlust. In: Spex Nr. 347, September 2013. Online unter: http://www.spex.de/2013/09/19/das-neue-spiel-prism-vs-kontrollverlust (Stand: 20.2.2014).

1.2 Journalisten in der Neuorientierung

Michael Praetorius, Publizist und Medienberater, äußerte sich beim *Forum Lokaljournalismus* der *Bundeszentrale für politische Bildung* im Januar 2014 zu den Herausforderungen im Medienbereich wie folgt: „Die Meldung an sich ist heute nichts mehr wert. Journalisten müssen Informationen aus unterschiedlichsten Quellen aufnehmen, aufbereiten, einschätzen, kuratieren, bewerten, verifizieren und das Ergebnis über viele Kanäle hinweg verbreiten.“ Dadurch, dass Informationen nicht mehr das Privileg von Nachrichtenagenturen, Zeitungen und Journalisten sind, sondern aus verschiedensten Medienkanälen stammen, ist das Prüfen von Informationen ein wichtiger Bestandteil der journalistischen Arbeit. Die Aufgabe von Journalisten ist es demnach, Themen zu setzen und in kürzester Zeit aus verlässlichen Informationen und belegten Quellen Berichte zu kreieren – und dies alles, ohne in der Informationsflut unterzugehen. Damit Journalisten in und mit der vorhandenen Informationsflut arbeiten können, müssen sie Tools benutzen können, mittels derer es ihnen möglich ist, relevante von irrelevanten Daten zu trennen, und die Daten im besten Fall bereits einer Vorab-Aufbereitung zu unterziehen, so dass aus Daten Informationen und aus Informationen Geschichten werden.

Es handelt sich demnach um die Faktoren Masse (Datenexplosion), Unübersichtlichkeit (vielfältige Kanäle, Archive, …) und Zeit (vermeintlich schnellerer Wertverfall von Nachrichten), die aufgrund der Technologisierungs- und Digitalisierungsprozesse des vergangenen Jahrzehnts handhabbarer gemacht werden müssen.

2 Definitionen

2.1 Information Retrieval

Information Retrieval (IR) befasst sich mit „computergestütztem Suchen nach komplexen Inhalten, [um] *bestehende* Informationen aufzufinden“.[3] Ziel ist dabei nicht, neue Strukturen zu entdecken, wie es im Bereich Knowledge Discovery (zu dem Data-Mining und Text Mining gehören) der Fall ist, sondern das Auffindbar- und Zugänglichmachen von Daten, Informationen und Wissen, die ohne IR nicht abrufbar wären.

Calvin Mooers, der als Erfinder des Begriffs Information Retrieval bezeichnet wird, definierte IR im Jahr 1951 als „the process or method whereby a prospective user of information is able to convert his need for information into an actual list of citations to documents in storage containing information useful to him. It is the finding or discovery process with respect to stored

[3] Vgl., auch im folgenden Satz, die Definition bei Wikipedia: http://de.wikipedia.org/wiki/Information_retrieval (Stand: 20.2.2014).

information“.[4] Information Retrieval unterstützt die Dokumentation und Organisation von Wissen. Die Fachgruppe IR der Gesellschaft für Informatik kommt in ihrer Definition auf die Unterschiede Wissensproduzent versus Wissensnachfragender zu sprechen, wenn sie schreibt: „Im Information Retrieval (IR) werden Informationssysteme in Bezug auf ihre Rolle im Prozeß des Wissenstransfers vom menschlichen Wissensproduzenten zum Informations-Nachfragenden betrachtet“ [5] – bezogen also auf das Verhältnis zwischen Sender und Empfänger.
Die Special Interest Group Information Retrieval der British Computer Society notiert, IR „is concerned with enabling people to locate useful information in large, relatively unstructured, computer-accessible archives. Much of IR research is aimed at finding ways to represent the information needs of users and to match these with the contents of an archive. In many cases, those information needs will be best met by locating suitable text documents, but in other cases it may require retrieval of other media, such as video, audio, and images“.[6]
Lewandowski weist auf die Unterschiede zwischen dem „klassischen“ und dem „Web Information Retrieval“ hin, die „hinsichtlich des zugrunde liegenden Dokumentenkorpus, hinsichtlich der Inhalte, der Nutzer und hinsichtlich der Eigenarten des IR-Systems selbst“ bestehen[7] – der Korpus ist bei letzterem unübersichtlicher, die Inhalte vielfältiger, die Nutzer unbekannter und die Algorithmen vielfältiger und unsichtbarer für den Nutzer.

2.2 Datum, Wissen, Information

Bevor die Anwendungsbereiche des Information Retrieval dargestellt werden, soll kurz die Unterscheidung zwischen *Datum*, *Wissen* und *Information* vorgenommen werden.
„Daten sind 'Einträge' mit einer bekannten syntaktischen Struktur.“[8] In der deutschen Wikipedia werden sie als „(maschinen-) lesbare und -bearbeitbare, in der Regel digitale Repräsentation von Information“ definiert, deren Inhalt „meist zunächst in Zeichen bzw. Zeichenketten kodiert (wird), deren Aufbau strengen Regeln folgt, der sogenannten Syntax“.[9] Die Syntax ist für IR-Methoden deswegen von Bedeutung, weil Daten gleichen Formats in gleicher Weise bzw. mit gleicher Methode gesucht werden können. XML-Dateien sind anders zu durchsuchen als Word-Dateien und Bild- oder Video-Dateien. Ein Datum als einzelnes Element hat, an irgendeiner Stelle in einer Datenbank abgelegt, (noch) keinerlei Wert. Ihm fehlt zur Einordnung die Semantik (Bedeutung) und die Pragmatik (Kontext). Durch Wissen werden Daten semantisch

[4] Zitiert nach: http://en.wikiquote.org/wiki/Calvin_Mooers (Stand: 20.2.2014).
[5] http://fg-retrieval.gi.de/startseite/information-ueber-die-fachgruppe.html (Stand: 20.2.2014).
[6] British Computer Society Information Retrieval Specialist Group: Constitution. Online unter: http://irsg.bcs.org/irsgdocs/IRSGconstitution.pdf (Stand: 20.2.2014).
[7] Dirk Lewandowski: Web Information Retrieval. Technologien zur Informationssuche im Internet. DGI-Schrift Informationswissenschaft 7, Frankfurt am Main 2005, S. 71.
[8] Thomas Gottron: Information Retrieval, Vorlesungsskript Sommersemester 2010, S.7.
[9] http://de.wikipedia.org/wiki/Daten#Daten_in_der_Informatik (Stand: 20.2.2014).

ergänzt und zu einer Art „Faktensammlung“.[10]

Über die Hierarchisierung und das Verhältnis von Wissen und Information gibt es unterschiedliche Ansichten: Während in der Wikipedia Wissen als „vernetzte Information“ definiert wird,[11] beschreibt Gottron Informationen als „Teilmenge des Wissens“ bzw. „nutzbares Wissen“.[12] Wissen wird durch Informationen um die Pragmatik, den Kontext, ergänzt. Systeme des Information Retrieval unterstützen die Suche nach Informationen und verwenden semantische und pragmatische Aspekte, um Daten zu kategorisieren und schaffen damit die Strukturgrundlage, um Daten schließlich durch gezielte Abfragen in sinnvoller Weise auffindbar zu machen.

3. Anwendungsbereiche

Die Anwendungsbereiche von IR-Methoden haben sich, nicht zuletzt durch das World Wide Web, in den vergangenen Jahren deutlich erweitert: „Während in der Vergangenheit vornehmlich große Firmen und Institutionen Anwender von IR-Systemen waren, ergibt sich heute vor allem durch das Internet, aber auch in Organisationen bis hin zum privaten Bereich ein wesentlich größeres Anwendungspotential, das nicht nur durch immer größere Datenmengen, sondern auch durch Vielfalt hinsichtlich Struktur und verwendete Medien gekennzeichnet ist und zudem stark durch soziale Interaktionen geprägt ist.“[13] Information Retrieval findet mittlerweile in unterschiedlichen Bereichen Anwendung – von klassischen Bereichen wie Bibliotheks- und Experteninformationssystemen über Datenbanken in großen Unternehmen bis hin zu Suchmaschinen im Internet und Desktop-Suchsystemen.[14] Die Suche nach Dokumenten und anderen Dateien kann sehr allgemein, aber auch sehr speziell sein – abhängig davon, woraus der Datenbestand besteht und auf welche Weise er strukturiert und damit zugänglich gemacht worden ist.

Methoden des Information Retrieval unterstützen nicht erst seit dem Aufkommen des World Wide Web bei der Suche nach Informationen. Sie sind lediglich seit der Informationsexplosion, die mit dem World Wide Web und den Entwicklungen in puncto Speicherkapazitäten einhergeht, zu einem größeren Forschungsfeld geworden und haben zu einer deutlich breiteren Anwendung geführt. Das Interesse, Methoden zu entwickeln, die in der Lage sind, in kürzester Zeit riesige Datenmassen nach bestimmten Informationen zu durchsuchen, ist in den vergangenen 10 Jahren deutlich gewachsen. Folge dessen ist nicht nur ein Wachsen des Forschungsfelds des IR,

[10] Gottron: Information Retrieval. 2010, S.7.
[11] http://de.wikipedia.org/wiki/Wissen (Stand: 20.2.2014).
[12] Gottron: Information Retrieval. 2010, S.7.
[13] http://fg-retrieval.gi.de/startseite/information-ueber-die-fachgruppe.html (Stand: 20.2.2014).
[14] Burkhardt Renz: Datenbanken und Informationssysteme. Information Retrieval – Konzepte und Beispiele, 2013. http://homepages.thm.de/~hg11260/mat/dis-ir-bh.pdf (Stand: 20.2.2014).

sondern ebenfalls Bemühungen, Verfahren zum Data Mining, dem Suchen nach Mustern in großen, unstrukturierten Datenmengen zu entwickeln, die jedoch nicht Thema dieser Arbeit sind. Auch die Nutzung von Blogs und Social Media Angeboten stellen „erhöhte Anforderungen an die Qualität und die Funktionalität von IR-Systemen, die nur durch den stärkeren Einsatz von wissensbasierten, computerlinguistischen und medienspezifischen Erschließungsverfahren erfüllt werden können“ [15] – ebenfalls ein Unterschied zur Suche in vordefinierten Datenbanken. Hinsichtlich der Verbesserung von IR-Verfahren kann aus der Beschreibung der Fachgruppe IR abgeleitet werden, dass im Zentrum weiterhin die Suche bzw. Auffindbarkeit und das Zugänglichmachen von

- großen Datenmengen,
- Daten vielfältiger Struktur und unterschiedlicher Formate,
- Daten unterschiedlicher Medienarten,
- Interaktionsdaten und
- Daten unterschiedlicher Sprachen

stehen.
Diese unterschiedlichen Daten zusammenzubringen, abrufbar und einer schnellen Auswertung zugänglich zu machen, ist weiterhin die Aufgabe von IR, wenngleich in zunehmend komplexerer Art.

3.1 Verfahren

In den Anwendungsbereichen des IR werden verschiedene Verfahren angewendet. Während sich manche Suchen vor allem auf die tatsächlichen Daten beziehen, gibt es komplexere Suchsysteme und Algorithmen, die relationale Suchen mit einbeziehen, sich Ontologien bedienen o. ä.. Im Folgenden werden zunächst zwei für die meisten Verfahren relevante Probleme aufgeführt, um im Anschluss daran einzelne Verfahren kurz zu skizzieren.

3.1.1 Heterogenität und Vagheit – Daten und Nutzer

Im World Wide Web sind bei der Suche nach Informationen – anders als bei der Befragung von Experten und der Recherche in einer Bibliothek bzw. in bereits als geeignet befundener Literatur – der Umgang mit einem deutlich größeren Datenbestand sowie die Heterogenität und die unterschiedliche Qualität der Informationen eine Herausforderung, der IR-Systeme mit qualifizierten und differenzierten Suchmethoden zu begegnen versuchen.[16]
Die Frage, ob eine Information zu einer Suche passt oder nicht, ist von verschiedenen Kriterien

[15] http://fg-retrieval.gi.de/startseite/information-ueber-die-fachgruppe.html (Stand: 20.2.2014).
[16] Gottron: Information Retrieval. 2010, S. 4ff.

abhängig. Zunächst spielt die Relevanz der gefundenen Informationen eine Rolle: Ob die Relevanz einer Information subjektiv, objektiv, situativ oder ob sie systemrelevant ist,[17] sollte im besten Fall bei Suchanfragen in die Berechnung der Treffer einbezogen werden, bedarf aber gegebenenfalls nicht nur einer präzisen Analyse der gefundenen Daten, sondern auch eines spezielleren Wissens über den oder die suchende Person bzw. Instanz. Die Heterogenität bezieht sich also in gewisser Weise nicht nur auf die vorhandenen Daten im World Wide Web, sondern auch auf die Nutzer. Während dem einen die Information eines Wikipediaeintrags genügt, beginnt bei einer anderen Person die Recherche erst bei den in der Wikipedia hinterlegten Quellen.

Ein möglichst detailliertes Wissen ist insofern sowohl auf Seiten der Informationen wie auch auf Seiten der Nutzer von Bedeutung: Je mehr von beiden Seiten bekannt ist und je präziser die Suche daraufhin formuliert wird, desto bessere Treffer sind (angenommen, die Algorithmen funktionieren entsprechend) zu erwarten.

Ein weiteres nutzer- wie datenseitiges Kriterium bzw. Problem, mit dem IR umgehen muss, ist die Vagheit, die für das Information Retrieval ein Problem ist, da sich „nicht immer exakt und situationsunabhängig festmachen lässt, worum es geht."[18] Die Fachgruppe „Information Retrieval" der Gesellschaft für Informatik präzisiert die Definition wie folgt: Vage Anfragen seien „dadurch gekennzeichnet, dass die Antwort a priori nicht eindeutig definiert ist. Hierzu zählen neben Fragen mit unscharfen Kriterien insbesondere auch solche, die nur im Dialog iterativ durch Reformulierung (in Abhängigkeit von den bisherigen Systemantworten) beantwortet werden können; häufig müssen zudem mehrere Datenbasen zur Beantwortung einer einzelnen Anfrage durchsucht werden."[19] Während die Fachgruppe dieses Problem der Unsicherheit bzw. Unvollständigkeit vor allem auf der Seite der Daten sieht und vertritt, dass es „meist aus der begrenzten Repräsentation von dessen Semantik (z.B. bei Texten oder multimedialen Dokumenten)"[20] herrührt, möchte ich behaupten, dass eine ähnliche Art der Unsicherheit vermutlich auch auf der Anwender- bzw. Nutzerseite zu beobachten ist. Während es lange Zeit üblich war, die Kategorisierung von Daten und die Rezeption selbiger auf unterschiedlichen Seiten zu verorten, kann mittlerweile davon ausgegangen werden, dass sich diese beiden Seiten überschneiden. Insbesondere bei Microblogging-Angeboten sind die Rezipienten der Informationen oftmals gleichzeitig auch Produzent_innen der Informationen – was für die weitere Arbeit bedeutet, dass sich die Begrifflichkeiten der Verschlagwortung und anderer IR-Methoden den potenziellen Suchanfragen annähern könnten. Wie und ob sich das auf die genannte Vagheit und Unsicherheit auswirken wird, könnte das Thema einer weiteren Arbeit sein.

[17] Zur Unterscheidung der Relvanzkriterien s. Gottron, ebd., S. 6.
[18] Gottron: Information Retrieval. 2010, S. 7.
[19] http://fg-retrieval.gi.de/startseite/information-ueber-die-fachgruppe.html (Stand: 20.2.2014).
[20] http://fg-retrieval.gi.de/startseite/information-ueber-die-fachgruppe.html (Stand: 20.2.2014).

Fehlerhafte Antworten bzw. ungenügende Angaben über den Inhalt von Dokumenten kommen beispielsweise durch Wörter mit unterschiedlichen Bedeutungen zustande (Bank: Geldinstitut und Sitzgelegenheit) oder bei Begriffen, die das gleiche beschreiben (Bank und Geldinstitut). Sowohl der/die Nutzer_in einer Datenbank als auch die Klassifizierung eines Dokuments nutzen insofern potenziell unpräzise Begriffe und schränken damit ungewollt die Zuordnungsmöglichkeiten ein. Konkretisieren lässt sich dieses Problem bei einem aktuellen Projekt, Classifying Microblogs For Disasters, das beabsichtigt, in Katastrophensituationen möglichst effektiv Informationen aus Microbloggingdiensten auszulesen, um so schnellere Hilfe organisieren zu können. Im Paper zum Projekt wird auf die Schwierigkeit hingewiesen, fehlerhafte, ironische oder in anderer Weise für das Katastrophengeschehen irrelevante Tweets durch Textanalyse herauszufiltern.[21] So verweisen die Autoren in ihrer Testung auf die hohe Anzahl an false positives, z.B. wenn von emotionalen Erdbeben die Rede ist, aber die Textanalyse „echte“ Erdbeben sucht.

Auch das Projekt GeoVisNews befasst sich mit den durch die Vagheit von Informationen bedingten Einschränkungen bzw. Schwierigkeiten für die Entwicklung von verlässlichen IR-Verfahren.22 Das Projekt befasst sich mit Geolokalisierung, für die zunächst jeder potenzielle Ort auf der Seite Geonames gesucht wird, um so die tatsächliche Geolocation des gesuchten Ortes herauszufinden und die Festlegung falscher Verortungen zu reduzieren. Dabei ist das Problem der Vagheit komplexer als bisher beschrieben: Die Probleme von GeoVisNews liegen darin, dass einerseits viele der in einem Dokument genannten Orte in keiner Weise von Relevanz sind und andererseits die für ein Dokument bedeutenden Orte oftmals gar nicht genannt werden: „For example, considering a document that describes the wedding of a celebrity, the document may contain some locations about the bride, such as where she was born and the hometown name of the bride, but these locations are not relevant enough in comparison with the location where the wedding is hold. The second fact is that several relevant locations may not appear in the documents.“[23] Die Schwierigkeit besteht insofern in der Bewertung der genannten sowie der möglichen Orte.

3.1.2 Teilgebiete des IR

Ein Teilgebiet des IR befasst sich mit der Suche in Dokumenten – die „Form und Art der Dokumente variiert dabei sehr stark“.[24] Bis heute handelt es sich dabei zumeist um

[21] Sarvnaz Karimi, Jie Yin, Cecile Paris, Classifying Microblogs For Disasters, ADCS '13: Proceedings of the 18th Australasian Document Computing Symposium, December 2013.
[22] Zechao Li, Jing Liu, Meng Wang, Changsheng Xu und Hanqing Lu: Enhancing News Organization for Convenient Retrieval and Browsing, in: Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP), Volume 10 Issue 1, December 2013.
[23] Ebd., S. 6.
[24] Gottron: Information Retrieval. 2010, S. 12f.

Textdokumente, wenngleich zunehmend häufiger auch andere Dokumentenarten darin enthalten sind, wie z.B. Abbildungen. Weitere Teilgebiete sind das sich aus dem *Text-IR* entwickelte *Hypertext-IR* bzw. *IR im Web*, bei dem Hypertexte durchsucht werden, sowie die *Multimedia-IR*, die sich mit der Suche in Bild-, Video- und Audiodateien befasst. Das Gebiet des *Question Answering* arbeitet daran, „nicht nur interessante Dokumente zur Befriedigung eines Informationsbedürfnisses zu finden, sondern daraus direkt die gewünschte Information zu extrahieren.“[25]

3.1.3 Aufgaben des IR

Bei den Aufgaben und Fragestellungen, mit denen sich die IR beschäftigt, sind zunächst die *Ad Hoc Anfragen* zu nennen, die die klassischste Aufgabe darstellen: Die Dokumentenmenge, in der gesucht wird, ist fix, eine Anfrage führt zu einer Ausgabe der relevanten Treffer. Weitere Aufgaben sind *Klassifikation* und *Clusteranalyse*: Während bei der Klassifikation feste Kategorien den Dokumenten zugeordnet werden, sind Cluster dazu da, Dokumente nach Ähnlichkeiten zu gruppieren. Clusteranalysen haben sich in unterschiedlicher Weise ausdifferenziert. Eine Möglichkeit ist beispielsweise die Arbeit mit sogenannten Tag-Clustern, einer semantischen Suche in Tag-Clouds, die bei unspezifischen Suchanfragen die weitere Navigation erleichtern und durch die Darstellung der Cluster in Form einer Tag-Cloud in der Lage sind, dem Nutzer einen schnellen Überblick über den Inhalt zu vermitteln.[26] Das Beispiel der Tag-Cloud zeigt, dass sich Information Retrieval immer auch mit der Darstellung der Suchergebnisse beschäftigen muss bzw. der Nutzwert der Suchergebnisse auch von der Art der Darstellung abhängt. Klassifikationsmethoden kombiniert mit Clusteranalysen kommen u.a. bei Zeitungsarchiven zum Einsatz, so beispielsweise im Pressearchiv der Süddeutschen Zeitung und einiger weiterer Zeitungen, welches seit 2004 unter dem Namen „Medienport“ über alle Quellen hinweg mittels Volltextsuche (kostenpflichtig) durchsuchbar ist. Zusätzlich gibt es die Möglichkeit, die Treffer über die Eingabe der Quelle, des Datums, des Ressorts oder der Rubrik weiter einzugrenzen oder über die Suche nach Sachthemen, Personen oder Firmen auf eine Dossierstruktur als zentralem Erschließungsobjekt zuzugreifen.[27] Die Dossiers sind untereinander verlinkt, die Verlinkung der Artikel erfolgt aufbauend auf die aktuelle Berichterstattung und hat dadurch eine dynamische Struktur. Nach einer automatischen Analyse und der Zuweisung von Klassifizierungsvorschlägen für jede Textdatei folgt in einem nächsten Schritt eine Clusteranalyse, bei der die Vorschläge „innerhalb ihres systematischen Umfelds mit allen horizontal und vertikal nahe verbundenen Dossiers analysiert“ werden, um dann die

[25] Ebd.
[26] Kathrin Knautz: Tag-Cluster – Semantische Suche in Tag-Clouds, in: B.I.T. Online 13(2010), Nr. 3.
[27] Im folgenden Absatz beziehe ich mich auf einen Artikel von Markus Schek: Automatische Klassifizierung und Visualisierung im Archiv der Süddeutschen Zeitung. Praxisforum 1/2005, S. 20ff.

Vorschläge, die aus unverbundenen Themenclustern stammen, aus der Vorschlagsliste zu löschen. Die verbliebenen Ergebnisse der Klassifizierung und der Analyse werden von den Dokumentaren angenommen, ergänzt oder verworfen. Andere Aufgaben von IR sind *Cross Language IR*, die Dokumente unterschiedlicher Sprachen durchsucht und Suchanfragen sowie Texte übersetzt, aber auch *Duplikat-* bzw. *Spamerkennung*, also Methoden, mittels derer Duplikate und irrelevante Spam-Dokumente gefunden und gemeldet werden können.

4. Ausblick

4.1 IR-Konferenzen

Die zumeist als Fachgruppen der jeweiligen nationalen Vereinigungen für Informatik organisierten Special Interest Groups Information Retrieval halten regelmäßig Konferenzen, Tagungen und Workshops ab. Neben der *Fachgruppe Information Retrieval in der Gesellschaft für Informatik* und der *Information Retrieval Specialist Group der British Computer Society (BCS IRSG)* ist in diesem Zusammenhang auch die *Special Interest Group der Association for Computing Machinery (SIGIR* der *ACM)* als größtem internationalem Forum zu nennen. Die Konferenzen sind ein wichtiger Ort, an dem neue Forschungen präsentiert und diskutiert werden.[28] Die Themen der Konferenzen reichen von speziellen Themen, wie z.B. Papers zu den Herausforderungen der Real Time Search im Newsbereich und bei Twitter über Cross-Language Search bis zu Papers über die Prognostizierbarkeit des Schwierigkeitsgrads spezieller Suchanfragen.[29] Mit der *ECIR 2014* findet in diesem Jahr in Amsterdam die *36. Europäische Konferenz zu Information Retrieval* statt. Ihrer Selbstbeschreibung zufolge handelt es sich dabei um das „main European forum for the presentation of new research results in the field of Information Retrieval“.[30] Die *Information Retrieval Specialist Group der British Computer Society (BCS IRSG)* bezeichnet die Konferenz als Forum für junge Forschende „with a vision to: provide a supportive forum for which their research can be presented, discussed and debated; provide feedback from seasoned Information Retrieval researchers; and facilitate a large student audience as well as established researchers by ensuring that the costs are minimal.“[31]
Weitere Konferenzen zu IR, die jährlich stattfinden, sind die International ACM/SIGIR Conference on Research & Development, die Joint Conference on Digital Libraries, die Conference on Information & Knowledge Management und die Conference on Web Search and Data Mining. Letztgenannte findet Ende Februar in New York statt. Die Schwerpunktsetzung der Konferenz ist hier, gemäß dem Titel eine, die IR-Methoden in Bezug auf Möglichkeiten des

[28] http://dl.acm.org/event.cfm?id=RE160 (Stand: 20.2.2014).
[29] http://ecir2014.org/ (Stand: 20.2.2014).
[30] http://ecir2014.org/ (Stand: 20.2.2014).
[31] http://irsg.bcs.org/ (Stand: 20.2.2014).

Data Mining betrachtet. Die Keynote wird gehalten über Data that Matter: Opportunities in Crisis Informatics Research.[32]

4.2 Offene Fragen und zukünftige Anforderungen an IR

Die hier dargestellten Bereiche des Information Retrieval skizzieren lediglich einzelne Facetten eines Felds, das um ein Vielfaches größer und komplexer ist. Eine weitergehende Untersuchung könnte sich detaillierter mit dem Problem der Vagheit, auf das ich in dieser Arbeit an verschiedenen Stellen zu sprechen gekommen bin, auseinandersetzen. Zu fragen wäre, welche Arten des Umgangs es bis heute damit gibt, um dann Analysen bzw. Vergleiche verschiedener Methoden des Information Retrieval vorzunehmen und so die Vor- und Nachteile für unterschiedliche Bereiche zu präzisieren. Offen bleibt nach wie vor die Frage, wie eine eindeutige Erkennbarkeit von Dokumenten ermöglicht werden kann. Wolfgang Hesse fragt in einem Beitrag zum Thema Ontologien diesbezüglich: „Können Ressourcen klar und eindeutig klassifiziert werden, z.B. in Dokumente, Daten, Metadaten, physische und virtuelle Aktoren (actors), physische Einheiten?“.[33] Um diese Frage beantworten zu können, muss zum einen nach weiteren Möglichkeiten des „Wie“ der Erschließbarkeit von Daten geforscht werden (das in seiner Frage nur implizit mitschwingt), zum anderen wäre an dieser Stelle zu fragen, inwieweit evtl. ein in dieser Arbeit bereits kurz erwähnter Aspekt von Bedeutung sein könnte: die sich eventuell einander näher gekommenen Rollen von Produzenten und Rezipienten durch die Benutzung von Microblogging-Diensten.

Lewandowski weist in einem Beitrag, in dem er die schlechte Erkennbarkeit bzw. schlechte Strukturierung von Dokumenten u.a. in Bezug auf die Schwierigkeiten bei der Web-Indexierung betont, meiner Ansicht nach bereits indirekt auf diese mögliche Richtung hin, indem er sagt, zwar gäbe es Strukturen in HTML-Dokumenten, die prinzipiell gut ausgelesen werden könnten und aussagekräftig wären, diese würden aber von den Autoren nicht bewusst genutzt.[34]

Untersucht werden könnte demnach, inwieweit ein größeres Wissen durch eine erweiterte Rollenverteilung und eine dialogisch orientierte Nutzung des Web 2.0 eine bessere Indexierung, eine bessere Struktur der Dokumente und eine Reduzierung der Unsicherheitsfaktoren zur Folge haben kann. Hinsichtlich der Frage des Journalismus im Web 2.0. spielen hierbei sicherlich auch Fragen von Autorschaft, Kuration und Dialog eine Rolle. Journalisten, die sich mit IR-Methoden auseinandersetzen, könnte dies perspektivisch eine gute Grundlage für das Vorgehen bei eigenen Recherchen bieten.

---

[32] http://www.wsdm-conference.org/2014/ (Stand: 20.2.2014).

[33] Wolfgang Hesse: Ontologien. In: Informatik Spektrum. Heft 6/2002. Online unter: https://www.gi.de/service/informatiklexikon/detailansicht/article/ontologien.html (Stand: 20.2.2014).

[34] Dirk Lewandowski: Web Information Retrieval. Technologien zur Informationssuche im Internet. DGI-Schrift Informationswissenschaft 7, Frankfurt am Main 2005

5. Literatur

British Computer Society Information Retrieval Specialist Group: Constitution. Online unter: http://irsg.bcs.org/irsgdocs/IRSGconstitution.pdf

Thomas Gottron: Information Retrieval, Vorlesungsskript Sommersemester 2010

Wolfgang Hesse: Ontologien. In: Informatik Spektrum. Heft 6/2002. Online unter: https://www.gi.de/service/informatiklexikon/detailansicht/article/ontologien.html

Sarvnaz Karimi, Jie Yin, Cecile Paris: Classifying Microblogs For Disasters, ADCS '13: Proceedings of the 18th Australasian Document Computing Symposium, Dezember 2013

Kathrin Knautz: Tag-Cluster – Semantische Suche in Tag-Clouds, in: B.I.T. Online 13(2010), Nr. 3

Dirk Lewandowski: Web Information Retrieval. Technologien zur Informationssuche im Internet. DGI-Schrift Informationswissenschaft 7, Frankfurt am Main 2005

Zechao Li, Jing Liu, Meng Wang, Changsheng Xu, Hanqing Lu: Enhancing News Organization for Convenient Retrieval and Browsing, in: Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP) , Volume 10 Issue 1, December 2013

Nina Maaßen: 60 Sekunden Internet. 600 neue Websites, 200 Millionen Emails. Bayerischer Rundfunk, 30.7.2013. Online unter: http://www.br.de/puls/themen/welt/morningstory-60-sekunden-internet100.html

Burkhardt Renz: Datenbanken und Informationssysteme. Information Retrieval – Konzepte und Beispiele, 2013. Online unter: http://homepages.thm.de/~hg11260/mat/dis-ir-bh.pdf

Markus Schek: Automatische Klassifizierung und Visualisierung im Archiv der Süddeutschen Zeitung. Praxisforum 1/2005

Michael Seemann: Das neue Spiel – Prism vs. Kontrollverlust. In: Spex Nr. 347, September 2013. Online unter: http://www.spex.de/2013/09/19/das-neue-spiel-prism-vs-kontrollverlust

Webseiten:

http://dl.acm.org/event.cfm?id=RE160

http://en.wikiquote.org/wiki/Calvin_Mooers

http://ecir2014.org/

http://fg-retrieval.gi.de/startseite/information-ueber-die-fachgruppe.html

http://irsg.bcs.org/

http://de.wikipedia.org/wiki/Daten#Daten_in_der_Informatik

http://de.wikipedia.org/wiki/Information_retrieval

http://de.wikipedia.org/wiki/Wissen