Automatische Klassifikationsmethoden

Autor(en): **Frei, H.-P.**

Objekttyp: **Article**

Zeitschrift: **Nachrichten VSB/SVD = Nouvelles ABS/ASD = Notizie ABS/ASD**

Band (Jahr): **53 (1977)**

Heft 6

PDF erstellt am: **21.07.2024**

Persistenter Link: https://doi.org/10.5169/seals-771436

Bibliographische Hinweise

Die Gedanken, Schlußfolgerungen, Lösungen und Fragen, die ich hier darzulegen versuchte, sind eher der Praxis denn der Literatur entnommen. Daher beschränke ich mich auch auf den Hinweis auf zwei allgemeine Werke zur Gesamtproblematik:

Harbeck, Rudolf (Hrsg.): Verteidigungs-Dokumentation. Beiträge zur Aufgabe, Organisation und Methodik der Dokumentation im Geschäftsbereich des Bundesministers der Verteidigung der BRD mit Hinweisen auf die Verteidigungs-Dokumentation anderer Staaten, München, Verlag Dokumentation, 1976.

Laisiepen, Klaus; *Lutterbeck,* Ernst; *Meyer-Uhlenried,* Karl-Heinrich: Grundlagen der praktischen Information und Dokumentation. Eine Einführung, München, Verlag Dokumentation, 1972.

Automatische Klassifikationsmethoden

H.-P. Frei
Institut für Informatik, ETH-Zentrum, 8092 Zürich

Abstract

Bei der automatischen Indexierung werden – obwohl problematisch – im wesentlichen statistische Methoden benutzt, welche auf der Ebene von einzelnen Wörtern operieren. Die Erfassung und Präzision bei der Suche werden durch die Art der Indexierung erheblich beeinflußt. Wegen der immensen Größe der in einer realen Situation entstehenden Files ist der File-Organisation besondere Aufmerksamkeit zu schenken.

1. Einleitung

Die Menge der produzierten Dokumente (Bücher, Artikel usw.) nimmt gegenwärtig exponentiell zu. Die Zeit hingegen, welche ein Benutzer (z. B. ein Wissenschafter) zur Lektüre aufwenden kann, bleibt konstant. Das bedeutet, daß ein immer kleinerer Teil der Literatur gelesen werden kann und daß ein Benutzer, will er informiert bleiben, seine Lektüre produktiver gestalten muß. Deshalb steigen die Anforderungen an die Dienstleistungen der Bibliotheken, was einen größeren Aufwand zur Folge hat. Neue Typen von Dokumenten bringen zusätzliche Schwierigkeiten mit sich.

Während man den Betrieb von Bibliotheken bezüglich Ausleihe und administrativer Kontrolle schon lange durch Automatisation zu rationalisieren versuchte, wird von automatischer Indexierung und Klassifikation erst seit wenig mehr als zehn Jahren gesprochen. Erst Mitte der sechziger Jahre sind

die ersten Projekte bekannt geworden, bei denen versucht wurde, mit Hilfe der elektronischen Datenverarbeitung die Indexierung und Klassifikation von Dokumenten effizienter zu gestalten. Bis heute gibt es jedoch nur wenige Leute, die ernsthaft an diesen Problemen arbeiten, an der Automatisation von Arbeiten also, die traditionellerweise von akademisch ausgebildetem Personal ausgeführt werden. Deshalb wäre es auch verfrüht, jetzt schon fertige Lösungen präsentieren zu wollen. Der Zweck dieses Beitrages ist es denn auch, Trends aufzuzeigen, welche bei solchen Projekten verfolgt werden. Zentral beleuchtet werden Überlegungen über die automatische Indexierung, über die Verbesserung von Präzision und Erfassung sowie über die automatische Klassifikation von auf diese Weise indexierten Dokumenten. Bei all diesen Problemen darf man die von G. Salton in seinem 1975 erschienenen Buch gemachte Bemerkung nicht vergessen, « . . . perfect performance, either manual or automatic, is not attainable in the library field.»

2. Automatische Indexierung

Ein Dokument zu indexieren bedeutet, den Sachinhalt dieses Dokumentes auf ein relativ kleine Anzahl von präzise definierten Begriffen abzubilden. Diese Arbeit wird normalerweise von qualifiziertem, akademisch gebildetem Personal ausgeführt.

Automatisch Indexieren bedeutet also mit Hilfe einer Maschine (Computer) den Sachinhalt eines Dokumentes zu erkennen und eine relativ kleine Anzahl von aussagekräftigen Begriffen zu finden, welche diesen Sachverhalt wiedergeben. Das Hauptproblem dabei ist, den Sachinhalt automatisch aus einem Dokument zu extrahieren. Man stößt dabei im wesentlichen auf dieselben Probleme, welche auch bei der automatischen Sprachübersetzung zu bewältigen sind. Der Zweck dieses Abschnittes ist es, einige Hinweise zu geben, wie das Problem der automatischen Indexierung im allgemeinen angegangen wird. Im weiteren sollen einige Grundbegriffe definiert und beschrieben werden, Begriffe, welche bei einem solchen Vorgehen eine zentrale Rolle spielen.

Die meisten in der Literatur erwähnten Indexierungsmethoden hängen nicht von der Art der dem Indexierungsvorgang zugrunde gelegten Daten ab. Je nach dem Aufwand und den technischen Möglichkeiten können diese Eingabe-Daten nur aus den Titeln der Dokumente, aus Abstracts, aus den Anfangs- und/oder Schlußsätzen jedes einzelnen Abschnitts des Textes oder im Extremfall sogar aus dem ganzen Text bestehen.

Die Häufigkeit F^k eines Wortes im Text der zu indexierenden n Dokumente ist ein grundlegender und immer wieder benötigter Begriff und definiert als:

$$F^k = \sum_{i=1}^{n} f_i^k$$

wo f_i^k: Häufigkeit des Wortes k in Dokument i

Für praktische Belange ist diese Größe als solche jedoch wenig brauchbar, da große Häufigkeiten von Wörtern eher auf informationsarme Begriffe hindeuten, kleine Häufigkeiten auf Begriffe, die später ohnehin wenig zum Suchprozeß beitragen werden. Die aus einem Dokument als nützliche Angaben zu extrahierenden Begriffe sollten also nicht einfach häufig vorkommende Wörter sein, sondern solche mit einer speziell hohen Aussagekraft. Weil Texte einerseits vom individuellen Stil des Autors geprägt sind, andererseits aber auch die Zeitperiode und Umgebung wiederspiegeln, in der sie geschrieben wurden, sind statistische Methoden alleine für die Erkennung des Sachinhaltes nicht geeignet. Ein Beiziehen von Wörter-Listen (z. B. Schlüsselwörter, informationsarme Wörter, Synonyme) ist deshalb unumgänglich. Trotz all dieser Einschränkungen und Nachteile ist bis heute das Auszählen und statistische Analysieren von Wörtern in einem Text bei den meisten Projekten, welche sich mit automatischer Indexierung befassen, die zentral angewandte Methode. Versuche haben gezeigt, daß dennoch die automatischen Produkte den manuellen nicht notwendigerweise unterlegen sind.

Die zwei wesentlichen Größen im Zusammenhang mit der statistischen Analyse sind das *Rauschen* und das *Signal* eines Wortes. Das Rauschen (engl. noise) ist definiert als

$$N^k = -\sum_{i=1}^{n} \frac{f_i^k}{F^k} \log\left(\frac{f_i^k}{F^k}\right), \quad \text{wo} \quad \sum_{i=1}^{n} \frac{f_i^k}{F^k} = 1$$

und ist ein Maß für die *Gleichverteilung* eines Wortes über die behandelten n Dokumente. Das Signal (engl. signal) andererseits ist definiert als

$$S^k = \log(F^k) - N^k$$

und ist ein Maß für die *Konzentration* eines bestimmten Wortes in einigen darin zugrunde gelegten Dokumenten. Betrachten wir zwei extreme Beispiele: Kommt ein Wort in jedem Dokument gleich oft vor, so ist das Rauschen maximal, das Signal hingegen ist o; kommt ein Wort hingegen nur in einem Dokument mit einer Häufigkeit F^k vor, so ist – wie man leicht nachrechnen kann – das Rauschen o, und das Signal hat einen maximalen Wert.

Mit Hilfe dieser Größen kann die Relevanz der aus einer Sammlung von Dokumenten herausgezogenen Wörter weit besser beurteilt werden als mit Hilfe des Begriffes der Häufigkeit alleine. Bei der maschinellen Indexierung werden neben den statistischen häufig auch noch andere Kriterien verwendet, wie die Position des Wortes im Text (Wörtern im Titel wird vielfach eine höhere Aussagekraft zugebilligt als solchen im Text). Auch die Distanz zwischen Wörtern, resp. die Distanz eines Wortes von einem bestimmten Schlüsselbegriff kann zur Analyse mitverwendet werden.

Obwohl man unschwer Beispiele konstruieren kann, bei denen eine auch recht elaborierte automatische Methode versagt, haben Untersuchungen gezeigt, daß automatische Methoden für mittelgroße Dokumentensammlungen durchaus praktikabel sind.

3. Erfassung und Präzision

Obwohl die Begriffe *Erfassung* (engl. recall) und *Präzision* (engl. precision) vor allem im Zusammenhang mit dem Auffinden von Dokumenten eine wesentliche Rolle spielen, werden sie an dieser Stelle erwähnt, da schon beim Indexieren und Klassifizieren das Wiederauffinden im Auge behalten werden muß. Deshalb werden in diesem Abschnitt die beiden Begriffe definiert und etwas näher betrachtet.

Die Erfassung ist definiert als

$$\text{Erfassung} = \frac{\text{produziertes relevantes Material}}{\text{relevantes Material}}$$

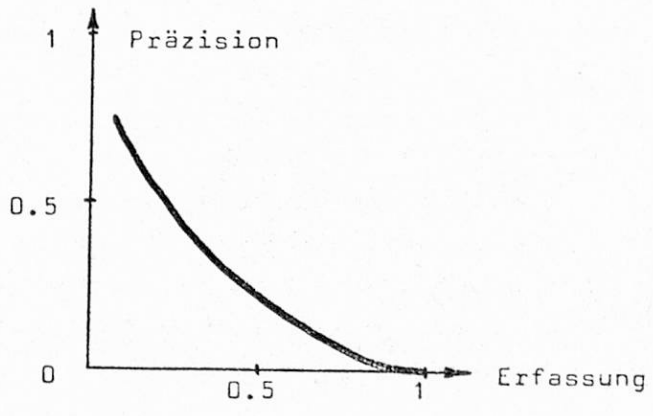
und gibt ein Maß für die Relevanz der bei einem Suchprozeß aufgefundenen Dokumente. Die numerischen Werte der Erfassung schwanken zwischen 0 und 1 und sind hoch, wenn wenige nützliche Dokumente bei der Suche übergangen wurden.

Die Präzision ist definiert als

$$\text{Präzision} = \frac{\text{produziertes relevantes Material}}{\text{produziertes Material}}$$

und gibt ein Maß für die Genauigkeit der getroffenen Auswahl. Auch diese numerischen Werte bewegen sich im Bereich zwischen 0 und 1 und sind um so höher *je weniger nutzlose* Dokumente beim Suchprozeß herausgezogen werden.

Bei existierenden automatischen Systemen ist die folgende Beziehung zwischen Erfassung und Präzision typisch:



Aus der Definition, wie auch aus obiger Figur, geht hervor, daß es recht einfach ist, (auf Kosten der Präzision) eine hohe Erfassung zu erzielen (man braucht lediglich sämtliche Dokumente auszuziehen und die Erfassung wird 1). Dementsprechend viele Methoden zur Verbesserung der Erfassung sind bekannt. Eine hohe Präzision zu erreichen ist hingegen sehr schwierig, und die in der Praxis erprobten und in der Literatur erwähnten Methoden zur Verbesserung der Präzision sind entsprechend problematisch.

Als allgemeiner Grundsatz zur Verbesserung der Erfassung gilt das Herbeiführen von zusätzlichen Übereinstimmungen zwischen den den Dokumenten zugehörigen Deskriptoren und den Frage-Deskriptoren. Dies kann erreicht werden durch die Benutzung von Ersatzbegriffen (aus Synonym-Tabelle oder Thesaurus), durch das Verändern von Begriffen, indem zum Beispiel Wörter zerschnitten oder Endungen weggelassen werden oder durch assoziative Techniken, bei denen untersucht wird, welche Begriffe häufig in einem gewissen Zusammenhang vorkommen. Eine Verbesserung der Präzision versucht man durch die Benutzung von stark spezifischen Deskriptoren (Spezial-Begriffe) oder durch die Benützung von Kombinationen von Deskriptoren zu erreichen.

4. Clustering

Die im Zusammenhang mit Bibliotheken entstehenden Files (Dateien) sind im allgemeinen so lang, daß wegen der daraus resultierenden riesigen Suchzeiten eine direkte File-Organisation kaum in Frage kommt. Indexierte Files wären wohl besser, sie sind jedoch wegen der vielen Eingänge und weil es schwierig ist, Indexierungsmerkmale zu ändern, auch nicht besonders empfehlenswert. Die am ehesten den Bibliotheken angepaßte File-Organisation ist das Clustering. Cluster-Files bestehen – wie der Name sagt – aus einer Menge von Clusters, die über einen Index erreichbar sind. Jeder Cluster besteht aus einer Menge von Elementen und wird durch ein Cluster-Profil (auch Zentroid genannt) repräsentiert. Diese Organisation unterscheidet sich von der normalen Bibliothekskartei dadurch, daß Clusters (die Klassen) automatisch erzeugt werden und gegenseitig überlappen können. Im Gegensatz zu vielen anderen File-Organisationen können Cluster-Files relativ einfach umorganisiert werden und die Suchzeiten sind recht akzeptabel.

Die Suche in einem in Clusters aufgeteilten File spielt sich folgendermaßen ab:

1. Der Index (i. a. bestehend aus den Zentroiden) wird durchsucht;
2. Die Dokumentenvektoren der gewählten Clusters werden inspiziert;
3. Die gefundenen Angaben werden in der Reihenfolge ihrer Ähnlichkeit zum Suchvektor ausgedruckt.

Unabhängig von der Organisation der Cluster selbst kann schon auf diesem Grad der Abstraktion die Suchtiefe gesteuert werden, indem nur das

beste, die zwei besten usw. Clusters berücksichtigt werden. Innerhalb eines Clusters kann die Suchtiefe auf analoge Art und Weise variiert werden.

Die in einem Cluster enthaltenen Elemente sind als Funktion ihrer Ähnlichkeit zueinander angeordnet. In anderen Worten: Neue Elemente werden um so näher beim Zentroid plaziert je ähnlicher zu diesem sie sind. Alle benachbarten Elemente sind einander sehr ähnlich, weit voneinander entfernte weisen größere Unterschiede auf. Diese Anordnung innerhalb eines Clusters garantiert, daß bei einer Suche automatisch zuerst diejenigen Elemente gefunden werden, die am ähnlichsten zum Cluster-Zentroid sind.

Das Zentroid dient als Identifikation des Clusters: Es wird gebraucht, um einen Cluster aufzufinden, damit darin nach Objekten gesucht oder neuankommende Objekte dem Cluster zugewiesen werden können. Zentroide sind entweder hypothetische oder reale repräsentative Objekte der zur Diskussion stehenden Klasse und werden – wie auch die Objekte selbst – durch Vektoren der Form

$$P = (p1, p2, \ldots, pn)$$

dargestellt. Die Komponenten p1 . . .pn des Vektors P geben – im einfacheren Fall – lediglich an, ob gewisse Deskriptor-Begriffe im repräsentierten Cluster vorhanden sind. Bei elaborierteren Organisationen enthalten diese Komponenten auch die Häufigkeit und/oder Gewichtung der beschreibenden Begriffe.

Eine Vielzahl von verschiedenen Clusteringmethoden sind entwickelt worden. Die relativ einfachen behandeln Objekt um Objekt und fügen jedes Objekt in einen passenden Cluster ein oder – falls keiner der existierenden Cluster ähnlich genug ist – beginnen sie mit dem Aufbau eines neuen Clusters. Aufwendigere Methoden benötigen eine Vielzahl von Objekten und Zentroiden und arbeiten mit sogenannten Ähnlichkeits-Matrizen, welche als Grundlage für den Aufbau der Clusters dienen. Die einfacheren Methoden haben den Nachteil, daß die aufgebaute Cluster-Struktur von der Reihenfolge der behandelten Objekte abhängt. Experimente hingegen zeigen, daß nicht wesentlich schlechtere Such-Resultate erzielt werden als mit den ausgeklügelteren Methoden.

5. Schlußbemerkungen

Die hier skizzierten Vorgänge sind keine fertigen Rezepte, sondern sollen den Trend der Forschungs- und Entwicklungsarbeiten auf dem Gebiet der Bibliotheksautomatisation aufzeigen. Trotzdem darf man nicht vergessen, daß die meisten dieser Methoden sich im praktischen Einsatz – wenn auch an kleinen Bibliotheken – bewährt haben, daß also gezeigt wurde, daß der eingeschlagene Weg zumindest gangbar ist. Vielfach wurde bei vergleichenden Messungen gefunden, daß relativ einfache Methoden die erforderten Dienst-

leistungen annähernd so gut erbrachten wie andere bedeutend aufwendigere.

Zudem wurde festgestellt, daß der Aufwand erheblich reduziert werden kann, wenn z. B. mit Hilfe von interaktiven Systemen, auch menschliche Eingriffe zugelassen werden. So können sich in Clusters geordnete Klassen recht schnell den Bedürfnissen der Benutzergemeinschaft anpassen, wenn man dafür sorgt, daß die von den Suchenden gelieferte Information (z. B. Zurückweisen von bestimmten Dokumenten) vom System ausgewertet wird. Solche implizite wie auch die expliziten vom Bibliothekspersonal vorgenommenen Einwirkungen gestatten dem System, sich wechselnden Bedürfnissen anzupassen.

Literatur

Cleverdon, C. et al.: ASLIB Cranfield Research Project, Cranfield 1966.

Licklider, J. C. R.: Libraries of the Future, MIT Press, Cambridge, Mass., 1965.

Salton, G., Ed.: The SMART Retrieval System, Prentice-Hall, Englewood Cliffs, N.J., 1971.

Salton, G.: Dynamic Information and Library Processing, Prentice-Hall, Englewood Cliffs, N. J., 1975.

Vickery, B. C.: Zur Theorie von Dokumentationssystemen, Verlag Dokumentation, München, 1970.

Manuelle Erschließung — ein Überblick

B. Glaus
ETH-Bibliothek, 8092 Zürich

Abstract

Bibliothekarische und dokumentalistische Erschließung heißt, standortgebunden oder -unabhängig Dokumente zu ihrer Nutzung formal und inhaltlich anbieten. Dies geschieht unmittelbar (Registratur, Präsenzbibliothek, Vorlage, Zirkulation) oder aber durch Erschließungsmittel (Bibliothekare und Dokumentalisten persönlich, Kataloge, Bibliographien).

1. Begriffliches

Sinnvollerweise kann unterschieden werden zwischen *Bestandserschließung* und *Literaturerschließung*. Jene umfaßt ein «System von Mitteln und Maßnahmen zur Orientierung über Bibliotheksbestände nach formalen und inhaltlichen Gesichtspunkten zum Zwecke ihrer gesellschaftlichen Nutzung».