Thomas Risse, Wolfgang Nejdl

Herausforderungen für die nationale, regionale und thematische Webarchivierung und deren Nutzung

Foto: privat

Thomas Risse

Foto: privat

Wolfgang Nejdl

Das World Wide Web ist als weltweites Informations- und Kommunikationsmedium etabliert. Neue Technologien erweitern regelmäßig die Nutzungsformen und erlauben es auch unerfahrenen Nutzern, Inhalte zu publizieren oder an Diskussionen teilzunehmen. Daher wird das Web auch als eine gute Dokumentation der heutigen Gesellschaft angesehen. Aufgrund seiner Dynamik sind die Inhalte des Web vergänglich und neue Technologien und Nutzungsformen stellen regelmäßig neue Herausforderungen an die Sammlung von Webinhalten für die Webarchivierung. Dominierten in den Anfangstagen der Webarchivierung noch statische Seiten, so hat man es heute häufig mit dynamisch generierten Inhalten zu tun, die Informationen aus verschiedenen Quellen integrieren. Neben dem klassischen domainorientieren Webharvesting kann auch ein steigendes Interesse aus verschiedenen Forschungsdisziplinen an thematischen Webkollektionen und deren Nutzung und Exploration beobachtet werden. In diesem Artikel werden einige Herausforderungen und Lösungsansätze für die Sammlung von thematischen und dynamischen Inhalten aus dem Web und den sozialen Medien vorgestellt. Des Weiteren werden aktuelle Probleme der wissenschaftlichen Nutzung diskutiert und gezeigt, wie Webarchive und andere temporale Kollektionen besser durchsucht werden können.

The World Wide Web is well established as a global information and communication medium. New technologies regularly come along which expand the forms of use and permit even inexperienced users to publish content or take part in discussions. For this reason the Web can also be seen as a good documenter of present-day society. The dynamism of the Web means that its content is, by its very nature, transitory, and new technologies and forms of use regularly present new challenges for the collection of web content for web archiving. Static pages still dominated in the early days of web archiving, whereas many dynamic types of content have now arisen which integrate information from different sources. There is now growing interest from various research disciplines in conventional domain-oriented web harvesting, in thematic web collections and in their use and exploration. This article examines a number of challenges and possible methods of collecting thematic and dynamic content from the Web and social media. Current problems which have arisen in academic use are discussed, and it is shown how web archives and other temporal collections can be searched more effectively.

soziale Medien als engmaschiges Netzwerk

Einleitung

Das World Wide Web – kurz das Web – ist für viele Menschen ein fester Bestandteil des täglichen Lebens geworden. Es gestattet, aktuelle Informationen zu verschiedenen Themen abzurufen, aber sich auch selbst in Diskussionen einzubringen (Diskussionsforen, Facebook, Twitter), eigene Arbeiten mit wenig Aufwand zu veröffentlichen (Blogs, Online-Publishing) oder gemeinsam an Projekten zu arbeiten (Open Source Software, Wikipedia). Das Web stellt somit eine unschätzbare Sammlung von Wissen über unsere heutige Gesellschaft dar.

konstanter Wandlungs- und Expansionsprozess

Das Web ist allerdings nicht statisch, sondern befindet sich in einem konstanten Wandlungs- und Expansionsprozess. So waren Ende 2014 fast 16 Millionen Domänen unterhalb der Top-Level-Domäne (TLD) ».de« registriert, und allein im Jahr 2014 gab es ca. 179.000 Neuregistrierungen (vgl. Denic 2015). Während das eine vergleichsweise überschaubare Zahl ist, sieht dies auf der Inhaltsebene ganz anders aus. Nachrichtenseiten veröffentlichen regelmäßig über den Tag neue Artikel und ändern dabei mehrfach ihre Eingangsseite. Kommentarfunktionen erlauben den Nutzern direkt auf Artikel zu reagieren, was jedes Mal zu einer neuen Version der Seite führt. Ntoulas et al. (2004) zeigen, dass das Web um 8 % pro Woche wächst und nach einem Jahr nur noch 40 % der Seiten unverändert sind. Mit dem Erfolg der sozialen Medien hat diese Dynamik weiter zugenommen. Die existierende Bandbreite der sozialen Medien erlaubt es jedem Individuum, sich auch ohne technisches Wissen im Internet zu präsentieren und eigene Artikel, Bilder, Videos etc. zu veröffentlichen. Facebook wird täglich von 890 Millionen Nutzern aktiv genutzt (vgl. Facebook 2015) und auf Tumblr.com sind derzeit über 220 Millionen Blogs registriert (vgl. Tumblr 2015).

Allerdings sollten soziale Medien nicht isoliert betrachten werden, sondern als ein engmaschiges Netzwerk über Anbietergrenzen hinweg. So bieten Blogging-Plattformen wie Wordpress, Blogger oder Tumblr viele Freiräume, um längere und komplexere Artikel zu veröffentlichen. Über Twitter und Facebook wird die Nachricht über einen neuen Artikel an den interessierten Leserkreis verbreitet. Diskussionen werden anschließend, je nach den individuellen Bedürfnissen nach Privatsphäre, auf dem Blog oder auf Facebook geführt. Aber auch Nachrichtenanbieter, Firmen oder öffentliche Einrichtungen nutzen die Kombination verschiedener Medien, um auf sich aufmerksam zu machen und um in direkten Kontakt mit Kunden zu treten. Die aktuelle Konzentration der sozialen Medien auf ein paar wenige und dominierende Anbieter führt auch zur allgemeinen Wahrnehmung, dass das Netz nicht vergisst. Die Anbieter unterstützen diese Annahme durch immer mehr zeitorientierte Navigationsmöglichkeiten, wie etwa die Facebook Timeline, die es erlauben, in die Vergangenheit zu reisen. Allerdings hat die Geschichte der Informationstechnik und des Internets gezeigt, dass nichts konstanter ist als der Wandel: Es wird immer wieder neue Dienst-

anbieter geben, die mit neuen Ideen und Technologien versuchen, einen Teil des Informationsmarktes für sich zu erobern. Während man kurzfristig von einem Fortbestehen der etablierten Anbieter ausgehen kann, sind langfristige Vorhersagen daher nicht möglich. Ein Beispiel ist GeoCities, die schon Mitte der 1990er-Jahre Webhosting für jedermann anboten und 1999 die drittmeist besuchte Website der Welt waren. Zehn Jahre später wurde GeoCities geschlossen. Teile der Inhalte sind heute nur noch dank einiger kurzfristiger Webharvestingaktivitäten erhalten. Auch wenn das kurzfristige Verschwinden etablierter Dienstanbieter recht unwahrscheinlich ist, so können sie dennoch ihre Dienste vom Netz nehmen und damit die Inhalte aus dem Web entfernen. Auch der Autor eines Beitrags hat immer die Möglichkeit, seine eigenen Inhalte zu ändern, zu löschen oder sich vollständig von einem Dienst zurückzuziehen. Sofern es sich um öffentliche Inhalte handelt, können damit interessante Beiträge, Meinungen und Beobachtungen verschwinden.

Ziel der Webarchivierung ist es, Webinhalte für die Dokumentation und spätere Nutzung zu bewahren. Seit der Gründung des Internet Archive in San Francisco im Jahre 1996 ist die Anzahl der weltweiten Webarchivierungsaktivitäten immer weiter angestiegen. Waren am Anfang vor allem Archive und Nationalbibliotheken mit nationalem Sammelauftrag die Hauptakteure, so sind im Laufe der Zeit auch fokussierte Sammlungsaktivitäten einzelner Organisationen hinzugekommen. Des Weiteren gibt es ein steigendes Interesse aus verschiedenen Anwendungsgebieten (Journalismus, Sozialwissenschaften etc.), Webinhalte zu sammeln und zu analysieren. Die immer enger werdende Verzahnung von Web und Social Web zusammen mit neuen Technologien führt zu immer neuen Herausforderungen, die mit existierenden Crawler-Ansätzen nur unzureichend erfüllt werden können. Im folgenden Kapitel wird, ausgehend vom nationalen und regionalen Webharvesting, insbesondere auf die Herausforderungen und Lösungsansätze für das thematische und integrierte Crawling von Web und sozialen Medien eingegangen. Aufgrund des anfänglichen Fokus auf die Sammlung und Bewahrung von Inhalten und auch rechtlicher Hindernisse bzgl. der Nutzung befinden sich die Analyse und die Exploration von Webarchiven noch am Anfang der Entwicklung. Der weitverbreitete Zugriff über die URL auf Archivinhalte mithilfe der (Open) Wayback Machine (vgl. IIPC 2015) und erste Volltextindizes sind noch weit von den Möglichkeiten entfernt, die Websuchmaschinen bieten. Eine besondere Herausforderung ist ferner die zeitliche Dimension und die Verfügbarkeit unterschiedlicher Versionen einer Seite innerhalb des Archivs. Nutzungsmöglichkeiten und eine verbesserte Herangehensweise an die temporale Exploration von Archiven werden im Kapitel »Exploration von Archiven« dieses Beitrags vorgestellt.

Die Herausforderungen für das Webharvesting

Das Ziel des Webharvestings ist die Erstellung von Webarchiven, um die Inhalte des World Wide Web für die Dokumentation und zukünftige Nutzung zu bewahren. Während das Internet Archive in San Francisco prinzipiell zum Ziel hat, das gesamte Web zu bewahren, werden in anderen Webharvestingaktivitäten nur Untermengen des WWW gesammelt. Dieses ist sowohl für domänenorientierte als auch für thematische Crawls der Fall, wie sie in verschiedenen Anwendungsgebieten wie etwa dem Journalismus oder den Sozialwissenschaften gewünscht sind. Der prinzipielle Ablauf ist in Abbildung 1 dargestellt und gliedert sich in die Phasen Crawl-Vorbereitung, Crawl-Durchführung, Crawl-Nachbereitung und die anschließende Nutzung bzw. weitere Bewahrungsaktivitäten. In der Crawl-Vorbereitung geht es insbesondere um die Übersetzung der Crawl-Intention in eine Crawl-Spezifikation, die zur Führung des Webcrawler in der Durchführungsphase genutzt wird. In der Crawl-Nachbereitung geht es um die Erstellung der Archivdateien, die Anreicherung mit Metainformationen und die Katalogisierung, um die Archive für die anschließende Nutzung vorzubereiten.

prinzipieller Ablauf von Crawls

Crawl-Vorbereitung

Die erste Herausforderung ergibt sich durch die Definition der Untermenge des WWW, die gesammelt werden soll, und deren Übersetzung in eine Crawl-Spezifikation. Am Anfang steht die Intention des Nutzers, beispielsweise die Sammlung nationaler oder regionaler Webinhalte oder die Dokumentation eines bestimmten Ereignisses (Bundestagswahl) oder Themas (Menschenrechte).

Intention des Nutzers

Die Beschreibung von Crawl-Intentionen mithilfe von Domänen ist insbesondere bei Nationalarchiven und Nationalbibliotheken, die mit der Bewahrung des nationalen Web beauftragt sind, die übliche Methodik. Für das deutsche Web wäre der Ausgangspunkt die Top-Level-Domain (TLD) ».de«. Mit der Erweiterung der TLD-Systems kommen auch thematische (z. B. ».haus«, ».reise«), regionale (z. B. ».cologne«, ».bayern«) und kommerzielle (z. B. ».bmw«, ».tui«) Domainen hinzu. Um eine nationale Websammlung beispielsweise für Deutschland zu erstellen, wird man mit der Nutzung der deutschen TLDs bereits eine sehr umfassen-

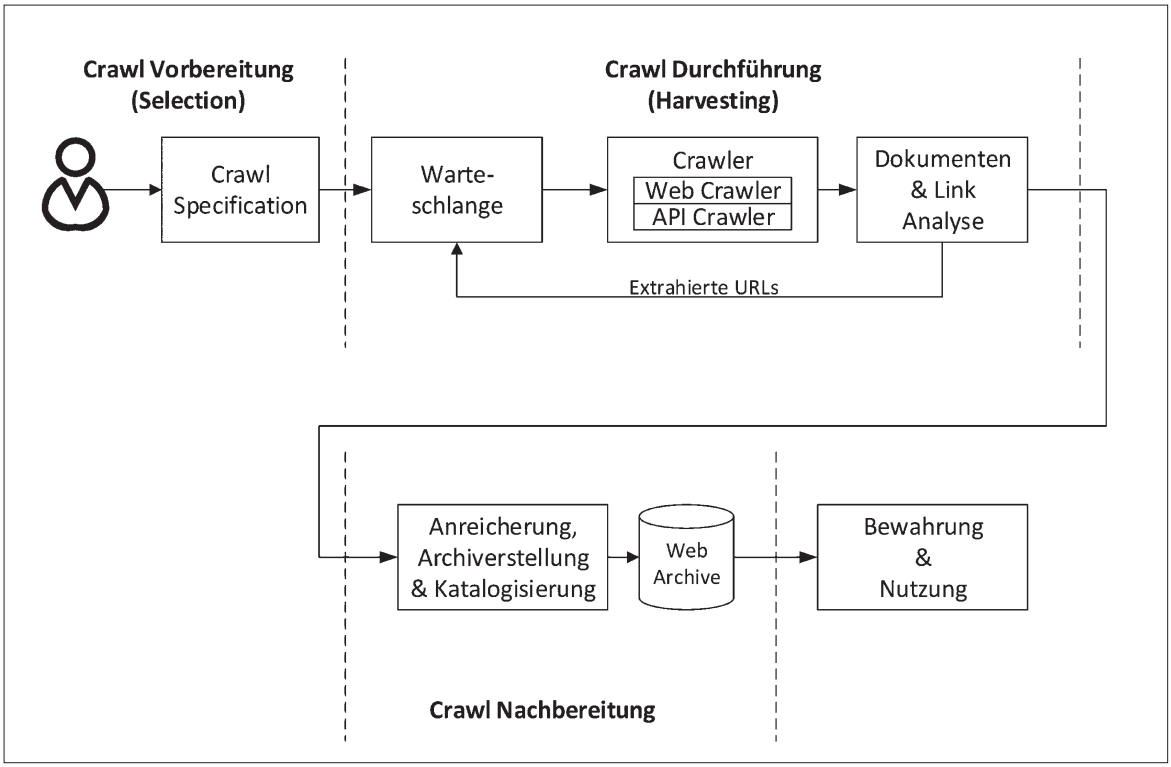

Foto: Risse / Nejdl

Abb. 1: Prinzip eines Web Archive Crawlers

de Sammlung erhalten. Allerdings ist eine zu strikte Fokussierung auf die TLD nicht unbedingt wünschenswert, da einzelne Seiten wiederum Inhalte von anderen Diensten einbinden, die nicht notwendigerweise in ein statisches Domänen-Schema passen (s.a. Abschnitte »Webcrawling und soziale Medien« und »Dynamische und interaktive Inhalte«).

Schlüsselwörter und explizite Entitäten

Während man für nationale Websammlungen mit dem Domänen-Schema gute Ergebnisse erzielt, sieht es auf der regionalen Ebene ganz anders aus. Dort ist es nicht ausreichend, nur die regionale Top-Level-Domäne zu nutzen, da (1) nicht alle Regionen eigene TLDs besitzen und (2) viele Inhalte unter der nationalen oder internationalen Domain abgelegt sind. Für eine umfassende Sammlung von regionalen Inhalten ist die explizite Auswahl von Websites unerlässlich. Eine Alternative für das regionale Webharvesting wäre, den Domänen-Ansatz mit einem themenorientierten Ansatz zu verbinden.

Themenorientiertes Webharvesting

semantische Beschreibung

Das themenorientierte Webharvesting hat zum Ziel, die Darstellung eines Themas oder Ereignisses im Web zu dokumentieren. Gerade bei Gedächtniseinrichtungen ist die manuell-intellektuelle Definition der Websites für das Harvesting nach wie vor die Methode der Wahl. Ein alternativer Ansatz ist anstelle einer langen Liste von URLs eine semantische Beschreibung der gewünschten Inhalte zu verwenden. In diesem Fall entscheidet der Crawler nicht auf Basis des Domänen-Namens, ob einer URL gefolgt wird, sondern anhand der inhaltlichen Überlappung einer Seite mit dem vorgegebenen Thema. Die Herausforderung ist auch hier die Beschreibung der Inhalte, die im Webarchiv abgelegt werden sollen. Der im Rahmen des europäischen ARCOMEM-Projektes entwickelte Web Archive Crawler (vgl. Risse et al. 2014) verwendet Schlüsselwörter und explizite Entitäten zur Beschreibung der inhaltlichen Intention. Um einen Crawl über die Region Hannover zu erstellen, wäre die Entität »Hannover« ein Teil der Beschreibung. Bei der alleinigen Verwendung von »Hannover« würde aber schon die Homepage der Nachbarstadt Laatzen nicht mehr Bestandteil des Crawls sein, da der Suchbegriff erst auf einer Unterseite genannt wird. Wenn der Crawler über die Hauptseite der Stadt Laatzen einsteigt, würde er diese als nicht relevant einstufen, da der Suchbegriff (Hannover) dort nicht vorhanden ist. Das umgekehrte Problem tritt auf, wenn der Crawler die Wikipedia-Seite von Hannover analysiert. Dort wird er auch die Disambiguierungsseite zu dem Begriff Hannover aufrufen und dort eine Reihe von Städten in der Welt finden, die den gleichen Namen in englischer Schreibweise haben, aber irrelevant bezüglich der Intention des Crawls sind. Das Hinzufügen weiterer Begriffe erlaubt eine exaktere Beschreibung der gewünschten

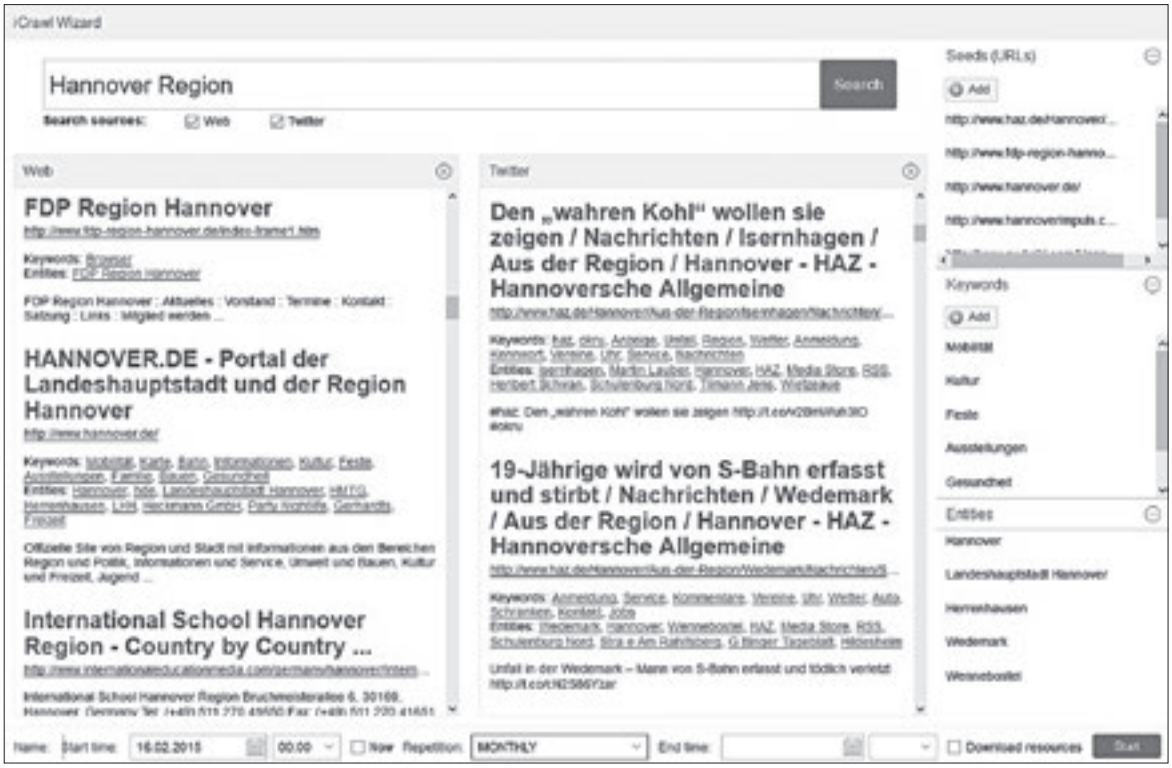

Foto: Risse / Nejdl

Abb. 2: iCrawl-Wizard

Inhalte, allerdings mit der Gefahr, den Crawl zu stark zu beschränken oder ungewollt einen zusätzlichen Fokus hinzuzufügen.

Ein etwas anderer Ansatz wird im iCrawl-Projekt des Forschungszentrums L3S verfolgt (vgl. iCrawl 2015). iCrawl nutzt eine Kombination aus Schlüsselwörtern, Entitäten und Referenzseiten zur Beurteilung der Signifikanz einer Seite. Die Referenzseiten sind ausgewählte Beispiele, die die gewünschten Inhalte besonders gut beschreiben. Für einen Crawl über die Region Hannover könnte eine Referenzseite die Wikipedia-Beschreibung der Region (http://de.wikipedia.org/wiki/Region_Hannover) oder die Homepage der Region (www.hannover.de) sein. Auf diesen finden sich viele Entitäten wie beispielsweise Firmen, Ortsnamen, besondere Wegepunkte oder wichtige regionale Politiker. Aus diesen Dokumenten werden die Entitäten, Begriffe und Wordkombinationen extrahiert und nach ihrer Häufigkeit gewichtet. Zusätzlich manuell spezifizierte Begriffe werden stärker gewichtet. Daraus ergibt sich ein Begriffsvektor, über den während der Crawl-Durchführung die Relevanz eines eingesammelten Dokuments bzgl. der Crawl-Spezifikation bewertet werden kann.

Die Kombination aus Referenzseiten, manuell spezifizierten Begriffen und zusätzlichen Einstiegspunkten (Seed-URLs) ergibt eine recht komplexe Beschreibung eines Crawls. Da die Qualität des resultierenden Archivs auch von der Qualität der Spezifikation abhängt, wurde der iCrawl-Wizard (vgl. Gossen et al. 2015) zur Unterstützung des Spezifikationsprozesses entwickelt. Die Benutzerschnittstelle ist in Abbildung 2 dargestellt. Nach der Eingabe einiger Begriffe (im Beispiel »Hannover Region«) leitet der iCrawl-Wizard diese an eine Websuchmaschine (im Beispiel an Bing) und Twitter weiter. Twitter wird als zusätzliche Quelle für Inhalte angefragt, um auch zeitnah auf Ereignisse reagieren zu können und Seitenempfehlungen aus der Community zu erhalten. Für die Darstellung der Suchergebnisse werden die URLs von Bing ebenso wie die in den Tweets genannten URLs gesammelt und die Schlüsselwörter und Entitäten der Seiten extrahiert. Der Nutzer kann anschließend Seed-URLs auswählen und die Crawl-Intention mithilfe der extrahierten Schlüsselwörter und Entitäten beschreiben. Bevor die Crawl-Spezifikation an den Webcrawler übergeben wird, können noch Start- und Endzeitpunkt sowie mögliche Wiederholungen angegeben werden.

Spezifikationsprozess

Crawl-Durchführung

Webcrawler

Webcrawler sind komplexe Programme, die einen einfachen Prozess umsetzen: Weblinks folgen und Seiten abrufen. In weitverbreiteten domänenorientierten Crawlern wie Heritrix (vgl. Mohr et al. 2004) kann dieses durch eine Vielzahl von Parametern wie Crawl-Tiefe, Dauer des Crawls, Domänen-Filter usw. beeinflusst

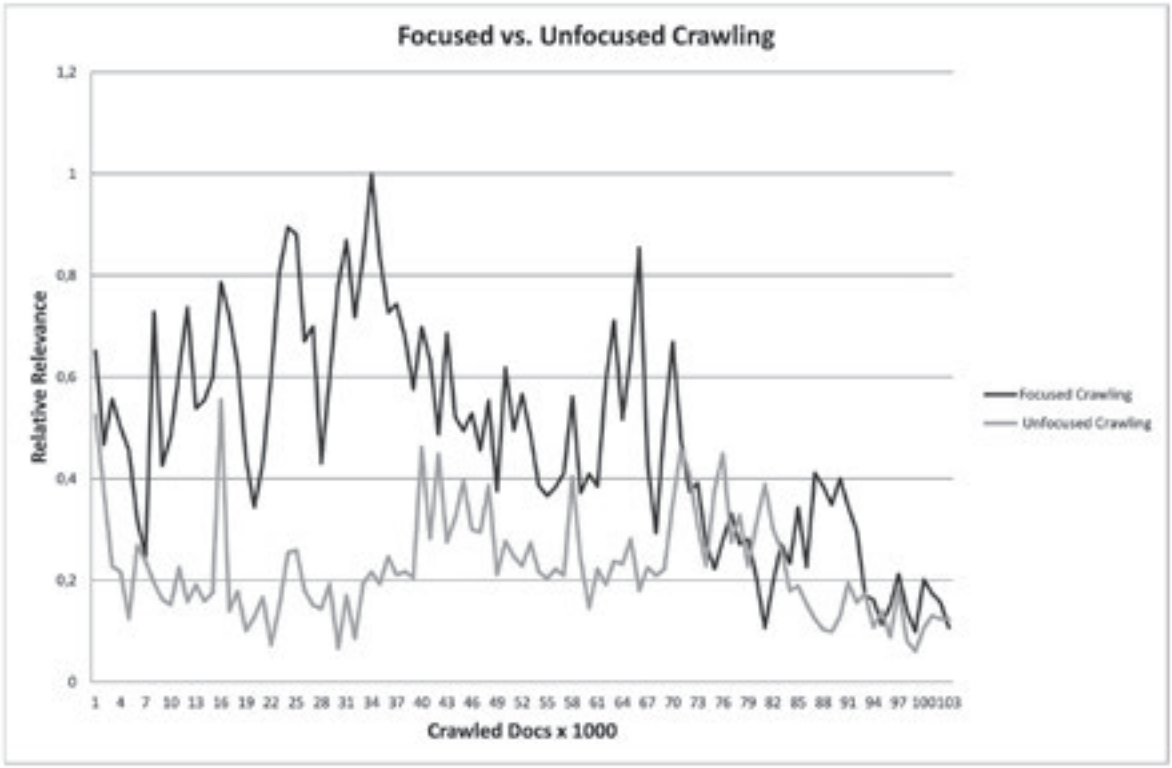

Foto: Risse/Siehndel

Abb. 3: Durchschnittliche Relevanz pro 1.000 Dokumente über die Zeit zum Vergleich von fokussiertem und unfokussiertem Crawler (Quelle: Plachouras et al. 2014; S. 535)

werden. Für thematische Crawls muss zusätzlich noch eine Relevanzbestimmung erfolgen.

Relevanzbestimmung

Die Relevanzbestimmung eines gesammelten Dokuments ist eine zentrale Funktionalität für das themenorientierte Webharvesting. Sie dient insbesondere der Steuerung des Crawlers. Während Crawler von Suchmaschinen zur Bestimmung der Relevanz einer URL auf den Ergebnissen vorheriger Crawls aufbauen können (z. B. ein existierender Link-Graph in Verbindung mit PageRank), ist dies einem Webarchive-Crawler in vielen Fällen nicht möglich. Stattdessen muss die Entscheidung über die Relevanz eines Links auf Basis des vorliegenden Dokuments gefällt werden. Nicht alle Links eines Dokuments sind relevant für das Thema (z. B. Seitennavigation, Werbung), weshalb diese herausgefiltert werden. Für die Bestimmung der Relevanz der übrigen Links wird der Link-Kontext betrachtet. Die Nutzung der Kontexteigenschaft basiert auf der Annahme, dass thematisch relevante Seiten auf weitere relevante Seiten verweisen und sich dieses im Kontext oder Ankertext des Links widerspiegelt. Für die Bestimmung der Relevanz eines Links werden die Relevanz des Dokuments und des Abschnitts, der den Link enthält, sowie der Ankertext des Links bzgl. ihrer Ähnlichkeit zum Referenzvektor zu einem Relevanzwert kombiniert. Die extrahierten Links eines Dokuments werden entsprechend ihres Relevanzwertes in die Warteschlange des Crawlers eingefügt. Die Relevanzwerte des URLs in der Warteschlange werden nach der Analyse neuer Dokumente aktualisiert. Wird ein Dokument von unterschiedlichen Seiten verlinkt, wird der Relevanzwert der URLs aufsummiert. Dieses stellt sicher, dass populäre Seiten mit höherer Priorität gesammelt werden.

Schwankungsbreite des fokussierten Crawls

Die Wirkung der themenorientierten Strategie zeigt sich in der Relevanzdarstellung eines Crawls über die Zeit. In Abbildung 3 wird der fokussierte ARCOMEM-Crawler (vgl. Plachouras et al. 2014) mit dem nicht fokussierenden Heritrix-Crawler verglichen. Beide Crawler beginnen mit derselben Seedliste zum Thema »Europäische Finanzkrise«. Im Verlauf des Crawls kann man beobachten, dass der fokussierte Crawler von Anfang an Dokumente mit höherer Relevanz findet, während diese beim nicht fokussierten Crawl schnell abfällt. Erst nach ca. 70.000 Dokumenten verhalten sich beide Crawler wieder ähnlich. Die größere Schwankungsbreite des fokussierten Crawls lässt sich dadurch erklären, dass der fokussierte Crawler immer wieder auf Seiten trifft, für die eine hohe Relevanz vorhergesagt wurde, die aber in Wirklichkeit eine niedrige Relevanz besitzen. Daraus resultiert eine Korrektur der Priorisierung der URL-Warteschlange, die andere relevante URLs höher priorisiert.

Die über die Zeit abfallende Relevanz der Seiten eröffnet auch die Möglichkeit einer alternativen Terminierungsbedingung eines Crawls. Üblicherweise sind

die Terminierungsbedingungen, also die Festlegung, wann ein Crawl beendet wird, für Webcrawls entweder über die Zeit oder die Menge definiert. Für thematische Crawls kann zusätzlich das Relevanzkriterium genutzt werden. Wenn die Relevanz für eine gewisse Zeit unter einen vorgegebenen Grenzwert fällt, kann dieses ein Signal zur Terminierung des Crawls sein. Wie sich in den Experimenten der Autoren gezeigt hat, variieren die Relevanzwerte in Abhängigkeit des Themas und der Spezifikation allerdings recht stark, weshalb weitere Untersuchungen notwendig sind, bevor Relevanzwerte für die Terminierung zuverlässig verwendet werden können.

Webcrawling und soziale Medien

Soziale Medien spielen eine bedeutende Rolle sowohl für die Nutzerschaft als auch für Betreiber von Websites. Sie ermöglichen Inhaltsanbietern, die Seitenbesucher mit Nachrichten und Informationen zu versorgen und gleichzeitig mit ihnen in direkte Interaktion zu treten. Die Betreiber der Seite versuchen durch die Einbettung von Social Media Feeds, die Diskussionen oder zugehörige Nachrichten dem Besucher sichtbar zu machen. Sie sind deshalb ein ebenso wichtiger Bestandteil einer Seite wie beispielsweise ein eingebettetes Bild oder Video. Entsprechend sollten diese mit derselben Priorität wie andere Komponenten einer Seite gesammelt werden. Allerdings unterscheiden sich Feeds von normalen Webseiten in ihrer Dynamik. Social Media Feeds stellen einen Strom von Nachrichten dar ohne eindeutige Terminierbarkeit. Die zur Darstellung genutzten eingebetteten Objekte sind Programme, die die dynamischen Inhalte über ein Application Programming Interface (API), also eine maschinelle Schnittstelle, beim Anbieter abrufen. Für den Webcrawler ergeben sich zwei Herausforderungen: einerseits das allgemeine Problem der Behandlung von eingebetteten Programmen (s. a. den nächstfolgenden Abschnitt), andererseits der Abruf der Inhalte selbst.

Sofern es sich um weit verbreitete soziale Medien wie Twitter oder Facebook handelt, können diese auch ohne Ausführung des eingebetteten Programmcodes identifiziert und die notwendigen Informationen für die API-Abfragen beim Dienstanbieter extrahiert werden. Da die Daten über eine API abgefragt werden und diese einen Datenstrom darstellen, können Standard-Webcrawler nicht genutzt werden. Aufgrund der unterschiedlichen Dateiformate (meistens JSON anstelle von HTML) werden auch unterschiedliche Analyse- und Extraktionstechniken benötigt. Diese Aufgabe wird von sogenannten API-Crawlern übernommen wie etwa Twittervane (2015). API-Crawler werden parallel zum Webcrawler ausgeführt und sammeln kontinuierlich den Strom von Nachrichten ein.

Ist eine hohe Authentizität der Inhalte gefordert, wie in den Sozial- und Geschichtswissenschaften, ist zudem eine möglichst gleichzeitige Sammlung von Webinhalten und verwandten Inhalten aus sozialen Netzen wünschenswert. Um dieses zu ermöglichen, ist eine enge Verzahnung von Webcrawler und API-Crawler notwendig. Aktuelle Systeme betrachten das Web und soziale Medien getrennt und haben keine direkte Verbindung. Werden im Web Hinweise auf soziale Medien gefunden, muss der API-Crawler separat konfiguriert und angestoßen werden. Im iCrawl-Projekt wird derzeit an einem fokussierenden Web Archiv Crawler gearbeitet, der Webcrawling und Social-Media-Crawling in einem System vereint und eng miteinander verzahnt. Dadurch wird es möglich, mit nur einer Crawl-Spezifikation beide Welten zu erschließen. Während des Sammelprozesses können Links und Feedinformationen automatisch zwischen den Crawlern ausgetauscht werden. Dies ermöglicht es, umfassendere und aktuellere Sammlungen von Webinhalten zu erstellen.

Social Media Feeds

Dynamische und interaktive Inhalte

Ähnlich der Einbettung von sozialen Medien in Webseiten werden auch andere Informationen (Börsenkurse, Bundesligatabelle) während der Aufbereitung und Darstellung im Browser dynamisch integriert. Dahinter verbergen sich im günstigsten Fall JavaScript-Programme, die einmalig oder bei Ereignissen aufgerufen werden. Auf Blogs oder Facebook werden damit auch gerne »endlose« Seiten generiert, indem beim Scrollen zum Seitenende automatisch weitere Einträge nachgeladen werden. Um alle Inhalte erfassen zu können, muss der Code in den Seiten ausgeführt und die Seiten für die Link-Extraktion vollständig im Speicher erzeugt werden. Um auch durch Benutzerinteraktionen generierte Inhalte zu erhalten (Dropdown-Menü) müssen die Aktionen des Benutzers simuliert und nach jedem Simulationsschritt und dem daraus resultierenden Seitenaufbau die neuen Inhalte und Links extrahiert werden. Dieses muss für jede einzelne Interaktionsfläche erfolgen, um eine möglichst vollständige Menge von Links zu erhalten. Um die Seiten zu generieren, nutzt Browser Monkey (Browser Monkeys 2006) den Firefox-Browser, während im Europäischen Projekt LiWA (vgl. Liwa 2008) WebKit dazu verwendet wurde. Leider sind bisher keine dieser Entwicklungen in die Standard Webarchive-Crawler integriert worden.

Aktionen des Benutzers simuliert

Eine andere Alternative ist es, die Struktur der Programme zu analysieren und spezielle Extraktoren zu entwickeln, die die Programmlogik imitieren, um an

die gewünschten Informationen zu gelangen. Aufgrund des manuellen Aufwandes ist dies aber nur für wenige weitverbreitete Websites und Programme sinnvoll (z. B. populäre soziale Medien, s. a. Abschnitt »Webcrawling und soziale Medien«). Des Weiteren werden die eingebetteten Programme regelmäßig verändert und weiterentwickelt, weshalb auch eine regelmäßige Anpassung der Extraktoren notwendig ist.

Anpassung der Extraktoren

Crawl-Nachbereitung

Nach Abschluss des Crawls müssen die finalen Archive erstellt werden. Die meisten Crawler erzeugen die Archive automatisch während der Crawl-Durchführung. Allerdings besteht die Möglichkeit, in der Nachbereitung die Archive zu bereinigen und mit weiteren Metainformation anzureichen. Der ARCOMEM-Crawler nutzt diese letzte Phase für eine umfassendere semantische Analyse der Inhalte, um Entitäten, Ereignisse, Themen und Schlüsselwörter zu extrahieren und das Sentiment einer Seite zu analysieren (vgl. Demidova et al. 2014). Diese Informationen können in einem letzten Filterschritt dazu verwendet werden, das Archiv stärker zu fokussieren. Zusätzlich werden die extrahierten Metainformation zusammen mit den Daten in den WARC-Archivdateien abgelegt. Applikationen, die mit diesen Archiven arbeiten, können diese Metainformation direkt für die Suche und Navigation in den Inhalten nutzen.

Nutzung von Primärquellen

Exploration von Archiven

Nach der Erstellung eines Archivs müssen geeignete Funktionalitäten bereitgestellt werden, um mit dem Archiv zu arbeiten. Durch die Allgegenwart von Suchmaschinen wie Google und Bing wird manchmal vergessen, wie schwierig Suche in großen Archiven eigentlich sein kann, wie aus der folgenden Diskussion deutlich wird.

Suche in großen Archiven

Wenn ein Nutzer von Google im Web schnell nach den bekanntesten Sehenswürdigkeiten einer Großstadt suchen möchte, reicht es üblicherweise, den Namen der Stadt und »Sehenswürdigkeit« einzutippen, und die ersten zehn Suchergebnisse, die dann angezeigt werden, sind meist eine gute Empfehlung für einen Touristen, der diese Stadt besucht. Hier hilft es, dass der Nutzer erstens weiß, welche Ergebnisse zu erwarten sind, und dass zweitens schon viele andere Personen erfolgreich nach diesen gesucht haben. Aus diesem Grund liefert auch die Abfrage nur mit dem Namen der Stadt sinnvolle Ergebnisse und gibt, ebenfalls meist mit den ersten zehn Suchergebnissen, einen guten Überblick über diese Stadt.

Anders verhält es sich, wenn die Suche in einem großen Archiv stattfindet, wo dem Nutzer nicht von vornherein bekannt ist, was darin gefunden werden kann. Interessante Arbeiten im geisteswissenschaftlichen Kontext sind die Artikel (vgl. Tibbo 2002, Duff et al. 2002, Duff et al. 2004), die die Arbeit von Geschichtswissenschaftlern mit (digitalen) Quellen diskutieren. Wichtig hierbei ist etwa die Nutzung von Primärquellen ebenso wie die Notwendigkeit, insbesondere in der ersten Phase der Auswertung dieser Quellen einen Überblick über die vorhandenen Materialien zu erhalten.

Im Folgenden wird kurz auf Ziele und Forschungsfragen im Projekt ALEXANDRIA eingegangen, einem Projekt, das die Arbeit mit Webarchiven mithilfe neuer Algorithmen radikal vereinfachen soll. Anhand eines Suchbeispiels im New York Times Archive wird erläutert, wie die Exploration in einem Archiv besser unterstützt werden kann.

ALEXANDRIA: Ziele und Forschungsfragen

Volltextsuche in Dokumentkollektionen und Archiven ist mittlerweile Standard, oft werden Open Source

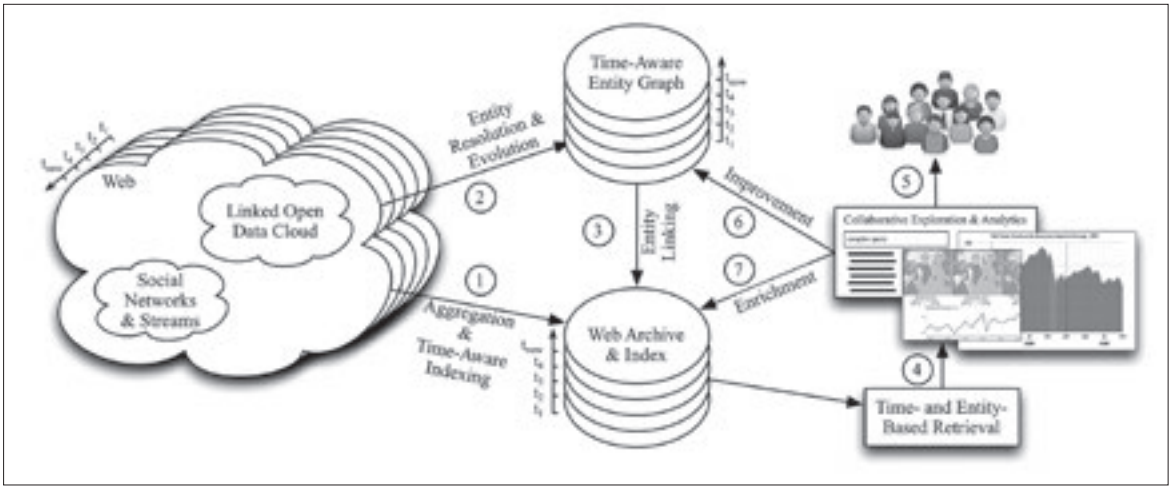

Foto: Risse / Nejdl

Abb. 4: Überblick über Architektur und Forschungsbereiche in ALEXANDRIA

Tools wie Lucene und Solr (vgl. Lucene 2015) genutzt, manchmal auch kommerzielle Systeme. Allerdings sind alle verfügbaren Tools und Algorithmen auf Kollektionen wie das Web und Dokumentkollektionen abgestimmt, in denen die Zeit keine große Rolle spielt. Diese Lücke schließt das ALEXANDRIA-Projekt (vgl. ALEXANDRIA 2014), ein ERC Advanced Grant über fünf Jahre, in dem neue Methoden und Algorithmen entwickelt werden, um (Web-)Archive radikal einfacher nutzbar zu machen. Das Ziel des Projektes ALEXANDRIA ist die Entwicklung von Modellen, Werkzeugen und Technologien, um die intuitive Suche nach und die Arbeit mit Inhalten in Webarchiven zu ermöglichen. Dazu werden semantische und zeitorientierte Indexmethoden entwickelt und mit semantischen Methoden kombiniert, um entitäts- und themenorientierte Suche und Analyse in Archiven zu ermöglichen, und damit auch komplexe und kollaborative Suchprozesse geeignet zu unterstützen.

Abbildung 4 gibt einen Überblick über Architektur und Forschungsbereiche in ALEXANDRIA. Das Web wird durch State-of-the-Art-Technologien archiviert, wie dies etwa durch das Internet Archive, die Deutsche Nationalbibliothek, die British Library und andere Bibliotheken geschieht, und wie dies in der Einleitung dieses Beitrags beschrieben wurde. Dabei wird auch das Social Web berücksichtigt, in geeignet aggregierter Form (1). Strukturierte und semi-strukturierte Informationen über Entitäten und Ereignisse werden aus dem Web extrahiert (2) und (unter Berücksichtigung zeitlicher Veränderungen) zur semantischen Indizierung der Archivinhalte genutzt (3). Gesucht werden kann dann mit spezieller Berücksichtigung der semantischen Entitäten und Ereignisse (4), mit Schwerpunkt auf komplexen semantischen Abfragen und Analysen (5). Dies liefert wiederum entsprechendes Feedback, um sowohl das Wissen über Entitäten und Ereignisse als auch die semantische Indizierung des Archivs zu verbessern (6, 7).

Zeit- und themenbasierte Exploration von Archiven

Erweiterte Explorationsmöglichkeiten sind notwendig, um den Überblick in einem Web- / Nachrichtenarchiv zu erhalten, wenn nicht nur nach bestimmten Artikeln gesucht wird. Das folgende Beispiel diskutiert eine entsprechende Suche im New York Times Annotated Corpus (vgl. LDC 2008), der die Artikel der New York Times von 1987 bis 2007 enthält. Ziel soll es sein, einen Überblick über die darin enthaltenen Artikel über Rudolph Giuliani zu erhalten, der von 1994 bis 2001 (also über zwei Amtsperioden) der 107. Bürgermeister von New York war. Die Anzahl der Artikel, die in diesem Zeitraum über Rudolph Giuliani veröffentlicht wurden, beträgt mehr als 20.000. Es ist daher nicht möglich, mit der üblichen chronologischen Anordnung von Artikeln, wie sie in Abbildung 5 zu sehen ist, einen Überblick über die Berichterstattung über Rudolph Giuliani zu bekommen. Die ersten zehn Artikel sind von 2007, die 100 besten Resultate decken nur die letzten fünf Jahre des durchsuchten Archivs ab.

semantische und zeitorientierte Indexmethoden

Etwas besser ist die in vielen Suchsystemen standardmäßig eingestellte TFIDF-Reihenfolge (term frequency, »Vorkommenshäufigkeit«, und inverse document frequency, »inverse Dokumenthäufigkeit«), die Artikel mit häufigem Auftreten des Suchbegriffs höher

TFIDF-Reihenfolge

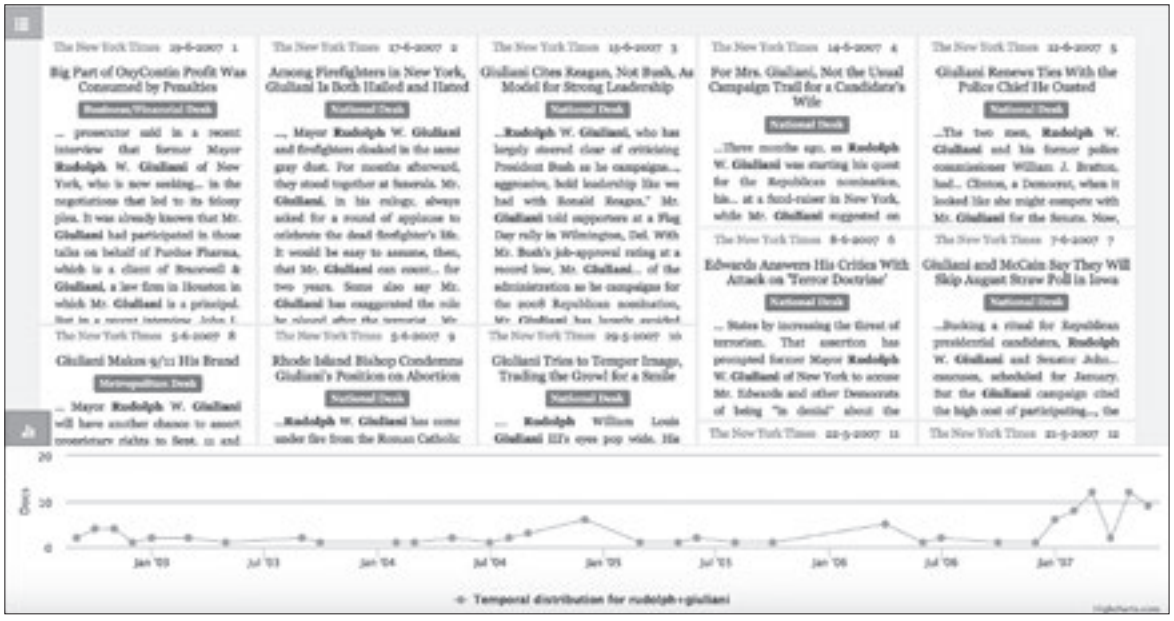

Foto: Risse / Nejdl

Abb. 5: »Rudolph Giuliani«, Ergebnisse chronologisch angeordnet

Foto: Risse / Nejdl

Abb. 6: »Rudolph Giuliani«, Ergebnisse nach TFIDF-Reihenfolge angeordnet

Foto: Risse / Nejdl

Abb. 7: »Rudolph Giuliani«, Ergebnisse nach HISTDIV-Reihenfolge angeordnet

bewertet als andere. Dies führt dazu, dass die 100 besten Artikel den gesamten Zeitraum des Archivs abdecken, wie aus der Zeitlinie in Abbildung 6 ersichtlich wird. Allerdings erscheinen auch hier die besten zehn Artikel mehr oder weniger zufällig ausgewählt, mit zwei Artikeln über den Tod von Rudolph Giuliani's Mutter im Jahr 2002 und fünf Artikeln über seinen ersten (erfolglosen) Wahlkampf 1989.

HISTDIV Algorithmus

Erst der im Rahmen des ALEXANDRIA-Projektes entwickelte HISTDIV (Historical Diversity) Algorithmus, der neben TFIDF sowohl die zeitliche als auch die thematische Abdeckung der Artikel berücksichtigt, ist in der Lage, einen besseren Überblick zu verschaffen (s. Abb. 7). Hier finden sich nun Artikel über alle drei Wahlkämpfe in New York (1989, 1993 und 1997), über die Krebserkrankung von Giuliani im Jahr 2000 und über seinen (erfolglosen) Wahlkampf zur Nominierung als Präsidentschaftskandidat im Jahr 2007. Weitere Details zum HISTDIV-Algorithmus sind in Singh et al. (2015) beschrieben.

Webarchive für die Digital Humanities

Die Forschungsfragen, die mithilfe von Webarchiven beantwortet werden, sind vielfältig. Dieser Abschnitt soll zumindest einige davon beispielhaft darstellen so-

wie Hinweise auf Workshops und Konferenzen geben, die eine gute Plattform für Diskussion und Austausch von Fragen und Ergebnissen darstellen.

Im Rahmen des L3S laufen im Kontext des Leibniz-Forschungsverbundes Science 2.0 Arbeiten zur Wissenschaftskommunikation in sozialen Medien, gemeinsam mit Wissenschaftlern aus dem Bereich der Soziologie (vgl. Hadgu et al. 2014), vorgestellt auf der Web Science Conference 2014, oder auch die Repräsentation von Weltliteratur im Web der letzten 20 Jahre. Weitere Arbeiten mit Wissenschaftlern aus dem Bereich der Politikwissenschaften zum Thema Politische Diskussion und Digitalisierung sind geplant.

Im Rahmen des BUDDAH-Projektes (Big UK Domain Data for the Arts and Humanities) (vgl. Buddah 2015), koordiniert von Jane Winters vom Institute of Historical Research in London, gemeinsam mit der British Library und dem Oxford Internet Institute, wurden mithilfe des von der British Library zur Verfügung gestellten UK Web Archives eine Reihe von Forschungsfragen aus unterschiedlichsten Disziplinen diskutiert. Darunter waren etwa die Fragen »Wie manifestiert sich britischer Euroskeptizismus im UK Web der letzten 20 Jahre?«, »Welche Online-Netzwerke und Plattformen gibt es für Dichtung und Poesie?«, oder auch die Untersuchung spezifischer Kollektionen wie z. B. des UK Parliament Web Archives, des Verteidigungsministeriums oder den sogenannten Shoebox Archives über Kriegsgefangene im Zweiten Weltkrieg.

Auch ein kürzlich erschienener Artikel von Jill Lepore, einer Geschichtswissenschaftlerin an der Harvard Universität, diskutiert in der Januar-Ausgabe des New Yorker (vgl. Lepore 2015) das Thema Webarchivierung. Ihre Beispiele dazu kommen aus dem Ukraine-Konflikt (Abschuss der Boeing 777 am 17. Juni 2014) ebenso wie aus einer Untersuchung von URLs im Harvard Law Review und weiteren rechtswissenschaftlichen Journalen, die bereits zu 70 % nicht mehr auf die ursprünglich verlinkten Seiten zeigen.

Im Juni 2014 fand an der Harvard Universität der WIRE-Workshop (»Working with Internet Archives for Research«) statt. Organisatoren waren Matthew Weber von der Rutgers University, David Lazer von der Northeastern University und Kris Carpenter vom Internet Archive in San Francisco, die das im Rahmen eines gemeinsamen Projektes entstandene ArchiveHub-Portal vorstellten, mit Sammlungen wie etwa über das Occupy Wall Street Movement (vgl. Weber et al. 2014) oder den Wirbelsturm Sandy. Zahlreiche Vortragende aus unterschiedlichen Disziplinen der Digital Humanities und der Informatik präsentierten weitere interessante Forschungsfragen und Ansätze für die Arbeit mit Webarchiven.

Auch die RESAW-Konferenz im Juni 2015 in Aarhus zum Thema »Web Archives as Scholarly Sources« sowie die jährlich stattfindenden ALEXANDRIA-Workshops in Hannover (der erste fand im September 2014 statt) waren und sind wichtige Treffpunkte für alle, die an Webarchiven und der Forschung mithilfe von Webarchiven interessiert sind.

Plattformen für Diskussion und Austausch

Diskussion

In den vorangegangenen Abschnitten wurde, ausgehend von einer immer größeren Vernetzung der Akteure im Internet, gezeigt, welche Herausforderungen und Ansätze es sowohl bei der Sammlung von Webinhalten als auch bei deren Nutzung und Exploration gibt. Neben dem Wachstum des Web stellen die zunehmende Integration von sozialen Medien und die Verwendung von dynamischen Inhalten, die eine programmbasierte Ausführung individueller Seiten notwendig machen, die Archivare des Web vor große Herausforderungen. Aufgrund der Komplexität dieser Programme wird es auch auf Dauer keine umfassende Lösung geben, man wird weiterhin auf Heuristiken angewiesen bleiben.

auf Heuristiken angewiesen

Auf der inhaltlichen Ebene hat in den letzten Jahren das Interesse an Webarchiven in verschiedenen Bereichen zugenommen, was wiederum zu neuen Anforderungen geführt hat. Mit dem Bewusstsein, dass eine allumfassende Archivierung des World Wide Web nicht möglich ist, steigt das Interesse an thematischen Webarchiven, wie sie beispielsweise von archive-it.org bereitgestellt werden. Crawler, die neben einer initialen Seedliste auch eine semantische Beschreibung der Inhalte nutzen und diese mit Inhalten aus sozialen Medien verbinden, erlauben es, mit geringerem Aufwand fokussierte Webarchive zu erstellen und auch kurzfristig auf Ereignisse zu reagieren.

integrierte Betrachtung

Ein bisher wenig beachteter Aspekt ist die integrierte Betrachtung von Webcrawling und sozialen Medien. Während die Nutzer ohne Hindernisse zwischen den Medien wechseln können, ist die Sammlung der Inhalte auf ein Medium beschränkt. Eine integrierte Betrachtung ermöglicht einerseits die Nutzung von sozialen Medien als Quelle von Verweisen auf relevante Inhalte im Web und andererseits die Sammlung relevanter Diskussionen zu Inhalten auf Webseiten aus den sozialen Medien. Die Entwicklung entsprechender Systeme befindet sich aber noch am Anfang.

Mit der Verfügbarkeit von Webarchiven ergibt sich natürlich auch ein wachsendes Interesse an deren Nutzung. Allerdings stellen die aktuell erstellten (W)ARC-Dateien praktisch nur große Datensammlungen mit minimalen Inhaltsangaben dar. Die weitverbreitete Wayback Machine (IIPC 2015) erlaubt keine

freie Suche, sondern nur den Zugriff über eine URL. Eine gezielte Nutzung bzw. Extraktion von Inhalten auf semantischer Basis ist daher aufwändig. Die Anreicherung von Archiven mit semantischen Metainformationen ist ein erster Schritt, um inhaltliche Beschreibungen von Archivdateien zu erhalten. Diese können auch zur Erstellung eines Schlüsselwortindexes genutzt werden, um eine den Websuchmaschinen ähnlicheren Zugriff auf die Inhalte zu ermöglichen. Allerdings kommen heutige Technologien aufgrund der großen Datenmenge schnell an ihre Grenzen.

Nutzungsszenarien von Webarchiven

Die Nutzungsszenarien von Webarchiven sind insbesondere in Wissenschaft, Journalismus und Marketing vielfältig. Explorative Suche in Kombination mit datengetriebenen Analysen ermöglichen sowohl aggregierte Sichten als auch, für die Detailanalyse, einen gezielten Zugriff auf einzelne Dokumente. So erlaubte die Dokumentation der sozialen Medien während des »Arabischen Frühlings« tiefere Einblicke in deren Nutzung. Der Arab Social Media Report (vgl. Mourtada 2011) kommt, nach der Analyse der Facebook- und Twitter-Kommunikation, zu der Schlussfolgerung, dass die sozialen Medien einen geringeren Einfluss auf die Proteste hatten, als bisher angenommen. Zwar sind die Benutzerzahlen in Ägypten und Tunesien 2011 rasant anstiegen, aber insgesamt zu dieser Zeit sehr gering. Soziale Medien wurden hauptsächlich als Informationsquelle genutzt und weniger zur Organisation von Protesten.

virtuelle Communities von besonderem Interesse

Virtuelle Communities sind in der Analyse von besonderem Interesse. Communities bilden sich zu einem Thema, wachsen über die Zeit und verschwinden ggf. zu einem späteren Zeitpunkt wieder. Die Datenjournalisten des Guardian haben in Zusammenarbeit mit den Universitäten Manchester, St Andrews und Leicester 2011 die Twitter-Kommunikation während der Ausschreitungen in London und anderen englischen Städten analysiert (vgl. Guardian 2011). Sie analysierten die Ausbreitung von Nachrichten und zeigten die Ergebnisse in interaktiven Visualisierungen. In ihren Untersuchungen haben sie u.a. gezeigt, wie sich Gerüchte verbreiteten (z. B. Angriff auf den Zoo von London) und wie diese von der Community widerlegt wurden.

Die vorgestellten Nutzungsszenarien basieren auf Analysen von verhältnismäßig kleinen Datenmengen (ca. 10 Millionen Tweets für den Arabischen Frühling und ca. 2,6 Millionen für die London Riots) und kurzen Betrachtungszeiträumen. Durch Analysen von zeitlich und mengenmäßig umfangreicheren Kollektionen können auch langsamere Evolutionsprozesse sichtbar gemacht werden. Linguisten könnten die Entstehung neuer Begriffe und deren Nutzung und Verbreitung beobachten. Historiker können die Nachwirkung von Ereignissen und die Änderung in ihrer Wahrnehmung über die Zeit untersuchen. Die öffentliche Wahrnehmung von Produkten und Firmen und ihrer zeitlichen Veränderungen gibt Marketingfirmen und -abteilungen wichtige Entscheidungshilfen.

Erste Projekte und Workshops zum Thema, diskutiert im vorangegangenen Kapitel dieses Beitrags, zeigen auch das hohe Potential von Webarchiven für die Forschung im Bereich der Digital Humanities. So gut wie immer stellen hier aber auch die unzureichend an Webarchive angepassten Such- und Explorationsmöglichkeiten sowie die noch weiter zu entwickelnden Untersuchungsmethoden Hindernisse dar, die noch zu überwinden sind.

Um das Potential von Webarchiven für die Forschung tatsächlich zu heben, sind neue Methoden notwendig, wie sie beispielsweise im ALEXANDRIA-Projekt entwickelt werden, wobei inhaltliche und zeitliche Aspekte sowohl bei der Erstellung von Webarchiven als auch bei ihrer Nutzung immer wichtiger werden, einhergehend mit einer verbesserten semantischen Suche in diesen Archiven. Und parallel zur Verbesserung der inhaltlichen Analyse und der Entwicklung verbesserter Nutzungsmöglichkeiten muss auch an der Skalierbarkeit der Technologien gearbeitet werden. Die Nutzung von Webarchiven ist definitiv ein Big-Data-Problem, das noch viele interessante Forschungsfragen beinhaltet.

Danksagung

Teile der vorgestellten Arbeiten wurden durch den European Research Council im ERC Advanced Grant ALEXANDRIA (ERC 339233) und durch die Europäische Kommission im 7. Forschungsrahmenprogramm in den Projekten ARCOMEM (ICT 270239) und LiWA (IST 216267) gefördert.

Literatur

(ALEXANDRIA, 2014) ALEXANDRIA – Foundations for Temporal Retrieval, Exploration and Analytics in Web Archives. [Zugriff am: 25.3.2015]. Verfügbar unter: http://alexandria-project.eu

(BROWSER MONKEYS 2006) Leverage browsers for link-extraction. [Zugriff am: 16.2.2015]. Verfügbar unter: https://webarchive.jira.com/wiki/display/SOC06/Leverage+browsers+for+link-extraction

(BUDDAH 2015) Big UK Domain Data for the Arts and Humanities. [Zugriff am: 25.3.2015]. Verfügbar unter: http://buddah.projects.history.ac.uk

(DEMIDOVA et al. 2014) DEMIDOVA, Elena, Nicola BARBIERI, Stefan DIETZE, Adam FUNK, Helge HOLZMANN, Diana MAYNARD, Nikolaos PAPAILIOU, Wim PETERS,

Thomas RISSE und Dimitris SPILIOTOPOULOS: Analysing and Enriching Focused Semantic Web Archives for Parliament Applications. *Future Internet*. 2014, **6**(3), S. 433–456

(DENIC 2015) Denic.de Statistiken: Domainentwicklung. [Zugriff am: 28.1.2015]. Verfügbar unter: www.denic.de/de/hintergrund/statistiken.html

(DUFF et al. 2002) DUFF, W. M. und C. A. JOHNSON: Accidentally found on purpose: information-seeking behavior of historians in archives. In: *The Library Quarterly*. 2002, S. 472–296

(DUFF et al. 2004) DUFF, W., B. CRAIG und J. CHERRY: Historians' use of archival sources: Promises and pitfalls of the digital age. *The Public Historian*. 2004, **26**(2), S. 7–22

(GOSSEN et al. 2015) GOSSEN, Gerhard, Elena DEMIDOVA und Thomas RISSE: The iCrawl Wizard – Supporting Interactive Focused Crawl Specification. In: *Proceedings of the European Conference on Information Retrieval (ECIR) 2015*, 2015

(FACEBOOK 2015) Facebook Newsroom. [Zugriff am: 11.2.2015]. Verfügbar unter: http://newsroom.fb.com/company-info

(GUARDIAN 2011) The Guardian: Twitter and the riots: how the news spread. [Zugriff am: 18.3.2015]. Verfügbar unter: www.theguardian.com/uk/2011/dec/07/twitter-riots-how-news-spread

(HADGU et al. 2014) HADGU, Asmelash und Robert JÄSCHKE: *Identifying and analyzing researchers on Twitter*. Intl. ACM Web Science Conference, Bloomington, IN, Juni 2014

(iCRAWL 2015) iCrawl: The Integrated Focused Crawling ToolBox. [Zugriff am: 13.2.2015]. Verfügbar unter: https://next.l3s.uni-hannover.de/projects/icrawl

(IIPC 2015) IIPC: OpenWayback. [Zugriff am: 19.2.2015]. Verfügbar unter: www.netpreserve.org/openwayback

(LDC 2008) LDC, The New York Times Annotated Corpus. [Zugriff am: 19.2.2015]. Verfügbar unter: https://catalog.ldc.upenn.edu/LDC2008T19

(LEPORE 2015) LEPORE, Jill: The Cobweb: Can the Internet be archived? In: The New Yorker, 26. Januar 2015

(LIWA 2008) LiWA – Living Web Archives. [Zugriff am: 16.2.2015]. Verfügbar unter: http://liwa-project.eu

(LUCENE 2015) Apache Lucene. [Zugriff am: 19.2.2015]. Verfügbar unter: http://lucene.apache.org

(MOHR et al. 2004) MOHR, G., M. STACK, I. RANITOVIC, D. AVERY und M. KIMPTON: Introduction to heritrix, an archival quality Web crawler. In: *Proceedings of the 4th International Web Archiving Workshop*, Bath, UK, 16. September 2004

(MOURTADA 2011) MOURTADA R. und F. SALEM: Civil movements: The impact of facebook and twitter. In: *Arab Social Media Report*. 2011, Vol. 1(2)

(NTOULAS et al. 2004) NTOULAS, A., J. CHO und C. OLSTON: What's New on the Web? The Evolution of the Web from a Search Engine Perspective. In: *Proceedings of the 13th International Conference on World Wide Web*, New York, NY, USA, 17–20 May 2004

(PLACHOURAS et al. 2014) PLACHOURAS, Vassilis, Florent CARPENTIER, Muhammad FAHEEM, Julien MASANÈS, Thomas RISSE, Pierre SENELLART, Patrick SIEHNDEL und Yannis STAVRAKAS: ARCOMEM Crawling Architecture. *Future Internet*. 2014, **6**(3), S. 518–541

(RISSE et al. 2014) RISSE, Thomas, Elena DEMIDOVA, Stefan DIETZE, Wim PETERS, Nikolaos PAPAILIOU, Katerina DOKA, Yannis STAVRAKAS, Vassilis PLACHOURAS, Pierre SENELLART, Florent CARPENTIER, Armin MANTRACH, Bogdan CAUTIS, Patrick SIEHNDEL und Dimitris SPILIOTOPOULOS: The ARCOMEM Architecture for Social- and Semantic-Driven Web Archiving. *Future Internet*. 2014, **6**(4), S. 688–716

(SINGH et al. 2015) SINGH, J., W. NEJDL und A. ANAND: History by Diversity: Helping Historians Search News Archives. In: *L3S Technical Report*. 01/2015

(TIBBO 2002) TIBBO, H. R.: Primarily history in America: How us historians search for primary materials at the dawn of the digital age. *American Archivist*. 2002, **66**(1), S. 9–50

(TUMBLR 2015) About Tumblr. [Zugriff am: 11.2.2015]. Verfügbar unter: https://www.tumblr.com/about

(TWITTERVANE 2015) Twittervane. [Zugriff am: 13.2.2015]. Verfügbar unter: http://netpreserve.org/projects/twittervane

(WEBER et al. 2014) WEBER, M. S., S. EVANS und K. DRISCOLL: *Seeking Structure in Anarchy: The Emergence of Organization in the Occupy Wall Street Movement*. Paper presented at the Annual Meeting of the National Communication Association. Chicago, Illinois, November 2014

Die Verfasser

Dr.-Ing. Thomas Risse, Stellv. Geschäftsführer des Forschungszentrums L3S & Forschungsgruppenleiter, Leibniz Universität Hannover, Forschungszentrum L3S, Appelstraße 9a, 30167 Hannover, E-Mail: risse@L3S.de

Prof. Dr. techn. Wolfgang Nejdl, Direktor des Forschungszentrums L3S, Leibniz Universität Hannover, Forschungszentrum L3S & Institut für Verteilte Systeme Wissensbasierte Systeme (KBS), Appelstraße 9a, 30167 Hannover, E-Mail: nejdl@L3S.de