Text-Mining in Adaptive READ

Stefan Agne, Armin Hust, Stefan Klink, Markus Junker, Andreas Dengel
Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)GmbH

Christoph Altenhofen
Institut für Arbeitswissenschaft und Technologiemanagement (IAT)

Jürgen Franke, Ingrid Renz
DaimlerChrysler AG

Bertin Klein
Insiders Information Management GmbH

Abstract

Adaptive READ ist ein vom BMBF gefördertes Verbundprojekt, welches an adaptiven Dokumentanalysetechniken für zukünftige Dokumenterschließungssysteme arbeitet. Eine zentrale Aufgabe hierbei ist die Unterstützung des Wissensmanagements durch Techniken zur Suche und Navigation in Textdokumenten. In Adaptive READ wird Informationssuche in Dokumenten als Prozeß gesehen. Daraus abgeleitet ergibt sich die Aufgabe, den Suchprozeß durch geeignete Werkzeuge ganzheitlich zu unterstützen. Aus der Auffassung von Informationssuche als Prozess ergibt sich auch die interessante Frage, ob sich beobachtete Suchprozesse zur Steuerung bzw. Verbesserung neuer Suchen verwenden lassen. Auch dieser Fragestellung wird in Adaptive READ nachgegangen. Die Arbeiten in diesen Bereichen lassen sich dem Text-Mining zuordnen.

1 Einleitung

Adaptive READ ist ein vom BMBF gefördertes und vom DFKI koordiniertes Verbundprojekt im Umfeld der Dokumentanalyse mit einem Gesamtprojektvolumen von etwa 40 Mio DM. Nach einer 9-monatigen Vorlaufphase startete das Hauptprojekt von Adaptive READ mit einer Laufzeit von 34 Monaten im März 2000.
Dokumentanalyse beschäftigt sich mit der inhaltlichen Erschließung insbesondere papiergebundener Dokumente – vom Einscannen und Verarbeiten der Bitmap bis hin zur Extraktion der relevanten Informationen und der inhaltlichen Suche in Dokumentbeständen. Die Installation eines Dokumenterschließungssystems ist in der Regel mit einem enormen manuellen Aufwand verbunden. Dies betrifft sowohl die Auswahl und Kombination einzelner Dokumentanalysekomponenten als auch die Anpassung dieser Komponenten an die konkrete Anwendung. Bisher waren Dokumenterschließungssysteme daher nur für Großanwendungen wie das automatische Anschriftenlesen bei der Post oder die automatische Erfassung von Überweisungsformularen rentabel.
Der Einsatz in Kleinanwendungen erfordert die Erarbeitung umfassender Konzepte für adaptive Dokumenterschließung. Adaptivität bedeutet hierbei zum einen, Dokumenterschließungstechniken so zu modularisieren, daß Systeme für neue Anwendungen möglichst einfach entwickelt werden können. Andererseits sollen die Mo-

dule selbst lernfähig gestaltet werden, was im Idealfall eine Anpassung des Systems während des laufenden Betriebs erlaubt. Durch die Erreichung dieser Ziele erhofft man zukünftig den rentablen Einsatz von Dokumentanalysetechniken in vielen neuen Anwendungen zu ermöglichen.
Dokumenterschließung ist eine Technologie zur Unterstützung des Wissensmanagements. Im Sinne des Wissensmanagements sind Dokumente Träger von Wissen, welches Menschen in ihren Arbeitsprozessen unterstützen kann. Eine zentrale Rolle bei der bedarfsgerechten Verfügbarmachung dieses Wissens kommt der technischen Unterstützung der Suche und Navigation in Dokumentsammlungen zu. Technische Lösungen hierzu sind bereits heute weit verbreitet: man denke nur an Technologien wie sie in Internet-Suchmaschinen verwendet werden. Adaptive READ hat sich auch dieses Themas angenommen. Zielsetzung ist hierbei insbesondere, durch den Einsatz von Lernverfahren die Suche nach Dokumenten und Textstellen zu beschleunigen und die Genauigkeit bei der Suche zu erhöhen. Vier der elf Partner in Adaptive READ beschäftigen sich mit Fragestellungen, die dem Bereich des Text-Minings zugeordnet werden können: die DaimlerChrysler AG, die DFKI GmbH, die Insiders Information Management GmbH und das Institut für Arbeitswissenschaft und Technologiemanagement (IAT) an der Universität Stuttgart.

2 Motivation und Aufgabenverteilung

Die Suche nach Informationen in Texten ist eine Aufgabenstellung, die im Information Retrieval (IR) betrachtet [8, 2] und von einer klassischen Aufgabenstellung dominiert wird: Ein Benutzer stellt eine Suchanfrage und möchte als Antwort möglichst alle relevanten Dokumente einer Kollektion erhalten. In Adaptive READ wird diese klassische Aufgabenstellung in zwei Richtungen erweitert betrachtet. Zum einen wird die Suche als interaktiver Prozeß aufgefaßt. Bei vielen Suchen und insbesondere bei der Exploration von Informationen in Texten wird eine Serie von Anfragen gestellt. Die Idee zu den jeweils neuen Anfragen ergibt sich dabei häufig aus zuvor vom Retrievalsystem gelieferten Antworten. Das abgeleitete Ziel von IR Systemen ist daher (neben dem Liefern möglichst aller relevanten Dokumente für eine Anfrage) insbesondere die ganzheitliche Unterstützung der Suche als Prozeß. Hierbei spielen Techniken eine Rolle, die einen schnellen Überblick über die Inhalte von Antwortdokumenten erlauben, die Vorschläge für Anfragemodifikationen erzeugen oder Feedback des Benutzers auf Dokumentebene zulassen. Die Aktualität derartiger Ansätze spiegelt sich in den Beiträgen zur Text Retrieval Evaluation Conference (TREC,[10]) der letzten Jahre wider.
In der klassischen Aufgabenstellung im IR hat das Retrievalsystem kein "Gedächtnis". Erfolge und Mißerfolge eines Benutzers bei der Suche nach bestimmten Informationen können nicht für den nächsten Benutzer, der vielleicht einen ähnlichen Informationsbedarf hat, nutzbar gemacht werden. Nimmt man an, daß einem Retrievalsystem derartige Informationen zur Verfügung stehen – ebenfalls eine Erweiterung der klassischen Retrievalaufgabe –, ergeben sich interessante Potentiale zur Verbesserung der Suchergebnisse und Steuerung der Suchprozesse. Unseres Wissens nach wurden auf Basis dieser Überlegung noch keine gezielten Forschungsarbeiten betrieben. Die Aufgabenstellung weist aber eine gewisse Verwandtschaft mit dem ebenfalls im IR aktuellen Collaborative Filtering [6, 4] auf.
Aufbauend auf den obigen Überlegungen werden von den einzelnen Partnern folgende Aufgaben schwerpunktmäßig bearbeitet:

- Die Firma Insiders beschäftigt sich mit ganzheitlichen Konzepten zur Unterstützung der Informationssuche als Prozess. DaimlerChrysler arbeitet an lernfähigen Methoden zur benutzerorientierten und dokumentübergreifenden inhaltlichen Zusammenfassung von Dokumenten. Derartige Methoden stellen

eine wichtige Komponente bei der Unterstützung des Suchprozesses dar, da sie dem Benutzer die Möglichkeit geben, schnell den Inhalt möglicherweise recht langer Dokumente zu überblicken.

- Das DFKI und das IAT arbeiten an Methoden zur Verbesserung der Suchmöglichkeiten über die Benutzerbeobachtung. Mit Hilfe in der Vergangenheit beobachteter Suchprozesse soll dabei ein Retrievalsystem an seine typischen Benutzer angepaßt werden. Hiervon wird sowohl eine Verbesserung der Genauigkeit bei einer einzelnen Suchanfrage als auch eine Unterstützung der Informationssuche als Prozeß erhofft. Die zu entwickelnden Ansätze sollen ebenfalls in das ganzheitliche Konzept zur Unterstützung der Informationssuche als Prozess einfließen.

Als gemeinsames Modell für die Mehrzahl der Arbeiten in den obigen Schwerpunkten wird das im IR populäre Vektorraummodell herangezogen [9]. Im Vektorraummodell werden Dokumente in Form von Vektoren $\vec{d} = (f_1, \ldots, f_n)$ dargestellt. Jedes f_i bezeichnet hierbei die Häufigkeit des Auftretens des Terms (Wortes) i im repräsentierten Dokument (alle Wörter einer Dokumentkollektion werden durchnummeriert). In der Regel werden die absoluten Worthäufigkeiten auf der Basis einer Dokumentkollektion durch Gewichte ersetzt, die angeben sollen, wie stark einzelne Wörter den Inhalt eines Dokuments repräsentieren. Anfragen $\vec{q}$ an das Retrievalsystem werden in analoger Weise repräsentiert. Eine Anfrage wird im Vektorraummodell mit den Dokumenten beantwortet, die eine hohe Ähnlichkeit zu ihr im Sinne eines über Vektoren definierten Ähnlichkeitsmaßes aufweisen.
Die folgenden Abschnitte erläutern die Ansätze und bisherigen Ergebnisse in den verschiedenen Schwerpunktgebieten.

3 Ganzheitliches Konzept zur Informationssuche

Die klassische Recherche beantwortet Suchanfragen mit einer Menge von Antwortdokumenten. Der Ansatz für die Aktivitäten der Firma Insiders in Adaptive READ ist die Beobachtung, dass Informationsbedürfnisse nicht immer mit einer Suchanfrage erschöpfend befriedigt sind. Suchende müssen dann weitere modifizierte Suchanfragen entwickeln. Dafür ist ganz wesentlich die Interpretation des Ergebnisses der vorher erzielten Antwortmenge und die Ableitung von Ideen, wie die vorhergehende Anfrage gezielt verbessert werden kann. Dabei ist es vollkommen dem Benutzer überlassen eine Suchanfrage zur Erzielung besserer Suchergebnisse sukzessive zu modifizieren.
Ausgangspunkt zur Unterstützung des Suchprozesses ist die klassische Recherche durch die sukzessive Modifikation von Anfragen. Nützlich hierbei ist es, vom System Vorschläge geliefert zu bekommen, mit welchen Suchbegriffen eine Anfrage aufgeweicht oder verfeinert werden kann. Teilweise fällt es Benutzern leichter, Antwortdokumente bezüglich ihrer Relevanz zu beurteilen, als die Anfrage selbst zu modifizieren. In diesem Fall sind Techniken hilfreich, die bewertete Antwortdokumente zur Verfeinerung der Suche verwenden. Hat man sehr viele Antwortdokumente, so ist darüberhinaus eine thematische Clusterung der Dokumente hilfreich. Sie erlaubt es direkt auf das Cluster zu fokussieren, welches am besten das Interesse repräsentiert.
Im Rahmen von Adaptive READ wurden die folgenden Funktionalitäten zur Unterstützung des Suchprozesses in Form einer SDK (Software Development Kit) unter dem Namen "*mindaccess*" implementiert:

- die Dokumentsuche über direkte Anfragen;

- die Berechnung von Termähnlichkeiten anhand einer Dokumentkollektion; sie wird verwendet, um sich zu Suchbegriffen verwandte Terme angeben zu lassen;
- die Neubewertung von Dokumenten anhand der Benutzerbewertung von Dokumenten, um eine Alternative zur direkten Manipulation der Anfragen zur Verfügung zu haben;
- das thematische Clustern von Antwortdokumenten zur zielgerichteten Navigation in großen Antwortmengen.

Die Funktionalitäten basieren auf Verfahren, die aus der Literatur entnommenen und angepaßt wurden (siehe z.B. [7] und [2]). Eine Hauptanforderung bestand darin, die Einzelverfahren in einem sauberen implementierungstechnischen Framework zusammenzufassen. Eine besondere Schwierigkeit ergab sich bei der Auswahl eines geeigneten Clusterverfahrens. Die aus der Literatur bekannten Verfahren des agglomerativen Clusterns oder $K-$Means stellten sich als nicht performant genug für die anvisierten Dokumentmengen von mehr als 100.000 Dokumenten heraus, sodaß eine Neuentwicklung erforderlich war.
Ein für den Benutzer nützliches Instrument zur schnellen Beurteilung gefundener Dokumente sind inhaltliche Textzusammenfassungen. Werden Antwortdokumente durch das Retrievalsystem in inhaltlichen Clustern zurückgeliefert, so ist es zur Unterstützung des Suchprozesses auch hilfreich, eine grobe Zusammenfassung des Inhalts aller Dokumente des betreffenden Clusters zur Verfügung zu haben. Darüberhinaus sind optimale Zusammenfassungen nicht statisch, sondern müssen sich am konkreten Benutzer und seinem Informationsbedürfnis orientieren. Ziel der schwerpunktmäßig bei DaimlerChrysler vorangetriebenen Arbeiten sind daher Methoden, die dokumentübergreifende und benutzerorientierte inhaltliche Zusammenfassungen generieren.
Die Idee auf der Basis des Vektorraummodells Textzusammenfassungen zu generieren, ist nicht neu. In [5] wird beschrieben, wie anhand von Termgewichtungen Gewichte für Sätze in Textdokumenten berechnet werden können. Hochgewichtete Sätze werden verwendet, um inhaltliche Zusammenfassungen von Texten zu erzeugen.
Beim Projektpartner DaimlerChrysler werden zur Berechnung von Dokumentzusammenfassungen Dokumentvektoren mit Buchstaben-Quadgrammen verwendet. Vorteil der Quadgramme ist, dass sie robuster gegenüber Tippfehlern oder Flexionsformen der Wörter sind. In einem zunächst implementierten Ansatz zur benutzerunabhängigen Erzeugung von Dokumentzusammenfassungen werden die Gewichte der Quadgramme benutzt, um Gewichtungen für die Wörter zu generieren. Anhand der Wortgewichte werden dann die Sätze der Dokumente gewichtet. Analog zu [5] bildet eine Auswahl der bestgewichteten Sätze die Zusammenfassung.
Einen interessanten Ansatz zur Erzeugung benutzeradaptiver Zusammenfassungen findet man in [1]. Dort werden hochgewichtete Sätze dem Benutzer zur Bewertung vorgelegt. Anhand der Benutzerbewertungen werden die Satzgewichtungen angepaßt und verbesserte Textzusammenfassungen generiert. Ein Nachteil dieses Verfahrens ist, daß es für den Benutzer mit einem erheblichen Aufwand verbunden ist. Der in Adaptive READ verfolgte Ansatz zur Generierung benutzerspezifischer Zusammenfassungen geht von miteinander vernetzten Dokumenten aus. Das benutzeradaptive Verfahren zur Generierung von Zusammenfassungen verfolgt in einem Dokumentbrowser den Klickpfad eines Benutzers. Es arbeitet mit zwei Mengen von Dokumenten, den für den Benutzer auf einem Klickpfad liegenden relevanten Dokumenten D_r und den nicht relevanten Dokumenten D_n.
Hierbei werden zwei alternative Methoden zur Ermittlung der benutzerrelevanten Dokumente D_r eingesetzt:

- D_r = alle Dokumente, für die der Benutzer ähnliche Dokumente angefordert hat;
- D_r = alle Dokumente, die auf dem Benutzer-Klickpfad liegen und eine hohe Ähnlichkeit zum zusammenzufassenden Dokument d aufweisen.

Die Menge D_n ergibt sich jeweils aus den restlichen auf dem Klickpfad des Benutzers liegenden Dokumente. Die benutzerangepaßte Gewichtung gb_t eines Wortes t zur Generierung einer Zusammenfassung des Dokumentes $\vec{d}$ wird berechnet mit

$$gb_t(\vec{d}) = g_t(\vec{d}) + \frac{1}{|D_r|} \sum_{\vec{d'} \in D_r} g_t(\vec{d'}) - \frac{1}{|D_n|} \sum_{\vec{d'} \in D_n} g_t(\vec{d'}),$$

wobei $g_t(\vec{d'})$ die herkömmliche Gewichtung des Wortes t im Dokument $\vec{d'}$ bezeichnet. Die beschriebenen Arbeiten fließen nicht nur in Adaptive READ ein, sondern sind auch integraler Bestandteil des bei DaimlerChrysler entwickelten Wissensmanagement-Werzeugs WIR [3](Weaving Intranet Relations). Zur Zeit werden dort dem Benutzer bereits verschieden generierte Zusammenfassungen angeboten (statisch oder anhand des Klickpfades adaptiert). Eine der nächsten Aufgaben besteht darin, Zusammenfassungen für Dokumentmengen zu generieren.

4 Verbesserung der Suchmöglichkeiten durch Benutzerbeobachtung

Am DFKI wird an Ansätzen zum vollautomatischen Lernen aus der Benutzerinteraktion mit einer Volltext-Suchmaschine gearbeitet. Ziel ist es, den Suchprozeß eines neuen Benutzers durch die Beobachtung früherer Suchprozesse auch von anderen Benutzern zu unterstützen. Hierzu wird mitprotokolliert, welche Anfragen die Benutzer stellen und wie sie die vom System zurückgelieferten Antwortdokumente bewerten.
In einem ersten Ansatz wird aus Vereinfachungsgründen lediglich die Menge aller an ein System gestellten Anfragen sowie die Bewertung der entsprechenden Anfragen einbezogen. Insbesondere wird die Zuordnung der Anfragen zu einzelnen Benutzern und der Suchprozeß selbst außer Acht gelassen. Ziel ist in diesem vereinfachten Szenario lediglich die Verbesserung der Antworten auf eine Anfrage, die Unterstützung des Suchprozesses wird also ebenfalls zunächst außen vor gelassen.
Gegenwärtig werden 3 Möglichkeiten verfolgt, um in diesem Szenario die Beantwortung neuer Anfragen zu verbessern. Sei zu ihrer Darstellung $A(\vec{q})$ die korrekte Antwortmenge zur Anfrage $\vec{q}$, $\{\vec{q_1}, \ldots, \vec{q_n}\}$ die Menge bekannter Suchanfragen und $\vec{q^*}$ eine neue Anfrage. Die Verfahren arbeiten jeweils

- über die Ähnlichkeit der neuen Anfrage zu bekannten Anfragen; Ausgangspunkt ist hierbei die Annahme des folgenden Zusammenhang: $\vec{q^*} \sim \vec{q_i} \rightarrow A(\vec{q^*}) \sim A(\vec{q_i})$.
- über die Darstellung der neuen Anfrage als Linearkombination aus bekannten Anfragen; Annahme: $\vec{q^*} \approx \sum_{i=1}^{n} a_i \vec{q_i} \rightarrow A(\vec{q^*}) = f(a_1, A(\vec{q_1}), \ldots, a_n, A(\vec{q_n}))$ $(a_i \in I\!R)$.
- über die Dekomposition alter Anfragen in einzelne Suchbegriffe und der Zuordnung von Antwortdokumenten zu diesen Suchbegriffen; Annahme: $D_t = \bigcup_{i=1}^{n} \{A(\vec{q_i}) |\ t \text{ ist Suchbegriff in } q_i\} \rightarrow A(\vec{q^*}) = f(t_1, D_{t_1}, \ldots, t_m, D_{t_m})$.

Alle diese Ansätze erlauben theoretisch die Verbesserung von Anfragen, mit denen das Retrieval-System zum ersten Mal konfrontiert wird. Die bisher erhaltenen

experimentellen Ergebnisse sind uneinheitlich. Bei verschiedenen Problemstellungen wurden sowohl statistisch signifikante Verbesserungen als auch signifikante Verschlechterungen der Retrievaleffektivität beobachtet.
Zu den Problemen und zukünftigen Schwerpunkten zählen:

- Es sollen Erklärungen für die bisherigen uneinheitlichen Ergebnisse gefunden und mögliche Verbesserungsansätze der einzelnen Methoden identifiziert werden.
- Die für die Experimente verwendeten Anfragemengen sind untypisch für Real-World Suchsysteme. Um realistischere Anfragemengen zu bekommen, arbeiten wir zukünftig verstärkt mit einem Internet-Suchmaschinenanbieter zusammen.
- Wir gehen im Kontext von Suchmaschinen davon aus, daß von einem Benutzer eine explizite Bewertung von Dokumenten bezüglich seiner Anfragen nicht verlangt werden kann. Indirekte Methoden, wie beispielsweise über das Registrieren des Öffnens eines Antwortdokuments oder das Messen der Lesedauer eines Antwortdokuments führen zu Trainingsmaterial geringerer Qualität. Es bleibt zu prüfen, wie robust sich die einzelnen Methoden hier verhalten.
- Unser bisher betrachtetes Evaluierungsszenario ist beschränkt in der Art, daß Suchprozesse und individuelle Benutzervorlieben aus Vereinfachungsgründen bisher außer Acht gelassen wurden. Mittelfristig sollen auch diese einbezogen werden.

Am DFKI wird bei den obengenannten Arbeiten davon ausgegangen, daß zur Suche nach Dokumenten lediglich die Textinformation in den Dokumenten zur Verfügung steht. Das IAT fokussiert im Gegensatz hierzu auf die Unterstützung von Dokumentsuchen über Meta-Informationen wie Autor, Einstelldatum oder eine durch Klassifikation gegebene thematische Einordnung der Dokumente. Ziel der Arbeiten ist die Entwicklung eines Verfahrens zur Optimierung der sogenannten Meta-Strukturen mit Informationen, die aus dem Nutzerverhalten abgeleitet werden können. Eine automatisierte Anpassung der Meta-Strukturen in einem Dokumentenmanagement-System (DMS) ermöglicht beispielsweise eine Art Rapid-Prototyping beim erstmaligen Einsatz eines DMS, da die aufwendige manuelle Erhebung des Nutzerverhaltens vor der Implementierung des DMS in eine automatische Anpassung während der Nutzungsphase verlagert werden kann. Ein solches Vorgehen ist mit den heute auf dem Markt befindlichen Werkzeugen nicht umzusetzen, da hier in der Regel starre Meta-Strukturen anzutreffen sind.
In Adaptive READ wurde zunächst eine allgemeine Konzeption für ein System mit adaptiven Meta-Strukturen erarbeitet sowie die grundlegenden Verarbeitungsprozesse beschrieben. So werden die Suchbegriffe, mit denen die Anwender im System nach Informationen suchen, protokolliert, wobei zur Suche natürlich auch Volltext-Recherche genutzt werden kann. Diese Protokolle werden später für alle Benutzer ausgewertet und, beispielsweise unter Nutzung eines Thesaurus, nach signifikanten Suchtermen durchsucht. Aus diesen Termen wird anschließend ein Vorschlag für die Anpassung der Meta-Strukturen des Systems an das Informationsbedürfnis der Anwender generiert, mit dem schließlich das System angepaßt werden kann.

5 Zusammenfassung

In Adaptive READ beschäftigen wir uns unter anderem mit der Suche und Navigation in Dokumenten. Schwerpunktmäßig arbeiten wir an der Unterstützung der Informationssuche als Prozeß, der als Ganzes unterstützt werden soll und durch

den Einsatz geeigneter Lernverfahren verbessert werden soll. Wir haben in diesem Beitrag die Arbeitsschwerpunkte und den aktuellen Stand der Arbeiten zusammengefaßt. Das Projekt bzw. die entsprechenden Arbeitspakete laufen noch bis Ende 2002. Weitere Informationen zu Adaptive READ lassen sich der Webseite `www.adaptive-read.de` entnehmen.

Literatur

[1] M. Amini. Interactive learning for text summarization. In *Workshop on Machine Learning and Textual Information Access, PKDD*, 2000.

[2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Pub. Co., 1999.

[3] U. Bohnacker and A. Schorr. Finding logically connected documents in a large collection of files. In *IAWTIC 2001 - International Conference on Intelligent Agents, Web Technology and Internet Commerce*, 2001.

[4] J.S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence*, 1998.

[5] J. Carbonell and J. Goldstein. Diversity-based reranking for reordering documents and producing summaries. In *Proceed. SIGIR*, Melbourne, Australia, 1998.

[6] H. Kautz, B. Selman, and M. Shah. Referral web: Combining social networks and collaborative filtering. *Communications of the ACM*, 40(3), 1997.

[7] C.D. Manning and G.Schütze. *Foundations of Natural Language Processing*. MIT Press, 1999.

[8] G. Salton. *Automatic Text Processing; the Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Publishing Company, 1989.

[9] Gerard Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[10] `http://trec.nist.gov/`.