Inhaltsbasiertes Information Retrieval: Die TextMining- Technologie

Dr. Sebastian Göser, IBM Deutschland Entwicklung GmbH

1 Inhaltsbasiertes Retrieval

Die *TextMining-Technologie* befaßt sich mit Extraktion, Repräsentation, Verfügbarmachung und allgemein mit der Verwendung von Dokumentinhalten. Derartige Inhalte sind das, was Menschen verstehen wenn sie den Dokumenttext lesen. Ein inhaltsbasiertes Retrieval-System ist ein System, das inhaltsbezogene Informationsbedürfnisse von Benutzern auf der Basis großer Dokumentmengen zu erfüllen versucht. Derartiges Information Retrieval ,,im weiteren Sinn" berücksichtigt ebenso das Vorwissen von Benutzern wie den konzeptuellen Inhalt von Suchanfragen und Texten.

Es ermöglicht Menschen bestimmte Aktivitäten, wenn sie den Inhalt eines Textes verstehen. So können sie Schlußfolgerungen aus diesem Inhalt ziehen, den Text zusammenfassen, ihn einer Kategorie oder einem Arbeitsgebiet zuordnen, ihn kritisieren oder seine Relevanz bezüglich einer vorgegebenen Fragestellung beurteilen. Größere Dokumentmengen können inhaltlich geordnet werden. Es ist unklar, inwieweit Dokumentinhalte überhaupt isoliert von solchen Aktivitäten betrachtet werden können. Ein inhaltsbezogenes Retrieval-System wird jedenfalls verschiedene Arten des Umgangs mit Dokumentinhalten unterstützen. So gewährleistet die TextMining-Technologie die Klassifikation und Kategorisierung von Dokumenten, die automatische Extraktion von informationstragenden Konzepten, die Formulierung und Reformulierung von Suchanfragen sowie die Zusammenfassung längerer Dokumente. Da eine fundamentale Aktivität jedoch das Auffinden relevanter Dokumente ist, steht im Zentrum der Technologie natürlich die Entwicklung von hocheffektiven Retrievalverfahren.

Im folgenden werden die Unterschiede des inhaltsbezogenen Retrieval zu einer eher stichwortorientierten Dokumentsuche herausgearbeitet. Dazu wurden zunächst zwei wesentliche methodische Ansätze in TextMining, nämlich das hybride Retrievalmodell und die Extraktion von bedeutungstragenden Termen anhand von Beispielen vorgestellt, und danach verschiedene Retrieval-Szenarien daraufhin untersucht, inwieweit inhaltsbezogene Technologien nützlich für sie sind. Schließlich wird anband der Architektur des Systems TextMiner die Erweiterung der klassischen zu einer inhaltsbezogenen Retrievalkonzeption erläutert.

2 Hybride Suche

Es gibt verschiedene konzeptuelle Retrievalmodelle, unter ihnen Boole'sche Modelle (siehe Frakes und Baeza-Yates 1992) und probabilistische wie beispielsweise Inferenznetzwerke (Turtle und Croft 1991). Um unseren Ansatz des hybriden Retrieval zu motivieren, möchten wir zunächst auf die komplementären Eigenschaften dieser beiden Modelle hinweisen. Hinsichtlich Schnelligkeit der einzelnen Retrievaloperation und Transparenz der Querysprache sind Boole'sche Retrievalsysteme schwer zu übertreffen. Auch sind die Boole'schen Operationen AND und OR durchaus intuitiv mit Bezug auf die de Saussure'sche Unterscheidung von paradigmatischen und syntagmatischen lexikalischen Relationen (Hull und Grefenstette 1996).

Andererseits gelten Boo1e'sche Retrievalsysteme als ineffektiv insbesondere bei Benutzung durch unerfahrene Sucher (siehe Belkin und Croft 1987). Im TRECKontext, wo Retrievalsysteme ausschließlich unter Effektivitätgesichtspunkten evaluiert werden, spielen Boole'sche Modelle keine Rolle. Andererseits haben Retrievalsysteme auf der Basis von Inferenznetzwerken für viele Dokumentkollektionen, auch in mehreren TRECs, sehr gute Effektivität gezeigt. Die empirischen Ergebnisse stimmen überein mit den vorteilhaften theoretischen Eigenschaften von Inferenznetzen.

Das hybride Retrievalmodell integriert ein Boole'sches und ein probabilistisches Retrievalmodell mit Bezug auf Querysprache und Dokumentrepräsentation. Das Boole'sche Retrievalmodell (siehe Programmer's Guide 1996) unterstützt neben Boole' sehen Operatoren verschiedene Adjazenzkonstrukte und linguistische Termexpansionen. Das probabilistische Modell (Maarek und Smajda 1989) kann als Spezialisierung eines Bayes'schen Inferenznetzwerk-Modells auf eine operatorfreie Freitext-Retrievalsprache betrachtet werden. Der Boole'sche Teil einer hybriden Suchanfrage wird als Filter der probabilistischen Suche vorgeschaltet, so daß die Schätzung der Relevanzwahrscheinlichkeit auf genau den Dokumenten basiert, die diese Boole'sche Anfrage erfüllen. Durch die kleinere Kandidatenmenge für das Schätzverfahren können erhebliche Performanzverbesserungen realisiert werden.

Inhaltlich ermöglicht das hybride Retrievalmodell Suchanfragen wie z.B.

- FREE (management in Japanese auto* industry)
- AND NOT(SYN(Toyota))

Diese Anfrage liefert nur solche Dokumente zurück, in denen nicht" Toyota" oder eines seiner Synonyme vorkommt. Lange Freitext-Argumente sind besonders wichtig angesichts der notorischen Ineffektivität "armer" Suchanfragen, wie sie beispielsweise für das World Wide Web charakteristisch sind.

3 Extraktion

Die partielle Extraktion von Dokumentinhalten gehört zum Kern des inhaltsbezogenen Retrievalansatzes. TextMining extrahiert und repräsentiert Teile der in Dokumenten "enthaltenen" Konzepte, die als Indikatoren des Dokumentinhalts betrachtet werden. Bekanntlich konnte der Nachweis verbesserter Retrievaleffektivität (im Sinne von TREC) bislang noch für keine der vielen Extraktionstechnologien erbracht werden. Unumstritten sind aber Extraktionsverfahren äußerst nützlich in vielen Benutzeraufgaben, die derzeit in Retrievalsituationen gelöst werden (siehe etwa Byrd, Ravin und Prager 1995).

Das TextMiner-System extrahiert Eigennamen, Mehrwortausdrücke, deren reguläre Abkürzungen und verschiedene numerische Informationen wie z.B. Daten, Preise und Netzadressen. Die Extraktion basiert auf einer "reichen" Tokenisierung und einer lexikalischen, insbesondere morphologischen Vorverarbeitung. Sie erfordert keinerlei Customisierung. Die Extraktionsverfahren sind derzeit auf amerikanisches und britisches Englisch beschränkt, aber grundsätzlich auf beliebige alphabetische Sprachen anwendbar. Im folgenden werden die Verfahren detaillierter beschrieben.

Die verschiedenen Extraktionsmethoden sind drei stufige Verfahren bestehend aus Patternmatching, der dokumentbezogenen Verarbeitung, und der Aggregierung. Für jeden Termtyp gibt es spezielle Mengen von Patterns (reguläre Ausdrücke), die deterministisch mit der Eingabe-Tokenliste abgeglichen werden. Ein Pattern für einen Organisationsnamen ist beispielsweise "<lnitiaLUpper Case>+ <Inc\>+ ", das einen Namen wie "Acme Software Development mc." characterisiert. Das Pattern "<Mr.> <FirstName> * <lnitial_UpperCase>+" würde einen Personennamen wie "Mr. Bill Clinton" treffen. Mehrwortausdrücke sind im wesentlichen noun compounds, die durch Wortart-Patterns wie "<Det> <Noun> +" (für z.B. "a water pipeline") charakterisiert werden (siehe Justeson und Katz 1994). Die Patterns enthalten den eigentlichen Term und charakterisierende Zusätze wie Mr., Dr., Inc., oder ein Artikelwort, wobei diese Zusätze häufig eine starke klassifizierende Evidenz liefern. Jeder erkannte Termkandidat erhält eine Typkennung.

Bei der dokumentbezogenen Verarbeitung werden gleichbedeutende bzw. koreferentielle Terme und teilweise deren Abkürzungen in Äquivalenzklassen zusammengefaßt. Jede Äquivalenzklasse wird durch einen kanonischen Term repräsentiert. Eine Häufigkeitsschwelle schließt zufällige Mehrwortterme aus. Die hier angewendeten Heuristiken basieren zumeist auf textlinguistischen Beobachtungen. Beispielsweise ist die erste Textreferenz auf eine Person, mit der sie als Diskursobjekt etabliert wird, zumeist reichhaltiger als die nachfolgenden, so z.B.: "President William Clinton "" Mr. Clinton ...".

- Die Aggregierung schließlich operiert über einer Menge von Dokumenten. Sie schließt Duplikate aus, also beispielsweise Terme, die bis auf Klein/Großschreibung gleich sind oder durch verschiedene Extraktoren gefunden wurden. Wiederum werden Häufigkeitsschwellen angewendet. Verschiedene kanonische Terme für dieselbe Termvariante führen zur Neuzuordnung des größeren kanonischen Terms.

Eine vorläufige Evaluation der Eigennamen-Extraktion anband von 88 Artikeln des Wall Street Journal, die das System zuvor noch nie gesehen hatte, erbrachte einen Recall (Ausbeute) von 97% bei einer Precision (Trefferquote) von 91 %. Weitere Evaluationen sind in Vorbereitung.

4 Szenarien

Inhaltsbezogene Retrievalsituationen sind dadurch charakterisiert, daß Benutzer unter Einsatz ihres Vorwissens mit Informationen aus größeren Dokumentkollektionen arbeiten (siehe Hersh 1996). Diese Kollektionen können im Sinne der Skalierbarkeit (siehe unten) persönlich, organisationsweit, wie z.B. eine Bereichsbibliothek, oder weltweit wie das World Wide Web sein. Informationen aus Kollektionen können Dokumente oder Teile davon, z.B. nur die Titel, sein, aber auch inhaltscharakterisierende Schlagwörter oder bestimmte, in einem Dokument behauptete Aussagen. Generell kann zwischen intensionaler, thematischer Information, die den Inhalt eines Dokuments beschreibt, und extensionaler, wahrheitsfunktionaler Information unterschieden werden. Ein Medizinwissenschaftler, der sich in ein neues Gebiet einarbeitet, wird eher thematisch suchen, während ein Kliniker, der eine Information über die Verträglichkeit zweier Medikamente sucht, eher faktenorientierte Strategien verwenden wird (Hersh 1996). In vielen Fällen dürfte ein inhaltsorientierter Retrievalprozeß mit ,,klassischen" Suchoperationen beginnen, die ein geeignetes Dokumentinventar bereitstellen.

Der Vortrag stellte Szenarien aus verschiedenen Fachbereichen vor und versuchte zu zeigen, daß konventionelle Retrievalstrategien hier weniger erfolgreich sein werden als inhaltsbezogene.

5 Architektur

TextMiner ist ein objektorientiertes TextMining-System. Seine Architektur unterscheidet eine Anwendungsschicht, die Schnittstellen zu verschiedensten Anwendungen darstellt, von den Schichten für die zugrundeliegende Technologie und für domänenabhängige Resourcen. Die Technologieschicht, die u.a. Such

maschinen und Extraktoren enthält, ist im wesentlichen unabhängig von speziellen Anwendungen oder inhaltlichen Domänen. Anwendungs- und Technologieschicht können installationsabhängig durch eine Client/Server- Interaktion verbunden werden.

Inhalte sind in der Anwendungsschicht als sogenannte Zielobjekte (targets) repräsentiert. Zielobjekte können beispielsweise eine "klassische" Resultatliste mit Relevanzwerten, die thematischen Terme eines Dokuments, oder bestimmte daraus extrahierte Fakten sein. Zielobjekte können auch selbst Texte sein, beispielsweise die Zusammenfassung eines Dokuments. Sie unterliegen einer partiellen Subsumptionsordnung, so daß mit denselben Klassenkonzepten Objekte zunehmender Komplexität repräsentiert werden können. Insgesamt ist die Architektur von TextMiner darauf ausgelegt, verschiedensten Applikationen in sehr flexibler Weise Zugang zu Dokumentinhalten zu verschaffen. Zahlreiche positive Erfahrungen in Kundenprojekten liegen inzwischen vor und bieten Anregung für weitere Entwicklungen.

TextMiner ist Teil des MediaMiner Lösungsangebots und kann damit (ab 19.12.1996) unter *http://service.software.ibm.com* als Beta-level code heruntergeladen werden.

Literatur

Belkin and Croft: Retrieval Techniques, in: Annual Review of Information Science and Technology, M. Williams, ed., New York, Elsevier 1987

Roy Byrd, Yael Ravin and John Prager: Lexical Assistance at the Information Retrieval User Interface, in: Proceedings 4th Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, Nevada, April 1995

Frakes, W., Baeza-Yates R.: Information Retrieval - Data Structures and Algorithms, Prentice-Halll992

Hersh, W.: Information Retrieval: A Health Care Perspective, Springer 1996 Hull, D. and Grefenstette, G.: Querying Across Languages: A Dictionary- Based Approach to Multilingual Information Retrieval, in: Proceedings of SIGIR-96, Konstanz, Hartung-Gorre 1996

Justeson, J.S. and Katz, S.M.: Technical terrninology: some linguistic properties and an algorithm for identification in text, Internal Report, mM Research

Maarek, Yand Smajda, F.: Software Library Construction from an IR Perspective, in: Proceedings of SIGIR 1991 TextMiner Programmer's Guide, mM Document Number SHI2-5987-00, 1996 Turtle, H. and Croft, B.: Evaluation of an Inference Network-Based Retrieval Model, in: ACM Transactions on Information Systems, vol. 9, no. 3, pp. 187-222