Information Retrieval Glossar

Alexandra Bünzli

11.04.2001

Answer-Extraction (Antwort-Extraktion) Antwortextraktionssysteme liefern aus Texten Sätze, die die natürlichsprachlichen Anfragen direkt beantworten.

Boolean Query (Boolesche Anfrage) Eine Kombination von Wörtern (→*Query*), die mit boolschen Operatoren (AND, OR, NOT) verknüpft sind.

Classification Der Prozess zur Einordung eines Dokuments in die richtige Kategorie.

Collection (Sammlung) Eine Menge von Dokumenten, in denen der Benutzer nach Informationen sucht.

Document Ein Stück Information, das der Benutzer finden will. Beispiele dafür sind eine Textdatei, eine WWW-Seite, ein Bild etc.

Fact Retrieval Systems Bei Fakten-Retrievalsystemen erhält man nicht Textstellen nachgewiesen, sondern die im Text repräsentierten Fakten.

Indexing (Indexierung) Der Vorgang eine Sammlung (→*Collection*) in eine Form zu bringen, die die Suche vereinfacht.

Information Extraction Ein verwandtes Gebiet des IR, das versucht semantische Strukturen und andere spezifische Arten von Information in Texten zu identifizieren.In vordefinierten Regeln wird genau festgelegt, welche Art von Information das System in der jewiligen Sammlung (→*Collection*) finden soll.

Information Retrieval (Informationswiedergewinnung) Das Fachgebiet, das sich mit der Erforschung von Systemen zur →*Indexing*, Suche und Wiedergewinnung von Information aus Texten und anderen unstrukturierten Daten beschäftigt.

Passage-Retrieval (Textstellenwiedergewinnung) Textstellenwiedergewinnungs-systeme liefern oft eine willkürliche Anzahl von Zeilen vor oder nach der Fundstelle, d. h. nicht ganze Dokumente wie in den klass. Information-Retrieval-Methoden.

Precision (Präzision) Ein Standardmass für die Evaluierung von IR-Systemen. Die Präzision gibt den Anteil der gefundenen (GEF) relevanten (REL) Dokumente an allen gefundenen Dokumenten (GEF) wieder.

$$\text{Precision: } p = \frac{|REL \cap GEF|}{|GEF|}$$

Beispiel: In einer Sammlung befinden sich 80 relevante Dokumente zu einem bestimmten Thema. Das System X liefert 60 Dokumente zurück, wovon 40 relevant sind. Dann ist die Präzision von X $p(X) = 40/60 \approx 67\%$. In einer idealen Welt wäre die Präzision 100%. Weil dies jedoch einfach zu erreichen ist (indem nur ein relevantes Dokument zurückgeliefert wird), versucht ein System sowohl Präzision als auch →*Recall* zu maximieren.

Probabilistic IR-Model Jedes IR-Modell, das die Wahrscheinlichkeit berücksichtigt, dass ein Term in einem Dokument auftaucht, oder dass ein Dokument dem Informationsbedarf des Benutzers genügt.

Query (Anfrage) Eine Anzahl Wörter, die die vom Benutzer gesuchte Information charakterisieren. Die Anfrage braucht nicht in natürlich- sprachlicher Form vorzuliegen.

Question-Answering Das Problem, die exakte Antwort zu einer natürlichsprachlichen Anfrage (→*Query*) eines Benutzers in einer grossen Sammlung zu finden.

Recall (Abdeckung) Ein Standardmass für die Evaluierung von IR-Systemen. Recall bezeichnet den Anteil der gefundenen (GEF) relevanten (REL) Dokumente in Bezug auf alle in der Sammlung vorhandenen relevanten Dokumente (REL).

$$\text{Recall: } r = \frac{|REL \cap GEF|}{|REL|}$$

Beispiel: In einer Sammlung befinden sich 80 relevante Dokumente zu einem bestimmten Thema. Das System X liefert 60 Dokumente zurück, wovon 40 relevant sind. Dann ist die Abdeckung von X $r(X) = 40/80 = 50\%$. In einer idealen Welt wäre die Abdeckung 100%. Weil dies jedoch einfach zu erreichen ist (indem alle Dokumente zurückgeliefert werden), versucht ein System sowohl Abdeckung als auch →*Precision* zu maximieren. Ein Problem bei der Berechnung dieser Formel stellt die Grösse (REL) dar. Wie findet man heraus, wieviele Dokumente zu einer Anfrage relevant sind? Man müsste alle Dokumente durchsehen, was bei grossen Sammlungen (→*Collection*) natürlich nicht möglich ist. Durch Stichproben oder Erweiterung der Suchanfrage versucht man die Menge aller relevanten Dokumente zu einer Anfrage (→*Query*) abzuschätzen.

Relevance Ein abstraktes Mass um zu definieren, wie gut ein Dokument den Informationsbedarf eines Benutzers befriedigt. Idealerweise sollte ein IR-System alle relevanten und keine nicht-relevanten Dokumente liefern. Leider ist die Bestimmung der Relevanz eines Dokuments subjektiv und deshalb schwierig zu quantifizieren.

Relecance Feedback Relevanz-Feedback ist ein Prozess, der die Resultate zu einer Anfrage (→*Query*) nochmals bearbeitet. Der Benutzer markiert, welche der gefundenen Dokumente für ihn relevant sind und das System sucht in diesen Dokumenten neue gemeinsame Terme, fügt sie zu der alten Anfrage hinzu und liefert mit dieser neuen Anfrage neue Dokumente. Dieser Vorgang kann mehrere Male wiederholt werden. (Manchmal wird dieser Vorgang auch als “find similar documents” oder “query by example” bezeichnet.)

Similarity (Ähnlichkeit) Die Similarity misst, wie ähnlich sich zwei Dokumente sind oder wie ähnlich sich die Anfrage (→*Query*) und ein Dokument sind. In einem Vektorraummodell (→*Vector space model*) wird es so interpretiert, dass je näher sich die beiden Vektoren, die die zu vergleichenden Dokumente oder Anfragen repräsentieren, sind, desto ähnlicher sind die Dokumente sich.

Stemming (Grunformenreduktion) Stemming nennt man den Vorgang, der die Präfixe und Suffixe eines Wortes in einem Dokument oder einer Anfrage (→*Query*) in die Form überführt, in der die Terme (→*Term*) im inneren Modell des Systems abgespeichert sind. Normalerweise wird das mit Wörtern gemacht, die die gleiche konzeptuelle Bedeutung haben.

Stopword Stoppwörter sind Wörter wie Artikel, Präpositionen etc. , die wenig semantischen Gehalt haben. Sie und auch Wörter, die in einer Sammlung (→*Collection*) häufig vorkommen und darum für die Unterscheidung der Dokumente wenig sinnvoll sind, werden in einer Stoppwortliste gespeichert. (Achtung: Diese Listen sind darum nicht unabhängig von den jeweiligen Sammlungen!) Diese Wörter werden nicht im inneren Modell des Systems verwendet.

Thesauri Thesauri sind eine geordnete Zusammenstellung von Begriffen eines Fachgebiets mit ihren (natürlichsprachlichen) Beziehungen (Relationen, relations). In ihnen können hierarchische Strukturen durch Beziehungen erzeugt werden. (Beziehungen wie allgemeiner (Oberbegriff, broader), spezieller (Unterbegriff, narrower), synonym, verwandt (related), Gegensatz (antonym) etc.)

Term Ein einzelnes Wort oder Konzept, das im Modell für ein Dokument oder einer Anfrage (→*Query*) verwendet wird. Term kann sich aber auch auf Wörter im Originaltext beziehen.

TREC (Text REtrieval Conference) Diese Gruppe stellt IR-Forschern eine grosse Testsammlung und ein Evaluierungssystem zur Verfügung, wodurch die entwickelten Systeme mit den gleichen Daten verglichen werden können.

Vector Space Model (Vektorraummodell) Das Vektorraummodell bildet die Dokumente und Anfragen (→*Query*) auf Vektoren ab. Die Eigenschaften der Vektoren sind normalerweise die Wörter eines Dokuments oder einer Anfrage, die auf eine Grundform abgebildet wurden (nach →*Stemming*) und aus denen die →*Stopwords* eliminiert wurden. Die Vektoren werden gewichtet, damit die Terme, die den Inhalt gut beschreiben eine stärkere Bedeutung bekommen. Bei der Suche werden alle Vektoren der Dokumente mit dem Anfragevektor verglichen. Die, welche dem Anfragevektor am nächsten sind, gelten als die ähnlichsten und somit sollten auch die Dokumente am ehesten die Frage beantworten (→*Similarity*).

Weighting (Gewichtung) Gewichten kann man auf verschiedene Arten. Als lokale Einflüsse bezeichnet man die Gewichtung nach der Häufigkeit des Terms (→*Term*) im Dokument, der Häufigkeit des Terms in Relation zum häufigsten Term im Dokument oder bei Dokumenten mit Struktur die Gewichtung nach Ort des Vorkommens (z.B. wird ein Wort im Titelfeld als wichtiger eingestuft). Auch die Gewichtung nach Position ist ein lokaler Einfluss: Wichtige Terme werden eher am Anfang des Dokuments plaziert (z.B. Nachrichtenmeldungen haben immer den gleichen Aufbau: 1. wesentliche Neuigkeit, 2. Hintergründe, 3. Details und Kommentare)

Ein globaler Einfluss wäre die Dokumenthäufigkeit eines Terms. Terme, die in vielen Dokumenten vorkommen, sind nicht geeignet zur Beschreibung und werden als Indexterme schwach gewertet. In der Regel werden lokale und globale Kriterien bei der Gewichtung kombiniert, woraus dann die verwendete Gewichtung des Terms resultiert.