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-