Internet-Suchwerkzeuge

Klassifikation und terminologische Kontrolle: Yahoo!, Open Directory und Oingo im Vergleich

In Password 11/2000 wurden durch einen Retrievaltest die qualitativ führenden Suchwerkzeuge im Internet bestimmt. In den nächsten Teilen unseres State of the Art - Berichts über Retrievalsysteme im World Wide Web beschreiben wir einzelne interessante Ansätze der Technik der Top-Suchwerkzeuge. Den Anfang machen die klassifikatorischen Verzeichnisse Yahoo! und das Open Directory-Projekt sowie das System Oingo, das im Rahmen eines "semantischen Retrievals" das Homonym- und Synonymproblem angeht.

Was hat Yahoo mit Gullivers Reisen zu tun?

Yahoo! ist eines der populärsten Suchwerkzeuge im In- und Ausland. Was zeichnet diese Beliebtheit aus, oder anders gefragt, was bietet Yahoo! besser als die Konkurrenz? Werfen wir einen Rückblick auf 1993 und die folgenden Jahre, auf jene Zeit, wo Yahoo! kreiert wird, für nur eine kurze Weile in den Kinderschuhen steckt, um sich dann letztendlich zu dem derzeitigen Suchsystem zu entpuppen. Jerry Yang und David Filo, zwei Studenten aus Palo Alto und Benutzer des WWW, erstellen anfangs für den Eigengebrauch Adressenlisten jener Sites, die sie besuchten und ihnen als interessant erschienen, um zu einem späteren Zeitpunkt eventuell noch einmal zu diesen Sites schnell zurückkehren zu können. Mit dem Zuwachs der Linksammlung vergrößert sich gleichzeitig das Problem der effizienten Suche nach diesen gesammelten Informationen. Eine Datenbank, in der die Verweise zu den WWW-Sites geordnet und klassifiziert werden, entsteht und wird zusammen mit ihrer Suchmaschine ins Web gestellt. Der erste Schritt vom privaten zum öffentlichen Nutzen ist damit getan. Viele andere Interessierte schliessen sich an und nehmen fortan die Möglichkeit wahr, bei der Ergänzung, Korrektur oder Streichung von Sites in Yahoo! mitzuwirken. Im rasanten Tempo werden neue Seiten der Yahoo!-Datenbank hinzugefügt. Innerhalb von zwei Jahren ist Yahoo! längst nicht mehr das Werkzeug einer Handvoll von Studenten. Unser Suchwerkzeug bekommt die URL der Yahoo Corporation, und ein Redaktionsstab wird mit der Katalogisierung und dem Updating eingereichter Eingänge beauftragt.

Yahoo! - Was bedeutet dies? Mag man bei dem Namen "Yahoo" vielleicht zunächst an die viehischen Wesen in Menschengestalt aus "Gullivers Reisen" von Jonathan Swift denken, steckt hingegen hinter dem "Yahoo" unseres Suchwerkzeuges auch das Akronym: **Y**et **A**nother **H**ierarchical **O**fficious **O**racle. Ob Filo und Yang sich als Yahoos (im Sinne Swifts) fühlen, sei dahingestellt.

Vom Erfolg überrannt?

Akzeptiert werden nahezu alle Typen von Sites, persönliche oder kommerzielle, nur persönliche Sites pornographischen Inhalts werden von Anfang an ausgeschlossen. 1999, nur sechs Jahre nach der "Gründungszeit" steht der Redaktionsstab vor dem Chaos einer nicht schaffbaren Bewältigung der Eingangsmassen. Danny Sullivan schreibt: "Submitting to Yahoo has long been one of the more frustrating experiences for many webmasters. Submissions go in, but no action may be taken for weeks or months, if at all" (Sullivan 1999). Um diesem Phänomen entgegenzuwirken, führt Yahoo! den Business Express Service ein, der dem Antragsteller für US-$199 zumindest ein Ja oder Nein garantiert, ob die Site innerhalb der nächsten sieben Arbeitstage aufgelistet wird. "Business Express allows webmasters to pay for guaranteed consideration of their sites, not for guaranteed listings" (ebd.). Bezahlung garantiert nicht, dass, wie (mit welcher Beschreibung) und wo (in welcher Kategorie) eine Site erscheint. In dieser Beziehung unterscheidet sich der Business Express in keiner Weise vom normalen Submissionprozess. Yahoo! - Redakteure haben keine Zeit, alle Eingänge zu bewältigen. Webmaster bekommen durch den neuen Service die Wahl, ob sie eine sofortige Antwort erhalten möchten oder nicht.

Yahoo! hält einen Datenspeicher mit rund einer Million indexierter Dokumente. Wenn man bedenkt, dass allein Google über rund 1,3 Milliarden Einträgen verfügt, so wird bei Yahoo! nur ein Bruchteil des Web (innerhalb des eigenen Verzeichnisses) vorgehalten. Eine ausschließliche Nutzung von Yahoo! bedeutet, auf über 99,9% der Inhalte des World Wide Web zu verzichten. Dies hat auch das Unternehmen erkannt und bietet zusätzlich ein Retrieval in einer Suchmaschine an. Partner war bis Mitte 2000 Inktomi, jetzt ist es Google.

Die Wiederentdeckung der Klassifikation

Charakteristikum von Yahoo! ist seine hierarchische Struktur. Im Eingangsbildschirm finden wir 14 Hauptklassen bzw. Hauptkategorien, von denen aus sich der Nutzer zu Subkategorien

mehrerer Stufen bis hin zu den Sites und - neuerdings - Nachrichten herunterklicken kann. Wir haben es mit einem polyhierarchischen Klassifikationssystem zu tun, d.h. eine Klasse kann durchaus mehr als einen Oberbegriff haben.

Wie funktioniert eine Suche bei Yahoo!? Das Yahoo!-Patent "Information Retrieval From Hierarchical Compound Documents" zeigt uns die eingesetzte Technik. "The Game of Go" lautet die Suchfrage. Der Server, der die Suchfrage mit einem Ergebnis beantwortet, ist an einen Dokumentenspeicher und einen Wortindex gekoppelt. Dieser Server unterzieht den Suchausdruck (durch Parsing) einer Sprachanalyse und versucht, Dokumente im Dokumentenspeicher passend zum gegebenen Suchausdruck abzugleichen. Yahoo!-Dokumente sind im Speicher logisch geordnet. Jedes Dokument repräsentiert entweder eine thematische Kategorie oder eine Site und steht als Nachweis im hierarchischen Zusammenhang mit anderen Nachweisen. Jeder Nachweis besitzt sowohl eine Identifikationsnummer als auch einen Inhalt. Auswahl und Benennung der Kategorien und Sites sowie Festlegung der hierarchischen Struktur setzt der Redaktionsstab fest. Abbildung 1 zeigt die logische Struktur einer (fingierten) hierarchisch aufgebauten Informationsmenge. Die linke Seite zeigt den Pfad über Recreation ("Dokument" mit der Yahoo!-Dokumentnummer 1), Games (2), Board Games (3) und Go (4) zu zwei URLs (5 und 6), die Go-Sites vertreten. Wenn wir den rechten Pfad anschauen, sehen wir ebenfalls die Zeichenfolge "Go", jedoch in einem völlig anderem Kontext als Unterklassen (20, 21 und 22) zu Restaurants (16). Unsere Beispielsuchfrage zielt natürlich nur auf die Treffer 4 bis 6 (obwohl hier "Games" als Zeichenfolge nicht vorkommt); die Nummern 20 bis 22 müssen ausgeschlossen werden. Eine UND-Verknüpfung der Zeichenfolge würde keine Treffer ergeben; folglich muss der Algorithmus cleverer vorgehen.

Die Elemente und Funktionen des Suchsystems stellt Abbildung 2 dar. Die alphabetisch geordneten Nachweise im Wortindex besitzen als Identifikationsmerkmal die Dokumentnummer, in denen das entsprechende Wort vorkommt. Ignoriert werden als Stoppwörter Allgemeinwörter wie "the" und "of". Mithilfe des Wortindex wird die Suche nach Dokumenten im Speicher beschleunigt. Die direkten Abgleichsnachweise kommen vom Wortindex und werden in einer vorläufigen Trefferliste zusammengetragen. Im Dokumentenspeicher haben die Datensätze die Felder:

- Dokumentnummer (in Abb. 2 fett gedruckt, z.B. 1)
- Nummer des letzten Unterbegriffs des jeweiligen Dokuments als "Unterbegriffszeiger" (in unserem Beispiel ist dies: 9, also Boating für Recreation)
- Nummer des Oberbegriffs als "Oberbegriffszeiger" (gibt es hier nicht)
- Text des Dokuments, repräsentiert durch den Nachweis (Recreation)
- Beschreibung
- Assoziationsbegriff (Fun)
- Hinweis auf Kategorie oder Site.

Wohlgemerkt: Dokumentennachweise bedeuten bei Yahoo! Referenzen für Kategorien oder auch Referenzen für Sites. Anhand des Zusammenspiels der Dokumentnummer, des Unter- und des Oberbegriffszeigers werden die Knoten und Verästelungen im Hierarchiebaum beschrieben. Für die Prüfung der Suchgeschwindigkeit, das Update, die Rangfolge der Unterbegriffe ("Intervall der Kinder") und die Gewichtung benutzt die Suchmaschine die Funktionen einer Dokumentenprofilordnung (in Abb.2 rechts unten). Für den Dokumentnachweis 1 gibt es demnach die Unterbe-

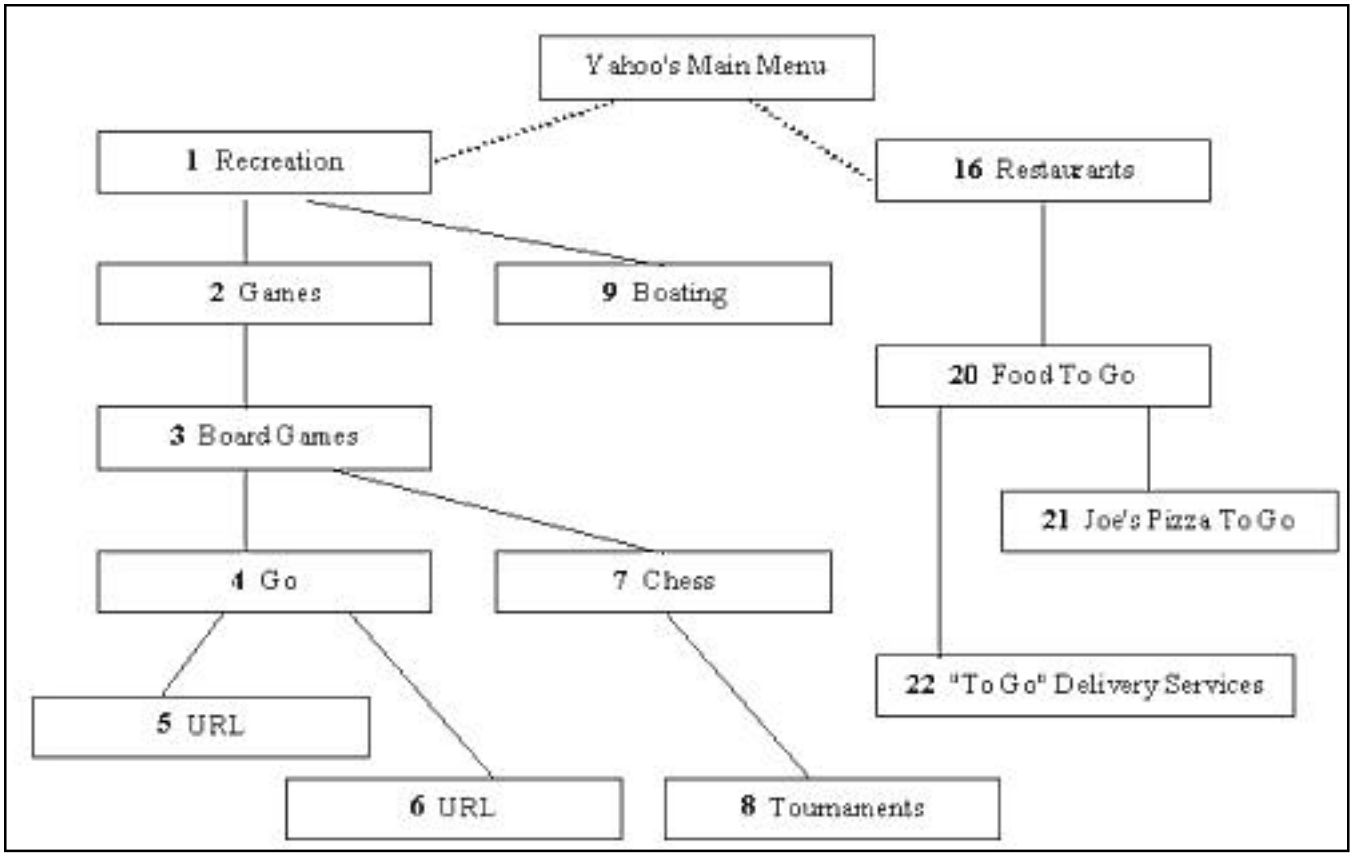

Abbildung 1: Kategorien (Klassen) und Dokumente in der Baumstruktur von Yahoo!
Quelle: Yahoo!; Patent Nr. US 5991756

griffskette ("Kinder") 2 bis 9. Die URL werden vom Redakteur gewichtet. So ist offenbar die URL 5 doppelt so hoch eingeschätzt wie URL 6.

Für Yahoo! ist der Unterschied zwischen einem "direkten" und einem "indirekten" Abgleich von zentraler Bedeutung. Beim direkten Abgleich trifft der Suchausdruck den Inhalt des Dokumentennachweises (direkte Treffer bei Go wären 4, 20, 21, 22). Beim Abgleich wird stets rechts trunkiert; so trifft Yahoo! bei der Eingabe von "Game" auch auf "Games". (Die automatische Trunkierung ist nur zu umgehen, wenn man den Suchbegriff - wie eine Phrase - in Anführungszeichen eingibt.) Um herauszufinden, ob der zu suchende Term z.B. nur in einem der Unterbegriffe des Dokumentennachweises versteckt ist, prüft und verbindet die Suchmaschine durch indirekte Abgleiche nach und nach alle Dokumentenketten. In den Abgleichslisten zu einem Treffer erscheinen die Dokumentnummer des jeweiligen direkten Abgleichs, das jeweilige Intervall des indirekten Abgleichs sowie "Null" als Ende der Liste. Diese Aufsplittung der Listen ist wesentlich, wenn mehrere Suchausdrücke (z.B. zu einer UND-Verknüpfung) miteinander verbunden werden sollen. Boolesche Operatoren werden auf die Abgleichslisten angewandt und als Resultat immer wieder neue Listen erstellt, solange, bis es nichts mehr zum Abgleichen gibt. Schleifenzähler sorgen dafür, dass die Suche als ständig neuer Abgleichsvorgang nicht infinitiv verläuft.

Ein direkter Treffer zu Games ist 2, indirekte Treffer sind 4 bis 8. Die Darstellung folgt der Form: Games: 2 ® 3-8 ® Null.

Die UND-Verknüpfung von zwei Suchargumenten wird erfüllt, wenn ein Suchargument ein direkter Treffer ist und das weitere Suchargument entweder im selben Dokumentnachweis oder in einem Unterbegriff als indirekter Treffer vorkommt. Unsere Suche nach "The Game of Go" eliminiert "The" und "of" als Stoppwörter; der Term "Game" führt zu den direkten Treffern 2 und 3; "Go" zu 4, 20-22. Die Schnittmenge ist leer. Die Unterbegriffe von "Games" (2) liegen in der Kette 3 - 8. In dieser Kette kommt Go (in 4) vor und wird damit zu einem indirekten Treffer. Mit den daran hängenden URLs 5 und 6 haben wir die Trefferliste komplett.

Was kann der Nutzer mit dem Suchergebnis letztlich anfangen? Wir geben als Suchbegriff "Frauen" ein und erhalten 85 Kategorien, 1420 Sites und 185 Nachrich-

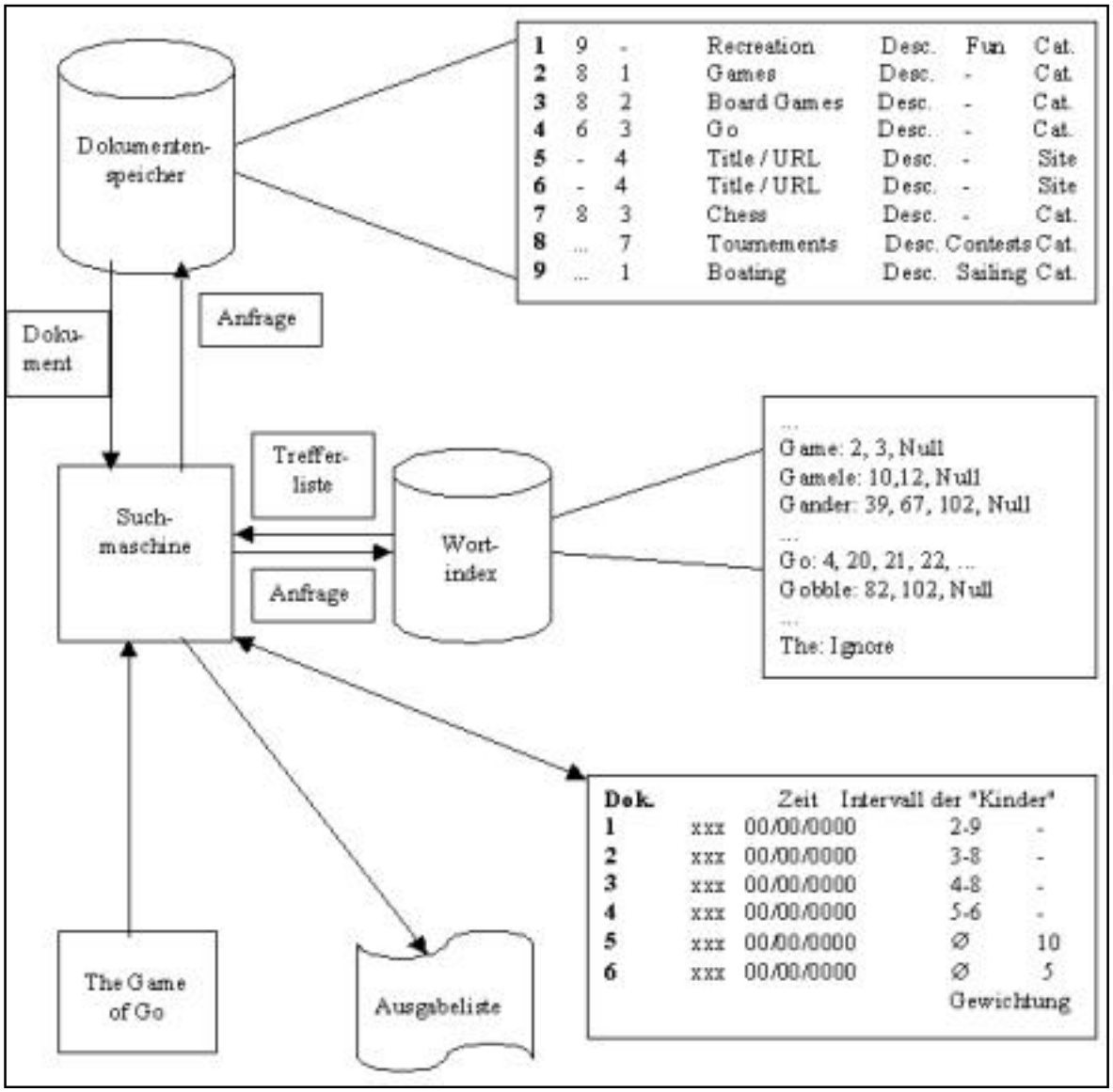

Abbildung 2: Indexstruktur bei Yahoo!
Quelle: Yahoo!; Patent Nr. US 5991756 (leicht modifiziert)

Abbildung 3: Trefferliste zu "Frauen" (Ausschnitt)

ten als Resultat (für die ersten Treffer siehe Abbildung 3). Bei der Sichtung der "Kategorienbäume" sieht man den "Wald" nicht mehr. Warum stehen nicht zumindest die Hauptkategorien geordnet untereinander. In Abb. 3 erblicken wir diverse Klassen, die alle mit der Bezeichnung "Frauen" beschrieben worden sind. Die Klassen sind jedoch nicht identisch, das Wort "Frauen" wird homonym verwandt: einmal für die Gesundheit der Frauen, zum anderen für ihre Stellung in Firmen oder als Sammlung von Frauensites in Österreich bzw. Berlin.

Das Klassifikationsgerüst von Yahoo! ist - für eine Klassifikation überraschend - an einigen Stellen polyhierarchisch organisiert. Die Klasse "Bibliotheks- und Informationswissenschaft" hat sowohl "Geisteswissenschaften" als auch "Bibliotheken" als Oberbegriffe.

Für das Herausgeberteam von Yahoo! besteht die Gefahr , dass ihnen die Systemordnung aus den Fingern gleitet. Wir konnten innerhalb unseres Beispieles aus Abb. 3 eine Stelle finden, wo die Relationen in eine nicht aufhebbare Schleife münden: "Frauen" (als Unterbegriff u.a. von "Gesellschaft und Soziales") hat als Unterbegriff "Frauen- und Geschlechterforschung", diese hat als Unterbegriff "Frauenangelegenheiten" und von diesen aus gelangt man wieder zu den "Frauen" usw. usf. Das leitende Ordnungskriterium bei Yahoo! ist uns verborgen geblieben. Wo ist der rote Faden?

Klassifikation zum Mitmachen: Das Open Directory-Projekt

Klassifikatorische Inhaltserschließung des World Wide Web hat seine praktische Grenze in der Kapazität der Indexer. Ein Unternehmen wie Yahoo! ist nicht in der Lage, eine umfassende Indexierung zu finanzieren. Der Ansatzpunkt des Open Directory-Projektes liegt darin, überhaupt keine festangestellten Indexer zu haben, sondern ehrenamtliche freie Mitarbeiter zu beschäftigen. Grundidee ist: Mit dem Internet wächst die Menge der "Netz-Bürger". Diese Nutzer können jeweils eine (kleine) Menge des Web beobachten und ihre Beobachtungsresultate dem Rest der Gemeinschaft präsentieren, wobei sie die - in ihrer Sicht - schlechten und nutzlosen Sites übergehen und nur den "besten Inhalt" erschließen. Die "Editors", die aktiv Seiten indexieren, sollten ausgewiesene Experten im jeweiligen Fachgebiet sein, so dass hier eine Art Peer Review vonstatten geht.

Ausgang des Projektes war ein gewisser Frust über Yahoo!, der Rich Skrenta und Bob Truel dazu führte, einen anderen Weg zu verfolgen: "Get the Web community itself to work on the directory!" (Sherman 2000, 44). Der erste Projektname 1998 ist "Gnuhoo", eine Kreuzung aus "Gnu" (der verwendeten UNIX-Plattform) und "Yahoo", gefolgt von "Newhoo". Mit der Übernahme des Projektes durch Netscape ändert sich die Bezeichnung offiziell in "Open Directory Project (ODP)", intern wird der Name "Mozilla Directory" in Anlehnung an den Codenamen des Netscape-Browsers gepflegt (deshalb die Homepage: dmoz.org).

ODP klassiert derzeit (Ende November 2000) gut 2,2 Millionen Sites in über 320.000 Klassen; der Mitarbeiterstab liegt bei rund 32.000. Wie bei Yahoo! wird nur ein Bruchteil des Web erfasst, allerdings ist hier durchaus mit noch beträchtlichen Steigerungen zu rechnen.

Das Klassifikationssystem ist stark präkombiniert aufgebaut. Die oberste Hierarchieebene zeigt Abbildung 4. Wie bei Yahoo! tauchen auch bei OPD identische Bezeichnungen in unterschiedlichen Klassen auf. Unser Beispiel zeigt dies für "Kultur und Unterhaltung". Diese Klassenbezeichnung ist u.a. bei diversen Städten vergeben worden, zudem gibt es eine "allgemeine" Klasse "Kultur und Unterhaltung". Jede dieser Klassen verweist auf unterschiedliche Dokumente.

Nach der zweiten - noch internationalen - Hierarchieebene kommen wir zu "deutsch" und arbeiten uns über "Kultur und Unterhaltung" zu den Kölner Karnevalsvereinen nach unten durch:

Top : World : Deutsch : Regional : Deutschland : Nordrhein-Westfalen : Städte und Gemeinden : K : Köln : Kultur und Unterhaltung : Karneval : Garden und Vereine.

Die Einführung einer Klassifikationsebene, die nach dem Alphabet sortiert (im Beispiel "K"), was beim OPD öfters vorkommt, kann wohl nur als Kapitulation vor den Problemen einer thematischen Ordnung gedeutet werden.

Sucht man bei ODP nach konkreten Klassen, so muss das komplette Suchargument in der Klassenbezeichnung vorkommen. Eine Anfrage nach "Kultur AND Unterhaltung" findet in der fortgeschrittenen Suche die entsprechende Unterklasse für Köln (unter 105 anderen), eine Frage nach "Köln AND Karneval" wird nicht fündig, liegt doch "Karneval" zwei Hierarchieniveaus unter "Köln".

Ein Problem beim ODP kann in der subjektiven Auswahl der zu indexierenden und der nicht auszuwertenden Dokumente liegen. Es wird berichtet, dass ein Mitarbeiter eines Unternehmens in seiner Funktion als Editor bei ODP die Seiten seiner Wettbewerber aus der Liste entfernt und dafür die seines eigenen Hauses eingefügt hat (vgl. Sherman 2000, 49). Als Sicherungsmechanismus gegen solchen Missbrauch hat ODP "Meta-Editors" eingesetzt, die für größere Bereiche des Klassifikationssystems zuständig sind und die die Arbeiten der Editors überwachen sollen.

Angesichts der Kooperationen mit ODP scheint das Projekt eine Erfolgsgeschichte zu werden. Wichtige Suchmaschinen, darunter Google und Oingo, bieten die ODP-Klassifikation zusätzlich zu ihren eigenen Services in ihrem Suchwerkzeug an. Die Probleme einer riesigen, überwuchernden präkombinierten Klassifikation bleiben jedoch erhalten.

Warum nicht auf etablierte Klassifikationssysteme zurückgreifen?

Es ist überraschend, dass sowohl Yahoo! als auch das Open Directory-Projekt bei der Erarbeitung ihres Klassifikationsschemas jeweils von vorne anfangen, wo doch etablierte Systeme vorliegen. Ein Rückgriff auf bibliothekarische Werke wie die Dewey Decimal Classification (DDC) oder auf Länder-, Aspekt- bzw. Produktcodes, wie sie vor Jahren von Predicasts entwickelt worden sind, läge eigentlich nahe.

Es gibt Beispiele von (kleinen) Suchwerkzeugen im Web, die bereits den Einsatz der DDC vorführen, etwa die Wolverhampton Web Library (WWLib), die seit 1995 britische Internet-Informationen klassifikatorisch indexiert. WWLib (vgl. Jenkins et al. 1998) indexiert die Web-Dokumente automatisch und ordnet jedem Dokument DDC-Notationen zu. Zunächst wurde ausschließlich auf die umgangssprachlichen Bezeichnungen der Klassen zurückgegriffen, z.B. auf:

641.568 Cooking for special occasions Including Christmas.

Abbildung 4: Eingangsbildschirm des Open Directory
Quelle: Open Directory Project; www.dmoz.org

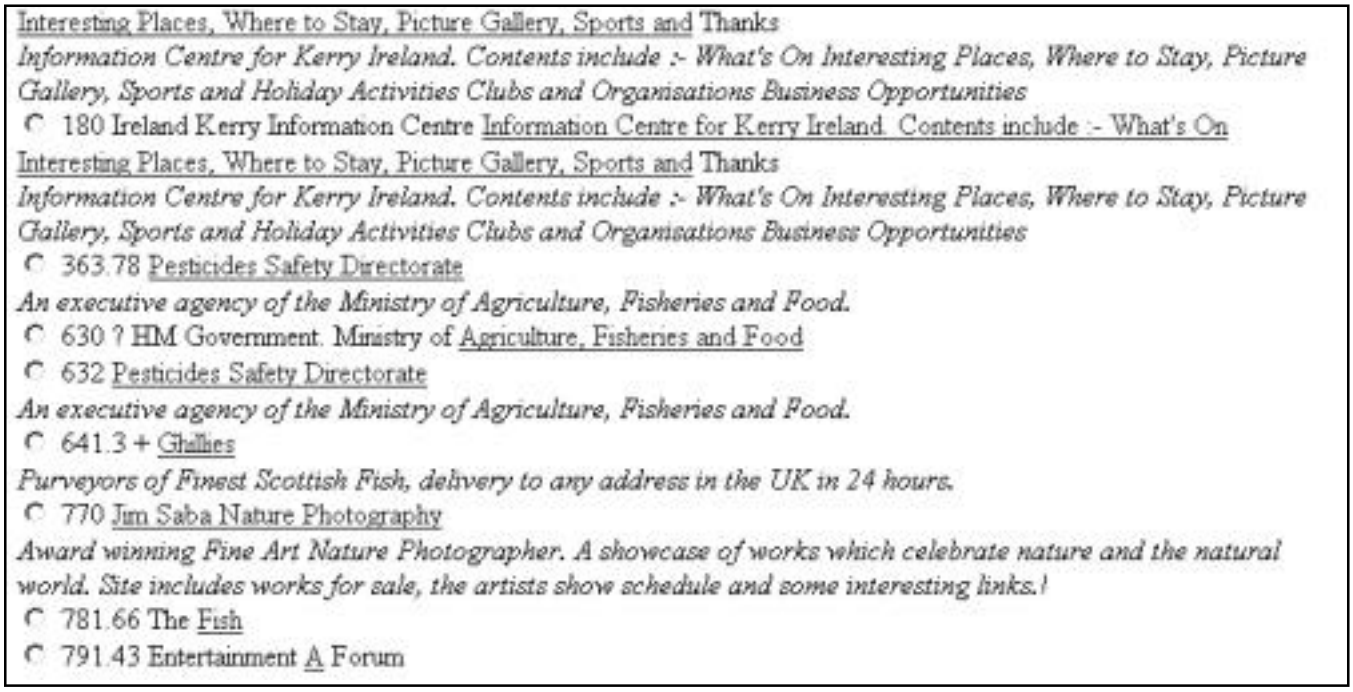

Abbildung 5: Resultate einer Suchfrage zu "fish" in einem DDC-Web-Katalog
Quelle: Wolverhampton Web Library (WWLib); URL: http://www.scit.wlv.ac.uk/wwlib/

Da die Maschine mit diesem Verfahren nur zu 40% auf zufriedenstellende Ergebnisse kommt, wird die Klassenbezeichnung jeder DDC-Klasse durch Listen von Schlagworten und Synonymen stark angereichert.

Eingesammelt werden die Dokumente durch einen Spider, der das World Wide Web automatisch durchsucht. Der Indexierungsprozess durchläuft zwei Phasen. Zunächst geschieht eine informationslinguistische und -statistische Analyse, in deren Verlauf gewichtete Stichworte aus der Vorlage gewonnen werden. Diese Menge an Termen wird in eine der zehn Hauptklassen der DDC eingeordnet. Hiernach wird die jeweils bestpassende Unterklasse gesucht usw., bis keine passende weitere Unterklasse zu finden ist. In diese Klasse wird das Dokument eingeordnet. Das Retrieval geschieht durch Eingabe einer DDC-Notation oder der natürlichsprachigen Bezeichnungen oder durch das Browsen durch die DDC-Hierarchien. Sucht man wortorientiert, so bekommt man Ergebnisse aus allen Klassen, die das Wort als Bezeichnung enthalten (siehe Abbildung 5 als Resultat einer Anfrage zu "fish").

Wir wollen hier nicht auf Details eingehen, sondern lediglich die WWLib paradigmatisch für zwei Aspekte anführen:

1. Eine klassifikatorische Erschließung von Web-Sites mittels etablierter Klassifikationssysteme (wie der DDC) ist möglich.
2. Diese klassifikatorische Erschließung kann auch - ggf. unterstützt durch intellektuelle Arbeiten - automatisch geschehen.

Insbesondere der zweite Punkt erscheint uns wesentlich, leiden doch die "großen" Web-Kataloge wie Yahoo! und das Open Directory unter der viel zu kleinen Zahl ausgewerteter Informationen. Mit der Kopplung an die Technik der automatischen Spider, wie dies WWLib vorführt, ist die Menge der Dokumente wesentlich nach oben zu verschieben.

Terminologische Kontrolle als Teil "semantischen Retrievals": Oingo

Oingo Inc., ein Unternehmen in Los Angeles, wird Ende 1998 von Adam Weissman und dem jetzigen CEO Gilad Elbaz gegründet. Ein Jahr später arbeitet die Suchmaschine im World Wide Web, angetreten mit dem Anspruch, "semantisches Retrieval" zu gewährleisten. "We know what you mean", ist Oingos Motto. Als Suchwerkzeug der Zweiten Generation baut Oingo auf vorhandene Aktivitäten auf und bietet auf deren Basis seinen informationellen Mehrwert. Kooperationspartner sind AltaVista und das Open Directory-Projekt. Die Kerntechnik von Oingo ist sein Bedeutungsraum (Meaning Space), ein Wörterbuch mit terminologischer Kontrolle, d.h. mit Homonym- und Synonymverwaltung, das zur Indexierung und zum Retrieval Verwendung findet. Oingos "Infostructure" hat drei Komponenten: ● die "Oingo Ontologie" (der Bedeutungsraum mit Lexikon und Relationen); ● Indexierung und ● Such-oberfläche.

Oingo arbeitet begriffs- und nicht wortorientiert. Die Einträge in der Terminologie sind demnach "Bedeutungen", identifiziert durch nicht-natürlichsprachige Codes (etwa ID 236). Den Codes sind die entsprechenden Wörter natürlicher Sprachen zugeordnet (z.B. "coffee" oder "café"). Derzeit arbeitet Oingo mit der englischen und der spanischen Sprache; weitere Sprachen sind geplant. Der Umfang der jetzigen Terminologiedatenbank liegt bei rund einer Million Einträgen.

Wie in einem Thesaurus stehen die Begriffe nicht isoliert zueinander, sondern innerhalb eines semantischen Netzes. Die Relationen zwischen den Begriffen stehen für semantische Ähnlichkeiten, wie sie durch die Alltagssprache ausgedrückt werden. In Abbildung 6 sehen wir die Bedeutung von "Java" (als Kaffee) in der semantischen Nähe von der Bedeutung von "Koffein"; die Bedeutung von "Java" (als Programmiersprache) steht an einer ganz anderen Stelle im Netz in der Umgebung der Bedeutung von "Perl". Begriffe, die durch Phrasen (z.B. "John Lennon") bezeichnet werden, sind zum Teil in der "Oingo Lingua" enthalten. Mit dieser terminologischen Kontrolle wird sowohl das Homonym- als auch das Synonymproblem gemildert.

Oingos (automatische) Indexierung setzt auf die Klassen und Dokumente des Open Directory sowie auf die Webseiten bei AltaVista auf. Die Indexierung ist eine Zuordnung der Begriffe (genauer: der ID-Codes) mit infor-

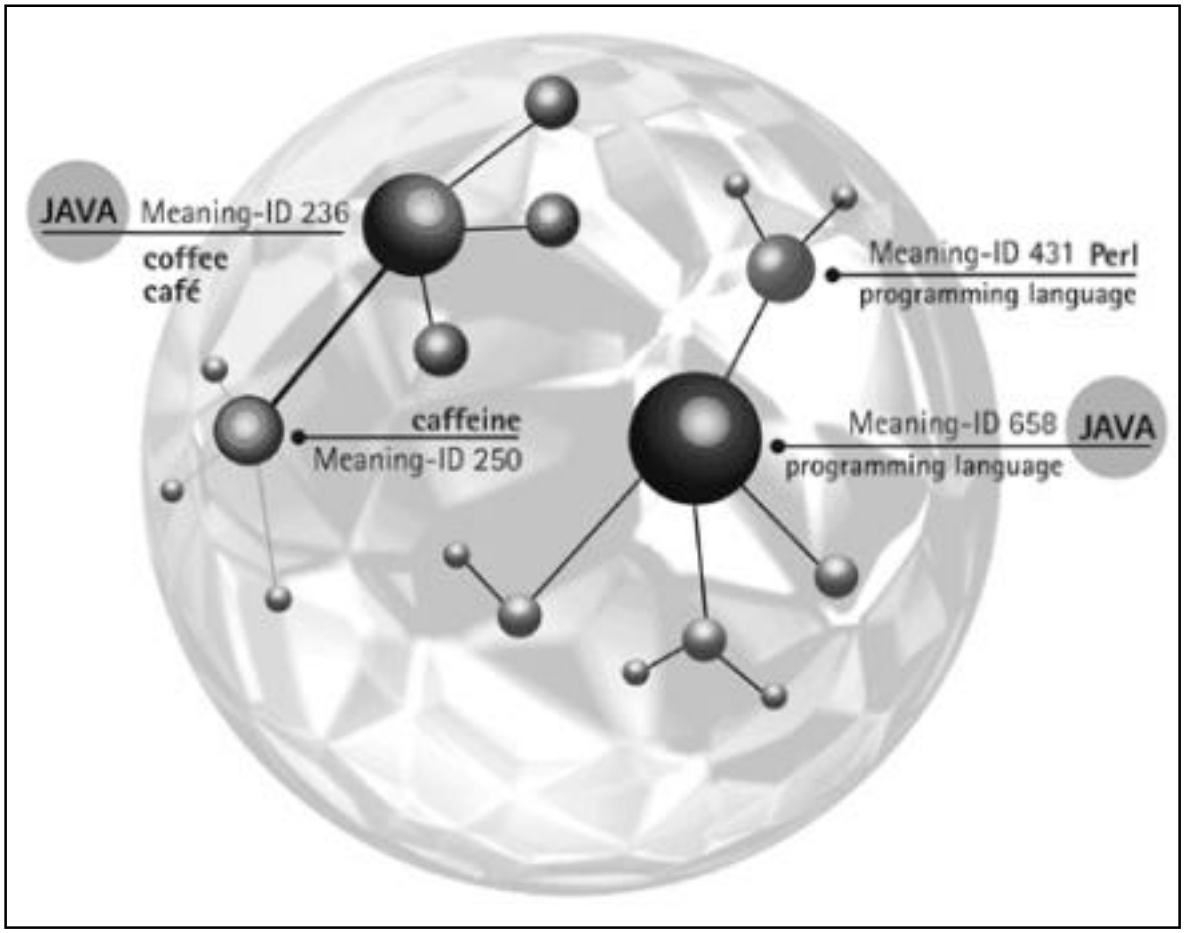

Abbildung 6: Semantisches Netz bei Oingo
Quelle: Oingo Inc.

mationsstatistisch errechneten Gewichtungswerten zu den Klassen bzw. Dokumenten. Zusätzlich durchsucht Oingo Webseiten nach graphischen Informationen (z.B. einem Bild mit Inschrift "Java Kaffee" oder einer einschlägigen Clip-Art-Graphik) und ordnet diese ebenfalls den Bedeutungs-IDs zu.

Das Suchsystem ist auf Rückkopplung eingerichtet. Wenn ein Nutzer ein Wort eingibt, das für mehrere Bedeutungen steht, wird zurückgefragt. Im Dialog wird dem Nutzer eine Liste der Oingo bekannten Homonyme angezeigt, aus der gezielt auszuwählen ist. In Abbildung 7 führen wir unser "Java"-Beispiel fort. Nach der Auswahl von "coffee (beverage)" werden (auf der linken Bildschirmhälfte) eine nach Gewichtung geordnete Rangliste der ODP-Klassen und (auf der rechten Hälfte) die Top 10 bei AltaVista angezeigt. Da nunmehr ausschließlich mit der angegebenen Bedeutung ("Kaffee") gesucht wird, verlieren wir natürlich das Spezifische des eigentlich gewünschten Java-Kaffees. Dieser Fall ist auch bei Thesauri bekannt und akzeptabel. Wenn wir auf einen Nicht-Deskriptor stoßen und via Benutze-Oberbegriff-Relation auf den Deskriptor verwiesen werden, geht es uns nicht anders. Innerhalb des Open Directory arbeitet Oingos Homonymkontrolle weitgehend zuverlässig und nachvollziehbar. Bei AltaVista treten jedoch an vielen Stellen Unstimmigkeiten auf. So klappt die "Java"-Suche mit dem Homonymzusatz "State" sehr wohl bei ODP, nicht aber bei AltaVista, das in diesem Fall den Homonymzusatz mißachtet.

Durch den Rückgriff auf die Bedeutungs-ID bei der Suche wird nach dem Begriff und nicht (nur) nach dem eingegebenen Wort recherchiert. Man kann davon ausgehen, dass das System die Synonyme der englischen Sprache (mehr oder minder) berücksichtigt. Suchen nach "films" bzw. "movies" bringen - bei ODP, nicht bei AltaVista - durchaus ähnliche, wenngleich nicht identische Ergebnisse. Für Oingo ist es also relevant, mit welchem Wort ein Nutzer einsteigt.

Fazit

Mit Yahoo! und dem Open Directory wird ein klassifikatorischer Ansatz zur Inhaltserschließung von Web-Dokumenten verfolgt. Zwei Riesenprobleme begleiten beide Unternehmen: Einer viel zu großen Menge an Klassen steht eine - gemessen am Web - viel zu kleine Menge an Dokumenten gegenüber.

Problemfall I bei Yahoo! wie beim Open Directory ist die Fülle an Klassen, die zudem extrem präkombiniert sind. Es läge nahe, hier durch "Anhängezahlen" oder durch "Facetten" entgegenzusteuern. So arbeiten Profi-Systeme mit facettierten Dokumentationssprachen ja durchaus erfolgreich. Das Predicasts-Codesystem benötigt drei Facetten (Produkt, Land, Aspekt), Profound arbeitet mit deren vier (Marktsegment, Land, Unternehmen, Aspekt). Unseren Web-Klassifikationen wäre mit zwei Facetten schon weitergeholfen. Eine Facette enthält die sachthematischen Bezüge, eine zweite die regionalen Bezüge. Durch zwei Eingabefenster am Bildschirm dürfte eine leichte Bedienbarkeit zu gewährleisten sein.

(Dieser Aspekt ist auch beim DDC-Einsatz wichtig. Über Schlüssel zusammengesetzte DDC-Notationen sind für den Web-Einsatz untauglich, weil vom Laien nicht verstehbar, und durch facettierte Teilsysteme zu ersetzen.) Web-Sites können bzw. - je nach Thema - müssen durch mehrere Notationen indexiert werden. Dieser Schritt von der Präkombination zur postkoordinierenden Vorgangsweise ist in der Informationspraxis (mit der Einführung der Thesauri) in den 60er Jahren des letzten Jahrhunderts gegangen worden. Für unsere Web-Directories ist dieser Schritt nunmehr überfällig.

Problem II betrifft die schmale Dokumentenbasis. Durch den Einsatz kostenlos arbeitender freier Mitarbeiter hat das Open Directory bessere Chancen, auf mehr klassifizierte Dokumente als Yahoo! zu kommen. Trotzdem bleiben wir im Promillebereich des Web. An dieser Stelle halten wir es für notwendig, über die intellektuelle Erschließung hinauszugehen, um auch automatisierte Varianten (wie bei der WWLib) einzusetzen.

Insbesondere das lästige Homonymproblem wird durch das "semantische Retrieval" von Oingo (allerdings bisher nur für das Englische und Spanische) stark gemildert. Über einen Dialog zwischen System und Nutzer wird die gewünschte Bedeutung des eingegebenen Wortes ermittelt und zur weiteren Suche verwendet. Zustimmend ist zu vermelden, dass Oingo einen klaren Schritt hin zur terminologischen Kontrolle gegangen ist, wie er u.a. in (Dokumentaren allseits bekannten) Thesaurusleitfäden oder -normen vorgeschlagen wird.

Eine weitere Annäherung der in der Praxis entstandenen, theoretisch nicht "vorbelasteten" klassifikatorischen Suchwerkzeuge auf der einen Seite und der informationswissenschaftlichen Klassifikationsforschung sowie der bibliothekarischen Klassifikationspraxis auf der anderen Seite dürfte für beide Teile erfolgversprechend sein.

Mechtild Stock &
Wolfgang G. Stock

Literatur

Yahoo!

David Filo; Jerry Yang: Yahoo! unplugged. Your Discovery Guide to the Web. - Foster City: IDG Books Worldwide, 1995.

Dan Lester: Yahoo! Profile of a Web Database. - In: Database 18 (1995), Nr. 6, S. 47-50.

Alan Neibaur: How to Do Everything with Yahoo! - Osborne McGraw-Hill, 2000.

Danny Sullivan: Yahoo Opens Express Submission Service. - In: The Search Engine Report März 1999. URL: http://searchenginewatch.internet.com/sereport/99/03-yahoo.html.

Jiong Wu: Information Retrieval from Hierarchical Compound Documents / Yahoo Inc. - Patent Nr. US 5991756 vom 23.11.1999.

The Open Directory

Greg R. Notess: Review of Open Directory.- 2000. URL: http://www.notess.com/search/dir/dmoz/index.shtml.

Chris Sherman: Humans do it better. Inside the Open Directory Project. - In: Online 24 (2000), Nr. 4, S. 43-50.

DDC

Charlotte Jenkins; Mike Jackson; Peter Burden; Jon Wallis: Automatic classification of Web resources using Java and Dewey Decimal Classification. - In: 7th International World Wide Web Conference. - Brisbane 1998. - URL: http://www7.scu.edu.au/programme/posters/1846/com1846.htm.

Oingo

Paula J. Hane: Beyond keyword searching - Oingo and Simpli.com introduce meaning-based searching. - In: Information Today 17 (2000), Nr. 1, S. 57.

Greg R. Notess: Up and coming search technologies. - In: Online 24 (2000), Nr. 3, S. 75-77.

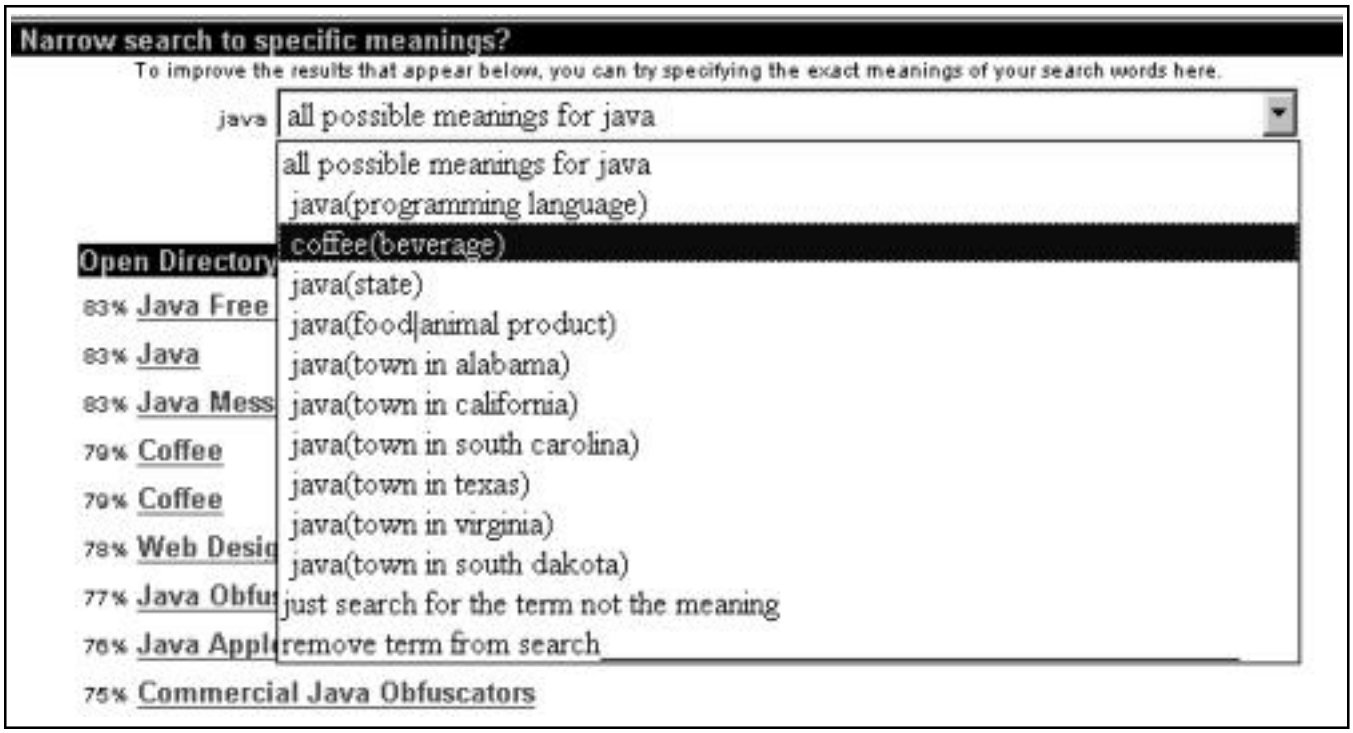

Abbildung 7: Dialog zur Homonymkontrolle bei Oingo
Quelle: www.Oingo.com