

Neue Entwicklungen im Informationsretrieval

**Zusammenfassung der Beiträge der Herbstschule der
Gesellschaft für Informatik in Schwerte**

Jörg Schlegel

Das konventionelle *Information Retrieval* (IR) setzt eine intellektuell aufbereitete Datenbasis voraus. Die Information Retrieval-Systeme sind in der Regel Boolesche Modelle. Für Suchanfragen, die entsprechend der Booleschen Logik gebildet sind, werden alle relevanten Dokumente gefunden.

Die intellektuelle Aufbereitung von Dokumenten ist aufwendig und teuer und deshalb bei der steigenden Zahl von Veröffentlichungen nicht auf Dauer realisierbar. Dafür müssen künftig maschinelle Verfahren eingesetzt werden. Dies erfordert neuartige Systeme für die Erschließung und das Wiederauffinden, die möglichst der Qualität der bisherigen Systeme entsprechen.

Auf der erstmals durchgeführten Herbstschule für das Information Retrieval, vom 27.9. -2.10.1998 veranstaltet von der Gesellschaft für Informatik, wurden neue Erkenntnisse zu den verschiedenen Retrieval-Verfahren vorgestellt.

Linguistische Methoden beziehen sich auf die Sprachlichkeit der Dokumente und Suchanfragen. Sowohl die Dokumente als auch die Queries werden vor einem Abgleich (Matching) mit linguistischen Methoden bearbeitet. Mittels morphologischer Verfahren können zunächst die Grundformen der Worte gebildet und aus selektierten Begriffen Ausdrücke zusammengesetzt werden. Erkenntnisse über die Wichtung der Begriffe liefern syntaktische Regeln. Der Vorgang der Indexierung wird durch eine Vielzahl von linguistischen Ressourcen unterstützt, dazu gehören Wörterbücher, Stopwortlisten, Thesauri, Synonymwörterbücher und Autorendateien. Auf Basis dieser Verfahren kann die Datenbasis auch kategorisiert bzw. Abstracts können automatisiert erstellt werden. Linguistische Methoden sind für das Information Retrieval heute noch sehr umstritten: Gegenwärtig fehlen noch entsprechende Evaluierungsmethoden, über die eine Effektivität nachgewiesen und eine Vergleichbarkeit mit anderen Systemen ermöglicht würde.

Mehrsprachiges Information Retrieval hat zum Ziel, unabhängig davon, in welcher Sprache die Suchanfrage formuliert ist, die relevanten Dokumente aus einem mehrsprachigen Datenfonds zu extrahieren. Für ein sprachübergreifendes Information Retrieval gibt es inzwischen erfolgversprechende Experimente. Die Notwendigkeit solcher Verfahren ist in der Tatsache begründet, daß im Internet der Zuwachs von nicht-englischsprachigen Dokumenten erheblich größer ist als der Zuwachs von englischsprachigen Dokumenten. Beim mehrsprachigen Information Retrieval wird sowohl die Suchanfrage als auch der Datenfonds unter Beachtung linguistischer Regeln in verschiedene Sprachen übersetzt. Für die automatisierte Spracherkennung gibt es inzwischen etablierte Verfahren. Das Extrahieren von Indexlisten erfolgt in ähnlicher Weise wie beim linguistischen Information Retrieval. Aus den Dokumenten werden Zeichenketten herausgelöst, diakritische Zeichen entfernt, eine Wortzerlegung und Normierung vorgenommen, sinntragende Begriffe und Zahlen bestimmt sowie Eigennamen erkannt und behandelt. Bei den mehrsprachigen Suchverfahren hat sich die Anwendung von Relevanz-Feedbackverfahren für die Verbesserung des Results als sinnvoll erwiesen. Bei Tests mit Prototypen solcher Systeme wurden ca. 27% der relevanten Dokumente gefunden. Diese Trefferzahl ist relativ gut, wenn berücksichtigt wird, daß bisher bei automatisierten Verfahren das Auffinden von 50-60% der relevanten Dokumente als Bestwert gilt.

Andere Ansätze gehen vom Einsatz von *Expertensystemen für das Information Retrieval* aus. Bereits in den 80er Jahren gab es Versuche, die Tätigkeit von Informationsvermittlern durch wissensbasierte Systeme zu unterstützen. Gegenwärtig werden die gewonnenen Erkenntnisse zur Entwicklung „autonomer Informationsagenten“ herangezogen, die selbständig in vernetzten Informati-

onssystemen agieren sollen. Grundlage für derartige Verfahren ist die Modellierung des semantischen und pragmatischen Wissens für ausgewählte Fachgebiete. Das Information Retrieval wird auf der Basis von komplexen kognitiven Interaktionsmodellen in Kombination mit automatisierten Präsentations-techniken durchgeführt.

Alle bisher genannten Techniken bezogen sich weitgehend auf das Textretrieval. Zunehmend spielen in Veröffentlichungen nicht-textuelle Daten in Form von Bildern, Audiosignalen und Videos eine Rolle. Auf das Information Retrieval bezogen bedeutet dies, daß mehrdimensionale Beziehungen gespeichert und für das Wiederauffinden aufbereitet werden müssen. Bei der Indexierung sind also neben dem Inhalt und der logischen Struktur auch Angaben zum Layout zu berücksichtigen. Zur Beschreibung von Dokumenten sind somit zusätzliche Attribute erforderlich. Ziel aller *Multimedia Information Retrieval-Modelle* ist eine weitgehend maschinelle Aufbereitung der Dokumente.

Im Audiobereich ist es möglich, für zeitlich begrenzte Intervalle charakteristische Merkmale wie Lautstärke, Tonhöhe, Helligkeit, Bandbreite und Harmonizität mit nachrichtentechnischen Verfahren zu bestimmen und in Form von numerischen Ausdrücken zu speichern. Die ermittelten Sequenzen bilden die Grundlage für die Indizierung. Beim Retrieval kann somit nach ähnlichen Klängen gesucht werden.

Beim Sprachretrieval sind Spracherkennungsmethoden zur Identifizierung von Inhalten vorgeschaltet. Die Suche in der Datenbasis erfolgt dann mit üblichen Textretrieval-Methoden.

Im Bereich der Musik ist es nötig, die Melodien nach Methoden der „melody transcription“ zu bearbeiten und das Ergebnis als Folge von Noten zu speichern. Beim Retrieval werden Ähnlichkeitsbetrachtungen von Strings (in diesem Fall Noten) vorgenommen.

Für Bilder werden gegenwärtig verschiedene Verfahren zur automatischen Indexierung untersucht. In den meisten Fällen wird die Syntax des Bildaufbaus (Farbe, Oberfläche und Kontur) für eine Indexbildung benutzt. Für die Beschreibung von Objekten und Themen haben sich bisher noch keine automatisierten Verfahren als anwendbar erwiesen. Da sich aber 50% aller Anfragen an Bildarchive gerade auf diesen Bereich beziehen, kann man hier noch nicht auf eine intellektuelle Aufbereitung verzichten.

Parallel zur Untersuchung neuartiger Modelle für das Information Retrieval ist es notwendig, neue *Methoden für Benutzung und die Ergebnisdarstellung* einzusetzen. Obwohl sich graphische Benutzeroberflächen bereits für die PC-Benutzung durchgesetzt haben, arbeiten die meisten Retrievalsysteme noch kommandoorientiert. Zahlreiche Ansätze zeigen, daß auch für die Aufgaben

des Information Retrieval nach ergonomischen Prinzipien gestaltete Benutzeroberflächen und eine Visualisierung der Ergebnisse sinnvoll und möglich sind. Erste Schritte dazu sind in der Gestaltung der WWW-Schnittstellen der klassischen Hosts der elektronischen Fachinformation zu erkennen. Die Verlinkung von Ergebnis und Dokument und die Möglichkeit des Browsens in Registern und Inhaltsverzeichnissen stellt einen ersten Schritt zur Verbesserung dar. Die angewendeten Prinzipien gehen jedoch alle auf Methoden aus der Buchgestaltung zurück. Unter Beachtung der Regeln für die Softwareergonomie sowie die Berücksichtigung von Designergesichtspunkten lassen sich besser zu bedienende und leicht verständliche Oberflächen gestalten. Ein besonderes Anliegen für Anwendungen im Information Retrieval ist es, die Boolesche Logik so transparent zu machen, daß auch der ungeübte Endnutzer solche Systeme erfolgreich anwenden kann. Gleichzeitig wird angestrebt, Ergebnisse nach Ranking geordnet anschaulich zu präsentieren.

Bevor die aufgezeigten Methoden, Erkenntnisse und Entwicklungen in praktisch nutzbare Werkzeuge umgesetzt werden können, ist eine kritische Bewertung und realitätsnahe Evaluierung erforderlich. Für die meisten der genannten Verfahren hat sich noch kein festes Methodeninventar zur Qualitätsbestimmung etabliert. Doch es gibt eine Reihe von Kriterien, nach denen die Systeme bewertet und vergleichbar gemacht werden können. Um fundierte Qualitätsaussagen zu erhalten und eine Vergleichbarkeit der Systeme zu ermöglichen, ist die Arbeit mit großen Testkollektionen erforderlich. Solche werden durch TREC (Text Retrieval Conference) in den USA und künftig GIRT (German Information Retrieval Test) in Deutschland bereitgestellt.

Die hier in einer kurzen Zusammenfassung dargestellte Thematik wurde von folgenden Referenten während der Herbstschule ausführlich behandelt:

Dr. Marc Rittberger (Uni Konstanz): Einführung in das IR und Internet-Suche

Prof. Dr. Reinhard Schramm (TU Ilmenau): Klassische Methoden der Inhaltserschließung und IR am Beispiel der Patentinformation

Dr. Sebastian Goeser (IBM Deutschland): Linguistische Methoden im IR

Thomas Rölleke (Uni Dortmund): Retrievalmodelle

Prof. Dr. Fuhr (Uni Dortmund): Multimedia Retrieval

Prof. Dr. Jürgen Krause (IZ): Benutzeroberflächen und Visualisierung im IR

Prof. Dr. Peter Schäuble (ETH Zürich): Mehrsprachige Informationssuche

Dr. Ulrich Thiel (GMD Darmstadt): Kognitive Ansätze des Intelligenzen IR

Prof. Dr. Christa Womser-Hacker (Uni Konstanz): Evaluierung von IR-Systemen.

