

Expertenbestimmung durch Analyse von E-Mails

Amancio Bouza, Benjamin J. J. Voigt

Universität Zürich, Institut für Informatik

Zusammenfassung

Wenn eine Person bemerkt, dass die eigenen Fähigkeiten für die Lösung einer Aufgabe nicht ausreichen, besteht Handlungsbedarf. Besonders in wissensintensiven Bereichen, die hauptsächlich auf impliziten Wissen aufbauen, bietet das Wissensmanagement nur die Möglichkeit der Vermittlung von Experten an. Hier beginnt der Prozess der Expertensuche, die sich aus Expertenbestimmung, Selektion und Eskalation zusammensetzt. Expertenbestimmung ist demnach der erste Schritt. Es wird anhand einer Mailingliste aus dem Mozillaprojekt gezeigt, dass Experten durch E-Mail-Analyse bestimmbar sind und wie die E-Mail-Kommunikation zwischen Personen analysiert werden kann, um letztlich Experten zu bestimmen.

1 Einleitung

Das Erledigen von anspruchsvollen Aufgaben erfordert Expertenwissen. Bemerkt ein Mitarbeiter, dass die eigenen Fähigkeiten für die Bewältigung der Probleme nicht ausreichen, muss er die Hilfe eines Experten in Anspruch nehmen, der ihm bei der Bearbeitung und Interpretation helfen kann. Diese Situation wird auch Wissensanomalie¹ genannt (Belkin et al. 1982). Selbst in Fällen, in denen seine eigenen Fähigkeiten ausreichen, kann es sinnvoll sein, Experten hinzuzuziehen, wenn die Aufgabe für einen Einzelnen zu gross ist (Osman & Norshuhada 2004) und eine Aufteilung der Arbeit möglich ist. Wenn der Mitarbeiter selbst keinen geeigneten Experten kennt und die Aufgabe implizites Wissen erfordert, ist es Aufgabe des betrieblichen Wissensmanagement ihm einen Experten zu vermitteln.

Besonders in wissensintensiven Bereichen wie beispielsweise in der Forschung und Entwicklung ist Wissen meist nur implizit vorhanden und

¹ Anomalous State of Knowledge (ASK)

schwierig zu lokalisieren (Willis et al. 2002). Häufig sind aber Experten nicht allgemein im Unternehmen bekannt. Dies gilt besonders für Personen, die noch nicht lange in einer Organisation arbeiten. Aber auch für länger in einer Organisation befindliche Personen ist es schwierig den Überblick über die Fähigkeiten aller Personen zu haben (McDonald & Ackermann 1998).

McDonald und Ackermann (2000) beschreiben die Expertensuche als einen Prozess, beginnend bei der Expertenbestimmung über die Expertenselektion bis zur Eskalation, falls die Hilfe nicht oder nur teilweise für die Lösung des Problems hilfreich war. Es existieren bereits Lösungen für die Expertenbestimmung wie z. B. Expertendatenbanken oder Expertenbestimmung durch die Analyse von sozialen Netzwerken. Bei Expertendatenbanken werden Informationen über die Fähigkeiten einzelner Personen als Profil gespeichert. Das Problem hierbei besteht in der Aktualität, der Korrektheit und Adäquatheit der Profile, da Personen sich neue Fähigkeiten aneignen können, diese manchmal auch nur implizit vorhanden sind und nur Informationen über Fähigkeiten erfasst werden, nicht aber welcher Experte die beste Wahl für ein individuelles Problem ist (Osman & Norshuhada 2004). Anstatt sich auf möglicherweise unvollständige oder veraltete manuell gepflegte Expertendatenbanken zu verlassen, könnte es sinnvoller sein, die Experten automatisch aufgrund ihrer E-Mail-Kommunikation zu identifizieren. Da E-Mails tagtäglich Gebrauch finden, zeigen sie stets ein aktuelles Abbild der Organisation (Moreale & Watt 2003) und bieten sich für die Expertenbestimmung an, wenn Experten ihr Wissen auch kommunizieren (Lindvall & Rus et al. 2002). E-Mails sind aufgrund ihres teilstrukturierten Inhalts für die automatische Expertensuche gut geeignet, da sie bereits viele Metainformationen (RFC 822) zum unstrukturierten Textinhalt enthalten (Weinberger 1999). Es wird sogar vorgeschlagen nur den strukturierten Teil der E-Mails zu nutzen (Moreale & Watt 2003).

Im Folgenden wollen wir anhand eines konkreten Beispiels untersuchen, ob die Möglichkeit besteht anhand der Kommunikation durch E-Mail in einer Mailingliste Experten herauszuarbeiten und wie gut das Ergebnis einer solchen Suche ist. Wir verstehen den Begriff des Experten dabei wie folgt (Posner 1988): „Ein Experte zeichnet sich dadurch aus, dass er auf einem bestimmten Gebiet dauerhaft (nicht zufällig und nicht nur einzelne Male) herausragende Leistungen erbringt.“ Im nachfolgenden zweiten Kapitel stellen wir unsere Annahmen dar. Den verwendeten Datensatz und die Analysewerkzeuge präsentieren wir im folgenden dritten Kapitel. Das vierte Kapitel geht auf das konkrete Vorgehen bei der Analyse von E-Mails ein. Im fünften und sechsten Kapitel präsentieren und bewerten wir dann die Ergebnisse der E-Mail-Analyse.

Im abschliessenden siebten Kapitel ziehen wir dann unsere Schlussfolgerungen aus den Ergebnissen.

2 Hypothese

Es wird empfohlen durch natürliche Strukturierung (RFC 822) angereicherte Inhalte von E-Mails (Weinberger 1999) bezüglich eines Beitrags zum Bestimmen von Experten zu analysieren. Da E-Mails tagtäglich Gebrauch finden, zeigen sie stets ein aktuelles Abbild der Organisation (Moreale & Watt 2003) und bieten sich für die Expertenbestimmung an, wenn Experten ihr Wissen auch kommunizieren (Lindvall & Rus et al. 2002).

Wir verstehen den Begriff des Experten wie folgt (Posner 1988): „Ein Experte zeichnet sich dadurch aus, dass er auf einem bestimmten Gebiet dauerhaft (nicht zufällig und nicht nur einzelne Male) herausragende Leistungen erbringt.“.

Sowohl das Schreiben von Antworten als auch das Zusammenfügen von Informationen zu einer Antwort zeugt von Expertise und zeigt den Unterschied zwischen einem Experten und allen anderen auf (Ackermann & McDonald 1996). Wir nehmen also an, dass wir Experten auf einem Gebiet durch die Auswertung des E-Mail-Verkehrs anhand der Frequenz, mit der die Personen antworten, bestimmen können, da dass ständige Antworten auf Fragen einer dauerhaften herausragenden Leistungserbringung entspricht. Des Weiteren vermuten wir, ausgehend von der Expertenbestimmung, dass ein Wechsel von Experten festgestellt werden kann.

3 Verwendeter Datensatz und Werkzeuge

Um Nachvollziehbarkeit zu gewährleisten wurde ein öffentlich zugänglicher Datensatz aus der Open Source Community gewählt. Um aussagekräftige Ergebnisse zu erzielen haben wir den verhältnismässig grossen E-Mail-Verteiler zum Modul² „Build Config“ des Mozilla Projekt ausgewählt. Da der Verteiler seit längerer Zeit³ existiert und die

² Das Mozilla-Projekt besteht aus Modulen. Ein Modul besteht aus einer Sammlung von Quelldateien, die ein kohärentes Bündel bilden.

³ Die Mailingliste „Build Config“ existiert seit April 1998.

Teilnehmer sehr aktiv sind, erfüllt der Datensatz die Bedingungen, die notwendig sind um das Kriterium der dauerhaften Leistung bezüglich Wissensdemonstration zu erfüllen.

Der ursprüngliche Datensatz bestand aus 28019 E-Mails. Nach der Clusterbildung und der Auswahl eines Clusters waren noch 11428 E-Mails vorhanden. Nach der anschliessenden Klassifikation lagen uns 14129 zum Thema des ausgewählten Clusters passenden E-Mails vor. Davon waren 9904 Antworten auf eine Frage und stellten also potentiell explizites Wissen dar.

Da die Klassifikationssoftware am besten mit einzelnen Dateien denn mit einer grossen Datei arbeitet und der Datensatz im Mbox-Format vorlag, war es nötig die Daten mit dem Mbox-Maildir-Konverter Mb2Md in das Maildir-Format zu konvertieren.

Für die Clusterbildung haben wir die Software Crossbow (McCallum 1996) mit der Standard Konfiguration (Naïve Bayes, Ereignis Modell Wörter) verwendet.

Die Clusterbildung ist mit einem sehr hohen Rechenaufwand verbunden. Darum haben wir uns entschieden in diesem Fall ausschliesslich nach den Wörtern aus den E-Mail-Betreffs zu clustern. Ein Betreff ist letztlich eine knappe Beschreibung der Nachricht und enthält wichtige Schlagworte. Für eine Überprüfung unserer Hypothese ist dieses Vorgehen ausreichend, da die Machbarkeit im eigentlichen Sinne, und nicht die Performanz eines solchen Systems im Vordergrund steht.

Als Textklassifizierungssoftware kam Rainbow (McCallum 1996) zum Einsatz (mit den Optionen `-h --skip-header --istext-avoid_uencode` und dem Standard Pruning von etwa 524 Wörtern der Englischen Sprache). Die Klassifikation wurde auf den E-Mail-Inhalt, und nicht nur auf den Betreff angewendet. Als Klassifizierungsmethode haben wir den Naïve Bayes-Algorithmus verwendet. Die Retrieval-Performanz ist vergleichbar mit anderen Algorithmen wie z. B. dem Support Vector Machine, wenn über 300 Trainings-Dokumente (Yang & Liu 1999) verwendet werden, was in unserem Fall gegeben war. Ausserdem ist der Naïve Bayes-Algorithmus allgemein gebräuchlich und die Resultate sind somit leichter vergleichbar.

Zum trennen der Antwort E-Mails von den Frage E-Mails wurde u. A. das UNIX Werkzeug „grep“ verwendet.

4 Vorgehen

Da in Mailinglisten ein breites Themenspektrum abgebildet ist, besteht die Gefahr, dass man Personen als vermeintliche Experten aufgrund ihres breit gefächerten aber nicht spezialisierten Wissens bestimmt. Darum haben wir uns zur Analyse auf ein Diskussionsthema beschränkt. Aufgrund der Cluster-Analyse haben wir E-Mails, die dem grössten Cluster „problem linux error“ (11428 E-Mails) zugeordnet wurden, als Trainings-Daten für die Klassifizierung ausgewählt. Unser ausgewähltes Thema dreht sich somit um Linux und Probleme – im Kontext des Mozilla Moduls „Build Config“.

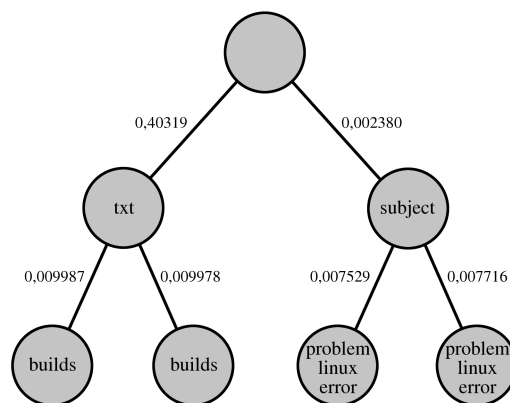


Abbildung 1: Cluster der E-Mail-Betreffs

In einem zweiten Schritt wurden die Inhalte des gesamten Datensatzes (inkl. der E-Mails, deren Betreffs zum Bilden des Clusters verwendet wurden) von dem trainierten Naïve Bayes-Klassifikationssystem analysiert. Als Ergebnis erhielten wir 14129 E-Mails die den Schwellenwert von 0.9 überschritten, d. h. innerhalb des Signifikanzniveaus von 0.1 lagen. Die Entscheidung für den relativ tiefen Schwellenwert begründen wir mit der möglichen Ambivalenz und vorhandenen Artefakten in dem Text selbst.

Abschliessend legen wir den Fokus nun auf die E-Mails, die als eine Antwort auf eine Frage-E-Mail geschrieben wurde. Das Schreiben von Antworten zeugt letztlich von der Expertise einer Person und macht den Unterschied zwischen Experten und allen anderen aus (Ackermann & McDonald 1996). E-Mails unterscheiden sich von gewöhnlichen Texten durch eine Codierung des Betreffs. Dabei werden beispielsweise

Antworten mit einem vorangestellten „Re:“⁴ (Crocker 1982) gekennzeichnet. Nach dieser Einschränkung bleiben uns noch 9904, die als Antwortmails anzusehen sind. Rückfragen oder -antworten auf Antworten werden mit einem doppelten „Re: Re:“ gekennzeichnet. Wir haben an dieser Stelle jedoch keine Unterscheidung zwischen einem „Re: “ und mehreren nachfolgenden „Re: “-Elementen gemacht, da in unserem Kontext der Expertenbestimmung eine Antwort auf eine Antwort gleichermassen als Expertenbeitrag angesehen wird.

5 Analyse und Ergebnisse

Zum Bestimmen eines Experten, müssen zwei Kriterien erfüllt sein: herausragende Leistung und dauerhafte Leistung. Die Dauerhaftigkeit können wir anhand der Verteilung der Antworten über die Zeit feststellen. Antwortmails erkennen wir als Leistung an.

In einem ersten Schritt haben wir alle Autoren festgestellt und aufgelistet. Wir haben sie aufgrund der Anzahl ihrer Antwort-E-Mails, die geschrieben wurden, sortiert.

Für die Darstellung in Abbildung 2, die die Verteilung der Autoren anhand der Anzahl ihrer Antworten zeigt, haben wir nur die ersten 30 Autoren verwendet. Die Anzahl der Antworten der übrigen Autoren nimmt stetig ab und konvergiert gegen 1.

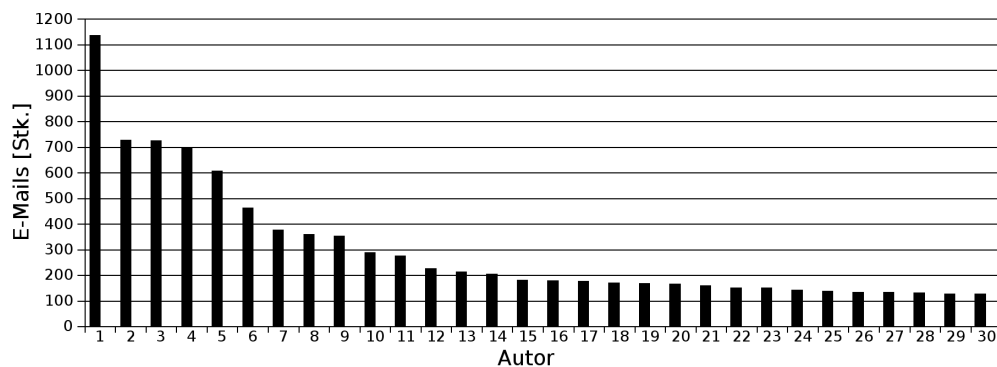


Abbildung 2: Histogramm von Antworten nach Autoren

⁴ Das Token „Re: “ bedeutet „Antwort auf “

Nach Auszählung der E-Mail-Frequenzen, haben wir die Antworten der einzelnen Autoren pro Zeitperiode kumuliert und dargestellt. Dazu haben wir uns auf die Top-10 Autoren beschränkt, da die Anzahl verfasster E-Mail-Antworten nach etwa den 10 Autoren mit den höchsten Frequenzen stark abfällt (vgl. Abbildung 2).

Unsere Analyse zeigt, dass die Anzahl der Antworten der Top-10 Autoren zu Beginn des Verteilers im Vergleich zum Ende der betrachteten Periode sehr viel höher ist und seinen Höhepunkt im August 1999 findet (vgl. Abbildung 3). Danach sinkt die Anzahl der Antworten markant und kommt erst 3 Jahre später wieder durch zwei der Top-10-Autoren auf ein vergleichbares Niveau.

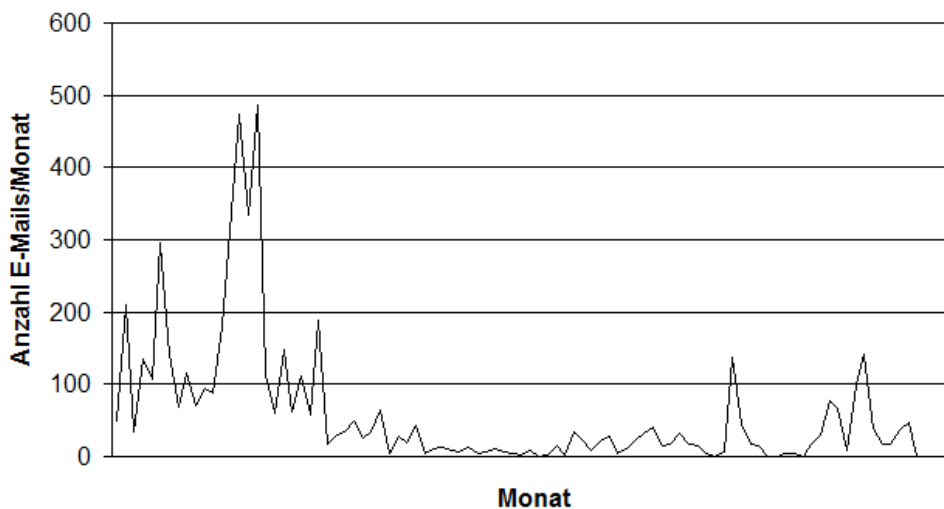


Abbildung 3: Verteilung der E-Mails je Monat vom April 1998 - November 2005

Abbildung 4 schliesslich zeigt die relative Verteilung der Antworten der Autoren pro Quartal. Die Anfangszeit präsentiert sich mit vielen aktiven und relativ gleichstarken⁵ Autoren. In den ersten 2 Jahren zeigt sich auch, dass die Community aus den domänenspezifischen Problemen anderer und deren Lösung durch die Beteiligung an der Mailingliste gelernt haben (Ackermann & McDonald 1996), was sich in der allmählichen Gleichverteilung der Antworten auf die Autoren ausdrückt. In Abbildung 4 wird vor allem die Dominanz von Autoren in einem bestimmten Quartal deutlich. So zum Beispiel dominiert Autor 1 die

⁵ Mit gleichstark ist hier die gleiche Anzahl an Antwortmails gemeint.

Anfangszeit der Mailingliste und wird dann von anderen Autoren abgelöst bis nur noch 2 Autoren die Antworten beherrschen, wobei Autor 4 der dominantere ist.

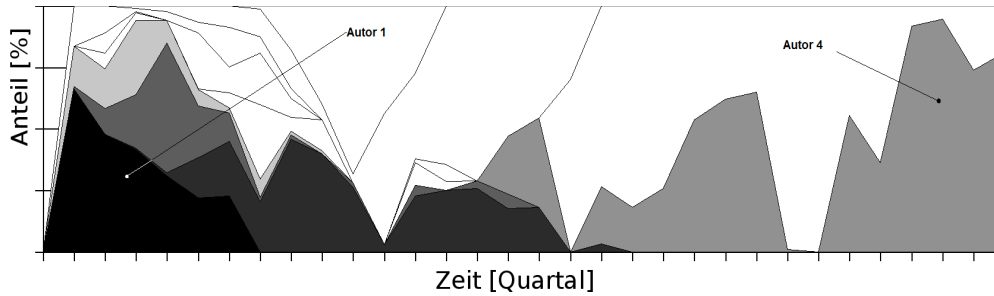


Abbildung 4: relative Verteilung der Antworten über die Zeit

6 Expertenbestimmung

Aufgrund der relativen Verteilung unter den Top-10 Autoren (vgl. Abbildung 4) kann bereits der Autor 1 als ein Experte der ersten Stunde und der Autor 4 als ein aktueller Experte für das ausgewählte Thema identifiziert werden.

Gemäss der Kriterien unserer Experten Definition erkennen wir überdurchschnittliche Leistung bei Autor 1 am Anfang und aktuell bei Autor 4. Beide Autoren haben im Durchschnitt über einen Zeitraum von 4 Quartale mehr Antworten als den Median plus Standardabweichung aus der Verteilung der E-Mail-Antworten verfasst. Dies ist konstant in den Quartalen 1-4 für Autor 1 der Fall und für die letzten 4 Quartale für Autor 4 der Fall.

Bei unserer Überprüfung der vermeintlichen Experten stellt sich heraus, dass der Modul-Owner nicht zu den Experten gehört. Allerdings steht Autor 4 dem Modul-Owner als Peer⁶ zur Seite. Mit Ausnahme von Autor 5 gehören alle vermeintlichen Experten dem engeren Kreis der Entwickler von Mozilla-Modulen an und haben dabei sogar eine tragende Rolle als Peer.

⁶ Ein Peer ist eine Person, den der Modul-Owner zur Unterstützung seines Modul vorgesehen hat.

7 Schlussfolgerungen

Wir haben gezeigt, dass sich Personen über die Frequenz von verfassten Antwortmails von der Masse abheben können. In unserem Fall hatte eine Person am Anfang die tragende Rolle und überliess nach einer gewissen Zeit anderen das Feld.

Somit konnten wir auch einen Expertenwechsel feststellen. Dies ist eine wichtige Erkenntnis, gerade für die Anwendung als Messkriterium, z. B. für erfolgreichen Wissenstransfer, da das Mass zeitnahe und adäquat ist. Des Weiteren können wir einen Beitrag leisten, indem wir feststellen, das E-Mails zur Expertenbestimmung, trotz der verschiedenen Einschränkungen, theoretisch möglich ist, und vor allem besser als die organisatorische Expertenbestimmung. Denn im vorliegenden Fall waren die „Peers“ und nicht der „Module Owner“ die Experten.

Da E-Mails und vermehrt auch Text-Nachrichten (z. B. via Skype) die Kommunikationsflüsse von Mitarbeitern in Unternehmen abbilden, ist eine Analyse dieser Daten eine wichtige Quelle für die Expertenermittlung. Wie wir zeigen konnten, ist dieses Vorgehen auch durchaus geeignet Experten zu identifizieren.

Für Systeme, in denen ausreichend kategorisierte Daten vorliegen, die mit Personen verknüpfbar sind, können wir die Experten in diesen Kategorien feststellen. Unser Beitrag stellt jedoch kein Verfahren zum aufbauen der Kategorien dar, vielmehr wurde die Clusterung als Hilfsmittel verwendet, da keine Kategorisierung des Datensatzes vorlag.

Unser Beitrag wird helfen existierende wissensvermittelnde Systeme effektiver zu gestalten, da er z. B. eine Überprüfung des Lernerfolges ohne explizite Tests und Übungen ermöglicht.

Da wir die Qualität der Antworten nicht ermittelt haben, müssen wir die Frage, ob sich die Expertenauswahl ändern würde wenn die Teilnehmer des E-Mail-Verteilers befragt würden, offen lassen. Eine entsprechende Studie wäre eine logische Ergänzung der vorliegenden Arbeit.

8 Literaturverzeichnis

Ackerman, M. S.; McDonald, D. W. McDonald (1996): Answer Garden 2: Merging Organizational Memory with Collaborative Help. Proceedings of CSCW '96, 97-105

Belkin, N. J.; Oddy, R. N.; Brooks, H. M. (1982): ASK for information retrieval: Part I.. In: Journal of Documentation 38(2), 61-71

- Crocker, D. H. (1982): Standard for the format of arpa internet text messages. RFC 822.
- Fahrmeir, L.; Künstler, R.; Pigeot, I.; Tutz, G. (1999): Statistik – der Weg zur Datenanalyse. 4. verbesserte Auflage, Springer.
- Lindvall, M; Rus, I.; Sinha, S. S. (2002): Technology Support for Knowledge Management. In: Lecture Notes in Computer Science, 2640 (2002), 94-103.
- McCallum, A. K. (1996): Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. [Abrufbar über: <http://www.cs.cmu.edu/~mccallum/bow/>], [Zugriff:11.12.2005]
- McDonald, D. W.; Ackermann, M. S. (1998): Just Talk to Me: A Field Study of Expertise Location. In: Proceedings of ACM Conference on Computer Supported Cooperative Work (1998), 315-324
- McDonald, D. W.; Ackermann, M. S. (2000): Expertise Recommender: A Flexible Recommendation System and Architecture. Computer-Supported Cooperative Work, ACM Press
- Moreale, E.; Watt, S. (2003): An Agent-Based Approach to Mailing List Knowledge Management. In: Agent Mediated Knowledge Management, 2926 (2003), 118-129
- Osman, G.; Norshuhada, S (2004): Expert-seeking Activity Framework. In: Journal of Advancing Information and Management Studies, 1-1 (2004), 63-73
- Posner, M. I. (1988): What is it to be an expert?. In: M.L.T Chi, R. Glaser & M.J. Farr (Hrsg.), The nature of expertise (pp. XXIX-XXXVI).
- Weinberger, D. (1999): Tacit Knowledge, KMWorld, 22nd.
- Watanabe, Y.; Sono, K.; Yokomizo, K.; Okada Y. (2004): A question answer system using mails posted to a mailing list. ACM Press.
- Willis, G.; Alani, H.; Ashri, R.; Crowder, R.; Kalfoglou, Y.; Kim, S. (2002): Design Issues for Agent-based Resource Locator Systems. Lecture Notes In Computer Science, 2569 (2002), 156-167.
- Yang, Y.; Liu, Y. (1999): A re-examination of text categorization methods. Proceedings ACM SIGIR (1999).