Robert Fugmann, Maria Isenberg, Jakob Hermann
Winter Hoechst AG, Frankfurt

Das Suchen nach verallgemeinerter Information

Beitrag Nr. 9 zur Theorie von Retrieval-Systemen

(The search for Generalized Information. Treatise IX on Retrieval System Theory.)

Fugmann, R., Isenberg, M., Winter, J.H.: Das Suchen nach verallgemeinerter Information. (The search for generalized information. Treatise IX on Retrieval System Theory).
(In German). Int. Classif. 12 (1985) No. 1, p. 7–10, 4 refs.

When using a mechanized information system, one always runs the risk of phrasing a query too specifically. If a search concept is contained in a stored text in only a slightly generalized variation, then in the traditional Boole'ean search logic this concept will not be retrieved as a response to the query. This is particularly detrimental when the failure to satisfy a search parameter would, in the eyes of the searcher, be more than offset by the occurence of another important and perhaps unexpected concept in the stored text.

In an earlier publication on this topic we described the device of "reverse retrieval". It would permit the retrieval of generalized information with conventional search techniques. This device, however, would be relatively expensive if *several* descriptors of the query and of a text in the store are to be compared in the mechanized matching process.

We now describe a device which makes it possible to retrieve generalized information of the aforementioned kind in a simpler and more versatile manner. It promises to be particularly effective in indexing languages with well developed hierarchies. During "hierarchical weighting" the machine program could assess the degree of generalization in which a search concept occurs in a stored text. It could also be made apparent in the printout of the responses which and how many search concepts occur in a stored text in only a generalized form. Depending on the degree of generalization which one is willing to tolerate for a response to a certain search concept, and depending on for how many and for which concepts one is willing to accept generalization or even entire absence, one could make one's subjective selection from a weighted arrangement of the search responses.

(Authors)

1. Einführung

Begibt sich ein Fachmann auf die Suche nach Literatur zu seinem Arbeitsthema, dann ist es für den einzuschlagenden Weg von großer Bedeutung, inwieweit er sich bezüglich der Begriffe, die in den gesuchten Texten vorkommen sollen, *im Voraus* festlegen kann, d.h. ohne daß er in die Texte potentiellen Interesses vorher hatte Einblick nehmen können. Ein Teil des Informationsbedarfs eines Forschers ist von der Art, daß er ihn überhaupt nicht im Voraus definieren kann. Vielmehr ist er hier *empfänglich* für jede, insbesondere auch unvorhergesehene Art von Information, die er – in ebenfalls unvorhersehbarer Weise – bei seinen Überlegungen verwerten kann. Diese Art von Information ist ihm nur auf dem Weg der *ungezielten Informationsbereitstellung* zugänglich.

Noch am wenigsten Definitionsarbeit wird ihm abverlangt, wenn er lediglich einen Text benennen muß, der für sein Suchthema einschlägig ist und sein Interesse an allen anderen Texten bekundet, die diesem Ausgangstext *in irgendeiner Weise* ähnlich sind, und wenn diese „Ähnlichkeit" darin zum Ausdruck kommen soll, daß in den anderen Texten sein Ausgangstext zitiert sein soll, bzw. daß sein Ausgangstext diese anderen Texte zitieren soll. Verfolgt man ein solches Netzwerk von Zitaten, so nutzt man die Literaturkundigkeit, die die Autoren dieser anderen Texte sich erarbeitet haben, zumindest so weit sie diese Literaturkundigkeit der Wissenschaftlichen Öffentlichkeit durch ihre Zitate zur Verfügung stellen können und möchten. Die Schwächen dieses Verfahrens zu erörtern liegt außerhalb des Themas dieses Aufsatzes.

Ein anderer Weg zur gesuchten Information besteht darin, daß man die Begriffe nennt, die in den gesuchten Texten auftreten sollen. Aber hierbei muß mit großer Sorgfalt abgewogen werden, welche Begriffe in diesen Texten unbedingt auftreten müssen, welche weiteren Begriffe in diesen Texten vollwertig durch andere Begriffe vertreten sein dürfen, und auf welche Begriffe man schließlich gänzlich verzichten könnte. Begriffe der letztgenannten Art werden gerne von den Fragestellern mehr beispielhaft und zur Veranschaulichung des Suchthemas benutzt, sind jedoch nicht als einschränkende Suchbedingungen zu betrachten.

Zuweilen fällt es schwer, überhaupt eine Rangordnung unter den einzelnen Suchbegriffen von der Art herzustellen, daß einige von ihnen unbedingt in den gesuchten Texten auftreten müssen, ein Rest jedoch durch andere vertreten sein kann. Anstelle des herkömmlichen „Boole'schen Retrievals" wäre dann der Weg des „gewichteten Retrievals" vorzuziehen, der Weg also, bei welchem man eine Reihe von prinzipiell gleichrangigen Suchbegriffen vorgibt und lediglich die Bedingung stellt, daß eine Mindestzahl von ihnen in den gesuchten Texten auftreten soll.

Beiden Arten von Retrieval ist es gemeinsam, daß man a priori, d.h. *bevor* man in potentiell interessante Texte hat Einblick nehmen können, eine Grenze oder Schwelle festlegen muß, jenseits welcher die Texte nicht mehr als Antwort auf die gestellte Frage angenommen werden, und zwar ohne daß der Fragesteller hierauf noch einen – evtl. korrigierenden – Einfluß nehmen kann. Dies ist der Preis, den der Fragesteller dafür bezahlen muß, daß er die Suche nicht selbst ausführen kann oder ausführen möchte und daß er sie an jemanden anders oder an einen Suchmechanismus delegiert.

All dies gilt unabhängig davon, ob man es bei den Frage- und Speicherdeskriptoren mit natursprachlichen Wörtern oder mit kunstsprachlichen Notationen zu tun hat.

In Abbildung 1 ist in den Fällen 1a bis 1c die Situation dargestellt, wie sie beim konventionellen Retrieval herrscht, etwa bei einer Fragestellung nach Literatur zur Korrosion bei Kupfer: Ein Treffer wird dann erzielt, wenn der Fragedeskriptor („Ko": Korrosion, „Cu": Kupfer) merkmalsärmer ist als der Speicherdeskriptor (Fall 1a) oder allenfalls die gleichen Merkmale aufweist wie der Speicherdeskriptor (Fall 1b).

Im Fall 1c trifft dies nicht zu. Deswegen wird in diesem Fall kein Treffer erzielt. Wir haben diese Suchstrate-

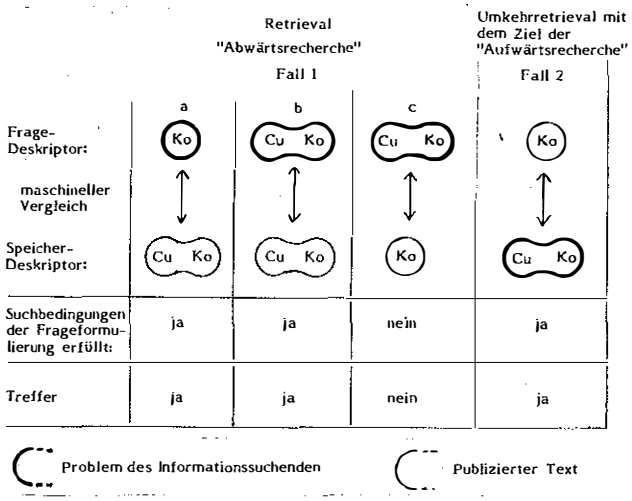


Abb. 1: Mechanismus von Retrieval und Umkehrretrieval

gie als „Abwärtsrecherche" bezeichnet, weil auf diesem Wege stets diejenigen Texte ermittelt werden, die in der Begriffshierarchie unterhalb des Fragedeskriptors stehen oder allenfalls auf gleicher Höhe.

2. Probleme bei der Bereitstellung von verallgemeinerter Information

Wir betrachten als Beispiel eine Fragestellung nach Literatur zum Thema Lochfraßerscheinungen an Kupferrohren. Wir müssen, bevor wir sie als Suchauftrag für eine Boole'sche Recherche formulieren, prüfen, ob sie nicht vielleicht allzu spezifisch ist, d.h. bedenklich viele begriffliche Merkmale enthält. Man darf als Antwort auf eine solche Fragestellung nämlich keine Texte erwarten, in denen anstatt von „Lochfraß" nur – allgemeiner, d.h. merkmalsärmer – von „Korrosion" die Rede ist. Vielleicht möchte der Fragesteller noch nicht einmal auf Texte verzichten, in denen – noch allgemeiner – nur von „Korrosion an kupfernen Werkstücken" jeglicher Art die Rede ist, also auch unter Einschluß beispielsweise von Kupferblechen. Sucht man gar in den Buchbeständen einer Bibliothek, so kann man nicht damit rechnen, daß der Inhalt der Bücher in so großer Detailliertheit indexiert ist, daß man derartig spezifische Fragen stellen kann. In diesem Falle würde man sich vorsorglich mit „Korrosion bei Buntmetallen" oder gar mit „Metallkorrosion" begnügen, denn bei einem so indexierten Buch kann man damit rechnen, daß dort auch spezielle Korrosionserscheinungen bei Kupfer abgehandelt sind.
bei einem so indexierten Buch kann man damit rechnen, daß dort auch spezielle Korrosionserscheinungen bei Kupfer abgehandelt sind.

Besonders interessant können für einen Fragesteller solche Texte sein, in denen einerseits ein verlangter Begriff nicht auftritt, andererseits aber wichtige Zusatzinformation geboten wird. So mag ein Chemiker Literatur über die Herstellung eines ganz bestimmten Stoffes suchen, würde aber zweifellos auch Texte gerne als Antwort entgegennehmen, in denen nicht die Herstellung, wohl aber stattdessen die Toxizität oder Explosivität des gesuchten Stoffes erwähnt ist. – Ähnlich ist die Sachlage in der Patentliteratur. Ein Chemiker interessiert sich vielleicht für die Herstellung eines ganz bestimmten Penicillinkörpers, z.B. von „Penicillin G", und in einem Patentanspruch ist die Herstellung von „Penicillin allgemein" geschützt. Wieder findet er seine ganz spezielle Suchbedingungen nicht vollständig erfüllt, und trotzdem ist ein solcher Patentanspruch für ihn von großem Interesse.

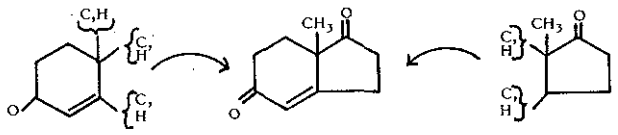
Um solche Informationsverluste zu vermeiden, beschreitet man in der Praxis zumeist den Weg, daß man intuitiv und mit viel Erfahrung diejenigen Begriffe und begrifflichen Merkmale aus einer Fragestellung herausläßt, von denen man vermutet, daß sie von einem interessanten Text möglicherweise nicht erfüllt werden könnten. Man setzt sie zumindest alternativ zu anderen Begriffen, deren Auftreten in einem Text man für erwünscht, jedoch nicht für unerläßlich hält.

Ein solches Verfahren kann aber immer nur eine Notlösung darstellen. Man kann nämlich immer nur mutmaßen, in welcher Richtung und wie sehr die interessierende Information verallgemeinert sein könnte und welcher Art die Begleitinformation sein könnte, die einen Text trotz alledem noch für den Fragesteller interessant machen könnte. Stellt man beispielsweise die Frage nach Herstellung des Stoffes X, dann könnte auch eine Literaturstelle interessant sein, in welcher wiederum nicht die Herstellung, auch nicht die Toxizität, sondern eine besonders kostengünstige Bezugsquelle oder die überraschend beobachtete Heilwirkung bei einer bestimmten Krankheit beschrieben ist. Anders ausgedrückt, kann man nicht im Voraus wissen, welche anderen Begriffe in einem Text einen bestimmten Suchbegriff, z.B. den der Herstellung, aus der Sicht eines Fragestellers vollwertig vertreten können. Der Stoff X seinerseits müßte durch Weglassen von zahlreichen begrifflichen Merkmalen nach den verschiedensten Richtungen hin und sehr weitgehend verallgemeinert werden, wollte man jeden nur denkbaren Verlust an möglicherweise interessanter (obwohl die gestellten Suchbedingungen nicht voll erfüllender) Information ausschließen. Man müßte, anders ausgedrückt, auf so viele Suchbegriffe gleichzeitig verzichten, daß das Suchergebnis meistens untragbar ballastreich wäre.

3. Möglichkeiten und Grenzen vom Umkehrretrieval

Im Vorstehenden haben wir die Zielsetzung der „Aufwärtsrecherche" betrachtet, mithilfe welcher (wenn auch mit unzulänglichen Hilfsmitteln) merkmalsärmere, z.B. verallgemeinert eingespeicherte Information aufgefunden werden soll. Wir haben für das gleiche Ziel kürzlich das Hilfsmittel des Umkehrretrievals beschrieben (1), (2). Im Fall 1c kommt es dadurch zum Treffer beim maschinellen Vergleich von Frage- und Speicherdeskriptor, daß der in einem gespeicherten Text angetroffene Deskriptor als Suchbedingung benutzt wird (Ko), und daß die Begriffe der Fragestellung in einen (vorübergehenden) Speicher eingebracht werden. Nach einer solchen Umkehr der Funktionen von Fragestellung und Speicher (Fall 2) ist nun wieder die Bedingung erfüllt, daß der „Frage"-Deskriptor merkmalsärmer ist als der „Speicher"-Deskriptor.

Das Umkehrretrieval findet in großem Umfang Anwendung bei der computergestützten chemischen Syntheseplanung (3). Hier ist dem Maschinenprogramm die Aufgabe gestellt, für eine spezifische Einzelverbindung, z.B. Verbindung A in Bild 2, Verallgemeinerungen in



B A C

Abb. 2: Analogieschluß von den Allgemeinbegriffen B und C auf A

der Literatur zu suchen, z.B. die Strukturen B oder C, für welche zugleich auch Herstellungsmethoden beschrieben sind. Man folgert auf dem Wege des Analogieschlusses, daß eine Herstellungsmethode, die für chemische Verbindungen aus der Klasse B oder C geeignet ist, wahrscheinlich auch für die Herstellung der Verbindung A eingesetzt werden kann, weil auch die Verbindung A den Klassen B und C angehört.

Der Einsatz des Umkehrretrievals zur Voraussage von pharmakologischen und toxikologischen Eigenschaften von Stoffen wurde kürzlich beschrieben (4).

Wenn nun der maschinelle Vergleich zwischen jeweils *mehreren* Fragedeskriptoren und *mehreren* Speicherdeskriptoren stattfindet, so kann es zu Komplikationen sowohl beim herkömmlichen, als auch beim Umkehrretrieval kommen. Es kommt hier für das Zustandekommen eines Treffers nämlich darauf an, daß *sämtliche*, in den Vergleich einbezogene Fragedeskriptoren merkmalsärmer sind als ihre Pendants unter den Speicherdeskriptoren, oder daß sie allenfalls die gleichen Merkmale aufweisen. Ist dies beim Retrieval nur *teilweise* der Fall, wie in Bild 3, Fall 3 dargestellt, so kann das Umkehrretrieval (Fall 4) nicht zum Ziel führen. Hier lautet das Problem des Informationssuchenden „Lochfraß bei Metallen", und es kommt beim maschinellen Vergleich deswegen nicht zum Treffer, weil der spezifische Begriff „Lochfraß" nicht von dem allgemeinen Begriff „Korrosion" erfüllt wird. Geht man nun zum Umkehrretrieval über (Fall 4 in Bild 3), so wird zwar diese Suchbedingung erfüllt, dafür schlägt aber der Vergleich von Kupfer (als Fragedeskriptor verwendet) mit dem Allgemeinbegriff „Metall" (als Speicherdeskriptor verwendet) fehl, denn der Speicherdeskriptor ist allgemeiner, d.h. merkmalsärmer als der Fragedeskriptor.

Abhilfe könnte in dieser Situation dadurch geschaffen werden, daß man jedes Thema eines Informationssuchenden in größtmöglicher Spezifität mit Speicherdeskriptoren beschreibt. Beispielsweise könnte der Allgemeinbegriff „Metall" so dargestellt werden, daß er jeden Fragedeskriptor nach einem speziellen Metall erfüllt. Man könnte also beispielsweise den Begriff „Metall" in mehrere Dutzend spezielle Metallbegriffe übersetzen (z.B. Kupfer, Eisen, Zinn, Zink usw.), wie es im Fall 5 von Bild 3 dargestellt ist. Hiermit wäre jedoch meistens ein beträchtlicher Aufwand verbunden.

Weiterhin müßten durch ein geeignetes Maschinenprogramm aus der Menge der Deskriptoren für einen gespeicherten Text einige entfernt werden, bevor diese Deskriptoren als Suchbedingungen verwendet werden könnten. Dies betrifft diejenigen Deskriptoren eines Textes, denen kein hierarchisch verwandter Deskriptor im Problem des Fragestellers gegenübersteht. Ist beispielsweise in einem publizierten Text außer von Kupfer und Korrosion auch von Kunststoff die Rede gewesen (Fall 6 in Bild 3), so wäre für das Umkehrretrieval der Deskriptor für „Kunststoffe" aus dem Satz der Fragedeskriptoren zu entfernen, weil ihm kein hierarchisch verwandter Deskriptor im Problem des Informationssuchenden gegenübersteht. Unterbleibt diese Elimination des Deskriptors „Kunststoff", so würde beim Umkehrretrieval der Vergleich an diesem (für den Fragesteller wenig wesentlichen, weil von ihm nicht genannten) Deskriptor scheitern.

Auch wenn man das Umkehrretrieval verwirklicht, bei welchem publizierte Texte als Fragedeskriptoren dargestellt werden und das Thema des Informationssuchenden mit Hilfe von Speicherdeskriptoren wiedergegeben wird, benötigt man in der Informationspraxis parallel hierzu stets das reguläre Retrieval. Hier ist in gewohnter Weise das Thema des Informationssuchenden als Fragestellung formuliert und publizierte Texte durch Speicherdeskriptoren. Man benötigt also für diese Kombination zwei grundverschieden organisierte Datenbanken, womit ein beträchtlicher Aufwand verbunden wäre. Aus diesem Grund und noch anderen Gründen haben wir die Variante des Umkehrretrievals zum Aufsuchen von verallgemeinerter Information vorerst nicht weiter vervollkommnet, sondern stattdessen einen anderen Weg näher untersucht, einen Weg allerdings, der eine Modifikation der herkömmlichen Suchtechnik erfordert und auf eine spezielle Variante des gewichteten Retrievals hinausläuft.

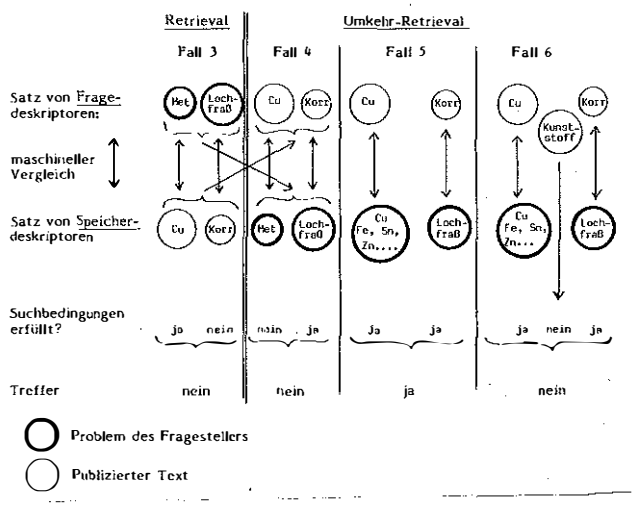


Abb. 3: Verlauf von Retrieval und Umkehrretrieval

4. Das „hierarchische Gewichten" von verallgemeinerter Information

Dieser Weg besteht darin, daß man für jeden einzelnen Fragedeskriptor in dem betreffenden System explizit definiert, durch welche allgemeineren, merkmalsärmeren Speicherdeskriptoren er gegebenenfalls erfüllbar sein soll. Es wird also, anders ausgedrückt, festgelegt, auf welche begrifflichen Merkmale vom Thema eines Informationssuchenden man unter Umständen zu verzichten bereit wäre. Beispielsweise kann auf diese Weise in entsprechenden hierarchischen Leitern festgelegt werden, daß „Kupfer" (unter bestimmten Umständen) auch durch „Metalle" erfüllbar sein soll, und daß „Lochfraß" auch durch den Allgemeinbegriff „Korrosion" erfüllbar sein soll. Findet nun das Frageprogramm einen nicht erfüllten Suchbegriff vor, so verwirft es nicht gleich den entsprechenden Speichertext als unzutreffend, sondern ent-

nimmt der hierarchischen Leiter diejenigen allgemeineren Suchbegriffe, die unter bestimmten Umständen an die Stelle der ursprünglichen, spezifischeren Suchbegriffe treten dürfen. Erst wenn keiner von diesen Allgemeinbegriffen erfüllt ist, würde das Maschinenprogramm den betreffenden gespeicherten Text als nicht einschlägig zurückweisen.

Dieses Verfahren ist auch insofern besonders flexibel, als Ausweichbegriffe verwendet werden können, die aus anderen Hierarchien stammen als der Ausgangsbegriff. Beispielsweise kann die Suchbedingung „Herstellungsverfahren" ersetzbar sein durch „Toxizität", oder auch „Explosivität". Dann findet man beim Suchen nach der Herstellung eines Stoffes auch einen Text, in dem nicht seine Herstellung, wohl aber seine Toxizität oder Explosivität beschrieben ist.

Läßt man es zu, daß ein spezifischer Suchbegriff durch einen gespeicherten allgemeineren Begriff aus der gleichen Hierarchie erfüllt wird (oder sogar durch einen Ausweichbegriff aus einer anderen Hierarchie), dann ist es allerdings nötig, daß unter den Antworten eine gewichtete Ordnung hergestellt wird: Je stärker ein gespeicherter Begriff gegenüber einem Suchbegriff verallgemeinert ist, und je mehr Suchbegriffe nur durch Verallgemeinerungen erfüllbar gewesen sind, desto weiter entfernt sich ein gespeicherter Text vom Thema des Fragestellers. In einem Speicher allerdings, in dem man es mit Buchtiteln oder mit Patentansprüchen zu tun hat, kann man auch relativ weitgehende Verallgemeinerungen akzeptieren. Wenn man für jeden Schritt aufwärts in der Hierarchie bei einem gespeicherten Begriff sozusagen einen Minuspunkt vergibt, dann ist die Zahl der Minuspunkte pro gespeichertem Text ein Maß dafür, wie weit sich dieser Text vom Thema der Fragestellung entfernt.

Für die Bewertung eines Antworttextes wird es einen Unterschied ausmachen, *welcher* besondere Suchbegriff nur in mehr oder minder stark verallgemeinerter Form erfüllt wurde oder überhaupt nicht. Im Druck der Antworten kann deutlich gemacht werden, mit wie vielen Minuspunkten ein nicht vollständig erfüllter Suchbegriff belegt worden ist und ob er überhaupt – wenn auch nur in sehr verallgemeinerter Form – erfüllt worden ist. Dann kann jeder Fragesteller das Ergebnis der Recherche in flexibler und subjektiver Weise interpretieren und auswerten. Wir schlagen für ein solches Verfahren die Bezeichnung „hierarchisches Gewichten" vor.

Im GREMAS-System zur Dokumentation von chemischen Verbindungen und Verbindungsklassen sind die Deskriptoren in eine fein abgestufte und ausgedehnte Hierarchie eingebettet. Erste Modellversuche in diesem System haben ergeben, daß das oben geschilderte Verfahren zumindest auf diesem Gebiet aussichtsreich ist.

Quellen.

(1) Fugmann, R., Winter, J.H.: Reverse Retrieval: Toward Analogy Inferences by Mechanised Classification. Intern. Classification 6 (1979), 85–91.

(2) Fugmann, R., Kusemann, G., Winter, J.H.: The Supply of Information on Chemical Reactions in the IDC-System. Information Processing & Management 15 (1979), 303–323.

(3) Winter, J.H.: Chemische Syntheseplanung. Springer-Verlag 1982.

(4) Kaufman, Joyce: Prediction of Toxicology and Pharmacology Based on Model Toxicophores and Pharmacophores Using the New TOX-MATCH-PHARM-MATCH Program. International Journal of Quantum Chemistry: Quantum Biology Symposium 10 (1983), 375–416.

Herrn Dr. Suhr von der Firma BASF AG danken wir für wesentliche Beträge zu diesen Überlegungen.

Address of Authors:
Hoechst AG, Postfach 800320, D-6230 Frankfurt 80