

Ulrike Steffens

Information Retrieval: Grundlagen, Systeme und Integration

Der Beitrag befaßt sich zunächst mit der Lösung von Rechercheproblemen mittels Information Retrieval. Die besonderen Anforderungen des Information Retrieval sowie Lösungskonzepte und Modelle werden skizziert. Die Systeme WAIS und INQUERY werden kurz vorgestellt. Schließlich werden die Vorteile einer Integration von Information Retrieval-Systemen in allgemeine Softwareumgebungen und der daraus resultierenden Kombination mit anderen Anwendungsdiensten betrachtet.

Information Retrieval: Fundamentals, Systems and Integration

The first section is concerned with the solution of retrieval problems using information retrieval algorithms. The special considerations that have to be made in information retrieval are presented as well as the adequate concepts and models. Furthermore, two actual systems, WAIS and INQUERY, are introduced. The last section will regard the advantages of the integration of information retrieval systems with general software development environments and of the resulting combination with other application services.

Information Retrieval: fondements, systèmes et intégration

D'abord l'article traite de la solution de problèmes de recherche à l'aide de l'Information Retrieval. Les considérations nécessaires pour l'emploi de l'Information Retrieval sont esquissées ainsi que les concepts de solution et des modèles. Les systèmes WAIS et INQUERY sont présentés brièvement. Enfin, les avantages de l'intégration des systèmes d'Information Retrieval dans les environnements de logiciel et de la combinaison avec d'autres services d'application qui en résulte sont mises en valeur.

1 Einleitung

Durch sinkende Kosten für digitale Speichermedien, weltweite Vernetzung sowie durch rechnergestütztes Arbeiten in Unternehmen und Organisationen steht eine große Zahl an Volltexten in elektronischer Form zur Verfügung. Diese Ablageform und die Größe der zur Verfügung stehenden Datenmasse legen es nahe, den Rechner nicht nur für die Speicherung der Texte, sondern auch für die Ausführung von Recherchealgorithmen zu nutzen. Dabei ist es wünschenswert, neben einfachen Anfragen nach Autor, Titel u.ä. auch eine Suche unter semantischen Aspekten durchführen zu können. Die rechnergestützte Auswertung von über einfache Datenbankabfragen hinausgehenden Anfragen an Textbibliotheken jeder Art wird Information Retrieval¹ genannt. Der vorliegende Text gibt Einblick in anerkannte Forschungsergebnisse auf dem Gebiet des Information Retrieval und verweist auf die entsprechende Literatur. Zwei in der Praxis verwendete Information-Retrieval-Systeme werden skizziert. Schließlich wird ein konkretes Forschungsprojekt im Bereich Information Retrieval vorgestellt.

2 Information Retrieval

Ein Information-Retrieval-Prozeß wird durch eine Person in Gang gesetzt, die ein bestimmtes Ziel verfolgt und hierfür ihr vorhandenes Wissen erweitern möchte (vgl. Abb. 1). Zu diesem Zweck nutzt sie ein System, das auf Textsammlungen zugreifen kann, die zuvor in rechneradäquate Repräsentationen umgewandelt worden sind. Die Anfrage der Person an das System wird nun ebenfalls in eine Rechnerrepräsentation umgeformt. Das System vergleicht Anfrage- und Dokumentenrepräsentatio-

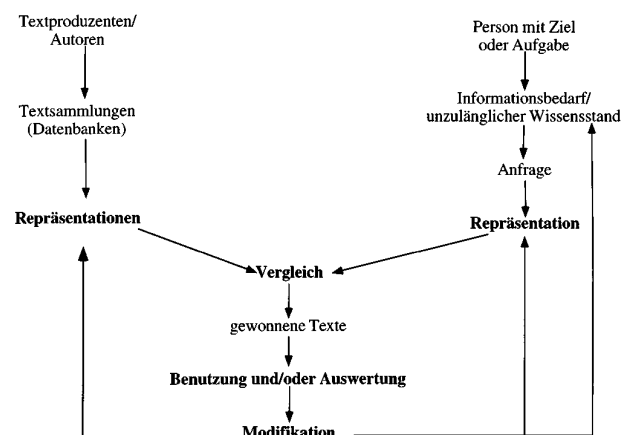


Abb. 1: Ablaufdiagramm des Information Retrieval nach [Belkin/Croft 92]

nen und liefert eine Menge von Texten zurück, die den Informationsbedarf des Anfragenden erfüllen sollen. Der Anfragende kann diese Dokumente verwenden, um die ihm gestellte Aufgabe zu lösen. Er kann zusätzlich auch die Qualität des Retrieval-Ergebnisses bewerten. Die Konsequenz der Bewertung kann zum einen sein, daß der Anfragende seine Anfrage präzisiert, um so ein treffenderes Ergebnis zu erzielen. Zum anderen kann die Bewertung jedoch auch genutzt werden, um die Dokumentenrepräsentationen so zu verändern, daß das Retrieval grundsätzlich zu effektiveren Ergebnissen führt. In einer rechneradäquaten Repräsentation der Textdaten liegt die Hauptanforderung beim Entwurf von Infor-

¹ Salton, G.; McGill M.J.: Introduction to Modern Information Retrieval, 1983.

mation-Retrieval-Systemen. Die natürlichsprachlichen Inhalte sowohl der in den Bibliotheken enthaltenen Dokumente als auch der Anfragen darauf sind formal nicht vollständig auf dem Rechner darstellbar².

So geht bei der Repräsentation von Dokumenten häufig der Kontext verloren, in dem bestimmte Begriffe verwendet werden. Der Begriff „Rolle“ beispielsweise kann in einem theaterwissenschaftlichen Text eine andere Semantik aufweisen als in einem historischen Artikel. Es ist daher wünschenswert, auch inhaltliche Zusammenhänge innerhalb von Dokumenten auf dem Rechner zu repräsentieren. Dabei muß jedoch beachtet werden, daß ergänzend gespeicherte Daten den Retrieval-Prozess erheblich verlangsamen können. Auf der Seite der Anfragen wachsen die Anforderungen an ein Information-Retrieval-System insofern, als daß der Anfragende nicht immer in der Lage ist, das ihn interessierende Themengebiet eindeutig abzugrenzen. Neben einer möglichst genauen Repräsentation der Anfrage ist es erforderlich, daß das System eine Benutzungsschnittstelle zur Verfügung stellt, die die Herangehensweise von Benutzern an ein Rechercheproblem berücksichtigt und somit eine intuitive Benutzung des Systems ermöglicht.

Um die Anforderungen an Information-Retrieval-Systeme möglichst weitgehend erfüllen zu können, wurden verschiedene Konzepte entwickelt, die mit den Eigenheiten der natürlichen Sprache befaßt sind³. So werden bei der Stammformreduktion Wortflexionen durch den entsprechenden Wortstamm repräsentiert. Bei der Stopwortlimitierung werden hochfrequente Wörter, wie z.B. „der“, „doch“ oder „von“, für das Retrieval nicht berücksichtigt. Die Begriffserkennung ermöglicht die Einbeziehung von Konzepten, wie das Erkennen von Firmen- oder Personennamen. Bei der Feldindizierung können die durchsuchten Dokumente in kleinere Abschnitte, wie z.B. Titel und Haupttext, unterteilt werden. Die phonetische Suche ermöglicht das Auffinden von Begriffen, die nicht in der korrekten Schreibweise angefragt wurden. Weitere, hier nicht genannte Konzepte sind vorstellbar.

Viele dieser Konzepte finden sich in Gesamtmodellen wieder, die festlegen, auf welche Weise die Repräsentation von Texten und die Auswertung von Anfragen insgesamt zu erfolgen hat⁴. Ziel solcher Modelle ist es, Information Retrieval so zu beschreiben, daß es in einer Form ausgeübt werden kann, die der menschlichen Auffassung dieses Prozesses zumindest annähernd entspricht.

3 Modelle des Information Retrieval

Die Modelle des Information Retrieval lassen sich in zwei Kategorien unterteilen (vgl. Abb. 2).

Exact-Match-Modelle legen für jedes Dokument einer Textdatenbank fest, ob es für eine Anfrage relevant ist oder nicht. Diese Vorgehensweise spiegelt die Realität jedoch nur lückenhaft wider: Ein Begriff kann in einem Dokument das zentrale Thema sein, während er in einem anderen nur am Rande erwähnt wird. Abstufungen dieser Art sind in Exact-Match-Modellen nicht vorgesehen.

Die Anfrageergebnisse von Best-Match-Modellen dagegen beinhalten für jedes Dokument einen Grad, zu dem

Information-Retrieval-Modelle				
Exact-Match-Modelle	Best-Match-Modelle			
- Dokumente sind relevant oder nicht - Abstufungen sind nicht möglich	- Grad der Relevanz für jedes Dokument			
	<table border="1"> <thead> <tr> <th>Vektorraum-Modelle</th> <th>probabilistische Modelle</th> </tr> </thead> <tbody> <tr> <td> - Repräsentation durch Vektoren - Abstufung über Distanz der Vektoren </td> <td> - stochastische Verfahren - Abstufung über Wahrscheinlichkeiten - z.B. Inferenznetze </td> </tr> </tbody> </table>	Vektorraum-Modelle	probabilistische Modelle	- Repräsentation durch Vektoren - Abstufung über Distanz der Vektoren
Vektorraum-Modelle	probabilistische Modelle			
- Repräsentation durch Vektoren - Abstufung über Distanz der Vektoren	- stochastische Verfahren - Abstufung über Wahrscheinlichkeiten - z.B. Inferenznetze			

Abb. 2: Modelle des Information Retrieval im Überblick

dieses Dokument der Anfrage genügt. Die Differenzierung, die diese Modelle vornehmen, entspricht dem intuitiven Blick auf das Information Retrieval eher als die zweiwertige Logik der Exact-Match-Modelle.

Best-Match-Modelle lassen sich nach der Art der Repräsentation und Auswertung von Texten und Anfragen weiter aufgliedern in Vektorraum- und probabilistische Modelle.

3.1 Vektorraum-Modelle

In Vektorraum-Modellen werden Dokumente und Anfragen als Vektoren im n-dimensionalen Raum repräsentiert. Jede Dimension stellt dabei eine Eigenschaft der Texte dar. Eine einfache Eigenschaft ist das Vorkommen eines bestimmten Begriffes. So ist in Abb. 3 die Eigenschaft „enthält den Begriff ‚Großbritannien‘“ gefordert. Es lassen sich jedoch auch beliebige andere Eigenschaften modellieren. Der Wert eines Vektors für die Dimension k beschreibt den Grad, zu dem der entsprechende Text die Eigenschaft k erfüllt. Für das Beispiel in Abb. 3 könnte beispielsweise die Vorkommenshäufigkeit von ‚Großbritannien‘, gewichtet mit der Textlänge,

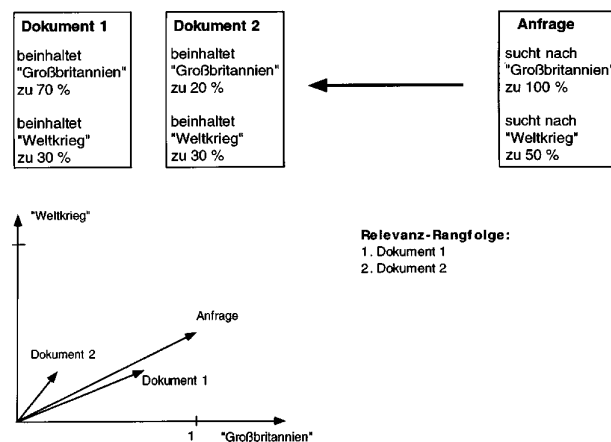


Abb. 3: Information Retrieval im zweidimensionalen Vektorraum

- Fuhr, Norbert: Information Retrieval, Skriptum zur Vorlesung im SS 93. Kapitel 2. Dortmund 1993.
- Broglio, J.; Callan, J.P.; Croft, W.B.: INQUERY System Overview. In: Proceedings of the TIPSTER Text Program (Phase 1), San Francisco 1994, S. 47-67.
- Turtle, H.R.; Croft, W.B.: A Comparison of Text Retrieval Models. In: Computer Journal 35 (3) 1992, S. 279-290.

herangezogen werden. Bei der Auswertung von Anfragen an ein solches System wird die Distanz zwischen dem Anfragevektor und den einzelnen Dokumentenvektoren errechnet. Von dem Dokument, dessen Vektor den geringsten Abstand zum Anfragevektor besitzt, wird angenommen, daß es für die Anfrage am relevantesten ist.

3.2 Probabilistische Modelle

Probabilistische Modelle bilden die semantische Unsicherheit des Retrieval-Prozesses ab, indem sie mit Hilfe von stochastischen Verfahren die Wahrscheinlichkeit ermitteln, daß ein bestimmtes Dokument für eine bestimmte Anfrage Relevanz besitzt⁵.

Als Beispiel soll hier das Modell der Inferenznetze dienen⁶. Inferenznetze stellen eine logische und stochastische Verknüpfung von Annahmen dar. Die Ausgangsannahmen dabei sind, daß eine Menge von Dokumenten der Reihe nach betrachtet wird (vgl. Abb. 4). Durch Verknüpfung läßt sich darstellen, daß bestimmte Terme durch einzelne Dokumente inhaltlich abgedeckt werden. So gibt es in Abb. 4 beispielsweise eine Verknüpfung des Begriffs „Großbritannien“ mit dem Dokument 1. Die Terme wiederum können in der Anfrage des Systembenutzers verknüpft erscheinen. Im Fall von Abb. 4 lautet die Anfrage z.B. „,Großbritannien‘ und ,Weltkrieg‘ ohne ,erster‘“. Durch sukzessive Verknüpfungen über die Terme, die sowohl die Anfrage als auch die Dokumente repräsentieren, entsteht ein Netz, in dem die Wahrscheinlichkeit, daß ein Dokument für eine Anfrage Relevanz besitzt, ermittelt werden kann.

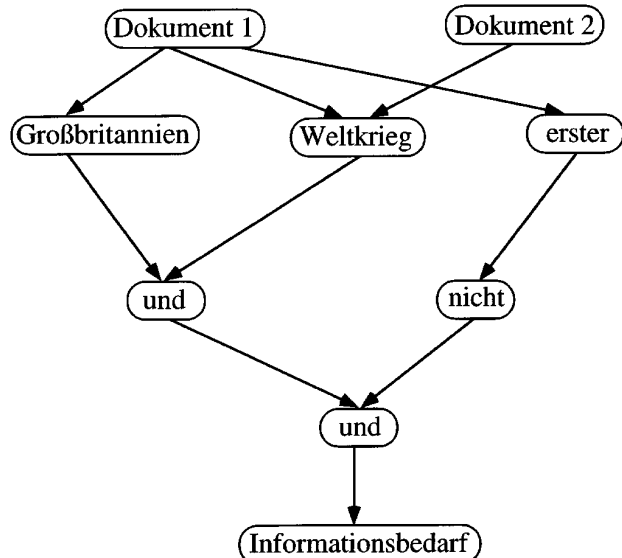


Abb. 4: Inferenznetz für die Anfrage „Großbritannien UND Weltkrieg UND NICHT (erster)“

4 Bewertungskriterien im Information Retrieval

Um die Qualität der verschiedenen Modelle beurteilen zu können, werden für das Information Retrieval die Bewertungskriterien des Rücklaufs und der Genauigkeit definiert⁷:

Der Rücklauf gibt die Fähigkeit eines Systems an, nützliche Dokumente zu finden. Hierzu wird der Quotient aus der Anzahl der gefundenen relevanten Dokumente und

der Anzahl der in der Textsammlung befindlichen relevanten Dokumente gebildet.

Die Genauigkeit stellt dar, wie konsequent „nutzloses“ Material ausgeschlossen wird. Dieses wird durch den Quotienten aus der Anzahl der gefundenen relevanten Dokumente und der Gesamtanzahl der gefundenen Dokumente zum Ausdruck gebracht.

5 Information-Retrieval-Systeme

Verschiedene Information-Retrieval-Systeme kommen in der Praxis zum Einsatz. Nachdem hier zunächst Systemen, die nach Exact-Match-Modellen arbeiten, der Vorzug gegeben wurde, werden sie gegenwärtig von Systemen, die dem Best-Match-Prinzip folgen, abgelöst. In dieser Kategorie sind unter anderem die Systeme WAIS⁸ und INQUERY zu nennen:

Das System WAIS wurde von der Firma Thinking Machines entwickelt. Es hat das Vektorraum-Modell zur Grundlage. Die Client-Server-Architektur des Systems ermöglicht eine verteilte Nutzung und damit auch die Suche auf Texten, die nicht lokal vorhanden sind⁹.

INQUERY wurde im Information Retrieval Laboratory der Universität von Massachusetts entwickelt. Es folgt dem probabilistischen Modell der Inferenznetze¹⁰. Inferenznetze besitzen den Vorteil, daß sie erweiterbar sind, so daß INQUERY mit verschiedenen Textrepräsentations- und Anfragekonzepten arbeiten kann¹¹.

6 Forschungsaspekte im Umfeld von Information Retrieval

Information-Retrieval-Systeme für sich genommen stellen solide Lösungen für Rechercheprobleme dar. Bisher gab es jedoch weitere, pragmatischere Anforderungen im Zusammenhang mit dem Information Retrieval. Diese rücken durch das wachsende Potential von Softwaresystemen in den Bereich des Möglichen¹². Solcherlei Anforderungen und mögliche Lösungsansätze werden am Arbeitsbereich Datenbanken- und Informationssysteme des Fachbereichs Informatik der Universität Hamburg untersucht¹³.

5 Fuhr, Norbert: Probabilistic Models in Information Retrieval. In: Computer Journal 35 (3) 1992, S. 243-255.

6 Turtle, H.R.; Croft, W.B.: Evaluation of an Inference Network-Based Retrieval Model. In: ACM Transactions on Information Systems, 9 (3) 1991, S. 187-222.

7 Frei, H.P.; Meienberg, S.; Schäuble, P.: The Perils of Interpreting Recall and Precision Values. In: Proceedings Information Retrieval GI/GMD-Workshop, 1991.

8 WAIS: Wide Area Information Server.

9 Pfeiffer, U.: HTTPs älterer Bruder: WAIS: Inhaltsorientierte Suche im Internet. In: iX (1) 1995, S. 120.

10 Callan, J.P.; Croft, W.B.; Harding, S.M.: The INQUERY Retrieval System. In: Proceedings of the Third International Conference on Database and Expert Systems, 1991.

11 Belkin, N.J.; Croft, W.B.: Information Filtering and Information Retrieval: Two Sides of the Same Coin? In: Communications of the ACM, 35 (12) 1992, S. 29-38.

12 W.B. Croft: What Do People Want From Information Retrieval? In: d-lib Magazine. Nov. 1995.

13 Steffens, Ulrike: Integration von Information Retrieval Funktionalität in eine offene, persistente Programmierumgebung. Informatik Mitteilung FBI-HH-M-257/96. Fachbereich Informatik, Universität Hamburg, 1996.

6.1 Integration von Information Retrieval und anderer Funktionalität

Information-Retrieval-Systeme bestanden bis jetzt unabhängig von anderen Systemen. Oft kann jedoch die Retrieval-Funktionalität sinnvoll mit den Diensten anderer Systeme ergänzt werden. Die angestrebte Lösung ist die Kombination von Information-Retrieval-Systemen mit verschiedenen Softwarekomponenten unter einer Benutzungsschnittstelle.

Bei vielen Information-Retrieval-Systemen stand bislang die Modellierung im Vordergrund. Ihre Oberflächen sind daher selten benutzerfreundlich gestaltet. Die Integration von Information Retrieval mit Fenstersystemen wäre hier von Vorteil, zumal sich dadurch Chancen ergeben, dem Anfragenden semantische Zusammenhänge graphisch verständlicher darzustellen.

Die Tatsache, daß durch Information Retrieval Texte direkt zur Verfügung stehen, erzeugt das Bedürfnis, diese Texte mit Hilfe desselben Softwaresystems auch verändern zu können. Hierdurch ist implizit eine Kombination von Information Retrieval und Texteditorsystemen gefordert.

Information-Retrieval- und Datenbanksysteme erfüllen, grob betrachtet, eine ähnliche Aufgabe, nämlich die Suche nach Daten. Daher läßt sich fordern, daß mit einem System sowohl nach Daten mit festgelegtem Feldformat, wie zum Beispiel den Daten eines Angestellten in einer Personaldatenbank, und nach Texten, die keiner zuvor festgelegten Form entsprechen, gesucht werden kann¹⁴.

6.2 Persistenz im Information Retrieval

Arbeitsschritte im Information Retrieval sowie ihre Ergebnisse sind häufig über eine Sitzung hinaus für den Benutzer interessant. Daher muß es möglich sein, Anfragen und ihre Ergebnisse dauerhaft (persistent) zu speichern. Hierdurch könnte beispielsweise ein Benutzer regelmäßig dieselben Anfragen an eine sich schnell verändernde Textsammlung stellen. Kataloge mit Standardanfragen an Textbestände wären ebenso denkbar wie eigene „Bibliotheken“ für einzelne Benutzer, in denen sie ihre Anfrageergebnisse ablegen und in denen sie auch wiederum suchen könnten.

6.3 Einbindung mehrerer verteilter Information-Retrieval-Systeme

Durch globale Netzwerke, wie z.B. das Internet, steht jedem Teilnehmer eine Vielzahl von Information-Retrieval-Systemen, die auf verschiedene Textsammlungen zugreifen, zur Verfügung. Es ist jedoch mühsam, für jede Sammlung eine unabhängige Anfrage abzusetzen. Die Lösung ist daher ein System, das nur einmal eine Anfrage entgegennimmt, und von sich aus alle ihm zugänglichen oder vom Benutzer ausgewählten Textdatenbanken befragt.

6.4 Ein Lösungsansatz für erweiterte Information-Retrieval-Anforderungen

Ein Ansatz, um die klassische Retrieval-Funktionalität um die oben genannten Aspekte zu erweitern, ist die

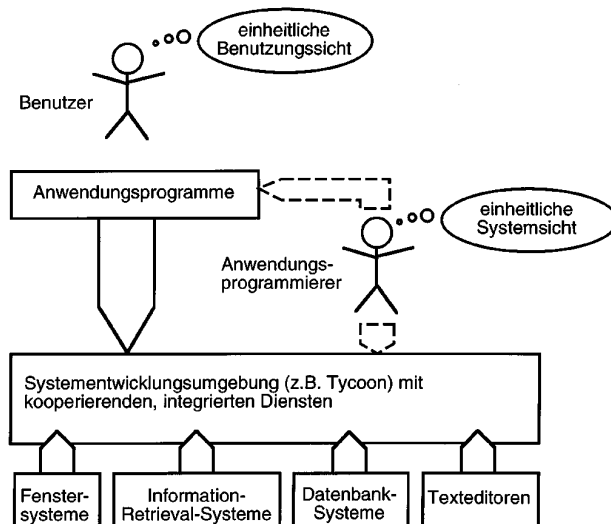


Abb. 5: Integration von Information-Retrieval-Systemen in allgemeine Systementwicklungsumgebungen

Einbettung von Information-Retrieval-Systemen in allgemeine Softwareentwicklungsumgebungen, wie z.B. das an der Universität Hamburg entwickelte Tycoon-System¹⁵. Auf diese Art können Systeme so zusammengefaßt werden, daß sie für Anwendungsprogrammierer sämtlich zugreifbar sind und dem Benutzer als ein Gesamtsystem erscheinen (vgl. Abb. 5). Die in diesem System integrierte Programmiersprache¹⁶ ermöglicht es, verschiedene Softwarekomponenten, wie z.B. Information-Retrieval- und Datenbanksysteme, zu verbinden. Zum einen stellt sie ein System von Datentypen zur Verfügung, das die Fehlerrobustheit bei der Integration verschiedener Dienste erhöht. Zum anderen ist dieses Typsystem anpassungsfähig und kann daher die Verbindung zwischen unterschiedlichen Komponenten herstellen.

Ein weiteres Konzept des Tycoon-Systems ist die persistente Speicherung von Programmen und Daten¹⁷, wodurch die Anforderung aus Abschnitt 6.2 erfüllbar wird. Weiterhin beinhaltet das System Ansätze, um mobile Softwareagenten von Rechner zu Rechner wandern zu lassen¹⁸. Diese sind für verteilte Information-Retrieval-Anfragen geeignet.

7 Zusammenfassung

Die inhaltliche Suche auf Textdaten mit Rechnerunterstützung ist aufgrund der Vielfalt natürlicher Sprache nur

14 Croft, W.B.: Integrating Text and Database Systems: Possibilities and Challenges. In: Proceedings of the Second East/West Database Workshop. Klagenfurt, Austria, 1994. Workshop in Computing, Springer-Verlag, 1995.

15 Tycoon: Typed communicating objects in open environments.

16 Matthes, Florian; Müßig, Sven; Schmidt, Joachim W.: Persistent Polymorphic Programming in Tycoon: An Introduction. Technical Report FIDE/94/106, Glasgow, 1994.

17 Matthes, Florian: Persistente Objektsysteme: Integrierte Datenbankentwicklung und Programmerstellung. Springer-Verlag, 1993.

18 Mathiske, Bernd: Mobilität in persistenten Objektsystemen. Dissertation, Fachbereich Informatik, Universität Hamburg, 1996.

mit Einschränkungen möglich. Es wurden jedoch verschiedene Konzepte und Modelle entwickelt, die eine zufriedenstellende Annäherung realisieren können. Aus dem konsequenten Einsatz von Informationstechnologie können sich aber noch weiterreichendere Vorteile für den Sektor des Information Retrieval ergeben. Die Integration mit anderen Softwarediensten, die Nutzung von Persistenz und die Durchführung von verteilten Retrieval-Prozessen sind Überlegungen in diese Richtung, die sich in Softwareentwicklungsumgebungen, wie sie momentan zur Verfügung stehen, realisieren lassen.

Anschrift der Autorin:

Ulrike Steffens
Arbeitsbereich Softwaresysteme
Technische Universität
Hamburg-Harburg
Harburger Schloßstraße 20
D-21073 Hamburg