SIGIR '94

Die diesjährige Konferenz der Special Interest Group Information Retrieval (SIGIR) der ACM fand vom 4. bis 6. Juli 1994 an der Dublin City University unter der Leitung von Alan Smeaton statt. Insgesamt kamen ca. 270 Teilnehmer aus 20 Ländern (inkl. Australien, Südkorea, Singapur, Japan, Slovenien, USA, Canada und den meisten europäischen Staaten). Ca. ein Drittel erhielt Mittel aus dem HCM-Förderfond der EU. SIGIR '94 wurde im Tagungsband von dessen Herausgebern Bruce Croft und Keith van Rijsbergen als die bisher größte SIGIR Konferenz angekündigt, da 35 reguläre Vorträge, zwei eingeladene Vorträge, zwei Panels, acht Tutorials, ein ACM-SIGIR Preisvortrag und mehrere Systemdemonstrationen stattfanden. Trotz dieser angekündigten Größe konnte man allerdings den Eindruck gewinnen, daß die inhaltliche Breite der Vorträge relativ gering war. Vor allem spiegelten sich neue Ansätze und IR-Paradigmen mit wenigen Ausnahmen (z.B. die Vorträge von Peter Ingwersen oder Matthias Hemmje) nur in den beiden Panel-Sitzungen zur Integration von IR und Datenbanken und zur Evaluierung Interaktiver Retrieval Systeme wider. Sowohl in vielen Vorträgen als auch Diskussionsbeiträgen wurde die Dominanz eines sehr traditionellen, engen Verständnisses von IR deutlich.

Enttäuschend war vor allem der eingeladene Vortrag von Jaime Carbonell, der sich im Gegensatz zum Vortragstitel nur sehr rudimentär mit IR und noch rudimentärer mit der Verzahnung IR und Natural Language Processing befaßte und im wesentlichen einen State-of-the-Art zum NLP im Bereich der Maschinellen Übersetzung behandelte. Der eingeladene Vortrag von Dennis Tsichritzis stellte vor dem Hintergrund der aktuellen Hardware-Entwicklung den Wandel des Informationsbegriffs innerhalb des Informationsprozesses von Zahlen, über Daten und Texte zur Vielschichtigkeit multimedialer Information dar und zeigte Bereiche auf, mit welchen sich IR-Leute aufgrund dieser Veränderungen werden auseinandersetzen müssen (z.B. die Erschließung dieser neuen Informationstypen). Insofern hätte dieser Vortrag als richtungsweisend für die Konferenz angesehen werden können, wären ihm weitere, mit ähnlich aktueller Thematik gefolgt.

Nun kurz zu den regulären Vorträgen der Tagung, die sich auf zwölf Sektionen verteilten, wovon einige parallel zueinander oder den Panel-Sitzungen stattfanden.

In der ersten Sektion zur Textkategorisierung fanden vier Vorträge statt. David Lewis und William Gale befaßten sich mit einem Trainingsalgorithmus zum Klassifizieren von Texten, wobei der Schwerpunkt auf dessen Effizienz lag. Yiming Yang stellte mit Expert Network einen Ansatz vor, der auf der Basis menschlicher Entscheidungen ein Netzwerk erstellte, welches die Links zwischen den Knoten aufgrund von Wort- und Kategorienverteilungen im Trainingsset berechnet. In der durchgeführten Evaluierung zeigte sich bzgl. recall und precision eine gegenüber anderen Methoden signifikante Verbesserung. Der Vortrag von Apt, Damerau und Weiss befaßte sich mit einem Textkategorisierungsmodell, das sprachunabhängig arbeitet. Die Experimente hierzu fanden auf der Basis deutscher und englischer Reuters- Texte statt. Rainer Hoch präsentierte das System INFOC

LAS, das im Bereich von Geschäftsbriefen unter Rückgriff auf verschiedene Wissensquellen (z.B. Wortfrequenzstatistik, morphologisches Wissen, typ-spezifische Wortlisten etc.) eine automatische Briefarchivierung vornimmt. In der Indexierungssektion stellte Chris Paice eine Evaluierungsmethode für Wortstammalgorithmen vor, die auch für die Optimierung der Algorithmen von Wert ist. David Eilis, Jonathan Furner-Hines und Peter Willett befaßten sich mit der Fragestellung, wie konsistent manuell vergebene Links in Hypertextbasen ausfallen, wobei besonders die Relation dieser Konsistenz zur Retrievaleffektivität herausgestellt wurde.

Im Vortrag von Ellen Voorhees ging es um die Erweiterung von Anfragen durch lexikalisch-semantisch ähnliche Wörter. Bzgl. Effektivität gemessen auf der Basis der TREC-Kollektion ergab sich, dass die Komplexität der Ausgangsanfrage eine entscheidende Rolle spielte. Die vier Vorträge in der Sektion Benutzermodellierung repräsentierten sehr unterschiedliche Sichtweisen und Ansatzpunkte. Bryce Allen untersuchte experimentell den Einfluß kognitiver Fähigkeiten wie der Wahrnehmungsgeschwindigkeit auf das Lernverhalten und die Suchperformanz von Endbenutzern und die Wirkungsweise dieser Mechanismen. Der Vortrag von Amanda Spink befaßte sich mit möglichen Quellen für Frageerweiterungen im Sinne einer Effektivitätssteigerung und zeigte Möglichkeiten auf, diese innerhalb von Relevance Feedback- Techniken zu integrieren. Brian Logan, Steven Reece und Karen Sparck Jones beschrieben ein theoretisches Modell über Informationsvermittlung basierend auf der KI - Technik der "belief revision". Der Vermittlungsprozeß und das Verhalten von menschlichen Informationsvermittlern wurde in Experimenten untersucht mit der Zielrichtung, Grundstrategien für eine Simulation menschlicher Informationsvermittler herauszuarbeiten. Peter Ingwersen schloß die Sitzung und damit den ersten Konferenztag mit seinem Vortrag über Polyrepräsentation von Benutzerbedürfnis und semantischen Einheiten innerhalb seiner kognitiven Theorie der IR-Interaktion, die durch das Zusammentreffen verschiedener kognitiver Strukturen und Transformationen bestimmt wird, wobei sich der individuelle Benutzer mit bestimmten kognitiven Ausprägungen im Einflußbereich verschiedener Modellelemente befindet.

Die erste Sitzung des 2. Konferenztages mit zwei Vorträgen galt der Theorie und Logik von IR. P.D. Bruza und T.W.C. Huibers stellten ein Verfahren zum Vergleich von Retrievalmechanismen vor, das nicht experimentell, sondern induktiv vorgeht und dabei die Axiome ausfiltert, die den Retrievalprozeß steuern. In der folgenden Präsentation von Fabrizio Sebastiani ging es um eine terminologische Logik zur Modellierung des IR-Prozesses, wobei durch die Einführung eines probabilistischen Elements die Relevanzbeziehung zwischen einem Dokument und einem Benutzerbedürfnis durch Wahrscheinlichkeitsgrade ausdrückbar ist.

Innerhalb der Sektion "Natural Language Processing" war der erste Vortrag von Christian Jacquemin und Jean Royaute der Termidentifikation gewidmet, wobei ein Verfahren beschrieben wurde, das durch partielle Parsing-Techniken aufgrund logischer Regeln und Metaregeln Basisterme und deren Wortformen auffindet. Mark Sanderson präsentierte Ergebnisse aus Studien, die untersuchten, ob korrekte Wortdisambiguierung eine Performanzsteigerung im IR-Prozeß bewirkt. Seiji Miike, Etsuo Itoh, Kenji Ono und Kazuo Sumita stellten ein japanisches Volltext-Retrievalsystem vor, das über eine dynamische Abstract Generierungsfunktion verfügt, wobei der Benutzer die Möglichkeit hat, die Textbereiche selbst festzulegen, die zusammengefaßt werden sollen. Parallel zur NLP-Sektion fanden drei Vorträge aus dem Bereich der statistischen Modelle statt. Der erste von Ijsbrand Jan Aalbersberg befaßte sich mit einem auf der Termfrequenz basierenden Retrievalmodell, wobei der Neuheitswert darin bestand, daß mehrere lokale Ähnlichkeiten zwischen Termrangnummer in der Anfrage und im Dokument in eine globale Ähnlichkeit transformiert werden.

Brian Bartell, Garrison Cottrell und Rik Belew zeigten, daß durch Kombination der Ergebnisse verschiedener Retrievalal-

gorithmen die Retrievalperformanz gesteigert werden kann, was daran liegt, dass unterschiedliche Methoden unterschiedliche Eigenschaften bei der Relevanzbestimmung favorisieren und damit unterschiedliche Schwerpunkte setzen. Weiterhin wurde gezeigt, wie die verschiedenen Methoden kombiniert werden können. Im Vortrag von Joon Ho Lee werden verschiedene, das Boolesche Retrievalmodell erweiternde Verfahren am Beispiel der AND- und 0R- Verknüpfung und der Anfragegewichtung untersucht und durch mathematische Eigenschaften angereichert, um die Retrievaleffektivität zu steigern. Zu Beginn der Evaluierungssektion stellten William Hersh, Chris Buckley und David Hickam eine Reihe von Retrievalexperimenten vor, die zeigten, daß unerfahrene Benutzer aus dem medizinischen Bereich mit einem auf dem Vektor-Raum-Modell basierenden Retrievalsystem gleich gute Ergebnisse erzielten wie erfahrenere Benutzer, die nur über ein Boolesches System verfügten. Aus den Ergebnissen konnte außerdem neues Testmaterial gewonnen werden. Kazem Taghva, Julie Borsack und Allen Condit verglichen in ihrer Studie, wie sich die Retrievaleffektivität verhält, wenn OCR-Texte ohne oder mit manueller Bearbeitung als Dokumentgrundlage verwendet werden. Die Autoren konnten keine signifikante Unterschiede im Hinblick auf recall und precision nachweisen. In dem Vortrag von Howard Turtle konnte unter Hinweis auf die methodischen Probleme gezeigt werden, daß natürlichsprachliche Anfragen im Vergleich zu Booleschen im Bereich juristischer Volltext-Dokumente eine Performanzsteigerung bewirken.

Die bei den regulären Vorträge aus dem Bereich probabilistischer Modelle befaßten sich zum einen mit der Verwendung der logistischen Regression als Dokumentrankingfunktion und deren Evaluierung (Fredric Gey), zum anderen mit Erweiterungen des 2-Poisson Modells durch zusätzliche Variablen und wiederum Modellevaluierung auf der Basis des TREC Testmaterials (S. Robertson und S. Walker). W.S.
Cooper erhielt den Triennial ACM SIGIR Award. Er stellte in seinem Vortrag die kritische Frage, ob der Aufwand, der eingesetzt wurde, um die probabilistische Theorie auf gesunde Beine zu stellen, nicht besser hätte in theoretisch weniger anspruchsvolle Untersuchungen explorativer Art investiert werden sollen. "Time will tell whether the theoretical baggage that accompanies the probabilistic method is more a benefit or an encumbrance", schloß Cooper.

Der 3. und letzte Konferenztag begann mit der Sektion Benutzerschnittstellen. Der erste Vortrag von Matthias Hemmje, Clemens Kunkel und Alexander Willett präsentierte mit LyberWorld ein 3D-System zur Visualisierung verschiedener Elemente des IR-Prozesses (räumliches Navigieren, Aufbau der Suchanfrage, Relevanzbestimmung etc.). Jack Conrad und Mary Hunter Utt berichteten über ein Zugriffs verfahren auf große Textdatenbanken mittels automatisch extrahierter domänenspezifischer Merkmale und deren Ähnlichkeitsbeziehungen, wobei jedoch der Schwerpunkt des Vortrags nicht auf der Gestaltung der Oberfläche lag.

Die Routing-Sektion beinhaltete drei Vorträge. Der Vortrag von Masahiro Morita und Yoichi Shinoda stellte die Ergebnisse verschiedener Experimente und einer Feldstudie dar, in welchen es darum ging, Benutzerinteressen zum Filtern neuer Information zu benutzen. Gegenstand des Vortrags von David Hull war das sog. Latent Semantic Indexing (LSI) als neuer IR-Ansatz, der versucht, die zugrundeliegende Termassoziationsstruktur durch eine Repräsentation der semantischen Faktoren zu modellieren. Experimente zeigen eine extreme Performanzsteigerung, wenn LSI in Verbindung mit statistischer Klassifikation eingesetzt wird. Chris Buckley, Gerard Salton und James Allan befaßten sich mit Relevanz Feedback im Kontext der TREC-Studien. Sie weisen nach, daß ein direkter Zusammenhang besteht zwischen der Effektivität und der Anzahl der aus relevanten Dokumenten hinzugefügten Termen.

In der Sektion Passage Retrieval ging James P. Callan auf eine differenzierte Definition und Rolle der Passagen ein. Experimentelle Ergebnisse wurden vorgestellt, die verschiedene Textsegmente (Paragra-

phen und unterschiedlich große Textfenster) als Passagen definierten sowohl für homogene als auch für heterogene Dokumentkollektionen. Ross Wilkinson ging in seinem Vortrag in die gleiche Richtung und zeigte, daß sich eine Effektivitätssteigerung erzielen läßt, wenn Wissen über die Dokumentstruktur vorhanden ist und verwendet wird. Elke Mittendorf und Peter Schäuble stellten abschließend einen neuen Ansatz vor, der auf der Basis des Hidden Markov Models für Passage Retrieval relevante Textfragmente generiert. Die letzte Sektion der Tagung war Systemimplementierungen gewidmet. Dabei ging es um die Update-Problematik bei IRS (Kurt Shoens, Anthony Tomasic, Hector Garcia-Molina), eine Methode für effizientes Ranking (Michael Persin) und um ein Volltext-IRS aus dem Toubleshooting-Bereich (Peter G. Anick).

Neben den angeführten Vorträgen fanden zwei interessante Panels über die Integration von IR und Datenbanken (Moderation: Norbert Fuhr) und die Evaluierung Interaktiver Retrieval Systeme (Moderation: Susan Dumais) statt. Die Tagungsbeiträge sind im Tagungsband (Croft, W.B., van Rijsbergen, C.J., SIGIR 94. Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Springer-Verlag) enthalten. Unbedingt erwähnt werden muß die hervorragende lokale Organisation von Alan Smeaton und seinen Mitarbeitern, das Conference Dinner im Royal Hospital Kilmainharn, wo die Tagungsteilnehmer bei Harfenklängen und Irish Folk zum Mitsingen und -tanzen animiert wurden, und nicht zuletzt die Fußballweltmeisterschaft, welche die Tagung bis zum Ausscheiden der Irischen Nationalmannschaft intensiv begleitete. Wenn es nach uns gegangen wäre, hätten die Iren den World Cup gewonnen!

Christa Womser-Hacker, Regensburg