URL:  
  
Category:   
Information Retrieval einfach erklärt  
KWs:   
Information Retrieval, information retrieval system, information retrieval definition, information retrieval beispiel  
Internal links:  
Yes  
Title:  
Information Retrieval: Wie Suchmaschinen Daten abrufen  
Description:  
Wie kommen Google & Co. an Informationen? Mit Information-Retrieval-Systemen! Wir erklären die Theorie, die hinter den Suchmaschinen steckt.  
Number of words: Target / Actual  
Min200- Max1000 / 2239  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Information Retrieval: Die große Suche nach Wissen

Wir leben in einer Informationsgesellschaft. Daten, Fakten und Wissen nehmen einen ungleich höheren Stellenwert ein als noch vor einem halben Jahrhundert. Gleichzeitig gib es dank des Internets immer mehr verfügbare Informationen. Doch diese müssen auch abgerufen werden – dabei helfen uns Suchmaschinen. Doch wie kommen diese wiederum an die Daten, die sie ausgeben? Die Erklärung ist die sogenannte Information Retrieval. Die Informationsbeschaffung – genauer: **Informationsrückgewinnung** – ist eine eigene Disziplin der Informatik und der Informationswissenschaften und vor allem für Suchmaschinen von großer Bedeutung. Anhandkomplexer Information-Retrieval-Systeme erkennen sie Intentionen, die hinter bestimmten Suchbegriffen stehen und machen relevante Daten zu Suchanfragen ausfindig.

## Zur Geschichte der Informationsbeschaffung

Bei der Information Retrieval geht es darum, **bestehendes Wissen zugänglich zu machen**. Das ist nicht erst seit Beginn des digitalen Zeitalters aktuell. Als einer der ersten, der sich ernsthaft darüber Gedanken gemacht hat, wie die Menschheit ihr geballtes Wissen angesichts einer immer unübersichtlicher werdenden Welt besser verfügbar machen kann, gilt der Wissenschaftler **Vannevar Bush**. Im Jahr 1945 hat er mit dem bahnbrechenden Artikel „As We May Think“ [Englischer Originalartikel auf der Website von The Atlantic] (https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/) eine Zukunftsvision der Informationsbeschaffung und -organisation vorgelegt.

Bush sah folgendes Problem in den Wissenschaften: Experten spezialisieren sich immer weiter und benötigen dafür mehr und mehr Informationen, die aber – eben wegen der Ausdifferenzierung – immer schwieriger zu finden sind. Dies war wohlgemerkt zu einer Zeit, in der Bibliotheken noch mit analogen Zettelkästen und großen Katalogen organisiert wurden. Eine Stichwortsuche war nur dann möglich, wenn sich ein fleißiger Bibliothekar zuvor die Mühe gemacht hatte, alle Werke manuell zu indexieren. Bush sah in den technischen Entwicklungen der Zeit, etwa dem Mikrofilm, eine Möglichkeit, Informationen besser verfügbar zu machen. Seine eigene Vision hieß **Memex**, eine Maschine so groß wie ein Schreibtisch, die als Wissensspeicher und Rechercheapparat dienen sollte. Memex wurde nie gebaut, aber die Technologie – der Benutzer springt von einem Artikel zum nächsten – kann als Vorläufer des Hypertexts gesehen werden.

In den 1950ern befasste sich vor allem der Informatiker Hans Peter Luhn mit den Aufgaben der Informationsbeschaffung und entwickelte Techniken, die auch heute noch von Relevanz sind: **Volltextverarbeitung, Auto-Indexierung und selektive Informationsverarbeitung (SDI)** gehen auf seine Forschung zurück. Diese Methoden waren für die Entwicklungen des Internets von großer Bedeutung, denn in der Informationsflut des World Wide Webs ist es unabdingbar, Information-Retrieval-Systeme einzusetzen. Ansonsten würden Sie niemals die Antworten erhalten, die Sie benötigen.

## Information Retrieval – eine Definition

Ziel der **Information Retrieval (IR)** ist es, maschinell gespeicherte Daten auffindbar zu machen. Anders als beim Data-Mining [Data-Mining: Analysemethoden für Big Data] (https://hosting.1und1.de/digitalguide/online-marketing/web-analyse/data-mining-so-lassen-sich-datenberge-erschliessen/), mit dem man Strukturen aus Datensätzen extrahiert, befasst sich IR damit, bestimmte Informationen aus einer Datenmenge zu filtern. Das typische Anwendungsgebiet ist eine Internet-Suchmaschine. Information-Retrieval-Systeme lösen hier vor allem zwei Probleme:

1. **Vagheit**: Die Anfragen der Nutzer sind oft ungenau, der eingegebene Suchbegriff lässt Interpretationsspielraum. Wer z. B. nach dem Begriff „Bank“ sucht, kann Informationen zum Bankwesen allgemein oder aber eine Wegbeschreibung zum nächsten Geldinstitut benötigen. Das Problem potenziert sich, wenn die Nutzer selbst noch nicht genau wissen, was für Informationen sie überhaupt finden möchten.
2. **Unsicherheit**: Die Inhalte der gespeicherten Informationen sind dem System mitunter nicht bekannt genug. Dadurch werden falsche Ergebnisse geliefert. Dies passiert z. B. bei Homonymen, also Wörtern, die mehrere Bedeutungen haben. So könnte der Nutzer gar nicht nach einem Geldinstitut suchen, sondern nach einer Sitzgelegenheit für seinen Garten.

Hinzu kommt, dass das Information-Retrieval-System die Informationen auch bewerten sollte, um dem Nutzer eine Reihenfolge der Daten anzubieten. Das erste Ergebnis sollte also im Idealfall die beste Antwort auf die Frage des Nutzers liefern.

## Vorstellung verschiedener Modelle

Für Information Retrieval bestehen verschiedene Modelle, die sich allerdings nicht zwangsläufig gegenseitig ausschließen, sondern miteinander kombinierbar sind. Es gibt inzwischen viele solcher Modelle, die sich teilweise nur in Details unterscheiden. Sie lassen sich allerdings grob in drei Kategorien unterteilen:

* **Mengentheoretische Modelle**: Ähnlichkeitsbeziehungen werden durch Mengenoperationen ermittelt (Boolsches Modell).
* **Algebraische Modelle**: Ähnlichkeiten werden paarweise ermittelt; Dokumente und Suchanfragen lassen sich dabei als Vektoren, Matrizen oder Tupel darstellen (Vektorraummodell).
* **Probabilistische Modelle**: Diese Modelle stellen Ähnlichkeitsbezüge her, indem sie die Datenmengen als mehrstufige Zufallsexperimente ansehen.

Im Folgenden stellen wir die drei archetypischen Modelle innerhalb dieser Kategorien vor. Bei den darüber hinaus existierenden Modellen handelt es sich vor allem um Mischformen der drei Typen. So hat das erweiterte Boolsche Modell sowohl Eigenschaften der mengentheoretischen als auch der algebraischen Modelle.

### Boolesches Modell

Die bekanntesten Suchmaschinen im Web basieren auf dem Booleschen Prinzip. Dabei handelt es sich um **logische Verknüpfungen**, mit denen Nutzer die Suche verfeinern und genauer bestimmen können. Mit UND, ODER oder NICHT (AND, OR, NOT) bzw. den entsprechenden Symbolen ∧, ∨ oder ¬ lässt sich eine Anfrage spezifizieren, wenn z. B. zwingend beide Begriffe im Ergebnis auftauchen oder Inhalte mit einem bestimmten Begriff ausgeblendet werden sollen. Nach diesem Prinzip funktionieren auch die Operatoren bei Google [Effektiv suchen mit den Google-Operatoren] (https://hosting.1und1.de/digitalguide/online-marketing/suchmaschinenmarketing/effektiv-suchen-mit-den-google-operatoren/). Der Nachteil dieses Systems ist, dass es keinerlei Rangordnung der Ergebnisse vorsieht. Sinnvoll ist eine Ordnung nach Nützlichkeit, die Methode liefert aber eine zufällige Reihenfolge.

### Vektorraummodell

In einem mathematischen Zugang lassen sich Inhalte auch als Vektoren darstellen. Im Vektorraummodell werden Begriffe (*terms*) als Koordinatenachsen abgebildet. Sowohl Dokumente als auch Suchanfragen erhalten spezifische Werte in Bezug zu dem Begriff und sind deshalb als Punkte oder **Vektoren innerhalb eines Vektorraums** darstellbar. Anschließend werden beide Vektoren miteinander verglichen. Der Vektor (also der Inhalt), der dem der Suchanfrage am ähnlichsten ist, sollte in der Rangliste der Ergebnisse an erster Stelle auftauchen. Der Nachteil hierbei ist, dass sich ohne Boolesche Operatoren keine Begriffe ausschließen lassen.

### Probabilistisches Modell

Das probabilistische Modell greift auf die **Wahrscheinlichkeitstheorie** zurück. Jedem Inhalt wird ein Wahrscheinlichkeitswert zugeordnet. Die Ergebnisse werden abhängig von der Wahrscheinlichkeit, mit der sie zur Suchintention passen, sortiert. Wie hoch die Chancen stehen, dass ein bestimmter Inhalt dem Wunsch des Nutzers entspricht, ermittelt das Modell durch sogenanntes Relevance-Feedback. Dabei werden Nutzer z. B. dazu aufgefordert, die Ergebnisse manuell zu bewerten. Bei der nächsten gleichlautenden Anfrage zeigt das Modell eine andere (und vielleicht bessere) Ergebnisliste. Nachteil dieses Verfahrens ist, dass es von zwei Anforderungen ausgeht, die nicht mit Sicherheit gegeben sind: Zum einen setzt das Modell voraus, dass die Nutzer gewillt sind, an dem System durch sein Feedback mitzuarbeiten. Zum anderen geht die Theorie davon aus, dass Nutzer die Ergebnisse unabhängig voneinander betrachten, also jeden Inhalt so bewerten, als wäre es der erste, den sie bezüglich der Suchanfrage lesen. In der Praxis schätzen Suchende die Nützlichkeit einer Information aber immer basierend auf bereits gesichteten Inhalten ein.

## Funktionsweisen der Informationsbeschaffung

Beim Information Retrieval kommen – unabhängig von den Modellen – verschiedene Methoden und Arbeitstechniken zum Einsatz. Deren Ziel ist es immer, dem Nutzer die Informationssuche zu vereinfachen und relevantere Ergebnisse zu liefern.

### Term Frequency-Inverse Document Frequency

Mit der Kombination aus Vorkommenshäufigkeit von Begriffen und der inversen Dokumenthäufigkeit wird die Wichtigkeit eines Begriffs für eine Suchanfrage berechnet. Der Wert wird abgekürzt als **tf-idf**.

* **Term Frequency**: Die Suchwortdichte gibt an, wie häufig ein Begriff in einem Dokument auftaucht. Die reine Vorkommenshäufigkeit kann allerdings kein alleiniges Indiz für die Relevanz des Textes sein. Denn in einem langen Dokument kommt der Suchbegriff u. U. häufiger vor als in einem kurzen. Deshalb sollte die Häufigkeit in Relation zum Umfang eines Dokuments gesehen werden. Dafür wird die Häufigkeit des Suchbegriffs durch die Häufigkeit des höchstfrequenten Wortes (z. B. „und“) geteilt:
* **Inverse Document Frequency**: Für idf betrachtet man nicht nur ein einzelnes Dokument, sondern einen kompletten Textkorpus. Wörter, die nur in sehr wenigen Dokumenten zu finden sind, in diesen aber wiederum sehr häufig, haben eine höhere Relevanz als Begriffe, die in nahezu allen Texten vorkommen. So hat z. B. der Begriff „Inverse Dokumenthäufigkeit“ einen deutlich höheren Wert als „und“.

Durch die Verbindung der beiden Tests können Information-Retrieval-Systeme bessere Ergebnisse liefern, als wenn sie allein eingesetzt würden: Wäre nur die Term Frequency von Bedeutung, dann würde die Suchanfrage „Die Sendung mit der Maus“ diejenigen Dokumente hoch einschätzen, in denen die Wörter „die“, „mit“ und „der“ häufig vorkommen. Das ist aber offensichtlich wenig hilfreich. Wird hingegen die Inverse Document Frequency hinzugezogen, bekommen „Sendung“ und „Maus“ eine sehr viel größere Bedeutung für die Suche und werden als die eigentlichen Suchbegriffe erkannt.

### Query Modification

Ein großes Problem der Informationsbeschaffung sind die Nutzer selbst: Durch zu ungenaue oder gar fehlerhafte Anfragen erhalten sie falsche oder unzureichende Informationen. Um dies zu vermeiden, haben Informationswissenschaftler die Query Modification eingeführt. Hierbei verändert das System selbstständig die eingegebene Suchanfrage. So werden z. B. Synonyme eingesetzt, die bessere Ergebnisse liefern. Dafür greift das System u. a. auf **Thesauri und Nutzer-Feedback** zurück. Um nicht auf die Mitarbeit des Nutzers angewiesen zu sein, kann man auch ein sogenanntes Pseudo-Feedback anwenden. Bei dieser Methode liest das System verwandte Begriffe aus den besten Suchergebnissen aus und schätzt diese als relevant für die entsprechende Suche ein. Anfragen können u. a. durch diese Techniken erweitert oder verbessert werden:

* **Stoppworteliminierung**: Als Stoppwörter bezeichnet man solche Ausdrücke, die nicht oder nur unwesentlich zum Inhalt des Textes beitragen. Es ist sinnvoll, Wörter wie „und“ oder alle Artikel nicht als repräsentativ für den Inhalt des Dokuments anzusehen.
* **Mehrwortgruppenidentifizierung**: Gruppierungen von Wörtern müssen als solche erkannt werden. Diese Identifizierung sorgt dafür, dass die Suchmaschine auch Teile von zusammengesetzten Wörtern als relevant ansieht.
* **Grund- und Stammformreduzierung**: Um effektiver suchen zu können, müssen Wörter auf ihre Wortstämme reduziert werden. Flexionsformen eines Wortes würden ansonsten nicht korrekt in den Suchergebnissen auftauchen.
* **Thesaurus**: Neben den im entsprechenden Dokument auftauchenden Begriffen sollte ein Information-Retrieval-System auch Synonyme des Wortes als relevant ansehen. Nur so lässt sich sicherstellen, dass Nutzer auch das finden, was sie wirklich suchen.

### Recall & Precision

Die Effektivität eines Information-Retrieval-Systems wird gemeinhin mit den Faktoren Trefferquote (*recall*) und Genauigkeit (*precision*) berechnet. Beide werden als Quotienten dargestellt.

* **Recall**: Wie vollständig sind die Suchergebnisse? Dafür wird die Anzahl der gefundenen, relevanten gegenüber der Anzahl der nicht gefundenen, relevanten Dokumente gestellt. Der Quotient gibt also an, wie wahrscheinlich es ist, dass ein relevantes Dokument auch gefunden wird:
* **Precision**: Wie genau ist das Suchergebnis? Dafür wird die Anzahl der gefundenen, relevanten gegenüber der Anzahl der gefundenen, irrelevanten Dokumente gestellt. Der Quotient gibt also an, wie wahrscheinlich es ist, dass ein gefundenes Dokument relevant ist:

Beide Werte liegen dabei grundsätzlich zwischen 0 und 1, wobei 1 ein perfekter Wert wäre. Zudem schließen sich perfekte Ergebnisse bei beiden Quotienten in der Praxis aus. Wer die Vollständigkeit des Suchergebnisses erhöht, tut dies auf Kosten der Genauigkeit und umgekehrt. Als weiterer Wert kann zudem der Fallout (also die Ausfallquote) berechnet werden: Dieser Quotient gibt die Falsch-Positiv-Rate wieder; er wird bestimmt aus dem Verhältnis der gefundenen, irrelevanten Dokumente zu den irrelevanten Inhalten, die nicht gefunden wurden. Darstellen lassen sich Recall und Precision in einem Achsendiagramm, bei dem jeder der beiden Werte je eine Achse belegt.

## Information Retrieval: Beispiel einer Suche

Jede Internetsuchmaschine basiert auf Information Retrieval. Somit wären Google, Bing und Yahoo prominente Beispiele für die computergestützte Informationsbeschaffung. Um zu zeigen, wie IR in der Praxis funktioniert, ist es aber sinnvoller, ein einfacheres, eigenes Beispiel zu nehmen. Dabei gehen wir von einer Suchmatrix in einer (sehr kleinen) Bibliothek für Kinderbücher aus. In allen Büchern kommen Tiere vor, doch wir möchten nur solche Bücher finden, in denen Elefanten und Giraffen eine Rolle spielen, aber keine Krokodile. Eine Suchanfrage mit der Booleschen Methode würde demnach so aussehen: Elefant UND Giraffe NICHT Krokodil. Das Ergebnis der Suche kann immer nur 1 oder 0 sein: Kommt der Begriff vor oder kommt er nicht vor?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Tim & Olli im Zoo | Susi auf Safari | Michael und der verrückte Zirkus | Katja und Rex: Abenteuer beim Gassi gehen | Dampferfahrt auf dem Nil |
| Elefant | 1 | 1 | 1 | 0 | 1 |
| Giraffe | 1 | 0 | 1 | 0 | 1 |
| Krokodil | 0 | 1 | 0 | 0 | 1 |
| Hund | 0 | 0 | 1 | 1 | 0 |
| Löwe | 1 | 1 | 1 | 0 | 0 |
| Katze | 0 | 0 | 0 | 1 | 0 |

Das Ergebnis der Suche wäre also „Tim & Olli im Zoo“ und „Michael und der verrückte Zirkus“. Damit ist aber noch keine Gewichtung der Ergebnisse gegeben. In welchem Buch geht es mehr um Elefanten und Giraffen? Dafür kann das System die Term Frequency und die Inverse Document Frequency bestimmen:

|  |  |  |
| --- | --- | --- |
| Tim & Olli im Zoo | | |
|  | Elefant | Giraffe |
| Term Frequency | 36 | 20 |
| Inverse Document Frequency | 0,1 | 0,2 |
| tf-idf-Maß | 3,6 | 4 |
| Michael und der verrückte Zirkus | | |
| Term Frequency | 10 | 11 |
| Inverse Document Frequency | 0,1 | 0,2 |
| tf-idf-Maß | 1 | 2,2 |

„Tim & Olli im Zoo“ ist also wahrscheinlich besser für die Suche nach einem Text mit Giraffen und Elefanten geeignet als „Michael und der verrückte Zoo“ und sollte deshalb an erster Stelle der Suchergebnisse auftauchen. Die Methode, die wir hier angewandt haben, funktioniert nur, wenn die Suchbegriffe festgelegt sind (**kontrollierte Indexierung**). Dies kann z. B. in Fachdatenbanken der Fall sein, bei dem die Nutzer in der Verwendung der Suchmaske geschult sind. In unserem Beispiel wäre eine Query Modification sinnvoll: Außer „Elefant“ würden auch eine Suche nach „Dickhäuter“ sowie grammatische Varianten dieser Wörter positive Ergebnisse liefern.

[Tipp: Neben Google gibt es noch viele weitere Suchmaschinen im World Wide Web. Die Alternativen zu Google [Google-Alternativen: Die große Suchmaschinen-Liste] (https://hosting.1und1.de/digitalguide/online-marketing/suchmaschinenmarketing/google-alternativen-im-ueberblick/) achten beispielsweise oftmals stärker auf den Datenschutz.]

**Teaser**

In unserem Alltag suchen und erhalten wir unzählige Informationen. Durch das Internet hat sich die Suche nach Daten und Fakten enorm gewandelt. Suchmaschinen helfen uns, in der riesigen Fülle der verfügbaren Informationen diejenigen zu finden, die wir brauchen. Doch wie funktioniert eigentlich die Informationsbeschaffung von Google und Co.? Hinter den großen Search Engines, aber auch kleineren Bibliotheksdatenbanken stecken sogenannte Information-Retrieval-Systeme. Wir erklären, was es damit auf sich hat.