

Das COSIMIR-Modell für Information Retrieval mit neuronalen Netzen

Thomas Mandl

Informationswissenschaft
Universität Hildesheim
Marienburger Platz 22 - 31141 Hildesheim
Tel.: ++49-5121-883-837, e-mail: mandl@rz.uni-hildesheim.de

Zusammenfassung

Das in diesem Artikel vorgestellte COSIMIR-Modell (Cognitive Similarity Learning in Information Retrieval) besteht aus einem neuronalen Netzwerk, das auf dem Backpropagation-Algorithmus beruht. COSIMIR lernt, die Ähnlichkeit zwischen Anfrage und Dokument anhand von Benutzerurteilen zu berechnen, und vermeidet so die heuristische Auswahl einer Ähnlichkeitsfunktion. Dadurch versucht es, den Kern des Information Retrieval (IR) Prozesses kognitiv adäquat zu modellieren. Zwar basieren zahlreiche IR Systeme bereits auf neuronalen Netzen, jedoch unterscheiden sie sich kaum von verbreiteten IR Modellen und nutzen nicht alle Stärken neuronaler Netze aus.

Abstract

The COSIMIR-Modell (Cognitive Similarity learning in Information Retrieval) presented in this paper, consists of of neural network based on the backpropagation algorithm. COSIMIR learns to calculate the similarity between query and document using users' judgements. Thus, it avoids the heuristic choice of a similarity function. COSIMIR intends to model the core of the Information Retrieval process in a way cognitively adequate. Numerous IR systems are already based on neural networks, however, they resemble common IR models and do not exploit all possibilities of neural networks.

1 Einleitung

Neuronale Netze zeichnen sich besonders durch ihre tolerante Verarbeitung von Information aus. Im Information Retrieval (IR) werden sie bereits vielfach eingesetzt, auch bei den TREC Konferenzen (Text-Retrieval Conference cf. Harman 1996, Voorhees/Harman 1997/98), einer großen Evaluierungsstudie wurden einige Verfahren erprobt. Die meisten bestehenden Verfahren nutzen aber die subsymbolische Mächtigkeit neuronaler Netze, wie sie etwa der Backpropagation Algorithmus bietet, nicht aus, sondern beschränken sich auf Spreading-Activation-Netze, die formal sehr stark dem Vektorraum Modell ähneln. Sie integrieren das Wissen, das während der Indexierungsphase gewonnen wurde, indem sie die Verbindungen des Netzwerkmodells mit den Termgewichten aus der Dokument-Term-Matrix gewichten.

Das COSIMIR-Modell (Cognitive Similarity learning in Information Retrieval) versucht auf einfache Weise, den zentralen Prozess eines IR-Systems in einem Backpropagation Netz zu implementieren. IR Systeme vergleichen die Repräsentation einer Anfrage mit Repräsentationen von Dokumenten und berechnen jeweils die Ähnlichkeit. COSIMIR bestimmt die Ähnlichkeit mit einem Backpropagation-Netz. Als Lerndaten gehen menschliche Urteile über Ähnlichkeiten zwischen Anfragen und Dokumenten ein. Input-Daten sind die im Indexierungsprozess gewonnenen Gewichte. Damit steht

Das COSIMIR-Modell für Information Retrieval mit neuronalen Netzen

im COSIMIR-Modell mehr Wissen für das System zur Verfügung als in den üblichen IR-Systemen auf der Basis neuronaler Netze.

Das nächste Kapitel fasst kurz die Grundlagen neuronaler Netze zusammen und gibt einen Überblick über den Stand der Forschung im Bereich neuronale Netze im IR. Das folgende Kapitel stellt den Backpropagation Algorithmus vor und führt darauf aufbauend das COSIMIR-Modell ein.

2 Neuronale Netze im IR

Zwar gibt es kaum kommerzielle Information Retrieval Systeme, die neuronale Netze einsetzen, aber die verbreiteten Spreading Activation Netzwerke im IR haben einen hohen Reifegrad erreicht und werden in realistischen Umgebungen getestet. Dieses Kapitel führt kurz in die Grundlagen der neuronalen Netze ein und stellt darauf aufbauend das am häufigsten für IR eingesetzte Modell vor. Dieses sogenannte Spreading Activation Netzwerk ist ein einfaches Modell mit meist zwei Schichten, das z.B. von Ruge 1995

Weitere neuronale Netzwerkmodelle, die im Bereich IR zum Einsatz kommen, sind die selbstorganisierenden Kohonen-Karten für Clustering (Merkl 1995) und Assoziativspeicher für fehlertolerante Suchen (Bentz et al. 1989).

2.1 Grundlagen neuronaler Netze

Künstliche neuronale Netze beruhen auf dem Vorbild des menschlichen Gehirns und zeichnen sich ebenfalls durch massive Parallelität aus. Ein neuronales Netz besteht aus zahlreichen einfachen Verarbeitungseinheiten oder Units, die rein lokal ihren Aktivierungswert berechnen und diesen über gewichtete Verbindungen weiterleiten. Rechnen besteht damit in der Ausbreitung von Aktivierungszuständen. Die Aktivierung einer Unit wird an alle mit ihr verbundenen Neuronen weitergeleitet und diese beziehen sie in ihre Berechnung mit ein. Meist ergibt sich die Aktivierung, die am Ende einer Verbindung ankommt, als Funktion von Ausgangs-Aktivierung und Verbindungsgewicht. Das Gewicht gibt also die Durchlässigkeit der Verbindung an. Die Aktivierungsfunktion der Neuronen kann eine Schwellenwertfunktion sein, d.h. das Neuron „feuert“ erst, sobald eine gewisse Grenze von Aktivierung im Input erreicht ist.

Teilmengen von Neuronen dienen als Input und Output und damit als Schnittstelle zur Außenwelt. Bei Betrachtung und Interpretation von Input und Output ergeben sich in der Regel sinnvolle Funktionen, die bei Betrachtung einzelner Neuronen nicht ersichtlich sind.

Die wichtigste Eigenschaft neuronaler Netze ist die Lernfähigkeit durch Veränderung der Verbindungsgewichte. Es gibt überwachte und unüberwachte Lernverfahren. Die unüberwachten Methoden reagieren nur auf den Fluss von Aktivierung im Netzwerk, während überwachte Verfahren einen extern vorgegebenen, erwünschten Output berücksichtigen. Eine ausführliche Darstellung neuronaler Netze bietet Zell 1994.

2.2 IR als Spreading Activation

Bereits seit längerem werden neuronale Netze in IR-Systeme integriert. Bereits kurz nach der Renaissance der Neuroinformatik Mitte der 80er Jahre hatten sich die Modelle von Belew (1989) Kwok (1989) in der IR-community etabliert. Einen guten Überblick über diese Entwicklung bieten Doszkocs et al. (1990). Die Modelle wurden weiterentwickelt und einigen Gruppen gelang eine erfolgreiche Teilnahme an TREC und damit der Schritt von experimentellen Systemen zu realen Massendaten aus dem Bereich Zeitungstexte (Kwok/Grünfeld 1996, Boughanem/Soulé-Dupuy 1998).

Als Grundmodell steht hinter aller erwähnten Ansätzen das sogenannte Spreading Activation Netzwerk. Dabei handelt es sich um bidirektionale, symmetrische Netzwerke mit Aufteilung in Schichten. Typischerweise tauschen zwei Schichten untereinander Aktivierung aus, wobei eine die Dokumente und die zweite Terme repräsentiert. Die Gewichte der Verbindungen werden mit der Dokument-Term-Matrix aus der Indexierung initialisiert. Das bedeutet, das Gewicht w_{td} zwischen Dokument-Neuron d und Term-Neuron t erhält den von einem Indexierungs- und Gewichtungsalgorithmus bestimmten Wert von Term t für Dokument d . Zahlreiche Verbindungen erhalten das Gewicht 0, da die entsprechenden Terme in den jeweiligen Dokumenten nicht vorkommen.

Beim Retrieval formuliert der Benutzer des IR-Systems wie gewohnt seine Anfrage. Das System aktiviert dann die gewählten Terme und die Aktivierung breitet sich im Netz aus. Zunächst werden die Dokument-Neuronen aktiviert, mit denen die Anfrage-Terme indexiert sind. Alle aktivierten Dokumente senden im zweiten Schritt Aktivierung an alle Terme, mit denen sie verknüpft sind. Bereits nach diesem Schritt sind in der Regel Terme aktiviert, die in der ursprünglichen Anfrage nicht enthalten sind, womit die Term-Expansion als inhärente Eigenschaft dieses Modells auftritt. Nach einer bestimmten Anzahl von Schritten oder nachdem ein bestimmter Aktivierungswert erreicht ist, endet die Aktivierungs-Ausbreitung und die am stärksten aktivierten Dokumente werden dem Benutzer als Ergebnis präsentiert.

Abbildung 1 zeigt ein beispielhaftes Netzwerk während eines Retrievalprozesses. Dabei werden Aktivierung und Verbindungsstärke durch die Dicke der Linien angedeutet.

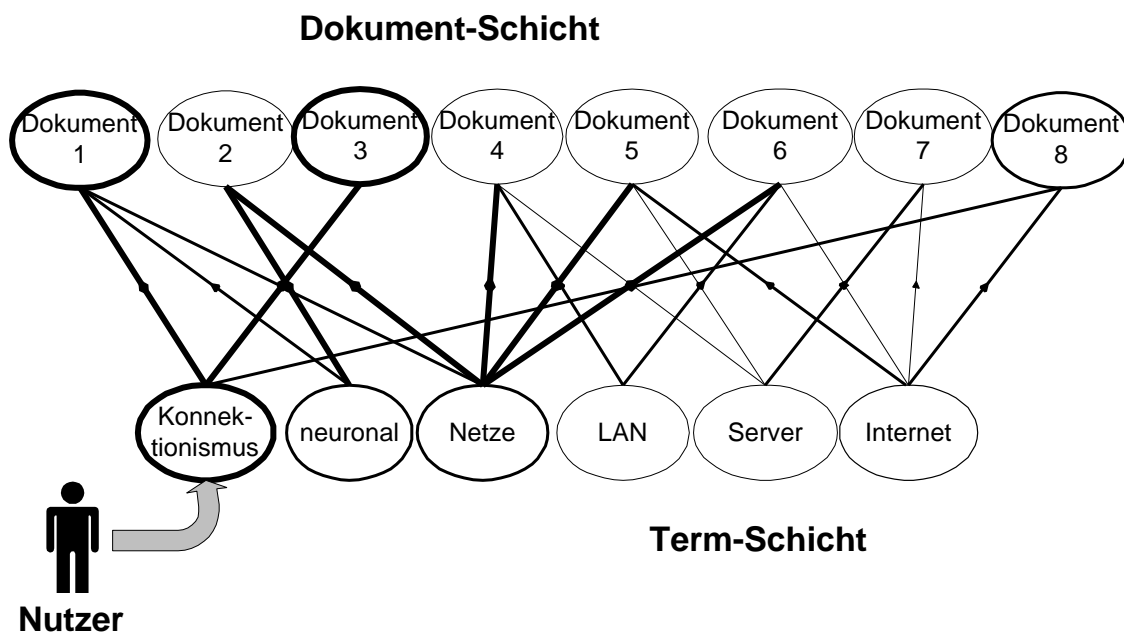


Abb. 1: Information Retrieval als Aktivierungsausbreitung

Der Benutzer hatte den Term *TCP/IP* eingegeben und inzwischen sind mehrere Dokumente und die mehrfach mit *TCP/IP* in einem Dokument vorkommenden Terme *network*, *server* und *client* aktiviert worden. Interessant ist, dass auch ein Dokument-Neuron Anfrage sein kann und die Funktionalität des Netzes identisch bleibt.

Dieses Grundmodell wurde mehrfach modifiziert. Eine naheliegende Erweiterung ist die Einbeziehung von relevance feedback. Nach einer gewissen Anzahl von Aktivierungsschritten evaluiert der Benutzer die Ergebnisdokumente. Je nach Relevanz kann ihre Aktivierung erhöht oder verringert werden. Beim weiteren Verlauf der Aktivierungsausbreitung beeinflussen diese Vorgaben des Benutzers das Ergebnis.

Bereits die frühen Modelle nutzten die Stärke neuronaler Netze und bezogen Lernen mit in ihre Systeme ein. Sowohl bei Kwok 1989 als auch bei Belew 1989 ist relevance feedback nicht nur wie oben beschrieben ein Ausgangspunkt für neue Aktivierungsausbreitung, sondern führt zu Veränderungen der Verbindungsstärken. Der vom Benutzer gesetzte Relevanzwert wird mit dem vom System errechneten Wert verglichen und die beteiligten Verbindungen werden so modifiziert, dass das beurteilte Neuron eher den gewünschten Wert erreicht.

Belew 1989 führt eine weitere Schicht ein, die Autoren repräsentiert und mit den Dokumenten verbunden ist. In seinem Modell sind auch Verbindungen innerhalb von Schichten möglich, also etwa assoziative Beziehungen zwischen Termen. Auch Boughanem/Soulé-Dupuy 1998 setzen in ihrem in TREC getesteten System MercureO2 assoziative Verbindungen ein, sehen aber nur drei Aktivierungsschritte vor. Damit läuft die Aktivierung einmal von den Termen zu den Dokumenten, einmal zurück und noch einmal zu den Dokumenten.

Das zweite in TREC eingesetzte System PIRCS (Kwok/Grunfeld 1996) basiert auf Kwok 1989 und sieht dementsprechend Lernen vor. Eine dritte Schicht repräsentiert Anfragen und steht mit den Termen in Verbindung.

Zahlreiche Autoren betonen die Ähnlichkeit zwischen den Spreading-Activation IR-Modellen und dem Vektorraum-Modell. Die Initialisierung der Gewichte mit den Werten aus der Dokument-Term-Matrix deutet dies bereits an. Mothe 1994 weist theoretisch und empirisch nach, dass ein Spreading Activation Modell nach einem Aktivierungsschritt äquivalent zum Vektorraum-Modell ist. Zwar bieten die Spreading Activation Netze mit der Aktivierungsausbreitung eine plausible Metapher für den IR-Prozess, jedoch stellen sie kein prinzipiell neues IR-Paradigma dar. Zudem schöpfen sie die Möglichkeiten neuronaler Netze im Bereich Lernfähigkeit nicht aus.

3 Das COSIMIR-Modell

Das COSIMIR-Modell (COgnitive SIMilarity Learning in Information Retrieval) versucht, die Schwächen der Spreading-Activation Modelle zu überwinden, indem in einem Backpropagation-Netzwerk zahlreiche Benutzerurteile zum Lernen genutzt werden.

3.1 Backpropagation Netzwerke

Das Backpropagation-Netz und der mächtige, dazugehörige Lernalgorithmus ist das wohl am meisten eingesetzte neuronale Netz. Beim Backpropagation-Netz sind die künstlichen Neuronen in Schichten angeordnet, wobei die Aktivierung immer nur in eine Richtung fließt. Jede Unit ist mit allen Units der nächsten Schicht verknüpft. Zwischen Input- und Output-Schicht befindet sich eine oder mehrere versteckte Schichten, die keine symbolische Entsprechung besitzen. Mit ihnen erhöht das Netz seine formalen Rechenfähigkeiten und kann laut Smolensky 1988 einen „intuitive processor“ implementieren, der intuitives Expertenverhalten besser modelliert als regel folgende Systeme.

Wie in den oben vorgestellten neuronalen Netzen senden die Units Signale an die Units der nächsten Schicht, die aus den ankommenden Impulsen ihre Aktivierung errechnen. Auf diese Weise wird aus einem Vektor, der an der Eingangsschicht anliegt, über Zwischenschichten hinweg die Aktivierung in einer Ausgangsschicht berechnet. Bei einem Lernvorgang wird die Aktivierung am Ausgangs-Layer mit einem Zielvektor, der mit dem Eingangsvektor assoziiert werden soll, verglichen. Die Differenz zwischen der tatsächlichen und der gewünschten Aktivierung wird als Fehler im Netz zurückgegeben. Dabei werden die Verbindungsstärken so verändert, dass beim nächsten Durchlauf der Fehler geringer ist. Nach sehr vielen Lernschritten mit vielen verschiedenen Input-Output-Paaren generalisiert der Backpropagation-Algorithmus häufig und nähert die gewünschte Funktion so gut an, dass das Netz nun auch auf neue Muster mit einer sinnvollen Ausgabe reagiert.

Ein großer Nachteil neuronaler Netze ist die fehlende Erklärungsfähigkeit. Die massiv parallelen Vorgänge lassen sich in der Regel nachträglich nicht interpretieren.

3.2 Die Grundlagen von COSIMIR

Das COSIMIR-Modell implementiert den zentralen Prozess im IR, den Abgleich zwischen Anfrage- und Dokument-Repräsentation in einem einfachen Backpropagation-Netzwerk. Dadurch nutzt es subsymbolische Repräsentationsmechanismen aus und kann formal mehr Klassen von Funktionen implementieren als ein Spreading-Activation-Netzwerk.

Das COSIMIR-Modell benutzt als Input eine Query und ein Dokument, die beide an der Eingangsschicht angelegt werden. Über eine versteckte Schicht wird die Aktivierung bis zur Ausgangsschicht propagiert, die nur aus einem Neuron besteht und die Relevanz bzw. Ähnlichkeit repräsentiert. Im Training wird die Relevanz von verschiedenen Kombinationen von Dokumenten und Anfragen gelernt. Dazu müssen Relevanzurteile von Benutzern gesammelt werden. So kann das COSIMIR-Modell eine kognitive Ähnlichkeitsfunktion implementieren.

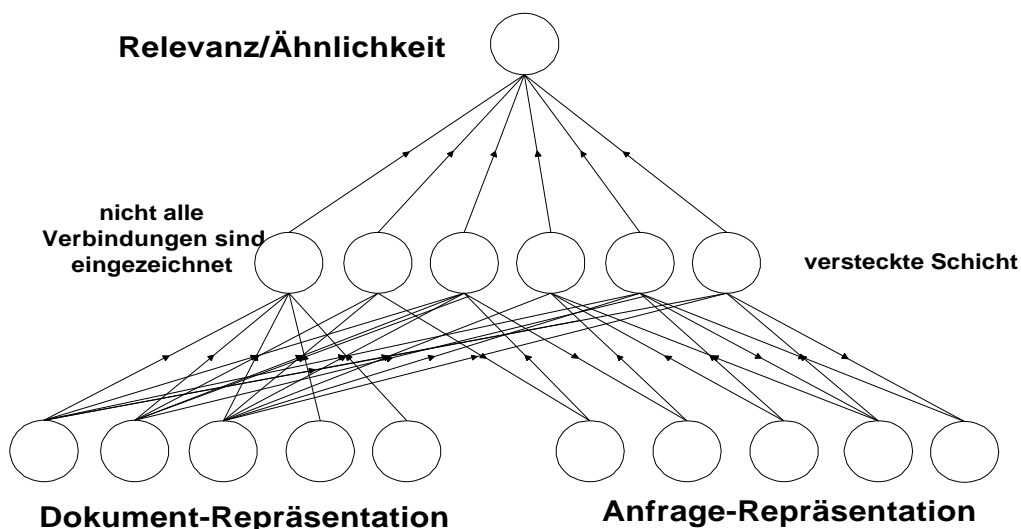


Abb. 2: Das COSIMIR-Modell

COSIMIR-Netze werden sehr groß, da für jeden Term zwei Input-Neuronen benötigt werden. Entsprechend benötigt COSIMIR relativ viele Trainingsdaten. Die Verbindungen, die von der Input-Schicht ausgehen, werden nur hinreichend trainiert, wenn z.B. auch jeder Term mindestens einmal in einem Trainingsbeispiel vorkommt. Allerdings müssen auch Trainingsbeispiele für Nicht-Relevanz eingebaut werden, da das System ansonsten lernt, immer eine hohe Relevanz zurückzuliefern. Dadurch steigt die Zahl der potentiell verwendbaren Daten erheblich, da alle nicht als relevant eingestuftene Dokumente als negative Beispiele benutzt werden können.

Eine ausführliche Darstellung von COSIMIR findet sich in Mandl 1998a.

3.2 Die Vorteile von COSIMIR

COSIMIR verfügt über einige Vorteile, die es von anderen IR-Systemen abgrenzen:

- Während die Boolesche Systeme eine äußerst einfache Ähnlichkeitsfunktion benutzen, arbeiten die statistischen Verfahren mit mathematischen Ähnlichkeitsfunktionen wie dem Kosinus. Diese einfachen Formeln können die Komplexität der menschlichen Ähnlichkeitsbeurteilung nur unzureichend abbilden. Wie etwa Tversky 1977 zeigt, sind menschliche Ähnlichkeitsurteile nicht immer

symmetrisch oder transitiv. Mathematische Modellierungen von Ähnlichkeit verfügen jedoch fast immer über diese Eigenschaften. Bei COSIMIR können sich je nach Trainingsdaten durchaus nicht symmetrische oder transitive Urteile ergeben.

- In vielen IR-Systemen fällt die Entscheidung für eine bestimmte Ähnlichkeitsfunktion auf rein heuristischer Basis, ohne dass konkrete Eigenschaften der Funktion auf Anforderungen des Systems bezogen werden. In COSIMIR entfällt diese Heuristik, da die Funktion vom Netz implementiert wird.
- Viele IR-Modelle erfordern aus formalen Gründen die paarweise Unabhängigkeit von Termen, eine Forderung, die in den allermeisten Fällen von den Daten offensichtlich nicht erfüllt wird. COSIMIR kann auf diese Annahme verzichten und modelliert im Idealfall die komplexen Zusammenhänge und Abhängigkeiten im Backpropagation-Netzwerk.
- COSIMIR erhält mehr Wissen als die Spreading-Activation-Netzwerke. Zu den Daten aus dem Indexierungsprozess kommen die Benutzerurteile hinzu und bilden den Kern des Modells. In Spreading-Activation-Netzwerken können solche Daten zwar auch eingesetzt werden, sie modifizieren jedoch nur nachträglich die Parameter (Kwok 1989). Damit überschreiben sie die Indexierungsdaten, während COSIMIR strikt zwischen den Ausgangsdaten und der Ähnlichkeitsfunktion trennen kann.
- COSIMIR benötigt keine Annahmen über die Gleichförmigkeit von Dokument- und Anfragevektor. Beide können auch in verschiedenen Repräsentationsschemata oder -sprachen vorliegen, solange nur genügend Trainingsdaten vorliegen. Damit eignet sich COSIMIR auch für Retrieval in heterogenen Umgebungen (Mandl 1998b).

4 Fazit und Ausblick

Inwieweit sich COSIMIR auch in der Praxis bewährt, können nur experimentelle Retrievaltests zeigen. Die größte Schwäche von COSIMIR wurde schon angesprochen, die Größe der entstehenden Netzwerke und die daraus resultierende Menge von Trainingsdaten. Für erste Experimente mit Text-Korpora soll daher eine Komprimierung der Daten vorgeschaltet werden. Dafür bietet sich Latent Semantic Indexing (LSI) an, das den Termraum auf ca. 100 bis 300 LSI-Dimensionen einschränkt (Deerwester et al. 1990).

Die Implementierung des Modells ist relativ einfach, da dazu auf Standard-Software zur Programmierung neuronaler Netze zurückgegriffen wird. Die Vor- und Nachbearbeitung der Daten ist allerdings aufwendig.

Momentan laufen Experimente mit der Cranfield-Kollektion, da diese für die Anzahl der Dokumente relativ viele Anfragen enthält.

Literatur

- Belew, R. (1989): Adaptive Information Retrieval: Using a Connectionist Representation to Retrieve and Learn about Documents. In: Belkin/Rijsbergen 1989. S. 11-20.
- Belkin, N. J.; Rijsbergen, C.J. van (Hrsg.) (1989): SIGIR89. 12th International Conference on Research and Development in Information Retrieval. (SIGIR89) Cambridge, MA, USA. June 25-28. New York.
- Bentz, Hans; Hagström, Michael; Palm, Guenther (1998): Information Storage and Effective Data Retrieval in Sparse Matrices. In: Neural Networks 2(4). S. 289-293.

- Boughanem, M.; Soulé-Dupuy, C. (1998): Mercure at trec6. In: Voorhees/Harman 1998.
- Deerwester, S.; Dumais, S. T.; Harshman, R. (1990): Indexing by Latent Semantic Analysis. In: Journal of The American Society For Information Science 1990. vol. 41 (6). S. 391-407.
- Doszkocs, T.E.; Reggia, J.; Lin, X. (1990): Connectionist Models and Information Retrieval. In: Annual Review of Information Science and Technology (ARIST), vol. 25. S. 209-260.
- Harman, Donna (Hrsg.) (1996): The Fourth Text Retrieval Conference (TREC-4). NIST Special Publication 500-236. National Institute of Standards and Technology. Gaithersburg, Maryland, 1.-3.11.1995. http://trec.nist.gov/pubs/trec4/t4_proceedings.html
- Kwok, K. L. (1989): A Neural Network for Probabilistic Information Retrieval. In: Belkin/Rijsbergen 1989. S. 21-30.
- Kwok, K.L.; Grunfeld, L. (1996): TREC-4 Ad-Hoc, Routing Retrieval and Filtering Experiments using PIRCS. In: Harman 1996.
- Mandl, T. (1998a): Das COSIMIR Modell: Information Retrieval mit dem Backpropagation Algorithmus. ELVIRA-Arbeitsbericht 10, IZ Sozialwissenschaften, Bonn.
- Mandl, T. (1998b): Vague Transformations in Information Retrieval. In: Zimmermann, H.; Schramm, V. (Hrsg.): Knowledge Management und Kommunikationssysteme: Workflow Management, Multimedia, Knowledge Transfer. Proc. 6. Int. Symposium für Informationswissenschaft. (ISI '98). 3.-7.11.98, Karlsuniversität Prag, UVK: Konstanz.. S. 312-325.
- Merkl, Dieter (1995): Content-Based Document Classification with Highly Compressed Input Data. In: Proceedings of the International Conference on Artificial Neural Networks ICANN 95. Paris. October 9-13 1995. vol. 2. S. 239-244.
- Mothe, J. (1994): Search Mechanisms Using a Neural Network Model. In: Intelligent Multimedia Information Retrieval Systems and Management. Proceedings of the RIAO 94 (Recherche d'Information assistée par Ordinateur). Rockefeller University. New York. S. 275-294.
- Ruge, Gerda (1995): Gedächtnis und Termassoziation. In: Kuhlen, Rainer; Rittberger, Marc (Hrsg.) (1995): HIM'95. Hypertext, Information Retrieval, Multimedia. Synergieeffekte elektronischer Informationssysteme. Konstanz. 5.-7. April 95. Konstanz. S. 243-257.
- Smolensky, P. (1988): On the Proper Treatment of Connectionism. In: Behavioral and Brain Sciences vol. 11. S. 1-74.
- Tversky, A. (1977): Features of Similarity. In: Psychological Review vol. 84 (4). S. 327
- Voorhees, Ellen; Harman, Donna (Hrsg.) (1997): The Fifth Text Retrieval Conference (TREC-5). NIST Special Publication 500-238. National Institute of Standards and Technology. Gaithersburg, Maryland, 20.-22.11.1996. http://trec.nist.gov/pubs/trec5/t5_proceedings.html
- Voorhees, Ellen; Harman, Donna (Hrsg.) (1998): The Sixth Text Retrieval Conference (TREC-6). NIST Special Publication 500-240. National Institute of Standards and Technology. Gaithersburg, Maryland, 19.-21.11.1996. http://trec.nist.gov/pubs/trec6/t6_proceedings.html
- Wilkinson, R.; Hingston, P. (1992): Incorporating the Vector Space Model in a Neural Network used for Document Retrieval. In: Library HiTech News vol. 10 (1-2). S. 69-75.
- Zell, A. (1994): Simulation Neuronaler Netze. Bonn et al.