nova-net Werkstattreihe

Innovation in der Internetökonomie

# Internetbasierte Expertensuche

Fabian Kaiser
Holger Schwarz
Mihály Jakob

Stuttgart 2006

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Herausgeber: Fabian Kaiser, Holger Schwarz, Mihály Jakob
Verlag: Fraunhofer IRB Verlag
Nobelstaße 12, 70569 Stuttgart

ISBN: 3-8167-7042-8

Erscheinungsjahr: 2006

Auslieferung und Vertrieb: Fraunhofer IRB Verlag
Nobelstraße 12
70569 Stuttgart
Telefon +49 (0) 711/9 70-25 00
Telefax +49 (0) 711/9 70-25 08
www.irb.buch.de
www.publica.fhg.de

## Inhaltsverzeichnis

# 1 Motivation und Rahmenbedingungen

Die Entwicklung von innovativen Produkten und Dienstleistungen bestimmt seit jeher den ökonomischen und kulturellen Fortschritt der Menschheit. Doch nicht nur Innovationen an sich wurden und werden entwickelt, auch der zu diesen führende Prozess musste sich im Laufe der Geschichte entsprechend an die kulturellen und sozialen Rahmenbedingungen anpassen. Mit Erfindung des Internets und der sich mittlerweile etablierenden Internetökonomie haben sich auch die Rahmenbedingungen von Innovationsprozessen stark geändert bzw. sind durch besondere Merkmale gekennzeichnet (vgl. Fichter et al. 2005):

- Dynamisierung von Innovationsprozessen
  Bedingt durch immer schnellere und besser vernetzte Informations- und Kommunikationstechnologien (IKT) steigen Verfügbarkeit und Verbreitung von Informationen und Wissen. In einem zunehmend weltumspannenden Handel von Produkten, Dienstleistungen und Informationen sinkt dabei die Halbwertszeit von Wissen kontinuierlich. Hoch dynamische Bewegungen verändern die Rahmenbedingungen wirtschaftlichen Handelns derart schnell, dass einmal erlangtes Wissen über Kontexte schnell veraltet sein kann, so dass es keine Antworten auf neue Fragestellungen mehr zu liefern vermag. In diesem Umfeld des rapiden Wandels verkürzen sich entsprechend Produkt- und Innovationszyklen. Um sich auf dem Markt behaupten zu können, müssen Unternehmen daher immer schneller neue, innovative Produkte bzw. Dienstleistungen entwickeln und neue Märkte erschließen.

- Steigende Komplexität des Innovationsmanagements
  Neben der fortschreitenden Dynamisierung zeichnen weitere Rahmenbedingungen verantwortlich für eine zunehmende Komplexität des Innovationsmanagements. Flexible Unternehmens- und Projektorganisationen führen zu Bildung von komplexen Akteuersnetzwerken, bei denen die Mitglieder oft zusätzlich räumlich und zeitlich entkoppelt sind. Virtuelle Unternehmen und Communities of Practice bilden sich projektbezogen und lösen sich nach getaner Arbeit wieder auf. Die Identifikation und Integration geeigneter Partner stellt dabei sowohl an das Management als auch an die Informations- und Kommunikationstechnik (IKT) entsprechend hohe Anforderungen.

Neben der herkömmlichen Telefonkommunikation nimmt das Internet bzw. seine derzeit bedeutendsten Dienste WWW und E-Mail mittlerweile die wichtigste Position in der Unternehmens-IKT ein, wobei sich auch die Telefonkommunikation im Zuge von Voice over IP (VoIP) zunehmend des Internets bedient. Dank der immer weiter steigenden Speicherkapazitäten wird das Internet dabei immer mehr zu einer umfangreichen Informationsbasis, die Unmengen an Wissen enthält.

Wissen und Informationen stellen dabei mit die wichtigsten Erfolgsfaktoren in Innovationsprozessen dar und die Koordinierung dieser Ressourcen ist aufgrund der zunehmenden Komplexität der Informationsbasen eine immer größere Herausforderung.
Wettbewerbsrelevante Informationen finden sich dabei in unterschiedlichen Quellen. Neben hochstrukturierten unternehmensinternen Datenbanken mit Daten zu Prozessen und Entwicklungen tun sich mit dem Internet mehr und mehr umfangreiche, dafür meist unstrukturierte Informationsbasen auf. Mittel und Wege zu finden, diese zu erschließen, verspricht einen entscheidenden Wissensvorsprung und damit immense Wettbewerbsvorteile. Neben diesem in technischen Ressourcen verfügbaren Wissen existiert eine zweite, ungleich wichtigere Wissensquelle – der Mensch.

Damit ist die gängige Unterscheidung in explizites und implizites Wissen angesprochen, die vor allem auf Polanyi (1985) zurückgeht. Explizites Wissen besteht in formal niedergeschriebener Form bzw. kann grundsätzlich verbalisiert und damit „expliziert“ werden. Diese expliziten Wissensbestandteile, die auch als „knowing what“ beschrieben werden, sind jedoch immer untrennbar mit implizitem Wissen verbunden, dem „knowing how“. Das implizite Wissen umfasst die habitualisierten Schemata des Wahrnehmens, Urteilens und Verhaltens. Als Regeln des sozialen Lebens, sind sie von jedem Teilnehmer erlernt worden und werden jeder Kommunikation unterstellt sowie im abweichenden Vollzug verändert. Diese impliziten Wissensbestandteile, welche die Ebene der Interaktion formen, sind nur begrenzt und immer nur in Ausschnitten durch Reflexions- und Diskursprozesse explizierbar (Rammert 2001).

Daraus folgt auch, dass die oben genannten technischen Informationsquellen immer nur  explizite Wissensteile darstellen können, d.h. Wissen, das in einer Form festgehalten wurde, die von Personen oder, mit Einschränkungen, durch automatisierte Systeme verarbeitbar ist. Für eine Umsetzung in Handlungen und Entscheidungen allein reichen diese expliziten Wissensteile jedoch nicht aus, sondern sie bedürfen immer ergänzender (impliziter) Regeln, wie sie anzuwenden sind. Menschen und Organisationen müssen immer auch „wissen“, wie bestimmte Informationen zu benutzen sind. Daraus folgt auch, dass Wissen eher als eine Kompetenz etwas zu tun, d.h. als Handlungsvermögen (Stehr 1994) angesehen werden sollte, denn als mobiles „Gut“ (vgl. Rammert 2003). Wettbewerbsentscheidend sind dann aber nicht nur die einzelnen expliziten Informationen, sondern gleichsam ihre impliziten Anwendungs- und Interpretationsregeln (vgl. Rammert 2003)

Eine Möglichkeit, dieses implizite Wissen in einen Innovationsprozess zu integrieren, ist die Kooperation und Interaktion mit Experten. Experten zeichnen sich eben nicht nur durch ihr explizites Wissen aus, sondern auch durch ihre Fähigkeit, intuitiv die richtigen Hinweise oder Anhaltspunkte zu liefern. Grundlage hierfür sind die umfangreichen Kompetenzen eines Experten auf seinen jeweiligen Fachgebieten und der damit einhergehende Überblick.

Dieses Wissen bzw. die Experten mit diesem Wissen müssen aufgrund der Komplexität und des Umfangs entsprechend organisiert werden, um einen effizienten Zugriff zu ermöglichen. Diese Aufgabe wird gemeinhin als Wissensmanagement bezeichnet. Die konkrete Zielsetzung beim Wissensmanagement ist es, möglichst große Teile des expliziten Wissens eines Unternehmens zu katalogisieren und zugänglich zu machen, Ideen sowie eine Infrastruktur anzubieten, um implizites Wissen in einem zunehmend wissensintensiven Wettbewerbsumfeld nutzbar zu machen (Probst et al.1999). In erster Linie zielt das Wissensmanagement dabei auf unternehmensinternes Wissen ab, also auf die Nutzung von Wissen der Mitarbeiter. Unter anderem kommen dabei verschiedene Arten von „Yellow Pages"-Systemen zum Einsatz, in denen Mitarbeiter und ihre Kompetenzen indexiert sind.

Bei Innovationsvorhaben ergibt sich jedoch oft die Problematik, dass unternehmensintern nicht ausreichend Wissen über den Kontext der geplanten Innovation vorhanden ist (Sauer 1999). Dabei können Unsicherheiten unterschiedlicher Art den Erfolg der potentiellen Innovation gefährden: technologisches Neuland, soziale Komponente, Marktkenntnis etc. bzw. der Blick auf die Innovation erfolgt durch die Brille des Unternehmens und ist damit potentiell eingeschränkt.

Um solchen Problemen zu begegnen, wird bei derartigen Vorhaben oft unternehmensexternes Wissen in Form von Unternehmensberatungen oder einzelnen ausgesuchten Experten hinzugezogen. Dadurch lassen sich hochspezialisierte Wissensquellen projektbezogen nutzen und das Risiko einer eingeschränkten Sichtweise auf die aktuelle Fragestellung wird reduziert (vgl. Jakob et al. 2005), Handlungsalternativen können aufgezeigt werden. Sollen in einem Innovationsvorhaben einzelne Experten themenbezogen hinzugezogen werden, dann besteht die entscheidende Herausforderung darin, die richtigen Experten ausfindig zu machen. Verschiedene Quellen können hierzu genutzt werden: Bücher, Fachzeitschriften, Bibliothekskataloge, persönliche Kontakte und nicht zuletzt das Internet mit seiner umfangreichen und stetig wachsenden Informationssammlung.

Das vorliegende Papier konzentriert sich auf die Suche von Experten im Internet. Im Folgenden werden generelle Probleme diskutiert, die sich bei der Expertensuche mittels allgemein gebräuchlicher Techniken wie Suchmaschinen ergeben. Abschnitt 2 stellt verwandte Arbeiten aus Industrie und Forschung vor, die speziell die Problematik der Expertensuche behandeln. In Abschnitt 3 werden Kernpunkte einer im Projekt *nova-net* realisierten, prototypischen Umsetzung der internetbasierten Expertensuche präsentiert. Kapitel 4 gibt schließlich einen Ausblick auf weitere geplante Arbeiten.

## 1.1 Methodische und technische Probleme und Anforderungen

Eine informationstechnische Unterstützung der Expertensuche im Internet ist mit verschiedenen Schwierigkeiten konfrontiert, die sich unter anderem aus den oben genannten Rahmenbedingungen ergeben. Die Anforderungen an eine technische

Umsetzung lassen sich auf die im Folgenden genannten Kernprobleme zurückführen.

### 1.1.1 Fehlende Unterstützung für eine automatisierte Verarbeitung von Inhalten

Das Internet stellt mittlerweile eine der größten Informationsbasen dar, die für jedermann ohne großen Aufwand zugänglich ist, wobei eine Grenze des Wachstums derzeit nicht absehbar ist. Mit zunehmendem Umfang nimmt offensichtlich auch der Umfang der potentiell relevanten gespeicherten Informationen zu. Insofern ist das stete Wachstum des Internets und seiner Akzeptanz zu begrüßen. Andererseits stellt eben dieses Wachstum den Suchenden vor zunehmend höhere Hürden. Zum einen entwickelt es sich zu einem immer aufwendigeren Unterfangen, bei der Suche entsprechendes Material zu sichten, da der Umfang des verfügbaren Datenmaterials mehr und mehr zunimmt. Zum anderen geht mit dem Wachstum des Internets auch eine Tendenz einher, alles und jedes im Netz verfügbar zu machen. Dadurch sinkt das Verhältnis von potentiell relevanten zu irrelevanten Dokumenten deutlich, was den Aufwand und die Schwierigkeit für eine Klassifikation entsprechend erhöht. Die Heterogenität des Internets verschärft dieses Problem: es existiert eine Vielzahl unterschiedlicher Dienste und Plattformen, die alle ihren eigenen Regeln für die Verarbeitung und Präsentation von Inhalten folgen. Diese sind dabei meist auf die Verarbeitung durch menschliche Benutzer hin optimiert. Eine Orientierung hin zu maschinell verarbeitbaren Strukturen wird zwar im Rahmen verschiedener Semantic-Web-Aktivitäten (vgl. Koivunen und Miller 2001) angestrebt, konnte sich bisher aber nur in einzelnen speziellen Bereichen durchsetzen, beispielsweise im B2B-Sektor. Diese mangelnde Unterstützung einer maschinellen Verarbeitung betont wieder eines der Kernprobleme: die Klassifizierung vorhandener Quellen und die Bewertung ihrer Relevanz für das betrachtete Themengebiet. Ohne Hilfsstrukturen wie den für das Semantic Web geforderten Metadaten lassen sich Suchanfragen, wie beispielsweise „finde Autoren auf dem Themengebiet Innovationsmanagement“, nicht ohne weiteres automatisiert beantworten. Solche Autoren finden sich in online verfügbaren Artikeln zum Innovationsmanagement, in Archiven von Mailinglisten oder Newsgroups als Verfasser von E-Mails, in Diskussionsforen oder Blogs und vielen anderen Kontexten. All diese Quellen erfordern aber aufgrund ihrer unterschiedlichen Struktur eine andere Vorgehensweise um zum einen das Vorliegen einer solchen Art von Quelle zu erkennen und im Anschluss die richtigen Techniken zur Identifikation des Autors eines Beitrags anzuwenden.

### 1.1.2 Geringe Suchkompetenz der Benutzer

Während das Internet als Informationsquelle stetig wächst, scheinen sich entsprechende Verbesserungen bei der Suchkompetenz des Anwenders von Standard-Suchmaschinen nicht abzuzeichnen. Silverstein et al. (1999) legen in Ihrer Studie dar, dass die Mehrzahl der Suchanfragen an eine Suchmaschine sich auf drei Begriffe beschränkt. Von den zurückgemeldeten Ergebnissen werden meist nur

die ersten zehn betrachtet. Eine Verfeinerung der Suchanfrage findet nur in relativ wenigen Fällen statt. Selbst wenn Suchmaschinen theoretisch den Informationsbedarf des Benutzers decken könnten, so hat dieser dennoch oft Schwierigkeiten, diesen Informationsbedarf hinreichend genau zu spezifizieren. Vor diesem Hintergrund und der oben geschilderte Komplexitätssteigerung des Internets ist aber eine entsprechend höhere Suchkompetenz auf Seiten des Suchenden gefordert. Es kann weder erwartet werden, dass sich Standard-Benutzer in die Feinheiten der Suche einarbeiten, noch kann davon ausgegangen werden kann, dass für die Aufgabe der Expertensuche in einem Unternehmen ein speziell qualifizierter Mitarbeiter existiert. Diese Defizite müssen entsprechend durch Softwareunterstützung kompensiert werden. Allein anhand einer exakten Spezifikation des Informationsbedarfs lassen sich zwar noch keine Experten identifizieren, da auf dieser Spezifikation aber der gesamte Suchprozess aufbaut, stellt sie eine wichtige Grundlage dar.

### 1.1.3 Geringe Indexierung des Deep-Webs durch Suchmaschinen

Ein weiteres Problem bei der Internetsuche im Allgemeinen und der Expertensuche im Speziellen stellt das sogenannte Deep Web dar (Bergman 2000). Unter Deep Web versteht man Ressourcen, die für Suchmaschinen nicht ohne weiteres erreichbar sind und deshalb von diesen nicht indexiert werden. Dies hat wiederum zur Folge, dass diese Ressourcen in Suchanfragen nicht berücksichtigt werden können. Dazu gehören beispielsweise diverse Schnittstellen zu Datenbanken. Auf die in den Datenbanken gespeicherten Informationen erhält der Benutzer durch das Ausfüllen von Formularen Zugriff. Wenn auch Anstrengungen zur Erschließung solcher Inhalte existieren (vgl. Wu et al. 2004), so sind aktuelle Suchmaschinen dieser Aufgabe bisher dennoch nicht gewachsen, da die Semantik der auszufüllenden Formularfelder der Suchmaschine nicht bekannt ist, und diese somit keine Möglichkeit hat, sie mit entsprechenden Werten zu füllen. Zudem ist fraglich, ob ohne ein umfangreiches Domain-Wissen die Ergebnisse einer solchen Anfrage sinnvoll verarbeitet werden können. Eine weitere Klasse der Deep-Web-Ressourcen sind dynamisch generierte Seiten, deren Inhalt vom jeweiligen Benutzer abhängt (Personalisierung). Zugriff auf diese erhält man oft nur durch eine Anmeldung am betreffenden System. Auch einfache Cookie-basierte Verfahren zur Personalisierung resultieren für Suchmaschinen ebenfalls in einer schlechten Verarbeitbarkeit. Ansätze um diese Art von Quellen für die Suche zugänglich zu machen finden sich beispielsweise in (Raghavan und Garcia-Molina 2001). Eine dritte Klasse von Deep-Web-Ressourcen ist auf den ersten Blick nicht als solche erkennbar. Es handelt sich um Ressourcen, die vom Betreiber der Website explizit von der Indexierung durch Suchmaschinen ausgenommen wurden. Derartiges findet sich relativ häufig im Bereich von Diskussionsforen und ähnlichen Quellen. Insbesondere Diskussionsforen stellen aber für die Expertensuche eine ergiebige Quelle dar, weshalb hier Mittel und Wege gefunden werden sollten, diese für die Suche nutzbar zu machen. Ansätze wie in der genannten Literatur beschrieben

sind folglich unabdingbar für die Expertensuche im Internet und müssen daher in das zu entwickelnde Gesamtsystem integriert werden.

### 1.1.4 Unzureichende Softwareunterstützung

Aufgrund des umfangreichen Datenmaterials, das zur Bestimmung der Kompetenzen eines potentiellen Experten untersucht werden muss, ist die manuelle Durchführung einer solchen Untersuchung entsprechend komplex. In einem Beispielszenario werden mehrere Personen identifiziert, die in verschiedenen bereits untersuchten Quellen Aussagen zum fraglichen Themengebiet gemacht haben. Diese Aussagen treffen möglicherweise nicht genau den Punkt, legen aber nahe, diese Personen näher zu betrachten. Die darauf folgende Arbeit ist ohne spezielle Softwareunterstützung monoton und äußerst aufwendig, da sie sich jedes Mal wiederholt: Es müssen weitere Quellen gesucht werden, die mit der jeweiligen Person in Verbindung gebracht werden können, es muss das Verhältnis dieser Person zur entsprechenden Quelle ermittelt werden und schließlich bestimmt werden, ob diese Quelle relevant für das Thema ist. Finden sich im Rahmen dieses Prozesses weitere potentielle Experten, so lässt sich der Prozess beliebig rekursiv fortsetzen. Neben einer – zumindest teilweisen - Automatisierung dieses Prozesses können Metadaten wie Personennetzwerke oder Kommunikationsstrukturen (vgl. Absatz 2.1) weitere wertvolle Informationen zur Expertise der betrachteten Personen liefern. Sie lassen sich jedoch mit vertretbarem Aufwand nur automatisiert ermitteln.

Eine weitere Problematik wird erst deutlich, wenn bereits eine Unterstützung der Expertensuche durch zusätzliche Software stattfindet und ist quasi die Folge aus oben genannten Anforderungen bzw. Problemen: in der Praxis hat sich gezeigt, dass eine vollständige Automatisierung der Expertensuche nicht zielführend ist. Selbst versierten Benutzern gelingt es dabei nicht, den Informationsbedarf so exakt zu spezifizieren, dass ein vollautomatisches System darauf aufbauend hochwertige Ergebnisse liefert. Das liegt unter anderem an den unterschiedlichen Sichtweisen, die der Suchende und potentielle Experten auf ein Themengebiet haben, ebenso an unterschiedlichen Vokabularen, die sich nicht immer durch den Einsatz von Synonymen aufeinander abbilden lassen und nicht zuletzt daran, dass aktuelle Softwaresysteme noch nicht in dem Maße fähig sind, die Semantik eines natürlichsprachlichen Textes zu verstehen, wie es zur exakten Bestimmung der Kompetenz eines potentiellen Experten nötig wäre. Eine wesentliche Anforderung an ein Expertensuchsystem ist daher, dass es dem Benutzer Routinearbeiten abnimmt, möglichst exakt Texte klassifiziert und Metadaten wie beispielsweise Kommunikationsnetze berechnet. Bei all dem muss dem Benutzer aber stets die Möglichkeit gegeben werden, lenkend oder korrigierend in den Suchverlauf einzugreifen, um dem System aus wenig Erfolg versprechenden Suchräumen zu verhelfen und es entsprechend in eine andere Richtung zu leiten.

# 2 Existierende Anwendungen

Die Problematik der Expertensuche ist nicht neu, folglich existieren auch verschiedene Konzepte und Anwendungen in diesem Bereich, über die in (Yiman und Kobsa 2000) und (McDonald und Ackerman 2000) eine gute Übersicht gegeben wird. Die wichtigsten Arbeiten entstanden im Kontext des Wissensmanagements und fokussieren die intraorganisationale Suche nach Experten. Dies stellt von der Zielsetzung her einen wesentlichen Unterschied zum hier betrachteten Anwendungsfall dar. Allerdings werden auch dort Techniken angewandt, die sich bei der Suche im WWW ebenfalls gewinnbringend einsetzen lassen. Im Folgenden sollen daher einige der bereits existierenden Ansätze und Systeme vorgestellt und kurz auf ihre Stärken und Schwächen eingegangen werden.

## 2.1 Xpertfinder

Der Xpertfinder (Heeren2001) wurde am Fraunhofer IPA entwickelt und hat das Ziel, innerhalb von Unternehmen oder Unternehmensnetzen Experten zu identifizieren. Es ist daher im Bereich der klassischen Wissensmanagement-Systeme angesiedelt. Xpertfinder setzt dabei auf die automatische Analyse von E-Mail- und News-Verkehr sowie serverseitig vorgehaltener Dokumente. Allen beobachteten Personen werden somit automatisch Eigenschaften und Fähigkeiten zugeordnet, was letztlich die manuelle Pflege von „Yellow Pages“-Systemen überflüssig macht bzw. die Pflege unterstützt. Im hier betrachteten Anwendungsfall ist dies jedoch nicht zielführend, da nicht davon ausgegangen werden kann, dass die gesuchte Kompetenz überhaupt unternehmensintern vorhanden ist. Zum einen ist daher eine Beschränkung auf das Intranet bzw. den direkten Informationsaustausch mit externen Institutionen (E-Mail) nicht angebracht. Des Weiteren schließt der hier vorhandene Informationsbedarf ein derartiges Vorgehen praktisch aus. Es werden unbekannte Personen zu einem bekannten Themengebiet gesucht wohingegen der Xpertfinder zuerst bekannte Personen identifiziert und anhand deren Kommunikation ihre bisher unbekannten Expertisen bestimmt. Es ist mit dem Xpertfinder also nicht möglich, Personen zu identifizieren, die sich nicht bereits in einem überwachten Kontakt mit dem Unternehmen befinden.

## 2.2 Matchmaker-Systeme

Unter Matchmaker-Systemen versteht man Anwendungen, die Personen mit gleichen oder ähnlichen Interessen zusammenführen. Diese Problematik findet sich oft in Forschungseinrichtungen, wo viele Personen an sich teilweise überlappenden Themen arbeiten. Um doppelte Arbeit zu minimieren und bei Bedarf Ansprechpartner zu finden, können mittels derartiger Systeme andere Personen identifiziert werden, die sich ebenfalls mit dem gesuchten Themengebiet beschäftigen und daher wahrscheinlich eine gewisse Kompetenz darin besitzen. Beispielhaft für eine derartige Vorgehensweise sind Ansätze, die in (Foner 1997), (Solé und Serra

2001), oder (Vivacqua 1999) vorgestellt werden. Bei diesen Ansätzen handelt es sich um Agentensysteme, in denen jedem Benutzer ein Agent zugeordnet ist. Diese Agenten analysieren die öffentliche Website des jeweiligen Benutzers und/oder weitere Dokumente, leiten daraus die Kompetenzen des Benutzers ab und machen diese Informationen den anderen Agenten zugänglich.Auch hier besteht aber wieder das Problem, dass alle potentiellen Experten dem System a priori bekannt sein müssen, womit ein derartiger Ansatz für die Expertensuche im Internet praktisch ausscheidet.

## 2.3 Netzwerkanalyse in Diskussionsforen

In der Forschung über virtuelle Gemeinschaften (virtual communities) werden unter anderem Techniken zur Analyse von Kommunikationsstrukturen eingesetzt, um den Grad der Interaktion zwischen den Beteiligten zu messen. Primär wird dies nicht zur Expertensuche eingesetzt sondern zum Monitoring der Entwicklung solcher virtueller Gemeinschaften. Beispielhaft hierfür ist ein an der TU Berlin entwickeltes System zur Visualisierung von Wissensnetzwerken (Trier 2005). Aufbauend auf Wrappern für diverse Webforen können die Kommunikationsstrukturen des untersuchten Forums analysiert werden. Im vorliegenden Anwendungsfall ist dies jedoch nur eingeschränkt hilfreich. Zum einen muss der Benutzer das zu untersuchende Forum zuvor selbständig identifiziert haben, zum anderen sind die Themen eines solchen Forums oft so weit gefasst, dass nicht einzelne spezielle Fragestellungen behandelt werden. Da derartige Analyse-Tools aber lediglich die Netzwerkstruktur ermitteln und keine inhaltliche Differenzierung auf Basis der einzelnen Beiträge durchführen, können hier keine Aussagen über die tatsächliche Expertise einer Person auf dem gesuchten Themenfeld gemacht werden, da die beobachtete Kommunikation im Allgemeinen zu unspezifisch ist. Schließlich muss auch bedacht werden, dass zu vielen Fragestellungen im Bereich von Innovationsvorhaben keine öffentlichen Diskussionsforen existieren. Anders als für Consumer-Produkte gibt es beispielsweise im Bereich von Fertigungsverfahren oder Werkstoffeigenschaften wenige bis keine solcher Foren, da das Interesse der breiten Öffentlichkeit daran zu gering ist. Die Verbindung zwischen zwei Personen im Netzwerk kann also oftmals nicht über eine gemeinsame Beteiligung an einer Diskussion ermittelt werden sondern muss sich weiterer Kriterien bedienen (gemeinsame Veröffentlichungen, Zitate, Verweise etc).

## 2.4 Online-Expertennetze

Im WWW haben sich mittlerweile diverse Plattformen etabliert, die zu verschiedenen Themengebieten Informationen und Experten auflisten. Eine solche Plattform ist beispielsweise http://www.brainguide.de, auf der sich Personen unter bestimmten Voraussetzungen als Experten registrieren können. Informationssuchende haben mittels eines hierarchisch gegliederten Themenbaums die Möglichkeit, relevante Personen zu identifizieren und sich anhand gelisteter Veröffentlichungen ein Bild von deren Expertise zu machen. Da derartige Plattformen meist kommerziell

orientiert sind, verlangen sie für die Eintragung eines Experten entsprechende Gebühren. Zwar existieren oftmals auch kostenlose Basiseinträge, deren Funktionalität und die Möglichkeit zur eigenen Präsentation sind jedoch meist stark eingeschränkt. Dies impliziert, dass sich hauptsächlich Personen beim System eintragen, die Ihr Wissen explizit über derartige Plattformen vermarkten wollen. Das wiederum bedingt eine deutliche Einschränkung für den Suchraum, da hiermit faktisch eine Vielzahl potentieller Experten beispielsweise aus der Forschung oder dem privaten Sektor ausgeschlossen wird. Mit dem Prinzip der Selbsteinschätzung und der Fokussierung auf kommerziell orientierte Experten geht auch eine Art „Spamming“ einher. Viele der gelisteten Experten ordnen sich zusätzlichen Kompetenzfeldern zu, die ihre Kernkompetenzen nicht oder nur am Rande berühren. Dahinter steckt die Absicht, entsprechend mehr Verkehr auf ihr Angebot zu locken und höhere Abschlussraten zu erzielen. Derartige Fälle lassen sich schwerlich komplett ausschließen, da es immer Mittel und Wege gibt, sich selbst in einem Kontext zu präsentieren, der den tatsächlichen Gegebenheiten nicht entspricht. Im vorliegenden Fall jedoch wird, bedingt durch die kommerzielle Orientierung, einer solchen großzügigen Bewertungen der eigenen Kompetenzen entsprechend Vorschub geleistet.

## 2.5 Google Answers

Der Frage-Antwort-Dienst von Google, „Google Answers“ (http://answers.google.com) geht das Problem der Einbindung unternehmensexternen Wissens direkter an. Hier konzentriert man sich nicht auf das Finden von Experten die kompetent Auskunft geben können, sondern auf die direkte Beantwortung gestellter Fragen. Dabei wird die Zuordnung zwischen Frage und Experte nicht automatisiert sondern die Frage wird prinzipiell an jedermann gerichtet. Derjenige Experte, der meint eine Frage beantworten zu können, tut dies, womit der komplexe Prozess der Expertensuche an den Experten selbst delegiert wird. Problematisch dabei ist, dass der antwortende Experte nicht zwangsläufig ein Experte auf dem gewünschten Themengebiet ist, sondern lediglich ein Suchexperte. Für detaillierte Fragestellungen ist es zudem sehr wahrscheinlich, dass keiner der bei Google registrierten Experten auch Expertise auf dem betrachteten Themengebiet besitzt. Ein weiterer Schwachpunkt ist, dass Fragen öffentlich gestellt werden müssen und Antworten auf diese in erster Linie ebenso öffentlich abgegeben werden. Insbesondere im Bereich von Innovationsvorhaben ist dies aus Gründen des Wettbewerbs oftmals nicht möglich und bereits aus der öffentliche Formulierung einer Frage kann ein eklatanter Wettbewerbsnachteil entstehen.

Das Hauptproblem der existierenden Ansätze liegt in der Fokussierung auf eng umgrenzte Datenquellen wie beispielsweise unternehmensinterne Datenbanken. Nur für diese ist mit vertretbarem Aufwand eine vollständige Indexierung möglich, was die Grundlage für die auf diesen Techniken basierenden Ansätze bildet. Eine Übertragung allgemein auf das Internet ist dabei nur eingeschränkt möglich, da die Indexierung des gesamten Webs nicht als gangbarer Weg angesehen werden

kann. Prinzipiell können Techniken wie Textklassifikation und Analyse von Kommunikationsnetzen auch bei der internetbasierten Expertensuche zum Einsatz kommen. Die große Herausforderung ist jedoch, in einem frei definierbaren Themen und Dokumentenumfeld Wissensinseln und Strukturen zu identifizieren, anhand derer Personen identifiziert werden können, die Expertise in einem diesem Themenfeld besitzen. Ansätze und Überlegungen dazu sollen im Folgenden näher untersucht werden.

# 3 Kernelemente einer internetspezifischen Expertensuchmaschine

Im Folgenden sollen verschiedene Konzepte erläutert werden, auf denen ein Softwaresystem aufgebaut werden kann, das den oben beschriebenen Anforderungen genügt bzw. die dort angesprochenen Probleme angeht. Zum Teil wurden diese Konzepte bereits im Rahmen des Projekts *nova-net* prototypisch im Rahmen der in Abbildung 1 dargestellten Expertensuchmaschine EXPOSE realisiert. Ihre einzelnen Komponenten bzw. die dahinterstehenden Konzepte sollen im Folgenden beschrieben werden. Ziel ist, eine modulare Plattform zu schaffen, in die die nachfolgend beschriebenen Methoden und Techniken eingebettet werden können, um je nach Kontext den Benutzer möglichst effizient bei der Expertensuche zu unterstützen.

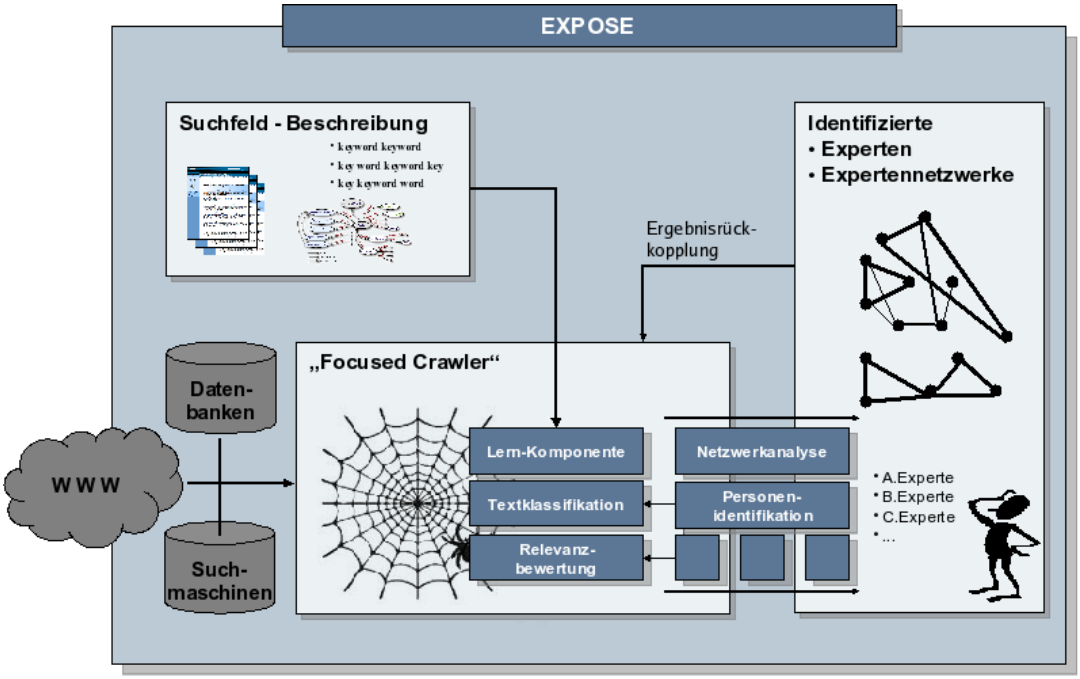


*Abbildung 1: Übersicht Expertensuchsystem EXPOSE*

Ausgangspunkt einer Expertensuche ist meist die mehr oder weniger detaillierte Beschreibung eines Themengebietes anhand derer passende Personen gefunden werden sollen. Nun besteht das Internet aber nicht aus Personen und deren Beschreibung sondern aus verschiedenen Arten von Ressourcen – HTML-Seiten, PDF- und Word-Dokumenten, Mails etc., auf denen Personen in verschiedenen Rollen genannt sind: als Autoren, Gesprächspartner, (Literatur-)Referenzen usw. Diese Ressourcen sind durch Hyperlinks miteinander verknüpft, wobei das Vorhandensein eines solchen Hyperlinks zwischen zwei Ressourcen im Allgemeinen darauf hindeutet, dass zwischen diese Ressourcen ein inhaltlicher oder personeller Zusammenhang besteht. Um also Personen zu einem bestimmten Themenfeld

zu finden, führt der naheliegende Weg über die Suche nach Ressourcen zu eben diesem Themenfeld und der anschließenden Extraktion von Personeninformationen aus den gefundenen Ressourcen.

## 3.1 Sucheinstieg mittels Standard-Suchmaschinen

Der klassische Weg um Ressourcen zu einem bestimmten Themengebiet zu finden ist es, das Themengebiet durch wenige Stichworte zu charakterisieren und die Suche mit einer Standard-Suchmaschine durchzuführen. Problematisch dabei ist, dass viele Sites (insbes. Foren) das Indexieren durch Suchmaschinen verbieten und somit große Teile dieser Daten nicht über Suchmaschinen zugreifbar sind. Ein weiteres Problem dieses Ansatzes liegt in der oftmals mangelnden Suchmaschinen-Kompetenz des Benutzers (vgl. Abschnitt 1.1).
Da sich mit dieser Technik der Informationsbedarf unter dem Gesichtspunkt der Expertensuche nicht decken lässt, sind weitere Verfahren notwendig. Dennoch ist der Einsatz von Standard-Suchmaschinen in einem ersten Schritt sinnvoll. Wenn auch einfache Anfragen oftmals nicht die benötigten hochwertigen Dokumente zurückliefern, so lässt sich doch anhand der Suchergebnisse zumindest ein Teilaspekt des zu untersuchenden Themengebietes erfassen. Zwar wird meist aufgrund von Wort-Mehrdeutigkeiten und nicht zuletzt bedingt durch Spam nur ein Teil der Ergebnisse das Thema auch nur im weitesten berühren. Dennoch kann bereits mit wenigen Treffern eine Grundlage geschaffen werden, auf der aufbauend, weiterführende Techniken eingesetzt werden können.

## 3.2 Spezifikation durch Beispiele

Typischerweise ist es kein triviales Unterfangen, den Informationsbedarf anhand einiger weniger Stichworte zu spezifizieren (vgl. Abschnitt 1.1.2). Deutlich einfacher ist es hingegen, Beispiele für Sachverhalte oder Themen zu geben. Oftmals ist intuitiv klar, dass ein bestimmter Text relevant ist, ohne dass man in der Lage wäre, die exakten Stichworte des Textes zu bestimmen, die ihn relevant erscheinen lassen. Anhand von Beispielen lässt sich ein Informationsbedarf zwar nicht exakt spezifizieren (wobei das auch mittels Stichworten meist nicht möglich ist), aber dennoch hinreichend genau umschreiben, wie sich im Weiteren zeigen wird. Auch solche Beispieldaten müssen jedoch zuerst gesammelt werden. Dieses Problem stellt sich in der Praxis jedoch als deutlich geringer dar, da solche Beispieldaten oftmals bereits vorhanden sind oder wie im vorangehenden Abschnitt beschrieben mittels Suchmaschinen gefunden werden können. Im Fall des Suchmaschineneinsatzes genügt meist eine relativ einfache Suchanfrage verbunden mit geringem Aufwand, um zumindest einige Beispieltexte zu finden, da in diesem Fall kein gesteigerter Wert auf Exaktheit und Umfang des gesamten Suchergebnisses gelegt werden muss. Benötigt wird also ein System, das ausgehend von Beispieltexten weitere Dokumente findet, die eine möglichst große inhaltlich Nähe zu den Beispielen aufweisen.

Dies wirft zwei neue Fragestellungen auf:

1. Wie lässt sich die inhaltliche Nähe zweier Texte automatisch ermitteln?

2. Wie lässt sich eine Suche gestalten, die nicht primär auf dem Einsatz von Stichwörtern und Standard-Suchmaschinen aufbaut?

Die Frage der inhaltlichen Nähe von Texten wird intensiv im Forschungsfeld des Information Retrieval behandelt (vgl. beispielsweise Ferber 2003). Grundlage ist dabei meist eine Vektordarstellung der Texte. Dabei repräsentiert ein n-dimensionaler Vektor genau ein Dokument mit *n* verschiedenen Wörtern (hier: Features). Der Wert eines solchen Features hängt dabei in erster Linie von der Häufigkeit seines Vorkommens im Text ab. Aufbauend auf dieser Darstellung lassen sich in erster Linie auf syntaktischer Ebene Ähnlichkeiten feststellen, indem herkömmliche Methoden zur Ähnlichkeitsmessung von mathematischen Vektoren eingesetzt. Ein Beispiel hierfür ist die Messung des Winkels zwischen zwei Vektoren. Andere Ansätze basieren auf Erkenntnissen und Methoden der Wahrscheinlichkeitstheorie. Beispiele hierfür sind Bayes-Filter oder „Support Vector Machines“ (vgl. Boser et al. 1992), die anhand von Worthäufigkeiten Wahrscheinlichkeiten für die Zugehörigkeit eines Dokuments zu einem Themengebiet berechnen. Um neben der syntaktischen Nähe auch semantische Aspekte zu berücksichtigen, lassen sich Techniken, wie beispielsweise das „Latente Semantic Indexing“ (LSI, vgl. Deerwester et al. 1990) einsetzen. Hierbei wird neben der reinen Häufigkeitsbetrachtung einzelner Begriffe auch deren gemeinsames Auftreten untersucht, wodurch sich inhaltlich ähnliche Dokumente besser identifizieren lassen.

Die zweite aufgeworfene Fragestellung impliziert, dass mit herkömmlichen Suchverfahren das Problem nicht zu lösen ist. In den folgenden Abschnitten „Focused Crawling“ und „Integration existierender Suchmaschinen“ wird diese Frage diskutiert.

Eine Alternative zur Spezifikation mittels Beispieldokumenten ist die Eingrenzung des Themengebietes mittels Begriffsnetzen oder allgemeinen Ontologien. Im Bereich der reinen Suche, unabhängig von der Expertenidentifikation, wurden hier bereits einige Arbeiten durchgeführt (vgl. beispielsweise Sizov et al. 2002, Maedche et al. 2002). Allerdings bietet sich dieses Vorgehen weniger für einen Ad-Hoc-Einsatz an, da entsprechende Vorarbeit in die Entwicklung einer themenbeschreibenden Ontologie investiert werden muss.

## 3.3 Focused Crawling

Herkömmliche Suchmaschinen erwarten vom Benutzer die Eingabe von Schlüsselwörtern und liefern diejenigen Dokumente als Ergebnis zurück, die diese Schlüsselwörter enthalten, evtl. unter Berücksichtigung bestimmter Verknüpfungen zwischen den Suchbegriffen. Durch die oben angeregte Änderung der Suchspezi-

fikation, weg von Schlüsselwörtern, hin zu Beispieldokumenten, entfällt somit aber auch die Eingabe für Standard-Suchmaschinen. Neben ihrer mangelnden Eignung für die Spezifikation durch Beispiele ist die mangelnde Indexierung von Ressourcen des Deep Webs ein weiterer Schwachpunkt herkömmlicher Suchmaschinen (vgl. Abschnitt 1.1.3). Daher muss ein Teil der Suchmaschinen-Arbeit eigenständig implementiert werden.

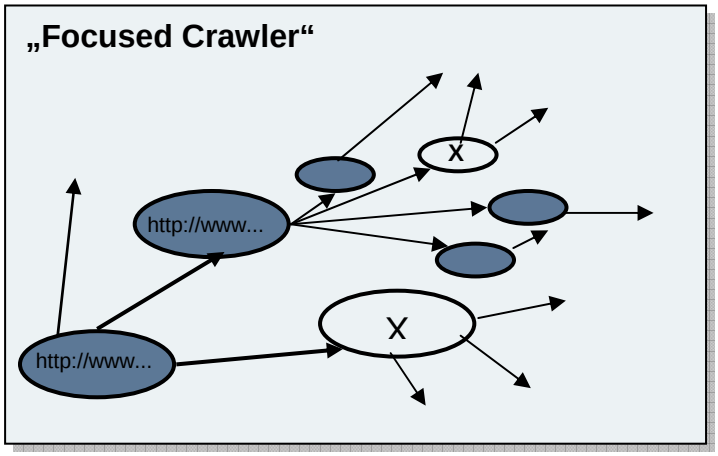


*Abbildung 2: Lauf eines Focused Crawlers*

Diesen genannten Problemen begegnet das sogenannte „Focused Crawling“ (Chakrabarti et al. 1999). Beim Focused Crawling werden ausgehend von einer Menge an Dokumenten die in diesen Dokumenten enthaltenen Verweise (Hyperlinks) mittels eines Crawlers verfolgt und die sich dahinter verbergenden Dokumente geladen. Daraufhin erfolgt ein thematischer Vergleich der geladenen Dokumente mit den Beispieldokumenten (vgl. Abschnitt 3.2). Abhängig vom Ausgang des Vergleichs und unter Berücksichtigung diverser weiterer Kriterien (vgl. Diligenti et al. 2000) werden die Dokumente als Ergebnis markiert und die darin enthaltenen Verweise rekursiv weiterverfolgt (vgl. Abbildung 2). Die Idee dahinter ist, dass Dokumente, die inhaltlich verwandt sind, oft über Verweise miteinander verknüpft sind, wobei solche Verweise nicht von allen Dokumenten eines Themas zu allen anderen desselben Themas gehen müssen, sondern diese auch über mehrere Zwischenschritte verbinden können. Im Unterschied zu Standard-Suchmaschinen werden somit lediglich sehr kleine Teile des Webs durchsucht, wobei der Fokus auf der themenrelevanten Teilmenge des Webs liegt. Irrelevante Dokumente liegen lediglich hinter mehr oder weniger vielen Verweisen aus einem relevanten Dokumente bzw. dienen beim Einsatz von Heuristiken zur Überbrückung von nicht direkt verbundenen Themeninseln.

Der Einsatz eines eigenen Crawlers besitzt auch unter dem Gesichtspunkt der Einbeziehung von Ressourcen des Deep Webs einen weiteren Vorteil: es lassen

sich Konnektoren für themenrelevante Deep-Web-Ressourcen anbinden, was mit Standard-Suchmaschinen nicht möglich ist. Zudem können Websites in die Suche einbezogen werden, die vom Betreiber von der Indexierung durch Suchmaschinen mittels des „Robots Exclusion Standard“ (RES, http://www.robotstxt.org/wc/norobots.html) ausgeschlossen wurden. Streng genommen ist ein Focused Crawler ebenfalls vom Robots Exclusion Standard betroffen, handelt es sich doch um ein System zum automatisierten Laden von Web-Ressourcen. Im Gegensatz zu herkömmlichen Suchmaschinen werden von Focused Crawlern aber im Allgemeinen nur wenige Seiten eines Servers angefordert, was letztlich dem Verhalten eines normalen Benutzers sehr nahe kommt. Durch eine zusätzliche programmtechnische Beschränkung der Anfragen an einen Server pro Zeiteinheit (bspw. höchstens einmal in 30 Sekunden) lässt sich verhindern, dass auf dem Zielserver eine höhere Last als durch einen normalen Benutzer entsteht. Eine freiere Auslegung des RES scheint daher in diesem Fall vertretbar zu sein, wobei diese freie Auslegung als optional implementiert werden sollte, d.h. ein Suchlauf (Crawl) muss auch unter vollständiger Berücksichtigung des RES möglich sein und geschützte Bereiche auf Wunsch ignorieren können.

Offensichtlich ist die Nutzung eines Focused Crawlers nicht mit dem Einsatz einer Suchmaschine vergleichbar. Während eine Suchmaschine in der Regel. innerhalb von Sekundenbruchteilen ein Suchergebnis präsentiert, so kann ein Focused Crawler durchaus mehrere Minuten bis Stunden laufen. Ursächlich hierfür ist, dass bei einer Suchmaschine das Crawling, also das Durchlaufen des Internets sowie der Aufbau des Index, zeitlich vor der Benutzeranfragenverarbeitung erfolgt bzw. seine Aktualisierung gänzlich unabhängig von der Benutzeranfragenverarbeitung ist. Ein Focused Crawler hingegen kann erst dann mit der Suche starten, wenn er die Spezifikation für seine Fokussierung kennt, andernfalls hätte er nicht die Eigenschaft „focused“. Er kann dabei so entworfen werden, dass die Spezifikation in Form von bereits vorliegenden Dokumenten erfolgt, aus Dokumenten besteht, die mittels herkömmlicher Suchmaschinen gefunden wurden, oder auf Basis themenbezogener Ontologien erfolgt (vgl. Abschnitt 3.2).

## 3.4 Modifikation der Suchspezifikation

Eine einzelne Person hat oft eine eingeschränkte Sichtweise auf ein Themengebiet. Ebenso wird das in den Beispieldokumenten oder der Suchanfrage genutzte Vokabular lediglich eine Teilmenge des in diesem Themenbereich verwendeten Vokabulars sein (Stichwort Synonyme). Ein System, das den Benutzer in dieser Hinsicht unterstützt, sollte daher in der Lage sein, relevante Begriffe in den Dokumenten zu erkennen und die Relevanzprüfung damit entsprechend anzupassen. Neben dem sogenannten „Spamming“ (vgl. Henzinger et al. 2003), auf das hier nicht näher eingegangen wird, resultiert aus der Mehrdeutigkeit von Suchanfragen oftmals eine große Schwierigkeit. Viele Begriffe wie z.B. „Bank“ oder „Jaguar“ haben in unterschiedlichen Kontexten eine gänzlich unterschiedliche Bedeutung. Suchanfragen, die solche Begriffe enthalten, produzieren daher offensichtlich ne-

ben den gewünschten auch völlig irrelevante Ergebnisse. Wünschenswert wäre hier eine Software, die dies automatisch erkennt, den Benutzer auf diesen Sachverhalt hinweist und ihn bei der Konkretisierung der Suchanfrage unterstützt. Im Falle von Suchmaschinen kann das beispielsweise eine Modifikation der Suchbegriffe bedeuten. Beim Focused Crawlings dagegen lässt es sich durch eine angepasste Selektion der Beispieldokumente realisieren. Eine Möglichkeit dies zu unterstützen liegt in der Visualisierung von Suchergebnissen und deren Zusammenhängen. Beispielsweise lassen sich aus den zurückgemeldeten Ergebnissen Themencluster ermitteln, die grafisch dargestellt und durch relevante Stichworte ausgezeichnet werden können. Anhand dieser Übersicht kann dann eine korrigierende oder lenkende Einflussnahme des Benutzers erfolgen.

## 3.5 Personenextraktion

Mit der Identifikation relevanter Ressourcen ist die Grundlage für Extraktion von Experteninformationen geschaffen. Um jedoch in diesen Ressourcen auch zuverlässig relevante Personen zu identifizieren, bedarf es umfangreicher syntaktischer und linguistischer Analysen, die die Semantik des Textes erfassen. Kann jedoch eine gewisse Unsicherheit und damit einhergehend eine gewisse Fehlerhäufigkeit in Kauf genommen werden, so lässt sich der Aufwand in dieser Phase deutlich reduzieren. In praktischen Untersuchungen (Mikheev et al. 1999, Palmer und Day 1997) haben sich folgende vereinfachte Verfahren als zielführend herausgestellt. Dabei waren im Vergleich zu komplexen Verfahren der Verarbeitung natürlichsprachlicher Texte sowohl der Aufwand bei der Implementierung als auch bei der Durchführung stark reduziert.

1. Vergleich aller Wörter eines Textes mit einer Namensdatenbank. Ein großer Teil aller Personennennungen kann mittels dieses einfachen Verfahrens bereits identifiziert werden, sofern einige Rahmenbedingungen, wie beispielsweise die häufige Abkürzung von Vornamen („*K.* Müller“), unterschiedliche Wortendungen („K. Müller*s* Kompetenzen“) oder Änderungen der Schreibweise („*Müller, Karl*“) berücksichtigt werden.

2. Suche nach Wörtern im Umfeld von Schlüsselbegriffen, die eine Personennennung andeuten. Beispiele hierfür sind „*laut* Müller“, „Müller *meint/sagt/schreibt/erklärt/behauptet/...*“, „*Frau* Müller“, und ähnliche. Auf diese Art lassen sich auch einzelne Vor- oder Nachnamen identifizieren, die andernfalls aufgrund von Mehrdeutigkeiten nicht sicher als Name identifiziert werden könnten. Der Begriff „*Müller*“ existiert beispielsweise nicht nur als Name sondern auch als Berufsbezeichnung.

Alle Personen die in einem Text nach dem oben genannten Verfahren identifiziert wurden, werden in erster Näherung als „Experten“ zum betreffenden Thema betrachtet. Die Motivation dafür ist, dass die genannte Person eine von vier Rollen in diesem Text einnimmt: sie ist der Autor bzw. einer der Autoren, es wird über sie

geschrieben (Gesprächsprotokoll, Vorstellung der Person etc.), sie dient als themenbezogene Referenz oder wird aus einem irrelevanten Grund erwähnt, d.h. sie hat möglicherweise mit dem Thema überhaupt nichts zu tun (beispielsweise in einer Werbeanzeige oder als Webmaster). In den ersten drei Fällen ist die Annahme, es handle sich tatsächlich um einen Experten damit sicherer, wohingegen sie im letzten Fall eher irrig ist. Die anschließende Aufgabe ist also, in der Menge der erkannten Personen diejenigen zu identifizieren, die tatsächlich eine Nähe zum Thema besitzen und daher mit einer gewissen Wahrscheinlichkeit über entsprechende Kompetenzen verfügen. Diese Thematik wird in Abschnitt 3.6 behandelt.

Die beiden vorgestellten Techniken zur Erkennung von Personennennungen basieren auf Heuristiken und besitzen daher eine gewisse Fehleranfälligkeit. Zwei Arten von Fehlern lassen sich dabei unterscheiden:

1. Personen werden trotz Nennung im Text nicht erkannt.

2. Es werden vermeintliche Personennennungen identifiziert, d.h. Text wird als die Nennung einer Person betrachtet, obwohl dort keine Person bzw. nicht die vermeintlich erkannte Person genannt wurde.

Der erstgenannten Problematik lässt sich, wenn auch nicht abschließend, so doch tendenziell, durch eine Vergrößerung der zugrunde liegenden Namensdatenbank oder einer Erweiterung der unter (2) genannten Regeln begegnen. Das zweite Problem stellt sich als ungleich schwieriger heraus. Hier muss erkannt werden, dass ein allgemeiner Begriff bzw. eine Begriffskombination keine Person repräsentiert. Mit bis dahin eingesetzten Techniken kann dies offensichtlich nicht erreicht werden, da gerade diese Techniken ja zur Fehlerkennung geführt haben. Ein alternativer Ansatz ist daher, die vermeintliche Person in einer ersten Näherung tatsächlich als Person zu betrachten und im Folgenden ihre vermeintlichen Kompetenzen zu analysieren (siehe dazu Abschnitt 3.6). In vielen Fällen wird sich dabei herausstellen, dass die „Person“ nicht als relevant einzustufen ist, d.h. eine weitere Betrachtung nicht notwendig ist.

## 3.6 Expertenidentifikation und Netzwerkanalyse

Die bei der oben beschriebenen Personenextraktion gewonnenen Daten bilden lediglich den Grundstock für eine nachfolgende Expertenidentifikation und enthalten noch wenige direkt verwertbare Informationen. Im Folgenden sollen daher Methoden betrachtet werden, die eine Bestimmung der Experten ermöglichen und hierbei auf den nach obigen Verfahren ermittelten Daten aufbauen,. Abbildung 3 verdeutlicht diesen Zusammenhang.

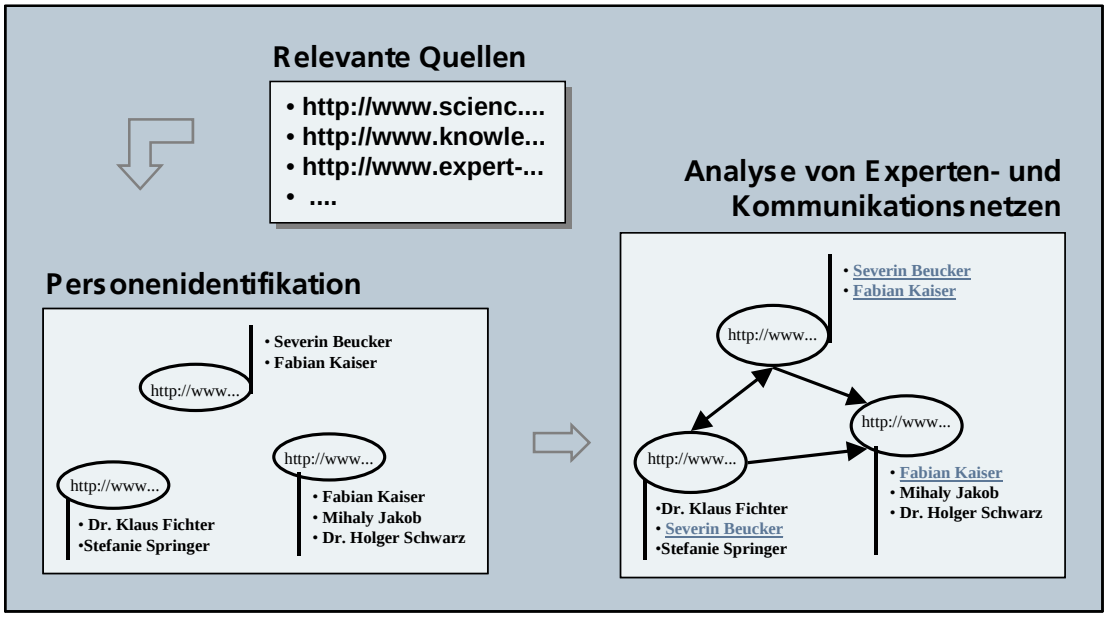


*Abbildung 3: Extraktion von Personendaten*

### 3.6.1 Häufigkeit und Kontext von Personennennungen

Auf dieser Datenbasis lassen sich dann verschiedene Berechnungen und Analysen durchführen. Ein wichtiges Indiz dafür, dass eine identifizierte Person tatsächlich Expertise in dem gesuchten Themenfeld besitzt ist, dass sie in diesem Zusammenhang häufiger in Erscheinung tritt. Wenn eine Person also in mehreren als relevant klassifizierten Dokumenten identifiziert wird, so lässt sich daraus folgern, dass sie tatsächlich gewisse Kenntnisse auf dem betrachteten Gebiet besitzt. Ferner liegt es nahe, solchen Personen eine höhere Bewertung zuzuordnen, die in zentralen Bereichen des Textes identifiziert wurden. Zentral ist hier sowohl im Sinne der Textformatierung, als auch in semantischem Sinne zu verstehen. Ersteres ist leicht zu ermitteln, indem die Position des Vorkommens relativ zu Textanfang und -länge bewertet wird. Die Motivation für dieses Kriterium ist, dass die Struktur der meisten Webseiten mehr oder weniger einem Schema folgt, bei dem an den Seitenrändern oftmals Navigationsleisten oder Werbebanner integriert sind, wo-

hingegen sich der eigentliche Inhalt im inneren Bereich der Seite befindet. Inwiefern ein Name semantisch gesehen zentral für einen Text ist, lässt sich ohne aufwendige linguistische Analysen nicht abschließend ermitteln (vgl. Abschnitt 3.5). Eine einfache und dennoch gute Ergebnisse liefernde Approximation allerdings besteht darin, die Position der Personennennung relativ zum Auftreten relevanter Begriffen des Themenfeldes zu ermitteln und diese Nähe zu bewerten (vgl. beispielsweise Song et al. 2004). Damit lässt sich einem Phänomen begegnen, das von herkömmlichen Printmedien in diesem Maße nicht bekannt ist: Im Web finden sich vielfach unterschiedliche Inhalte in einem gemeinsamen Dokument. Die Seite, wie sie von Printmedien her bekannt ist, hat nicht mehr den starken klassifizierenden Charakter, aus dem geschlossen werden kann, dass thematisch zusammen gehört, was auf derselben Seite steht. Ohne die Betrachtung der Nähe von Personennennung und relevanten Begriffen würden auf derartigen Seiten, die sich teilweise mit dem Thema beschäftigen, beliebige Personen als potentielle Experten klassifiziert werden, sofern sie an irgendeiner Stelle genannt würden.

### 3.6.2 Analyse von Kommunikationsstrukturen

Weitere Informationen über die vermeintliche Expertise einer Person kann über ihr Kommunikationsverhalten ermittelt werden. Beispielsweise deutet ein reger Austausch über einen längeren Zeitraum mit bestimmten Personen auf ein gesteigertes Interesse am Themengebiet hin und damit möglicherweise auch auf eine besondere Kompetenz. Dies kann zum einen an der Häufigkeit des Austauschs festgemacht werden, aber auch an den Kommunikationspartner selbst. Sind an einer solchen Diskussion Personen beteiligt, die bereits als Experten identifiziert wurden, so deutet dies darauf hin, dass auch die anderen Beteiligten gewisse Kompetenzen in diesem Bereich besitzen oder evtl. durch die Diskussion erlangt haben. Derartige Diskussionen lassen sich am besten im E-Mail-Verkehr oder Diskussionsforen analysieren. Beim E-Mail-Verkehr stellt sich aber offensichtlich das Problem, dass man im Allgemeinen keinen Zugriff auf diesen erhält, da es hier zum einen datenschutzrechtliche Probleme gibt und zum anderen auch technisch gesehen der E-Mail-Verkehr zweier Personen nicht beliebig von außerhalb eingesehen werden kann. Hier bleibt also letztlich nur der Weg über öffentlich zugängliche Mailinglisten und deren Archive sowie über öffentliche Newsgruppen. Dennoch stellt dieses Medium dort wo es genutzt werden kann eine wertvolle Informationsquelle dar. Insbesondere lassen sich Metadaten wie Autor, Empfänger und Betreff vergleichsweise einfach extrahieren und für die Expertensuche nutzen.

### 3.6.3 Nennung in verschiedenen Quellen und deren Zusammenhang

Bisher wurde betrachtet, welche Informationen aus dem gemeinsamen Auftreten mehrerer Personen in einem Text abgeleitet werden können, und welchen Einfluss die Position der Nennung auf die Bewertung einer Person haben kann. Weitere Anhaltspunkte dazu ergibt die Vernetzung der einzelnen Quellen. Diese lässt sich in zwei Klassen unterteilen, die direkte Verlinkung mehrerer themenrelevanter Sei-

ten und die Vernetzung über sogenannte „Hubs“, also Seiten, deren textueller Inhalt keine oder nur geringe Themenrelevanz besitzt, die jedoch auf mehrere relevante Seiten verweisen. Je dichter diese Vernetzung ist, desto stärker sind die betreffenden Seiten wahrscheinlich inhaltlich verbunden. Dieser Schluss ist naheliegend, da das Setzen eines Links im allgemeinen zusätzlichen Aufwand (Suche, Erstellung, Pflege) erfordert und somit meist nur dann durchgeführt wird, wenn auf bestimmte Sachverhalte hingewiesen werden soll (Zustimmung, weitere Informationen, Abgrenzung etc). Sind nun aber themenrelevante Quellen stark miteinander verknüpft, dann handelt es sich mit einer gewissen Wahrscheinlichkeit um eine Form von Wissenscluster zu diesem Thema und folglich besitzen die in diesem Kontext genannten Personen wahrscheinlich besondere Kenntnisse in diesem Bereich. Es lässt sich also aus der Struktur der Dokumentenvernetzung ein Rückschluss auf die Qualität und Relevanz der Dokumente ziehen (ähnlich dem PageRank-Verfahren, vgl. Brin und Page1998) und damit auf die Kompetenz der dort genannten Personen.

Aus den oben genannten Punkten wird klar, dass hier eine absolute Aussage über die Expertise der identifizierten Personen nicht gemacht werden kann. Es handelt sich bei allen Ansätzen letztlich um die Arbeit mit Wahrscheinlichkeiten. Ein darauf aufbauendes System wird daher immer eine gewisse Fehlerquote aufweisen. Diese Fehlerquote möglichst gering zu halten ist Ziel der Integration der hier beschriebenen Ansätze.

## 3.7 Suchmaschinen-Integration zur Optimierung des Focused Crawlings

Eingangs wurde auf die mangelnde Suchkompetenz der Standard-Benutzer von Suchmaschinen hingewiesen und erläutert, inwiefern dieses Problem dazu beiträgt, dass die Expertensuche mittels herkömmlicher Suchmaschinen oft nur unbefriedigend gelöst werden kann. Offensichtlich ist aber auch das „Focused Crawling“ keine Lösung frei von jedweden Nachteilen. Der gravierendste hierbei ist, dass nur Quellen gefunden werden können, die tatsächlich auf möglichst direktem Wege mittels Verweisen verknüpft sind, da der Focused Crawler bei fehlenden Verbindungen unabhängige Dokumenteninseln nicht erreichen kann. Hier mag man nun argumentieren, dass nach Albert et al. (1999) die mittlere Distanz zweier beliebiger Ressourcen weniger als 20 Klicks beträgt, sie bei verwandten Themen aber sogar noch deutlich darunter liegen dürfte. Demnach wäre es durchaus möglich, alle relevanten Dokumente zu finden. Dagegen sprechen aber zum einen neuere Untersuchungen (vgl. Broder et al. 2000), die deutlich höhere Distanzen ermittelt haben. Zum anderen sagen diese rein netztopologischen Betrachtungen nichts über den Inhalt der Verbindungsseiten aus. Ein Focused Crawler wird beispielsweise nur in sehr engen Grenzen (vgl. Diligenti et al. 2000) Verweisen folgen, die keinen Bezug zum Thema aufweisen, da er ansonsten seine „focused“-Eigenschaft verlieren würde und damit nicht mehr in akzeptabler Zeit Ergebnisse liefern könnte. Das im Information Retrieval immer wiederkehrende Problem des Auseinanderdriftens von Präzision & Rücklauf (Precision and Recall) zeigt sich al-

so auch hier.: Einer relativ hohen Präzision des Verfahrens steht ein geringer Rücklauf gemessen an allen theoretisch vorhandenen aber nicht gemeldeten Treffern gegenüber, wenn die relevanten Quellen nicht oder nur schwach verlinkt sind.

Dieses Problem lässt sich zum Teil mit der Integration von Standard-Suchmaschinen in das Focused Crawling angehen. Suchmaschinen wie Google oder Yahoo indexieren weitreichende Teile des öffentlich zugänglichen WWW (> $9*10^9$ Seiten, Quelle: Google, Stand November 2005) und machen diesen Index über Webschnittstellen und/oder APIs zugänglich. Damit besitzen sie umfangreiche Informationen über Inhalte und Struktur weiter Teile des WWW. Leider sind diese Daten nicht in ihrem vollen Umfang für Außenstehende nutzbar, da die angebotenen Schnittstellen lediglich einfache, auf Seiten der Suchmaschinenbetreiber wenig rechenintensive Anfragen ermöglichen. Dennoch lässt sich die Datensammlung der Standard-Suchmaschinen auch für das Focused Crawling gewinnbringend einsetzen – weniger um die Präzision zu erhöhen, als vielmehr zur breiteren Streuung der Ergebnisse und damit zur Steigerung des Rücklaufs. Für Spezialanwendungen gibt es bereits erste Ansätze, die in diese Richtung gehen (vgl. Kruger et al. 2000, Diligenti et al. 2000). Eine systematische und umfassende Integration von Standardsuchmaschinen und Focused Crawlern für beliebige Themenbereiche existiert aber unseres Wissens nach bisher nicht.

### 3.7.1 Suchanfrage nach Schlüsselwortextraktion

Betrachtet man die Worte eines Dokuments unter dem Gesichtspunkt, welchen Beitrag sie jeweils zum semantischen Inhalt liefern, so ist offensichtlich, dass einige wenige Begriffe das Themengebiet abstecken und die Restlichen zur inhaltlichen Konkretisierung und Differenzierung beitragen bzw. lediglich Füllwörter darstellen (Stichwort Stopwörter). Letztere sind für die Suche nicht von Interesse, und werden um den Rechenaufwand zu minimieren meist von weiteren Betrachtungen ausgenommen (Beispiel: sog. Stopwörter wie „als“, „an“, „bei“ etc.). Auch die Erfassung der Semantik eines Textes, hier als inhaltliche Konkretisierung und Differenzierung angesprochen, ist ein hochgradig rechenintensives Unterfangen, das heute noch nicht für allgemeine Texte ohne Einschränkungen möglich ist. Es kann auch davon ausgegangen werden, dass ein derart detailliertes Verständnis der untersuchten Quellen nicht nötig ist, um das hier angestrebte Ziel, Experten zu finden, zu erreichen. Es sei dabei wieder auf die latente Unsicherheit hingewiesen, die ohnehin den gesamten Suchprozess begleitet. Es bleibt also letztlich nur die Orientierung an den wenigen Begriffen, die das jeweilige Themengebiet skizzenhaft beschreiben. Wurden bereits relevante Quellen identifiziert, so lassen sich beispielsweise über die TF*IDF-Werte (vgl. Salton und McGill 1983) der darin enthaltenen Begriffe potentielle Schlüsselwörter extrahieren, um daraus eine Suchanfrage für Standard-Suchmaschinen zu generieren. Damit lässt sich dann genau jenes Problem umgehen, dass verwandte Quellen nicht zwangsläufig durch einen kurzen Verweis-Pfad verbunden sind. Die Suchmaschine kennt im Allgemeinen deutlich mehr Quellen, die die identifizierten Suchbegriffe enthalten, auch wenn

diese nicht oder nur über relativ lange Verweispfade miteinander verbunden sind. Über eine zusätzliche Anreicherung des Focused Crawlers mit Ergebnissen von Suchmaschinen können die Suche und Analyse auf einer viel breiteren Datenbasis durchgeführt werden, was entsprechend die Wahrscheinlichkeit erhöht, tatsächlich relevante Quellen und damit relevante Personen zu finden. Offensichtlich erhöht sich der Aufwand für die Prüfung der Suchergebnisse, da von den Suchmaschinen gemeldete Ergebnisse tendenziell mehr irrelevante Dokumente listen, als die eines Focused Crawlers. Da sich der Aufwand für diese (automatisierte) Nachprüfung aber in Grenzen hält, kann er meist in Kauf genommen werden, wenn dafür in der Summe der Rücklauf steigt.

Einen Spezialfall stellt in diesem Zusammenhang die Extraktion nicht nur beliebiger, häufig vorkommender Begriffe, sonder speziell die von Personennamen dar. So wie Suchmaschinen meist weitere Dokumente mit identifizierten Schlüsselwörtern indexiert haben, auch wenn diese nicht unmittelbar miteinander verknüpft sind, befinden sich auch mit einer gewissen Wahrscheinlichkeit weitere Dokumente in ihrem Index, die weitere Nennungen von identifizierten Personen enthalten. Unterstellt man, dass eine Person mit Kompetenzen auf bestimmten Gebieten nicht nur einmalig in diesem Kontext in Erscheinung tritt, so lassen sich auch über diesen Pfad zielgerichtet weitere Informationen gewinnen. Dieser Ansatz birgt also entscheidend mehr Semantik also die reine Fokussierung auf beliebige Begriffe, also rein syntaktische Einheiten.

### 3.7.2 Rückwärtsverweise

Das Crawling basiert allgemein auf dem Prinzip, Verweisen zu folgen, die in bereits bekannten Quellen gefunden wurden. Dabei ist ein Verweis stets eine gerichtete Kante zwischen zwei Knoten im Sinne der Graphentheorie. Gelangt man also von einem Knoten *A* (siehe Abbildung 4) über einen Verweis $V_{ab}$ zum Knoten *B*, so ist am Knoten *B* nicht unbedingt auch einen Verweis $V_{ba}$ zum Knoten *A* vorzufinden. Der Rückweg ist damit – wenn überhaupt - nicht zwangsläufig in einem Schritt möglich.

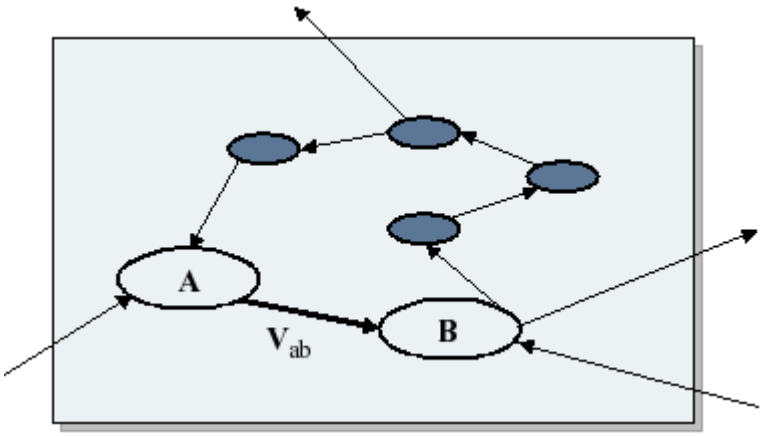


*Abbildung 4: Das WWW als gerichteter Graph*

Offensichtlich will man beim Focused Crawling diesen Rückwärtsschritt auch nicht durchführen, der Inhalt des Knotens *A* ist ja bereits bekannt. Wenn nun aber *A* noch nicht besucht sondern lediglich *B* über einen anderen Pfad erreicht wurde, dann ist nicht gewiss, ob *A* jemals erreicht wird, obwohl evtl. ein inhaltlicher Zusammenhang besteht und die Verweis-Entfernung von *A* nach *B* lediglich eins beträgt (Entfernung *B -> A* ist aufgrund der Kanteneigenschaft „gerichtet“ nicht bekannt). Auch dieses Problem lässt sich jedoch mit Hilfe von Standard-Suchmaschinen angehen. Ähnlich der oben genannten Suchwortanfragen, lässt sich auch hier das Wissen von Suchmaschinen wie Google oder MSN dazu verwenden, eben solche Rückwärtsverweise zu ermitteln (vgl. Diligenti et al. 2000). Suchmaschinen bieten oftmals die Möglichkeit, gezielt nach Seiten zu suchen, die einen Verweis auf eine bestimmte andere Seite gesetzt haben. In diesem Fall ist der Inhalt beider Seiten für die Suchmaschine völlig irrelevant. Konkret bedeutet dies im vorliegenden Anwendungsfall, dass sehr wahrscheinlich auch der Verweis von *A* nach *B* gefunden wird und somit Knoten *A* in die Analyse mit einbezogen werden kann. Wurde also eine relevante Quelle identifiziert, so lässt sich mittels Analyse der in dieser Quelle enthaltenen Verweise sowie dem Einsatz der hier beschriebenen Technik die direkte Nachbarschaft im Dokumentengraphen ermitteln, sofern diese der eingesetzten Suchmaschine bekannt ist.

### 3.7.3 Suche nach ähnlichen Dokumenten

Offensichtlich liegt das Wesen des WWW in der Vernetzung von Dokumenten, die einen wie auch immer gearteten Beitrag zu einem Thema liefern. Dennoch sind offensichtlich nicht alle Quellen zu einem Thema mittels kurzer Verweispfade miteinander verbunden. Um dennoch beim Crawlen eine einmal erreichte Wissensinsel verlassen zu können oder eine neue aufzufinden, bedarf es daher weiterer Informationen über die Netzstruktur und Einstiegspunkte in solche Themeninseln. Auch hier kann das Wissen der Suchmaschinen über Verweisstrukturen wieder zu besseren Ergebnissen verhelfen. Google beispielsweise bietet die Funktion „related:“ an. Mit deren Hilfe können Seiten gefunden werden, die inhaltlich verwandt zu einer in der Anfrage spezifizierten Seite sind. Als Ergebnis werden unabhängig von der Verknüpfung mit der Ausgangsseite diejenigen Quellen zurückgemeldet, deren inhaltlicher Abstand zur Ausgangsseite minimal ist. Zwar kann keine sichere Aussage darüber gemacht werden, wie beispielsweise Google dies technisch realisiert, die Vermutung liegt jedoch nahe, dass dazu die Vektordarstellungen der jeweiligen Seiten herangezogen werden und somit Differenz dieser Vektoren als Maß für deren Ähnlichkeit betrachtet wird (vgl. Abschnitt 3.2).

# 4 Zusammenfassung und Ausblick

In den vorangegangen Abschnitten wurde dargestellt, warum Unternehmen oftmals auf externes Wissen angewiesen sind und dass die Suche nach solchem Wissen, in Form von Experten, diese Unternehmen vor große Herausforderungen stellt. Mit Fokus auf dem Internet und insbesondere seinen beiden wichtigsten Diensten E-Mail und WWW wurden Probleme bei der Suche nach Experten besprochen, wobei sowohl auf Unzulänglichkeiten existierender Techniken und Anwendungen eingegangen wurde, als auch auf Probleme, die Benutzer bei deren Nutzung haben. Im Anschluss daran wurde eine Übersicht über verschiedene Techniken und Ansätze gegeben, mit deren Hilfe diesen Unzulänglichkeiten begegnet werden soll, um den Benutzern eine effiziente Unterstützung bei der Suche nach Experten im Internet zu bieten. Es wurde diskutiert, inwiefern die Kombination von Standard-Suchmaschinen und Focused Crawlern einen Mehrwert gegenüber dem isolierten Einsatz der jeweiligen Systeme darstellt.

Dennoch sei an dieser Stelle erneut darauf hingewiesen, dass auch diese Kombination bzw. Integration kein Allheilmittel bezüglich der aufgeführten Probleme darstellt. Insbesondere besteht weiter das Problem, dass aktuelle Suchmaschinen nur einen Teil des Webs abdecken und über den nicht abgedeckten Teil keinerlei Informationen besitzen und somit keine hilfreichen Aussagen darüber treffen können. Das liegt zum einen daran, dass Betreiber von Websites den Suchmaschinen untersagen, ihren Inhalt zu indexieren, zum anderen aber auch daran, dass ein großer Teil des WWW nur über Formulare zugänglich ist, welche die Crawler der Suchmaschinen derzeit noch nicht automatisiert ausfüllen können. Diese Teile des sogenannten „Deep-Web“ sind dem Benutzer daher nur unter erschwerten Bedingungen zugänglich und können in automatisierte Suchverfahren bisher nicht ohne weiteres eingebunden werden.

Im Rahmen des Projektes *nova-net* wurden bereits einige der diskutierten Techniken und Ideen prototypisch im Rahmen der Expertensuchmaschine EXPOSE implementiert. Erste Ergebnisse zeigen, dass die vorgestellten Konzepte nicht nur theoretisch realisierbar sind, sondern auch in der Praxis sowohl eine Arbeitserleichterung darstellen, als auch die Ergebnisqualität steigern. In einer nun folgenden Phase sollen diese Techniken und Ihre Verbindung weiter ausgearbeitet werden, um schließlich in einer erweiterbaren Plattform integriert zu werden, die Benutzer bei der Suche nach Experten im Internet unterstützt. Der Einsatz in der Praxis wird dabei eine wichtige Rolle spielen, um die Akzeptanz einzelner Techniken sowie deren Integration zu überprüfen.

# 5 Literaturübersicht


Albert, R.; Barabasi, A.-L.; Jeong, H. (1999): Diameter of the World Wide Web. Nature, 401(6749):130--131, September 1999

Bergman, M. (2000): The Deep Web: Surfacing the hidden value. http://www.brightplanet.com/pdf/deepwebwhitepaper.pdf (Stand 02.11.2005)

Boser, B.; Guyon, I.; Vapnik, V. (1992): An training algorithm for optimal margin classifiers, In: Fifth Annual Workshop on Computational Learning Theory, Seite 144-152, ACM, Pittsburgh

Brin, S.; Page, L. (1998): The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: Proceedings of the Seventh International Web Conference (WWW98)

Broder, A.; Kumar, R.; Maghoul, F.; Raghavan, P.; Rajagopalan, S.; Stata, R.; Tomkins, A.; Wiener, J. (2000). Graph structure in the Web, 9th International World Wide Web Conference, Amsterdam

Chakrabarti, Soumen; van der Berg, Martin; Dom, Byron (1999): Focused crawling: a new approach to topic-specific web resource discovery. In: Proceedings of the 8th International World-Wide Web Conference (WWW8)

Deerwester, Scott C.; Dumais Susan T.; Landauer, Thomas K.; Furnas, George W.; Harshman, Richard A. (1990): Indexing by Latent Semantic Analysis. In: Journal of the American Society of Information Science, Vol. 41, No. 4, Wiley

Diligenti, Michelangelo; Coetzee, Frans; Lawrence, Steve; Giles, C. Lee; Gori, Marco (2000): Focused Crawling using Context Graphs, In: Proceedings of the 26th. International Conference on Very Large Databases, VLDB 2000

Ferber, Reginald (2003): Information Retrieval. Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web. dpunkt-Verlag

Foner, Leonard N. (1997): Yenta: A Multi-Agent, Referral-Based Matchmaking System. In: Proceedings of The First International Conference on Autonomous Agents

Henzinger, M., Motwani, R., Silverstein, C. (2003): Challenges in Web Search Engines. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence

Heeren, Frank (2001): Der direkte Weg zum Wissen in den Köpfen. In: Wissensmanagement (2001), Nr. 4, S. 37-40

Jakob, Mihály; Kaiser, Fabian; Schwarz, Holger (2005): SEMAFOR: A Framework for an Extensible Scenario Management System. In: Proceedings of the IEEE International Engineering Management Conference (IEMC) 2005

Koivunen, M. R.; Miller, E. (2001): W3C Semantic Web Activity. In: Proceedings of the Semantic Web Kick-off Seminar in Finland, http://www.w3.org/2001/12/semweb-fin/w3csw (Stand 02.11.2005)

Kruger, Andries; Lee Giles, C.; Coetzee, Frans M.; Glover, Eric; Flake, Gary W.; Lawrence, Steve; Omlin, Christian (2000): DEADLINER: Building a New Niche Search Engine. 9th International Conference on Information and Knowledge Management, CIKM 2000, ACM Press

Maedche, A.; Ehrig, M.; Stojanovic, L.; Handschuh, S.; Volz, R. (2002): Ontology-Focused Crawling of Documents and Relational Metadata. In: Proceedings of the Eleventh International World Wide Web Conference

McDonald, David W.; Ackerman, Mark S. (2000) Expertise Recommender: A Flexible Recommendation System and Architecture. In: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work

Mikheev, A., Moens, M., Grover, C. (1999): Named entity recognition without gazetteers. In: Proceedings of EACL 1999, ACM Press

Palmer, D.; Day, D. (1997): A Statistical Profile of the Named Entity Task. In: Proceedings of the 5th Conference on Applied Natural Language Processing

Polanyi, M. (1985): Implizites Wissen. Suhrkamp, Frankfurt a. M.

Probst, Gilbert; Raub, Steffen; Romhardt, Kai (1999): Wissen managen : wie Unternehmen ihre wertvollste Ressource optimal nutzen. Gabler, Wiesbaden

Raghavan, Sriram; Garcia-Molina, Hector (2001): Crawling the Hidden Web. In: Proceedings of the Twenty-seventh International Conference on Very Large Databases (VLDB)

Rammert, Werner (2001): Nicht-explizites Wissen in Soziologie und Sozionik - Ein kursorischer Überblick. In: Abschlussbericht "Management

von nicht-explizitem Wissen - Noch mehr von der Natur lernen“. Ulm: Forschungsinstitut für anwendungsorientierte Wissensverarbeitung, S. 113-192.

Rammert, Werner (2003): Zwei Paradoxien einer innovationsorientierten Wissenspolitik: Die Verknüpfung heterogenen und die Verwertung impliziten Wissens. In: Soziale Welt 54 (2003), S. 483-508

Salton, Gerard; McGill, Michael J. (1983): Introduction to modern information retrieval. McGraw-Hill, Singapur

Sauer, Dieter (1999): Perspektiven sozialwissenschaftlicher Innovationsforschung – Eine Einleitung. In: Sauer, Dieter; Lang, Christa (Hg.): Paradoxien der Innovationen. Frankfurt (Main), New York: 149-174.

Silverstein, Craig; Henziger, Monika; Marais, Hannes; Moricz, Michael (1999): Analysis of a Very Large Web Search Engine Query Log. Association for Computing Machinery, Special Interest Group on Information Retrieval (ACM SIGIR) Forum Vol. 33, No 1, New York

Sizov, Sergej; Siersdorfer, Stefan; Theobald, Martin; Weikum, Gerhard (2002): The BINGO! Focused Crawler: From Bookmarks to Archetypes. In: Proceedings of the 18th International Conference on Data Engineering

Solé, Ramon Sangüesa; Serra, Josep M. Pujol (2001): Netexpert: A multiagent system for expertise location. In: Proceedings of the IJCAI-01 Workshop on Knowledge Management

Song, Ruihua; Liu, Haifeng; Wen, Ji-Rong ; Ma, Wei-Ying (2004): Learning Block Importance Models for Web Pages. In: Proceedings of the 13th World Wide Web Conference

Stehr, Nico (1994): Arbeit, Eigentum und Wissen: Zur Theorie von Wissensgesellschaften. Suhrkamp, Frankfurt/Main

Trier, Matthias (2005) A Tool for IT-supported Visualization and Analysis of Virtual Communication Networks in Knowledge Communities. In: Proceedings of the Wirtschaftsinformatik 2005. Physica-Verlag Heidelberg

Vivacqua, Adriana S. (1999): Agents for Expertise Location. In: Proceedings of the AAAI Spring Symposium on Intelligent Agents in Cyberspace 1999

Wu, W.; Yu, C.; Doan, A.; Meng, W. (2004): An interactive clustering-based approach to integrating source query interfaces on the deep web. In: Proceedings of the SIGMOD Conference 2004

Yimam, D.; Kobsa, A. (2000). Expert Finding Systems for Organizations: Problem and Domain Analysis and the DEMOIR Approach. Beyond Knowledge Management: Sharing Expertise. M. Ackerman, A. Cohen, V. Pipek and V. Wulf (Hrsg.). Boston, MA, MIT Press.