Zeitschrift: Arbido
Herausgeber: Verein Schweizerischer Archivarinnen und Archivare; Bibliothek Information Schweiz
Band: 17 (2002)
Heft: 10

Artikel: Die Zukunft der Informationssuche : Information nutzbar machen
Autor: Braschler, Martin / Schäuble, Peter
DOI: https://doi.org/10.5169/seals-768770

Die Zukunft der Informationssuche Information nutzbar machen

Martin Braschler
Head of Research

Peter Schäuble
CEO
Eurospider Information Technology AG, Zürich

1. Einleitung

Die Disziplin des *«Information Retrieval»* steht heute vor einem Umbruch. Entstanden ist das Fachgebiet vor mehr als 40 Jahren mit dem Ziel, *Dokumente automatisch nach inhaltlichen Kriterien zugreifbar zu machen.* Die anfänglich vorwiegend akademische Disziplin des Information Retrieval, dessen grundlegende Konzepte in den frühen 1960er-Jahren ihre Wurzeln haben, hat nun aber mit dem Übergang ins Informationszeitalter stark an kommerzieller Bedeutung gewonnen.

Die Verfügbarkeit geschäftskritischer Informationen hat aus verschiedenen Gründen signifikant zugenommen (World Wide Web, Mail, CRM- und ERP-Systeme). Hinzu kommt eine noch schneller wachsende Datenflut, die das Auffinden relevanter Information erschwert. Gleichzeitig ergaben sich neue, vielfältige Einsatzmöglichkeiten für Information-Retrieval-Techniken, die den Bedarf für entsprechende Systeme geweckt haben.

Klassisch liegt den Systemen und Techniken des Information Retrievals die Idee zugrunde, einen Informationssuchenden bei der Zusammenstellung eines Dossiers mit geeigneter Information zu unterstützen. Hierzu formuliert der Benutzer eine mehr oder weniger ausgefeilte Anfrage, auf welche das System ihm mit einer Liste geeigneter Objekte (Dokumente, Bücher etc.) antwortet. Verbreitung fanden solche Systeme unter anderem im Umfeld von Bibliotheken.

Diese Situation hat sich in den letzten Jahren grundlegend geändert: Information-Retrieval-Technologie soll heute für ein sehr grosses Spektrum an Aufgaben eingesetzt werden, wie

- Transaktionen auslösen (günstiges Ferienangebot buchen, SW kaufen etc.)
- Decision Support (Marktinformationen bereitstellen, Compliance Management etc.)
- Fakten Retrieval (Kontaktinformationen, Spezifikationen etc.)
- Interaktionen (E-Government, Online Bewerbungen etc.)
- Kooperation (Projekte, virtuelle Unternehmen etc.)

Es geht also heute nicht nur darum, nützliche Informationen bereitzustellen, sondern diese auch *personalisiert* und *aufgabenspezifisch* anzubieten.

Diese Erweiterung der Anforderungen führt zu Systemen, welche den Benutzer führen und unterstützen, sei es durch Dialog oder Aufbereitung der Information zwecks zielgerichteter Präsentation.

Ein Information-Retrieval-System besteht traditionell aus zwei Hauptkomponenten: Inhaltserschliessung (Indexing) und Vergleichen (Matching). Bei der *Erschliessung von Informationsobjekten* geht es darum, diese in eine Form zu bringen, die einen effektiven Vergleich ermöglicht. Dies ist deshalb erforderlich, weil Informationsobjekte a priori nicht direkt vergleichbar sind, zum Beispiel aufgrund der vielfältigen Möglichkeiten, in natürlicher Sprache denselben Sachverhalt zu beschreiben.

Die Erschliessung ermöglicht:

- Das Einbinden unterschiedlicher Informationsobjekte, wie Dokumente, Suchanfragen, Benutzer- oder Interessensprofile, Themen einer Taxonomie u. a.
- Das Einbinden verschiedener, heterogener Quellen.
- Die Extraktion eines Maximums an Information aus diesen Objekten für den Einsatz in späteren Vergleichen.

Beim *Vergleich von Informationsobjekten* soll ermittelt werden, ob eine inhaltliche (semantische) Beziehung zwischen den Objekten besteht. Das Resultat des Vergleichs kann ein Ja/Nein-Entscheid sein, aber auch ein numerischer Wert, welcher die Stärke der semantischen Beziehung ausdrückt. Das Vergleichen von Informationsobjekten kann für verschiedenste Zwecke genutzt werden:

- Der Vergleich einer Suchanfrage und eines Dokumentes soll ergeben, ob das Dokument zu der Anfrage relevant ist und demzufolge als Bestandteil des Suchresultates dem Benutzer präsentiert werden soll.
- Der Vergleich eines Benutzer-/Interessensprofils und eines Dokumentes soll ergeben, ob das Dokument an den Benutzer weitergeleitet werden soll.
- Der Vergleich eines Themas einer Taxonomie (Themenkataloges) und eines Dokumentes soll ergeben, ob das The-

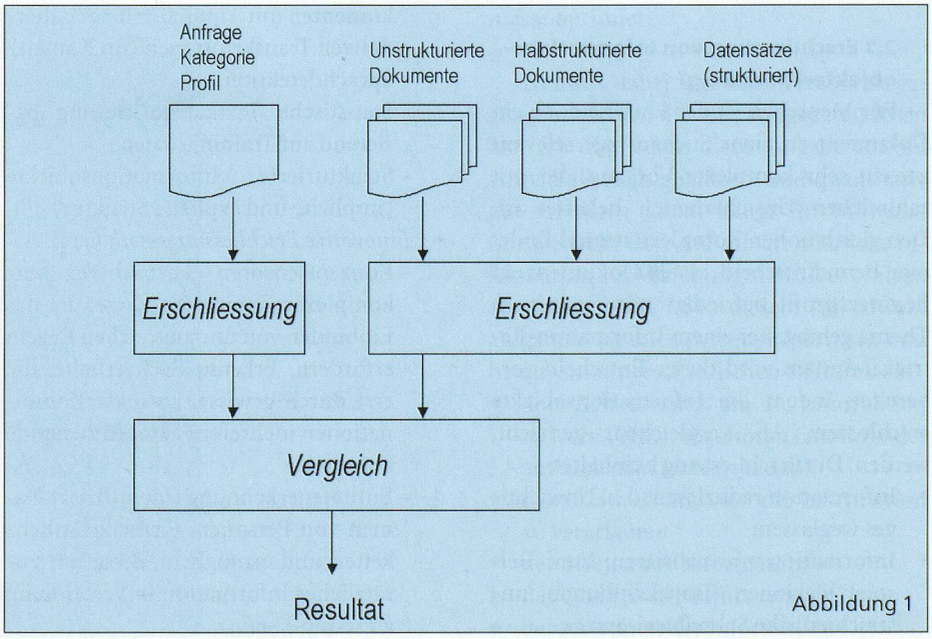

Abbildung 1

ma dem Dokument zugeordnet werden soll. Im Falle eines Katalogs wird dann das Dokument in der entsprechenden Kategorie gelistet.

In Zukunft werden im oben umrissenen, verallgemeinerten Bedürfnisrahmen zwei weitere Komponenten eine entscheidende Rolle spielen: Benutzerunterstützung (Guidance) und Informationsdarstellung (Information Presentation). Die Benutzerunterstützung leitet die Benutzer bei der Interaktion mit dem Retrieval-System so, dass sie im nächsten Schritt eine Aktion wählen, die einen möglichst hohen Nutzen bringt. Nach Ausführung der Aktion (z.B. Suche modifizieren, Suchresultat übersetzen und zusammenfassen lassen oder Relevanzrückkoppelung), muss die Information so dargestellt werden, dass wiederum ein grösstmöglicher Nutzen resultiert.

Im Folgenden werden zuerst die beiden traditionellen Hauptkomponenten Erschliessung und Vergleich näher detailliert. Darauf aufbauend werden dann die Benutzerunterstützung und Präsentation besprochen, und es wird dargelegt, wie ein modernes Information-Retrieval-System den erweiterten Bedürfnissen heutiger Benutzer gerecht wird.

2. Erschliessung und Vergleich

Für das Erschliessen und den Vergleich von Informationsobjekten wurden verschiedenste Methoden entwickelt. Information-Retrieval-Systeme können unterschieden werden anhand der Nutzung von einfachen, fortgeschrittenen oder innovativen Methoden für die Erschliessung und das Vergleichen.

2.1 Erschliessung von Informationsobjekten (Indexing)

Für Menschen ist der Entscheid, ob ein Dokument zu einer Suchanfrage relevant ist, ein sehr komplexer Vorgang, der mit zahlreichen Unsicherheiten behaftet ist. Den gleich hohen Komplexitätsgrad findet man beim Entscheid, ob ein Dokument ein Benutzerprofil befriedigt oder zu einem Thema gehört. Bei einem Information-Retrieval-System wird dieser Entscheid vorbereitet, indem die Informationsobjekte erschlossen, d.h. vergleichbar gemacht, werden. Die Erschliessung beinhaltet:

- Information reduzieren, d.h. Unwichtiges weglassen;
- Information normalisieren, zum Beispiel Flexionen, Transkriptionen, unterschiedliche Schreibweisen;

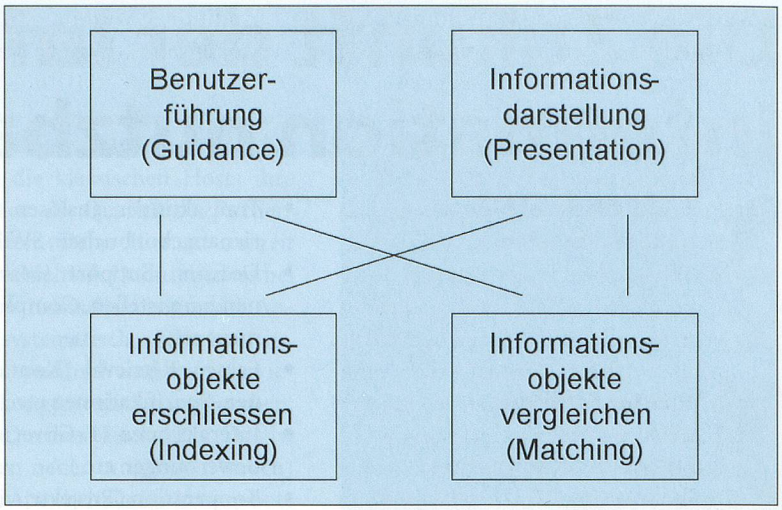

Abbildung 2

- Information anreichern: Themen zuordnen oder Entitäten erkennen.

Es hat sich als nützlich erwiesen, zwischen verschiedenen Graden an Komplexität betreffs Erschliessungsmethoden zu unterscheiden.

- *Einfache Erschliessungsmethoden:*
  - Konversion von Dokumentenformaten
  - Konversion von Zeichensätzen
  - Wortsegmentierung
  - Stoppwort-Elimination
  - Informationsobjekte parsen
- *Fortgeschrittene Erschliessungsmethoden:*
  - Wortnormalisierung (Finden verschiedener Flexionen eines Wortes)
  - Kompositazerlegung (Zerlegung zusammengesetzter Nomina im Deutschen)
  - N-Gramme (fehlertolerante Erschliessung durch Zerlegung der Wörter in kleinere, mehrbuchstabige Einheiten)
  - Fuzzy Matching (Auffinden von Dokumenten mit Tippfehlern oder alternativen Transkriptionen von Namen)
  - Sprachdetektion
  - Statistische Textkategorisierung basierend auf Trainingsdaten
  - Strukturierte Informationsobjekte (implizite und explizite Struktur)
- *Innovative Erschliessungsmethoden:*
  - Konzeptsensoren (Formulieren sehr komplexer Zusammenhänge, die das Einbinden von umfangreichen Regeln erfordern. Erkennt Sachverhalte, die erst durch gewisse, korrekte Kombinationen mehrerer Faktoren zustande kommen.)
  - Entitätenerkennung (identifiziert Namen von Personen, Firmen, Örtlichkeiten und ermöglicht, diese mit zusätzlicher Information in Verbindung zu setzen.)
  - Nominalphrasenextraktion (Phrasen, d.h. Mehrwortbegriffe, werden erkannt, und als Einheit weiterverarbeitet.)

2.2 Vergleich von Informationsobjekten

Nach der Erschliessung müssen die Informationsobjekte verglichen werden, um Suchanfragen, Benutzer-/Interessensprofile und Themen einer Taxonomie mit relevanter Information zu verknüpfen. Solche Verknüpfungen sind mit zahlreichen Unsicherheiten verbunden. Ob ein Dokument ein Benutzerinteresse befriedigt, hängt unter anderem davon ab, welches fachspezifische Wissen vorausgesetzt wird. Ein Spezialist kann anders als ein Laie entscheiden. Ein Information-Retrieval-System ist somit mit einem «unlösbaren» Problem konfrontiert, welches in möglichst vielen Fällen möglichst gut zu lösen ist. Im Folgenden sind die verschiedenen Methoden für den Vergleich von Informationsobjekten kurz beschrieben.

- *Einfache Vergleichsmethoden für Informationsobjekte:*
  - Boole'sches Retrieval (AND, OR, NOT)
  - Coordination Level Matching (Rangliste wird gemäss der Anzahl der gefundenen Suchbegriffe geordnet.)
- *Fortgeschrittene Vergleichsmethoden für Informationsobjekte:*
  - probabilistisch (Sortierung aufgrund von geschätzter Relevanz)
  - regelbasiert (Sortierung mit Hilfe einer Menge von Regeln)
  - Relevanzrückkoppelung (Benutzer können Suchresultate auf Relevanz bewerten, worauf das System automatisiert die Anfrage weiter verfeinert und bessere Suchresultate liefert.)

- Anfrageerweiterung (Automatische Anfrageerweiterung erweitert Suchanfragen um verwandte Begriffe und hilft damit, vollständigere Suchergebnisse zu erreichen.)
- Metadaten
- Subkollektionen (ein- und ausblenden)
- Duplikatelimination (Objekte in exakt oder beinahe identischer Form)
- Zugriffsbeschränkung (Unberechtigter Benutzer darf nicht durch eine «Hintertür» Kenntnis von oder gar Zugriff auf geheime Information erlangen.)

- *Innovative Vergleichsmethoden für Informationsobjekte:*
  - Sprachübergreifend (Organisationen und Unternehmen müssen heute zunehmend Dokumentenkollektionen erschliessen, welche Objekte in vielen verschiedenen Sprachen enthalten. Auf solche Kollektionen muss effizient mit nur einer Anfrage, in der vom Benutzer bevorzugten Sprache formuliert, zugegriffen werden können.)
  - Passagen-Retrieval (In längeren Informationsobjekten sind oft nur kurze Abschnitte relevant. Das System muss solche Abschnitte identifizieren können und entsprechend gewichten.)

Ein modernes Information-Retrieval-System muss in der Lage sein, diese Erschliessungs- und Vergleichsmethoden flexibel zu kombinieren, um damit der wachsenden Bandbreite von Bedürfnissen gerecht zu werden.

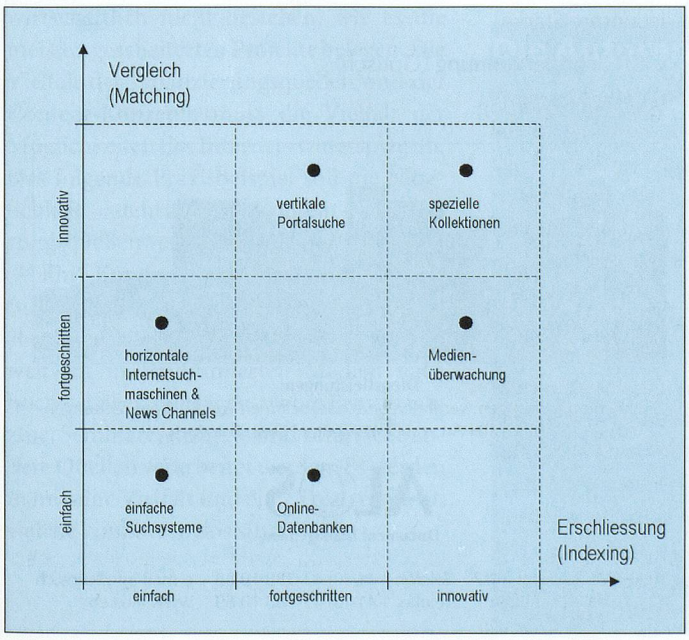

Abbildung 3

Traditionell wird die Effektivität dieser Methoden objektiv gemessen mithilfe der Masse *Ausbeute* und *Präzision*. Ausbeute drückt aus, welcher Anteil der gewünschten Information gefunden wird, während Präzision misst, wie viel der gefundenen Information relevant ist.

Es folgen einige Beispiele für die Kombination von spezifischen Methoden zur Erschliessung und zum Vergleich.

Beispiel 1: News Channels

Typische News Portale mit lernfähigen Channels basieren auf einer *einfachen* Erschliessung, ohne Sprachanalyse und Dokumentenstrukturen zu nutzen. Die lernfähigen Channels von derartigen News Portalen basieren auf *fortgeschrittenen* statistischen Methoden für den Vergleich (z.B. Bayes'sche Klassifizierer oder kNN), welche anhand von Trainingsbeispielen entscheiden, ob ein neues Dokument in einen Channel eingespiesen wird.

Beispiel 2: (Horizontale) Internet-Suchmaschinen

Internet-Suchmaschinen erschliessen eine sehr grosse Zahl extrem heterogener Webseiten, so dass auch die besten Suchmaschinen *einfache* Erschliessungsmethoden anwenden. Dafür kommen *fortgeschrittene* Vergleichsmethoden zum Einsatz, welche beispielsweise die Hyperlink-Struktur des Web berücksichtigen.

Beispiel 3: Spezielle Dokumentenkollektionen

Spezielle themenspezifische und homogene Dokumentenkollektionen erfordern eine *komplexe* Erschliessung mit der automatischen Erkennung von Referenzen/Metadaten und mit Einbezug des speziellen Fachvokabulars. Gleichzeitig ist auch eine *komplexe* Vergleichsmethode erforderlich, um z.B. einen sprachübergreifenden Zugriff auf relevante Dokumente zu gewährleisten, welche in einer anderen Sprache als die Suchanfrage formuliert sind.

Die Grafik unten links illustriert die Position der erwähnten und weiterer Beispiele.

3. Benutzerunterstützung und Präsentation

Das Informationszeitalter stellt neue und höhere Ansprüche an Information-Retrieval-Systeme, welche nicht mehr ausschliesslich durch die beiden traditionellen Komponenten Erschliessung und Vergleich befriedigt werden. In Fachkreisen werden neue Komponenten wie maschinelles Übersetzen, Informationsvisualisierung oder Bild- und Spracherkennung intensiv diskutiert. Der Nutzen dieser neuen Komponenten kann nicht alleine mit den traditionellen Leistungsmassen Ausbeute und Präzision gemessen werden. Diese Masse für Retrieval-Effektivität – und die mit ihnen verbundene Darstellung des Resultates als eine Rangliste, geordnet nach Wahrscheinlichkeit der Relevanz – werden an Bedeutung verlieren. Neue, aufgabenspezifische Nützlichkeitsmasse werden vermehrt eingesetzt und ein völlig neues Information-Retrieval-Paradigma ist erforderlich, welches diese neuen Nützlichkeitsmasse optimiert.

Es muss dabei von einer Person ausgegangen werden, die eine Aufgabe zu lösen hat. Ob die Person diese Aufgabe sowohl schnell als auch qualitativ und quantitativ gut erledigen kann, hängt im Informationszeitalter davon ab, ob ihr nützliche Information zur Verfügung steht. Der Nutzen der Information ist in diesem Fall sowohl aufgaben- als auch personenspezifisch und hängt von unterschiedlichen Faktoren ab:

- dem Zeitaufwand, um die Information zu lesen (Text), abzuhören (Audio) oder anzuschauen (Bilder, Video)
- dem Zeitaufwand, um die Information zu verarbeiten
- den Kosten der Information
- der Qualität der Information
- der Aktualität der Information

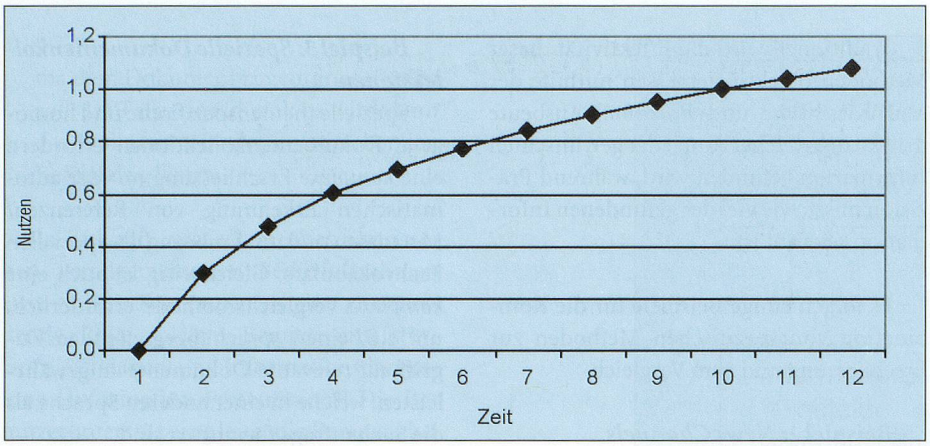

Abbildung 4

- der Vertrauenswürdigkeit der Information

Es ist davon auszugehen, dass zukünftige Information-Retrieval-Systeme nicht ausschliesslich Ausbeute und Präzision, sondern den Nutzen zu optimieren haben. Konkret heisst das, dass in möglichst kurzer Zeit aufgaben- und personenspezifisch ein möglichst hoher Nutzen zu erzielen ist.

In diesem neuen, verallgemeinerten Rahmen spielen wiederum zwei Komponenten eine entscheidende Rolle: Benutzerunterstützung (Guidance) und Informationsdarstellung (Information Presentation).

Die Benutzerunterstützung leitet die Benutzer bei der Interaktion mit dem Retrieval-System so, dass sie im nächsten Schritt eine Aktion wählen, die einen möglichst hohen Nutzen bringt. Nach Ausführung der Aktion (z.B. Suche modifizieren, Suchresultat übersetzen und zusammenfassen lassen, oder Relevanzrückkoppelung), muss die Information so dargestellt werden, dass wiederum ein grösstmöglicher Nutzen resultiert. Dieses Wechselspiel von Benutzerunterstützung und Informationsdarstellung soll abhängig von der Zeit eine möglichst schnell wachsende Nutzenfunktion ergeben (siehe Abbildung 4).

Die beiden neuen Komponenten korrespondieren mit den bekannten Komponenten Erschliessung und Vergleich. Damit ist der Weg für eine neue Generation von Information-Retrieval-Systemenvorgezeichnet.

Zur Benutzerführung gehören virtuelle Retrieval-Experten, welche dem Benutzer gezielte Empfehlungen unterbreiten. Solche Empfehlungssysteme können auf Methoden wie dem kooperativen Filtering oder dem Data Mining basieren. Auch die Darstellung von Information eröffnet äusserst interessante Perspektiven. Methoden, welche Dokumente zusammenfassen, automatisch übersetzen und Antworten extrahieren, werden in naher Zukunft für derartige Zwecke einsetzbar sein. ■

contact:

E-Mails:
martin.braschler@eurospider.com
peter.schaeuble@eurospider.com