Innovative Retrievalsysteme für die Praxis

GI - Workshop für Anbieter und Anwender veranstaltet von GI-Fachgruppe Information Retrieval, GMD-IPSI und Fachhochschule Darmstadt (Fb IuD) GMD Darmstadt ,13. und 14. November 1996

Gerhard Knorz,
Fachhochschule Darmstadt

1 Zusammenfassung

Keine gute Zeit für Tagungen - das ist die gegenwärtige Erfahrung so mancher Organisatoren mit dem Effekt, daß der Terminkalender vielbeschäftigter Personen in diesem Herbst mehr Freiräume bietet, als es die Ankündigungen im Frühjahr hatten erwarten lassen. Doch es gibt Ausnahmen! Wer hätte zu Beispiel gedacht, daß ein kleiner Hochschulverband für Informationswissenschaft zur ISI'96, Berlin offensichtlich mühelos dreimal mehr Tagungsteilnehmer zusammenbekommt, als er Mitglieder hat? Oder daß sich ohne große Werbekampagne Anbieter großer Systeme wie Fulcrum oder BRS Search zusammen mit 13 anderen Ausstellern/Exponaten und über 40 Teilnehmern zu einem sehr informativen und lebhaften zweitägigen Workshop in Darmstadt treffen? Initiativ geworden war die Fachgruppe Information Retrieval in der Gesellschaft für Informatik, die Anwender, Anbieter, Entwickler und Forscher zusammenbringen wollte, um an praktischen Beispielen und konkreten Projekten zu diskutieren, was gegenwärtig unter Innovation im Retrievalbereich zu verstehen ist und in welche Richtung aktuelle Anforderungen und Angebote weisen. Umgesetzt wurde der Anstoß von Ulrich Thiel (GMD) und Gerhard Knorz (FHD) , unter deren Federführung sich Darmstadt einmal mehr im Hinblick auf Geographie und Infrastruktur als ein sehr geeigneter Ort für eine Veranstaltung dieser Art bewies. Das Institut für integrierte Publikations- und Informationssysteme (IPSI) der GMD sorgte im technischen und organisatorischen Bereich für eine entspannt-anregende Workshop-Atmosphäre und steuerte substantiell zum inhaltlichen Programm bei. Studierende der FH werden einen Beitrag dazu leisten, die Dokumentation des Workshops auch im Bereich der Exponate vollständig und informationsreich zu publizieren.

Was den Workshop auszeichnete war ein gesprächsbereicher heterogener Teilnehmerkreis, der ein facettenreiches Angebot im Vortrags- und Ausstellungsbereich dazu nutzte, dazuzulernen, selbst eigene Erfahrungen einzubringen und neue Kontakte zu knüpfen.

2 Vortragsprogramm

2.1 *Übersicht*

Da sowohl Vorträge als auch die Systeme im Ausstellungsbereich im WWW unter der URL http://www-cui.darmstadt.gmd.de/mindlIR-WS.html mehrheitlich mit Abstracts dokumentiert sind, kann ich mich auf einen knappen Überblick beschränken und das aus meiner Sicht besonders Interessante punktuell hervorheben.

Der Workshop begann mit einem eingeladenen Vortrag von Jürgen Krause (Universität Koblenz/IZ Sozialwissenschaften) über Visualisierung und Information Retrieval und bot insgesamt 9 Beiträge und 2 (Podiums)Diskussionen an:

Konzepte von Retrievalsystemen vermittelten
Adrian Müller (GMD): "Das Multimedia Retrievalsystem MIRAKEL", ein GMD-Prototyp, der Abduktion als Grundlage für ein auf Logik-basiertes Retrieval verwendet

- A. Schappert (Siemens): "Personalized Information Filtering", die Aufgabe für einen firmenintern zu nutzenden Prototyp, der auf dem Vektormodell aufbaut und das Problem lösen soll, die Flut von Beiträgen etwa in Newsgroups für Mitarbeiter sinnvoll zu filtern.
- Ulrich Pfeifer: "FreeWAIS-sf', die an der Universität Dortmund betriebene Weiterentwicklung von WAlS mit einer Reihe interessanter Features und einer ganz erstaunlichen weltweiten Verbreitung.

Eher theoretische und grundsätzliche Aussagen zu Information Retrieval und adäquaten Ansätzen dazu lieferten:

- Gerhard Rahmstorf: "Semantisches Retrieval: Ansatz, Erfahrungen, Perspektiven". Themen in Anfrage und Dokumenten werden als Begriffe angesehen, deren differenziert zu bestimmende Relation darüber entscheidet, ob ein Dokument als sinnvolles Antwortdokument gelten kann.
- Robert Fugmann: "Durch Indexierungssprachen-Grammatik zu höherer Retrievalqualität". Eine differenzierte intellektuelle Erschließung, bei der im Prinzip Frame-Slots mit Einträgen aus einem kontrollierten Vokabular gefüllt

werden, wird als Voraussetzung für ein selektives Retrieval in großen Datenbeständen postuliert.

Einen Überblick über Entwicklungsrichtung und Entwicklungsstand bei der Anwendung computerlinguistischer Analysen für ein inhaltsorientiertes Retrieval beim IBM Entwicklungslabor Böblingen gibt

Sebastian Göser: "TextMiner, ein inhaltsbasiertes Information Retrieval-Systern".

Wie vielversprechend ein Transfer von der Hochschule in die Informationspraxis sein kann stellten R. Domenig und D. Linder dar:

. "Eine neue Dienstleistung: Aktuelle Rechtsinformationen Online" ist das Ergebnis der Zusammenarbeit eines Rechtsverlags mit einem Spin-off der ETH Zürich, die das dort entwickelte System Spider (P. Schäuble) in eine Produktentwicklung einbringt.

Ein strategischer Vortrag über die Rolle von Information Retrieval Funktionalität in neuen Architekturen eines Marktes mit konkurrierenden globalen Playem zog die Teilnehmer in seinen Bann.

Jürgen Harbarth (Software AG): "Information Retrieval und Datenbanken ein ungleiches Paar?"

Zwei Diskussionen ließen sich vom engen Zeitrahmen nicht beeindrucken:

Innovative IR-Systeme: Gehen Entwicklung und Praxis den richtigen Weg? Eine Diskussion, die die gegenwärtige Politik und die IR-Forschung folgerichtig ins Thema einbezog und die es dem Plenum überlies, sich sein eigenes Bild zusammenzupuzzeln, aus pessimistischer und optimistischer Analyse, die sicher beide nur Teile der realen Situation widerspiegelten. Es diskutierten auf dem Podium: Fugmann, Harbarth, Knorz, Krause, Rahmstorf, Schwantner, Thiel.

TREC (Text Retrieval Conference) hat unstrittig die IR-Szene nachhaltig verändert und beflügelt. Gerhard Knorz stellte die Ziele und die Struktur dieses "offenen Wettbewerbs für Retrievalsysteme" vor, die erstmals forschungsorientiere Ansätze und kommerzielle Ansätze aufeinander bezogen und in ihrer Leistungsfähigkeit vergleichbar gemacht hat. Michael Kluck vom IZ-Sozialwissenschaften berichtete im Anschluß von den weitgehend abgeschlossenen Vorbereitungen, mit Daten von SOLlS und FORIS eine Evaluierungsumgebung aufzubauen, die u. a. ein Retrieval unter Nutzung von manueller Erschließung als Vergleichsmaßstab mit einbeziehen kann. Unter Moderation von Dr. Reginald Ferber stieß die Thematik in der Diskussion auf großes Interesse.

2.2 *Einschätzungen und Schlußfolgerungen*

Den absolut gelungenen Auftakt verdankt der Workshop dem Vortrag von Jürgen Krause: Überaus pointiert und anschaulich arbeitete er heraus, welche Formen von (graphischen) Benutzerschnittstellen heute üblich und möglich sind, welche Kämpfe um unterschiedlichen Philosophien dabei ausgetragen werden (Metapher vs. visual formalism), wie unsicher und unvollständig unser gegenwärtiges Wissen darüber ist, wie "gute" Benutzerschnittstellen auszusehen haben, und inwieweit dieses Wissen auf Experten verteilt ist, die sich kaum kennen (Software Ergonomie – Gestaltung7Design).

Es wird so schnell wohl kaum ein Teilnehmer vergessen, mit welchem schlagenden Befund Krause demonstrierte, daß analytisches Denken die Empirie nicht ersetzen kann: Wenn eine optische Täuschung dem Menschen unterschiedliche Größen von Gegenständen vorgaukelt, wird sich dies natürlich beim Greifen nach diesen Gegenständen so auswirken, wie dies auch bei realen Größendifferenzen der Fall ist - so sollte man logischerweise meinen. Die empirische Nachprüfung beweist das Gegenteil: Die Hand greift richtig und nur das visuelle Bewußtsein wird getäuscht!

Gegenüber großen Tagungen hat ein Workshop seine spezifischen Vorteile. Einer davon ist, daß es nicht falsch verstanden wird, wenn Ideen und Systeme vorgestellt werden, die rur Impulse und Veränderungen noch offen sind. Und daß es ein legitimes Ziel eines Vortrags sein kann, in einer konstruktiven Diskussion abzuklären, wo die kritischen Punkte eines Entwurfs liegen. Wenn eine Diskussion sich die angemessene Zeit nehmen kann, auf einem unterschiedlichen Kenntnisstand und Erfahrungshintergrund das anfängliche Mißverstehen aufzuarbeiten, dann haben in der Regel alle Seiten dazugelernt. So zum Beispiel bei dem Ansatz der Informationsvermittlungsstelle im Geo-Forschungs-Zentrum Potsdam, das nicht mittels Indexierung von den Texten zu den Recherchefragen, sondern mittels kleinerer Spezialthesauri von den Fragen zu den Dokumenten kommen will. Oder bei dem Information Filtering Ansatz von Siemens, der zahlreiche Ideen im Plenum provozierte.

Das Thema "Innovation" ist sicher zu vielschichtig, als daß der Workshop den ernsthaften Versuch hätte wagen können, zu systematisieren und zu definieren. Schlagwörter aus Vortrag und Diskussion, konkretisiert durch Beispiele der ausgestellten Systeme und Prototypen spannen aber zumindest einen Erwartungsraum auf:

- Ranking ist mittlerweile eine selbstverständliche Option der meisten Systeme geworden (in welcher Qualität auch immer). Bei Relevance-Feedback- Ver-

fahren sieht es noch anders aus. Die überzeugend negativen Ergebnisse für Boolesche Retrievalverfahren im Rahmen der TREC-Evaluierung haben Wirkung gezeigt. Zwar wählt hierzulande wohl kein Anwender sein Retrievalsystem anhand der TREC-Proceedings aus (Schwantner), aber die Entwicklungsabteilungen kommen an den Ergebnissen nicht vorbei und verwenden selbstverständlich das Testmaterial, um sich hinsichtlich Retrievalqualität positionieren zu können. In den USA haben die TREC-Ergebnisse dazu geführt, daß probabilistische Systeme bei großen Anwendungsprojekten einen klaren Wettbewerbsvorteil besitzen. Sie sind dabei, zum Standard zu werden. *(S.* Göser, IBM)

Multilingualität ist für viele Marktsegmente und Systeme eine wichtige Voraussetzung. Dies bedeutet gegenwärtig mehrsprachige Bedienerführung, Zeichensätze und Stemming-Algorithmen (incl. Kompositazerlegung).
Die Verwaltung multimedialer Daten hat nicht nur einen hohen Aufmerksamkeitswert sondern auch praktische Relevanz.
Dokumente und Anwendungen haben sich von einfachen Standardfällen hin zu einer heterogenen Breite entwickelt. Ähnlich, wie die Verbreitung relationaler Datenbanktechnologie zum neuen Gebiet der Non-Standard-Datenbanken geführt hat, ist eine größere Vielfalt an praktisch und erfolgreich eingesetzten Retrievalmethoden zu erwarten (Knorz). Auch Anätze, die jenseits des Main Streams entstehen, werden zunächst für Spezialanwendungen ihre Chance haben.

- Wo und in welcher Form ein Informationsproblem als Retrievalproblem zu formulieren ist, ist keineswegs unveränderbar und vorgegeben. Die Strukturierung größerer Antwortmengen mit dem Ziel, ein verfeinertes Retrieval durch Navigation zu ersetzen, ist ein Beispiel dafür.
  Die effektive Unterstützung des Retrieval durch geeignete graphische Oberflächen bis hin zu 3D-Schnittstellen hat viele Ideen aber nur wenige überzeugende Grundformen hervorgebracht. Gegenwärtig muß man davon ausgehen, daß viele Ansätze entweder gar nicht oder aber nur für einen engen Ausschnitt von Retrievalproblemen geeignet sind (Krause). Visualisierung bleibt ein spannendes Thema.
- Information Retrieval kann nicht unabhängig von Datenbanken bzw. allgemeinen unternehmensweiten Architekturen für Informationssysteme gesehen werden. Dies verbindet Faktenextraktion mit Information Retrieval und wird im Markt solche Methoden und Anbieter nach vorne bringen, die sich als Third Party-Anbieter in marktführende Softwarearchitekturen einklinken (Harbarth).

Man wird auch für die Zukunft akzeptieren müssen, daß der einfache Informationszugriff durch den gelegentlichen Nutzer und die zuverlässige hochleistungsfähige Recherche durch Informationsspezialisten in vieler Hinsicht gegenläufige Anforderungen an Schnittstelle und Methodik von IR-Systemen stellen.

3 Ausstellung

Die Ausstellung, organisiert von Dr. K. Tzeras, war mit ihren 14 Exponaten naturgemäß die besondere Attraktion für die Workshop-Teilnehmer. Die eigens für Demonstrationen reservierte Zeit verging wie im Flug und hätte vermutlich noch großzügiger eingeplant werden können. Es werden in den Proceedings des Workshops auch Beschreibungen der ausgestellten Systeme und Prototypen integriert sein, so daß Interessierten eine angemessene Informationsquelle zur Verfügung gestellt werden wird. Ich will mich hier darauf beschränken, die einzelnen Systeme zu nennen und aus meiner ganz subjektiven Sicht heraus ausgewählte Demonstrationen kurz zu charakterisieren. Der Hintergrund für eine gewisse Zurückhaltung ist tatsächlich auch die Tatsache, daß es mir persönlich keineswegs möglich war, alle ausgestellten Systeme näher zu betrachten.

- Einige der Exponate waren direkt auf einen der Vorträge des Workshop-Programms bezogen:
- FreeWAIS-sf+SFgate (Ulrich Pfeifer, Universität Dortmund)
- Knowledge Browser (Anne Otto, Informationsvermittlungsstelle im Geo-Forschungs-Zentrum)
- MIRACLE (Adrian Müller, GMD-IPSI, Darmstadt) Rechtsinformationen Online (R. Domenig, Eurospider Information Technology AG Zürich) • WWW-Anwendung auf einem Teil der ZDF-Agenturdatenbank (Jürgen Harbarth, Software AG)
- Die bekannten kommerziellen Retrieval-Engines wurden auf dem Workshop in Kurzbeiträgen vorgestellt, was sich wohl an dieser Stelle erübrigt:
  - Fu1crum Search Server (Herr Kohlhepp, Fu1crum GmbH)
  - . BRS/Search und NetAnswer (Robert Romanski, Dataware Technologies GmbH; München)
- Aus Hamburg bzw. Overijse, Belgien wurde das Retrievalsystem Di-Ret präsentiert, das auf etliche Evaluierungen verweisen kann und das hauptsächlich auf einer Relevance Feedback-Technik basiert (Dr. Carlo Vemimb, Information Services Consultants, GmbH)

SpaCAM - Textretrieval und Textanalyse mit spärlich codierten Assoziativmatrizen, vertreten durch die Universität Hildesheim (Prof. Dr. H.J. Bentz, Dirk Hollenbach) und connex software GmbH Hildesheim (Michael Hagström). Der generische Ansatz, der verschiedensten Problemstellungen im Bereich Retrieval, Sprachverarbeitung und Mustererkennung flexibel angepaßt werden kann versucht, Objekte (im IR: Wörter) problemangepaßt durch wenige Merkmale eines großen Merkmalsraums zu beschreiben und ein geeignetes Ähnlichkeitsmaß zu definieren.

- Neuronales Information Retrieval- Dr. Friedhelm Schwenker, Neuroinformatik Universität Ulm
- Part-Libraries. Indexierung von Informationsinhalten für den Machinenbau. Dipl-Psych. P. Wingartz, Research+Development GbR, Kempen

Systeme, die in der GMD oder in Projekten mit der GMD entstanden bzw. die in der GMD selbst verwendet werden:

INQUERY (Center for Intelligen Information Retrieval, Amherst Mass. USA) LyberWorld - 3D- Visualisierung für Datenbank- und IR-Systeme

- (Matthias Hemmje, GMD-IPSI) Speak! Dialogue Modelling for Speech Generation in Multimodal Information Systems (Dr. Adelheit Stein, GMD-IPSI) Virgilio - Virtual Reality Scenes Generation for Large and Structured Data
- Objekts. Aldo Paradiso; D.I.S. Universita di Roma "La Sapienca" und GMD-IPSI. Das System generiert 3-D- Welten zur
- Darstellung von Antwortmengen, z.B. eine Gebäudemetapher für hierarchisch angelegte Dokumentenc1uster, in der der Nutzer navigieren kann.

4 Proceedings

Die Beiträge für die Proceedings werden - soweit nicht schon geschehen - bis Anfang Dezember endgültig vorliegen. Parallel dazu und in Abhängigkeit von Substanz und Möglichkeiten wird über die Art der Publikation (konventionell oder Netz, Verlag oder eigene Institution) entschieden werden. Anfragen können an die Veranstalter gerichtet werden.