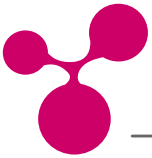
Technische Universität Dresden – Fakultät Informatik
Professur für Multimediatechnik, Privat-Dozentur für Angewandte Informatik

Prof. Dr.-Ing. Klaus Meißner
PD Dr.-Ing. habil. Martin Engelien
(Hrsg.)



GeNeMe '08

Gemeinschaften in Neuen Medien

an der
Fakultät Informatik der Technischen Universität Dresden

mit Unterstützung der

GI-Regionalgruppe Dresden
Initiative D21 e.V.
Kontext E GmbH, Dresden
Medienzentrum der TU Dresden
SALT Solutions GmbH, Dresden
SAP Research CEC Dresden
Saxonia Systems AG, Dresden
T-Systems Multimedia Solutions GmbH
3m5. Media GmbH, Dresden

am 01. und 02. Oktober 2008 in Dresden
http://www-mmt.inf.tu-dresden.de/geneme/
geneme@mail-mmt.inf.tu-dresden.de

D Arbeiten in virtuellen Unternehmen

D.1 Ein empirischer Zugang zur Ermittlung von Kompetenzprofilen in der Digitalen Wirtschaft

Sabrina Ziebarth, Nils Malzahn, Sam Zeini, Ulrich Hoppe
Universität Duisburg-Essen, Abteilung Informatik und Angewandte Kognitionswissenschaft

1 Einleitung

Das BMBF-Forschungs- und Entwicklungsprogram „Arbeiten - Lernen - Kompetenzen entwickeln. Innovationsfähigkeit in einer modernen Arbeitswelt“[1] stellt den Zusammenhang zwischen Arbeitsgestaltung, Kompetenzentwicklung und Innovationsfähigkeit in den Vordergrund. Insbesondere im High-Tech-Sektor, wie z. B. in der digitalen Wirtschaft, stellt der zunehmende Innovationsdruck eine nicht zu unterschätzende Herausforderung für die Kompetenzentwicklung bei den Beschäftigten dar. Mit einem Blick auf diese Beschäftigten in der digitalen Wirtschaft und insbesondere der Quereinsteiger, die vor dem Hintergrund das Fachkräftemangels eine zunehmend bedeutsame Rolle spielen, entwickelt das Forschungsprojekt KoPIWA - Kompetenzentwicklung und Prozessunterstützung in „Open Innovation“-Netzwerken der IT-Branche durch Wissensmodellierung und Analyse - (Förderkennziffer 01FM07067-72) ein ganzheitliches Konzept als Lösung auf die sich hieraus ergebende Frage nach einem softwaregestützten Kompetenzmanagement für innovationsgetriebene Arbeit in der digitalen Wirtschaft.

Damit verbunden ist ein Teilziel des Projektes KoPIWA[2], die Identifikation wichtiger Kompetenzen der IT-Branche wie auch ihrer Zusammenhänge untereinander sowie im Kontext mit Kompetenz-Profilen für bestimmte IT-Berufe. Kompetenzbasiertes Personalmanagement ist ein wichtiges Werkzeug für Personal- und Sukzessionsplanung sowie Personalentwicklung [Draganidis06]. Während etwa das HR-XML-Konsortium[3] “Kompetenz” als *„a specific, identifiable, definable, and measureable knowledge, skill, ability and/or other deployment-related characteristic (e.g. attitude, behavior, physical ability) which a human resource may possess and which is necessary for, or material to, the performance of an activity within a specific business context”* definiert, betrachten wir im Folgenden Kompetenzen lediglich als die zentralen auf die Eigenschaften von potentiellen Stelleninhabern bezogenen Begriffe, die zur

1 http://www.bmbf.de/pub/innovationsfaehigkeit_arbeitswelt.pdf (Zugriff am 13.05.2008)

2 Dieses Teilziel wird im Rahmen des KoPIWA-Teilvorhabens „Kompetenzmodellierung und Dynamisierung von Kompetenzprofilen (Förderkennzeichen 01FM07067) bearbeitet.

3 http://ns.hr-xml.org/2_5/HR-XML-2_5/CPO/Competencies.html (Zugriff am 13.05.2008)

Definition von Anforderungen in Stellenanzeigen verwandt werden. Hierfür sind keine Annahmen über Operationalisierung oder Messbarkeit erforderlich.
Die Ergebnisse sollen als Vorstufe für die Erstellung einer Kompetenz-Ontologie für die Digitale Wirtschaft dienen, welche unter anderem zum Kompetenz-basierten Personalmanagement, als Informationsquelle für Schulabgänger und Berufswechsler sowie zur Karierreberatung und –planung genutzt werden soll. Zur Exploration geeigneter Verfahren wurden dazu zunächst circa 3000 digitale Stellenanzeigen aus dem Bereich der IT-Branche erhoben und mit Methoden aus den Bereichen Data Mining und Information Retrieval analysiert. Data Mining sei hier verstanden als Teilschritt des Wissensentdeckungsprozesses in Datenbanken, dem *„nichttriviale[n] Prozeß der Identifikation gültiger, neuer, potentiell nützlicher und schlußendlich verständlicher Muster in (großen) Datenbeständen"* [Görz03]. Data Mining-Ansätze werden heutzutage zur Analyse großer Datenmengen zum Beispiel aus Unternehmen, Forschungsprojekten oder dem Internet eingesetzt. Zur Analyse der erhobenen Stellenanzeigen wurden die Data Mining-Verfahren Clustering und Assoziationsregel-Erkennung sowie das Information Retrieval-Verfahren Latent Semantic Indexing eingesetzt. Clustering-Verfahren ermöglichen die Gruppierung von Stellenprofilen hinsichtlich der in ihnen geforderten Kompetenzen. Durch die gemeinsame Nennung von Kompetenzen in Stellenanzeigen entstehen zwischen diesen Beziehungen, welche durch Assoziationsregel-Verfahren identifiziert werden können. Doch nicht nur diese „offensichtlichen" Beziehungen sind von Interesse, sondern auch solche, welche durch zufällige Wortwahl (Synonymie, Polysemie) verschleiert werden. Diese können mit der Methode des Latent Semantic Indexing bestimmt werden.

2 Auswahl und Vorverarbeitung der Daten

Stellenanzeigen spiegeln den aktuellen Personal-Bedarf der Wirtschaft wider. Sie beschreiben welche Kompetenzen Personalabteilungen bestimmten Berufen zuordnen und sind somit ein guter Indikator dafür, was von diesen Berufen erwartet wird. Es gibt eine große Anzahl von Online-Stellenbörsen, wie zum Beispiel Monster[4], StepStone[5] oder Jobscout24[6], welche Stellenanzeigen in digitaler Form zur Verfügung stellen. Digitale Stellenanzeigen können automatisiert verarbeitet werden, was einen großen Vorteil für große Datensätze darstellt.
Für diese Arbeit wurden zwei Datensätze durch das Auswerten von Stellenanzeigen der Online-Stellenbörse Monster erhoben. Die betrachten Anzeigen stammen alle aus dem IT-Umfeld und wurden im März 2008 erhoben. Der erste, kleinere Datensatz betrachtet 67 Stellenanzeigen aus vier NRW-Städten (Duisburg, Essen, Oberhausen

4 http://www.monster.de (Zugriff am 13.05.2008)
5 http://www.stepstone.de (Zugriff am 13.05.2008)
6 http://www.jobscout24.de (Zugriff am 13.05.2008)

und Dinslaken). Kleine Datensätze vereinfachen die Analyse durch geringeren Zeitaufwand für die Algorithmen, und die Interpretation der Ergebnisse dadurch, dass sie durch manuelle Inspektion der Stellenanzeigen besser nachvollzogen und validiert werden können. Allerdings können die Ergebnisse aus kleinen Datensätzen nicht ohne weiteres verallgemeinert werden, so dass sie eher der Plausibiliätsprüfung des gewählten Ansatzes dienen. Daher wurde anschließend ein zweiter Datensatz bestehend aus 2981 Stellenanzeigen der IT-Branche aus ganz NRW erhoben.

Unter Nutzung von Rapid Miner[7] [Mierswa06] – einem Open-Source-Tool für Wissensentdeckung und Data Mining – zusammen mit dem sog. Text-Plugin[8] wurden die Stellenanzeigen des kleinen Datensatzes in Wort-Vektoren zerlegt, welche wiederum zu einem alle Terme enthaltenen Vektor aggregiert wurden. Nach Entfernen von Stoppwörten wie Artikeln, Konjunktionen und Präpositionen, so wie aller Wörter mit nur einem Zeichen, enthielt der Wort-Vektor im Hinblick auf Kompetenzen noch viele uninteressante Terme, wie zum Beispiel Firmen- oder Ortsnamen. Diese wurde mit Hilfe der Web Services des Projekts Deutscher Wortschatz[9] eliminiert. Die übrigen Terme wurden durch den deutschen Stemming-Algorithmus auf ihre Grundform abgebildet. Um interessante Wortkombinationen zu entdecken, wurden Bi-Gramme, also Kombinationen von zwei Termen hinzugefügt. Die Terme und Term-Kombinationen wurden mittels des sog. TF-IDF Verfahrens gewichtet und entsprechend ihrer Gewichtung sortiert.

TF-IDF steht für *„Term Frequency – Inverse Document Frequency"* und ist definiert als

$$\textit{Frequenz für Term i in allen Dokumenten} * \log \frac{\textit{Anzahl der Dokumente}}{\textit{Anzahl der Dokumente, die Term i enthalten}}$$

Das Verfahren bevorzugt häufig vorkommende Terme, die nur in ausgewählten Dokumenten vorkommen gegenüber solchen Termen, die nur selten oder häufig aber in relativ vielen Dokumenten vorkommen. Aus diesem Grund wurden Terme mit einem sehr niedrigen TF-IDF Gewicht entfernt. Um die Qualität der Terme zu garantieren, wurde der Wort-Vektor schlussendlich manuell gefiltert. Nach der Vorverarbeitung enthielt der Wort-Vektor noch 201 Terme (siehe Tabelle 1).

7 http://rapid-i.com (Zugriff am 13.05.2008)

8 http://nemoz.org/joomla/index.php?option=com_content&task=view&id=43&Itemid=83 (Zugriff am 13.05.2008)

9 http://wortschatz.uni-leipzig.de/ (Zugriff am 13.05.2008)

sap	1
it	0.562
softwar	0.479
design	0.419
entwickl	0.347
crm	0.333
servic	0.324
berat	0.301
system	0.298
…	

Tabelle 1: Auszug aus dem gewichteten (normalisierten) Wort-Vektor.

Jede Stellenanzeige aus den beiden Datensätzen wurde hinsichtlich der 201 Terme als TF-IDF gewichteter Wort-Vektor dargestellt.

3 Identifikation von Profilen durch Clustering von Stellenanzeigen

Cluster-Techniken werden im Data Mining dafür eingesetzt, die Datensätze in natürliche Gruppen ähnlicher Beispiele zu zerlegen [Witten05]. Das Clustering der Stellenanzeigen sollte zeigen, ob es in den Stellenanzeigen Gruppen von Berufen gibt, die sich durch die für sie geforderten Kompetenzen ähneln und welche Kompetenzen für die Zuordnung zu einer der Gruppen entscheidend sind. Im Folgenden werden die Ergebnisse des Clusterings diskutiert.

K-Means ist ein klassisches Clustering-Verfahren, welches iterativ die gegebenen Beispiele entsprechend ihrer Distanz zu den Clusterkernen in k Cluster aufteilt [mehr siehe Witten05]. Die Ergebnisse der Cluster-Bildung des kleinen Datensatzes mit Hilfe des k-Means Algorithmus (mit k=5) sind in Tabelle 2 dargestellt.

Cluster 0	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
14 Anzeigen	11 Anzeigen	18 Anzeigen	17 Anzeigen	7 Anzeigen
sap: 1.0 erfahr: 0.71 berufserfahr: 0.64 projekt: 0.64 international: 0.57 it: 0.57 system: 0.57 onlin: 0.571 englisch-kenntniss: 0.5 berat: 0.5 studium: 0.5 abgeschloss: 0.5 word: 0.43 wort: 0.43 kund: 0.43 bw: 0.43 abgeschloss_ studium: 0.43 ph: 0.36 cod: 0.36 eee: 0.36 …	oracl: 0.64 erfahr: 0.64 softwar: 0.64 informat: 0.64 wirtschafts-informat: 0.55 it: 0.55 kund: 0.55 studium: 0.55 java: 0.45 engineering: 0.45 windows: 0.45 sql: 0.45 team: 0.45 software-entwickl: 0.45 entwickl: 0.45 design: 0.36 verantwort: 0.36 international: 0.36 englisch-kenntniss: 0.36 servic: 0.36 …	it: 0.10 entwickeln: 0.09 entwickl: 0.08 security: 0.07 fachinformat: 0.07 ausbild-fachinformat: 0.07 anwendungs-entwickl: 0.07 ausbild: 0.06 berat: 0.06 international: 0.06 netzwerk: 0.06 lern: 0.06 support: 0.06 anwend: 0.06 serv: 0.06 gut: 0.06 selbstand: 0.05 projektleit: 0.05 team: 0.05 englischkenntniss-wort: 0.05 …	offic: 0.71 kund: 0.71 technisch: 0.65 it: 0.59 system: 0.59 ms: 0.59 erfahr: 0.59 internet: 0.59 servic: 0.53 ms_offic: 0.53 produkt: 0.53 fahig: 0.53 anwend: 0.47 team: 0.47 dienstleist: 0.41 person: 0.41 berufserfahr: 0.41 professionell: 0.41 web: 0.41 onlin: 0.41 …	kreation: 1.0 mediengestalt_ kreation: 1.0 mediengestalt: 1.0 flash_ mediengestalt: 1.0 flash: 1.0 design_flash: 1.0 grafik_design: 1.0 grafik: 1.0 gestalt_grafik: 1.0 interfac_gestalt: 1.0 interfac: 1.0 design_interfac: 1.0 design_screen: 1.0 training_screen: 1.0 branding_training: 1.0 branding: 1.0 marketing_ branding: 1.0 media_marketing: 1.0 digital_medi: 1.0 kreativ_digital: 1.0 …
SAP-Berater	Anwendungs-Entwickler	IT-Fach-informatiker	Datenpflege, Customer Relationship	Medien-Designer

Tabelle 2:
Ergebnisse der Clusterung mit k-Means. Die Zahlen geben das durchschnittliche Vorkommen der Terme im Cluster an.

Cluster 0 fasst Berufe aus dem SAP-Berater Umfeld zusammen, für die vor allem Erfahrung mit SAP und (internationalen) Projekten so wie gute Englischkenntnisse, ein akademischer Abschluss und Office-Kenntnisse gefordert werden. In Cluster 1 befinden sich Berufe aus dem Bereich Datenbank- und Anwendungsentwicklung, für die besonders Fachkompetenzen wie Erfahrung mit Oracle und SQL im Bereich Datenbanken und Erfahrung mit Software-Engineering, Java und Design im Bereich Anwendungsentwicklung gefordert werden. Dazu kommen gute Teamfähigkeit, gute Englischkenntnisse und ein abgeschlossenes Studium im Bereich Informatik oder Wirtschaftsinformatik. Im Gegensatz zu den vorangegangenen zwei Clustern, bündelt Cluster 2 vor allem Stellenanzeigen, die eine abgeschlossene Ausbildung im Bereich der IT fordern und kein abgeschlossenes Studium. Bewerber sollten hier Erfahrung mit Anwendungsentwicklung, Support und System-Administration haben. Die Stellenanzeigen in Cluster 3 haben ihren Kompetenz-Schwerpunkt im Bereich Office-Anwendungen, Umgang mit Kunden und der Nutzung des Internets. Cluster 3 beinhaltet somit Berufe im Bereich Datenpflege und Customer Relationship unter Nutzung von SAP, sowie Projekt-Management. Cluster 4 bündelt schlussendlich die kreativen Berufe im Bereich Digitales Medien-Design, für die vor allem Erfahrung im Bereich Medien-Design (z.B. Grafiken, Flash) und Interface-Design gefordert sind.

Die Ergebnisse verdeutlichen, dass der gewählte Ansatz zu tragfähigen Ergebnissen zu führen scheint. Daher wurde das Verfahren auf eine größere Menge von Stellenanzeigen ausgeweitet.

Da das Clustering von großen Datenmengen mit höherem Rechenaufwand verbunden ist werden nur Auszüge des gesamten Datensatzes, sog. Sample geclustert. Für diese Arbeit wurden verschiedene Sample mit je 200 Stellenanzeigen mit verschiedenen Algorithmen (k-Means, k-Medoids, FarthestFirst [Hochbaum85]) und verschiedenen Parametern (k=5 und k=6) geclustert. Es entstandeni 310 Cluster. Um aus diesen die „besten“ Cluster zu bestimmen, wurden die 310 Cluster als neuer Eingabe-Datensatz betrachtet und erneut geclustert. Zu diesem Zweck wurden die Zentroide der jeweiligen Cluster aus den Durchschnittswerten der TF-IDF-Gewichte der Terme im Cluster gebildet. Das Clustering der Zentroiden erfolgte anschließend mit dem X-Means-Algorithmus, welcher einen erweiterten k-Means-Algorithmus darstellt, der unter anderem die beste Anzahl von Clustern in einem gegebenen Bereich bestimmen kann [Pelleg00].

Cluster 0	**Cluster 1**	**Cluster 2**	**Cluster 3**
50 Anzeigen	82 Anzeigen	42 Anzeigen	136 Anzeigen
senior : 0.29 it : 0.17 management : 0.17 security : 0.14 berat : 0.12 consultant : 0.08 betrieb : 0.07 projekt : 0.07 tatigkeitsfeld : 0.06 servic : 0.06 kund : 0.06 prozess : 0.06 system : 0.05 datenbank : 0.05 entwickl : 0.05 selbststand : 0.05 business : 0.05 mehrjahr_ berufserfahr : 0.05 berufserfahr : 0.04 fachlich : 0.04	kaufmann : 0.12 it : 0.1 servic : 0.1 dienstleist : 0.08 support : 0.07 administration : 0.07 belastbar : 0.07 abgeschloss-ausbild: 0.07 offic : 0.07 technisch : 0.07 personalvermittl : 0.07 installation : 0.06 teamfah : 0.06 personal : 0.06 projektarbeit : 0.06 hardwar : 0.06 ausbild : 0.05 berufserfahr : 0.05 serv : 0.05 windows : 0.05	sap : 0.55 sd : 0.12 logist : 0.09 consultant : 0.08 manag : 0.07 berat : 0.07 international : 0.06 management : 0.06 projekt : 0.05 system : 0.05 bw : 0.05 kaufmann : 0.05 fachlich : 0.05 engagement : 0.05 dienstleist : 0.04 kund : 0.04 kontinui : 0.04 it : 0.04 weiterentwickl : 0.04 sich : 0.04	entwickl : 0.12 java : 0.1 softwar : 0.1 softwareentwickl : 0.08 engineering : 0.07 business : 0.06 reporting : 0.06 it : 0.05 technologi : 0.05 management : 0.05 informat : 0.05 design : 0.04 onlin : 0.04 kund : 0.04 oracl : 0.04 team : 0.04 serv : 0.04 anwendungsentwickl : 0.04 unix : 0.04 projektleit : 0.04
IT-Manager	IT-Kaufmann	SAP-Berater	Entwickler

Tabelle 3:
Ergebnisse der Cluster-Clusterung mit X-Means. Die Zahlen geben den durchschnittlichen TF-IDF-Wert der Terme im Cluster an.

Das Clustern von Clustern ist ein verbreiteter Ansatz für Mustererkennungs-Probleme [Chan92]. Durch das Clustern der Ergebnisse verschiedener Clustering-Algorithmen kann die Robustheit und Qualität des endgültigen Clusterings signifikant verbessert werden [Gionis07]. Die Ergebnisse sind in Tabelle 3 dargestellt.
Das Ergebnis zeigt vier große Cluster: Wie beim Clustering des kleinen Datensatzes ergeben sich ein SAP-Berater-Cluster (Cluster 2) und ein Entwickler-Cluster (Cluster 3). Stellenanzeigen im SAP-Berater-Cluster sind durch Anforderungen im Bereich SAP-Beratung und Management, so wie Erfahrung mit (internationalen) Projekten geprägt, während Stellenanzeigen im Entwickler-Cluster Kompetenzen wie Java,

Engineering, Design und Datenbanken fordern. Das Mediengestalter-Cluster aus dem kleinen Datensatz bildet sich nicht aus. Die zugehörigen Stellenanzeigen werden auf die anderen Cluster verteilt. Dies könnte an der verhältnismäßig geringeren Anzahl von Stellenanzeigen in diesem Bereich liegen. Andererseits zeichnen sich zwei neue Cluster ab: Cluster 0 bündelt Kompetenzen für IT-Manager und -Berater, wie Erfahrung mit Projekten, Kunden und Prozessen, so wie spezielle Kenntnisse im Bereich von Entwicklung, Systemen und Datenbanken. In Cluster 1 zeigen sich vor allen Kompetenzen im kaufmännischen und im Support-Bereich.
Die Ergebnisse des Clusterings zeigen, dass sich auf Grund der geforderten Kompetenzen in den Stellenanzeigen Berufs-Gruppen bilden lassen. Allerdings müssen die Namen von Berufsgruppen noch manuell zu den Clustern zugeordnet werden. In zukünftigen Arbeiten werden wir versuchen die Benennung der Cluster ebenfalls aus den Stellenanzeigen heraus zu vergeben. Einen ersten Ansatz stellt die Auswertung der Titel der zu den Clustern gehörigen Stellenanzeigen dar.

4 Ermittlung von Kompetenz-Zusammenhängen in Stellenanzeigen mit Hilfe von Assoziationsregel-Verfahren

Algorithmen für Assoziationsregeln, wie zum Beispiel der Apriori-Algorithmus, können dazu verwendet werden Terme zu identifizieren, die häufig in verschiedenen Stellenanzeigen gemeinsam auftauchen. Daraus werden Assoziationsregeln abgeleitet, die bestimmte Schwellwerte für die Häufigkeit (Support) und Richtigkeit (Confidence) des Vorkommens erfüllen [Witten05]. Bei der Analyse von Stellenanzeigen sollen sie helfen Zusammenhänge zwischen Kompetenzen hinsichtlich ihrer gemeinsamen Nennung zu identifizieren.
Bei Anwendung des Apriori-Algorithmus auf den großen Datensatz wurden zunächst alle Kompetenzen betrachtet. Es zeigten sich starke Zusammenhängen zwischen zwei Gruppen von Termen; die erste Gruppen bestand aus den Termen Kaufmann, Personalvermittlung und Rechnungswesen und die zweite aus den Termen SAP BW, BI, Business Intelligence, Netweaver und Studium. Es zeigten sich also Zusammenhänge in der Nennung von fachlichen Kompetenzen.
In einem zweiten Versuch wurden nur nicht-fachliche Kompetenzen wie soziale und individuelle Kompetenzen betrachtet. Die Ergebnisse sind in Tabelle 4 dargestellt. Die Assoziationsregeln mit dem meisten Support ergaben sich zirkulär aus den Kompetenzen Belastbarkeit, Dynamik und Verantwortungsbewusstsein. Sie spiegeln somit den Aufbau einer typischen Stellenanzeige wider. Verantwortungsbewusstsein, Belastbarkeit und Flexibilität sind offenbar als Tripel gefordert. Der hohe Anteil an allen Stellenanzeigen zeigt zudem, dass diese Eigenschaften für nahezu alle Stellen wichtig sind. Es legt sogar den Schluss nahe, dass diese Anforderungen eigentlich eine Stellenanzeigenfloskel darstellen und daher nicht zur Spezifikation von Stellenanzeigen und damit zur Findung von passenden Bewerbern geeignet sind.

Bedingung	Folgerung	Support	Confidence
belastbar	dynamisch	0,733	0,830
dynamisch	belastbar	0,733	0,875
verantwort	belastbar	0,662	0,880
verantwort	dynamisch	0,647	0,860
belastbar, verantwort	dynamisch	0,564	0,851
dynamisch, verantwort	belastbar	0,564	0,871

Tabelle 4: Auszug der Assoziationsregeln für nicht fachliche Kompetenzen.

Schließlich kann mit Hilfe des Tertius-Algorithmus [Flach99] gezielt nach Termen gesucht werden, die nicht zusammen auftreten, um potenzielle Berufsprofile voneinander abzugrenzen. Der Tertius-Algorithmus sucht wie der Apriori-Algorithmus nach Regeln mit mehreren Bedingungen, unterscheidet sich aber darin, dass diese Regeln mit ODER und nicht mit UND verknüpft sind. Ein Auszug der Ergebnisse ist in Tabelle 5 dargestellt.

engineering = true ==> kaufmann = false
kaufmann = true ==> kreativ = false
projektleit = true ==> administration = false
telekommunikation = true ==> kreativ = false
netzwerk = true ==> betriebswirtschaft = false
netzwerk = true ==> kreativ = false
schulung = true ==> web = false
studium_informat = true ==> hochschulstudium = false
linux = true ==> projektleit = false
personalvermittl = true ==> engagement = false
senior = true ==> installation = false
einsatzbereitschaft = true ==> kreativ = false

Tabelle 5: Auszug aus den Assoziationsregeln mit Tertius.

Die Ergebnisse sind insofern interessant, als dass sie zeigen, welche Terme nicht zusammen in den Stellenanzeigen genannt werden. Bemerkenswert sind solche Regeln, die sich nicht einfach erklären lassen, wie zum Beispiel, dass die Terme „Personalvermittlung“ und „engagiert“ oder „kreativ“ und „Anwendungsentwicklung“ nicht zusammen genannt werden. Sie zeigen auf, welche Muster es in den Köpfen der Stellenausschreiber offenbar nicht gibt.

5 Bestimmung verborgener Strukturen in Stellenanzeigen durch Latent Semantic Indexing (LSI)

Latent Semantic Indexing (LSI) ist ein statistisches Verfahren aus dem Umfeld des Information Retrievals, welches unter anderem dazu entwickelt wurde um die Probleme der Synonymie (es gibt verschiedene Worte, die das gleich Objekt beschreiben) und der Polysemie (ein Wort kann verschiedene Bedeutungen haben) zu lösen [Deerwester90]. Das Verfahren basiert auf der Annahme, dass es eine latente Struktur in den Daten gibt, die durch die zufällige Wahl der Worte verschleiert wird. Unter Einsatz von Singulärwert-Zerlegung wird bei LSI der semantische Raum so angeordnet, dass wichtige Muster hervorgehoben und unwichtigere nicht beachtet werden.

Mit Hilfe des LSI-Verfahrens wurden in dem kleinen Datensatz Beziehungen zwischen Kompetenzen, zwischen Stellenanzeigen und zwischen Kompetenzen und Stellenanzeigen identifiziert. Diese wurden dann mit der online-Version des CoNaVi-Tools [Malzahnetal05] visualisiert (siehe Abbildung 1 und 2). Dieses erlaubt das graphische Browsen der Ergebnisse und ein gezieltes Suchen. Es kann nach Kompetenzen, die für ein bestimmtes Stellenprofil interessant sind, nach Kompetenzen, die häufig mit einer bestimmten Kompetenz genannt werden und nach Stellenprofilen, die auf Grund der in ihnen geforderten Kompetenzen ähnlich zu einem bestimmen Stellenprofil sind gesucht werden. Das System steht unter http://ziebarth.collide.info/kopiwa/ online zur Verfügung.

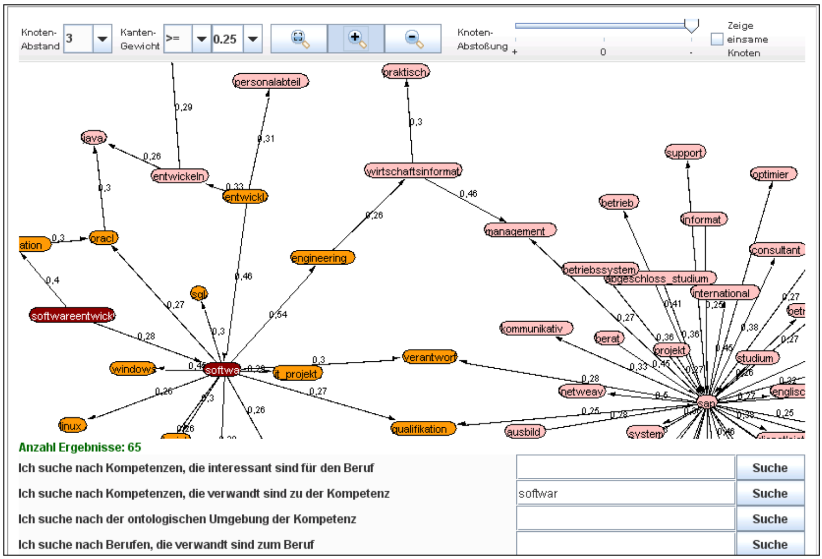


Abbildung 1: Visualisierung von Ausschnitten des Kompetenz-Kompetenz-Netzes basierend auf den LSI-Ergebnissen.

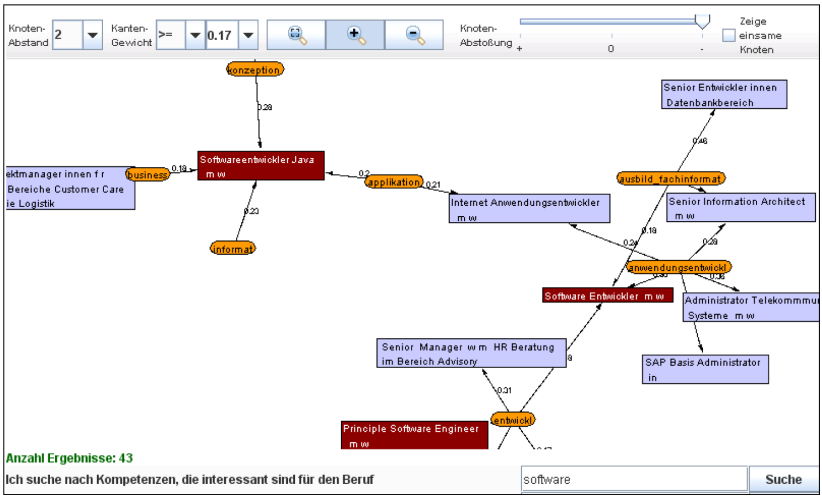


Abbildung 2: Visualisierung von Ausschnitten der Stellenprofil-Kompetenz-Beziehungen basierend auf den LSI-Ergebnissen.

Es zeigte sich, dass mit Hilfe des LSI-Verfahrens Beziehungen identifiziert werden können. Zum Beispiel werden Beziehungen zwischen „Java“ und den Termen „Softwarelösungen“, „Security“ und „Teamfähigkeit“ identifiziert. Das Gesamtsystem ermöglicht einen guten ersten Überblick über die Kompetenzen, die für bestimmte Profile gefordert werden (siehe Abbildung 2) und darüber welche Kompetenzen häufig zusammen genannt werden (siehe Abbildung 1) sowie für welche Stellenprofile ähnliche Kompetenzen gefordert werden.

6 Fazit und Ausblick

In einer ersten heuristischen Phase des KoPIWA-Projekts (BMBF, Förderkennziffer: 1FM07067-72) wurden circa 3000 Stellenanzeigen erhoben und mit verschiedenen Verfahren des Data Minings (Clustering, Assoziationsregeln) und Information Retrievals (Latent Semantic Indexing) analysiert mit dem Ziel Berufsprofile aus den Stellenanzeigen zu ermitteln. Die automatische Auswertung der Stellenanzeigen ermöglicht eine zeitnahe Eingrenzung der momentan am Markt geforderten Berufsprofile. Diese Ergebnisse können dann in die Weiterbildung einbezogen werden. Das Clustern der Stellenanzeigen entsprechend der in ihnen enthaltenen Kompetenzen lässt die Identifikation von Berufsprofilen zu. Durch Assoziationsregel-Verfahren und Latent Semantic Indexing konnten Beziehungen von Kompetenzen auf Grund von gemeinsamer Nennung in Stellenanzeigen identifiziert werden. Eine Analyse des Nicht-Ko-Auftretens von Termen zeigte einige interessante Resultate bezüglich Kompetenzen, die in Stellenanzeigen nicht zusammen genannt werden. Die Ergebnisse des LSI-Verfahrens (Beziehungen zwischen Kompetenzen, Stellenanzeigen sowie Kompetenzen und Stellenanzeigen) wurden als Graphen visualisiert und mit einem System verbunden, welches zum Beispiel Anfragen nach Kompetenzen in einem bestimmen Profil oder nach häufig zusammen genannten Kompetenzen beantwortet. Die Gesamtergebnisse geben einen guten Überblick über aktuelle Profile und Berufsgruppen sowie geforderte Kompetenzen in der IT-Branche. Sie stellen damit eine mögliche Vorstufe für den Aufbau einer Kompetenz-Ontologie der IT-Branche durch Experten dar.

Bisher wurde nur eine kleine Anzahl von Termen als mögliche Kompetenzen in Betracht gezogen. Um die Ergebnisse zu verallgemeinern muss vor der Analyse größerer Datensätze die Anzahl betrachteter Kompetenzen vergrößert werden. Schließlich sollten die gefundenen Kategorien von Stellenprofilen in Sub-Cluster zerlegt werden, um die Ergebnisse zu verfeinern und spezifischere Einblicke zu gewinnen. Durch die Analyse zeitbehafteter Daten sollen zukünftig modegetriebene Textbausteine und „Trends“ in den Stellenprofilen erkannt werden. Zusammen mit der schon eingesetzten grafischen Aufbereitung der Ergebnisse wird der Digitalen Wirtschaft, vertreten durch den „Bundesverband der Digitalen Wirtschaft“ (BVDW), damit ein Instrument zur Verfügung gestellt, dass geeignet ist, aktuelle

Kompetenzbedarfe der einschlägigen Unternehmen zu ermitteln und ggf. spezifische Weiterbildungsangebote zu konzipieren. Diese Angebote können dann zusammen genutzt werden, um dem Fachkräftemangel in der Branche entgegenzuwirken.

Literatur

[Chan92] K. P. Chan, Y. S. Cheung. Clustering of Clusters. In *Pattern Recognition*, Vol. 25, Nr. 2, Seiten 211-217, 1992

[Deerwester90] Scott Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman. Indexing by Latent Semantic Analysis. In *Journal of the American Society of Information Science*, Seiten 391-407, 1990

[Draganidis06] Fotis Draganidis und Gregoris Mentzas. Competency based management: a review of systems and approaches. In *Information Management and Computer Security*, Vol. 14, Nr. 1, Seiten 51-64, 2006

[Flach99] P.A. Flach und N. Lachiche. Confirmation-guided discovery of first-order rules with Tertius. In *Machine Learning*, Vol. 42, Seiten 61-95, 1999

[Gionis07] Astridis Gionis, Heikki Mannila, Panayiotis Tsaparas. Clustering Aggregation. In *ACM Transactions on Knowledge Discovery from Data*, Vol. 1, Nr. 1, Article 4, 2007

[Görz03] G. Görz, C.-R. Rollinger und J. Schneeberger. Handbuch der Künstlichen Intelligenz, Oldenbourg, 2003

[Hochbaum85] Hochbaum, Shmoys. A best possible heuristic for the k-center problem. In *Mathematics of Operations Research*, 10(2), Seiten 180-184, 1985

[Malzahletal05] Nils Malzahn, Sam Zeini, Andreas Harrer. Ontology Facilitated Community Navigation - Who Is Interesting for What I Am Interested in?. In *Modeling and Using Context, 5th International and Interdisciplinary Conference, CONTEXT 2005,* Seiten 292 - 303, Paris, Frankreich, 2005

[Mierswa06] Ingo Mierswa, Michael Wurst, Ralf Klinkenberg, Martin Scholz und Timm Euler. YALE: Rapid Prototyping for Complex Data Mining Tasks. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, Seiten 935-940, 2006

[Pelleg00] Dan Pelleg, Andrew W. Moore: X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In: *Seventeenth International Conference on Machine Learning*, Seiten 727-734, 2000

[Witten05] Ian H. Witten und Eibe Frank. Data Mining – Practical Machine Learning Tools and Techniques. Elsevier Inc, 2005