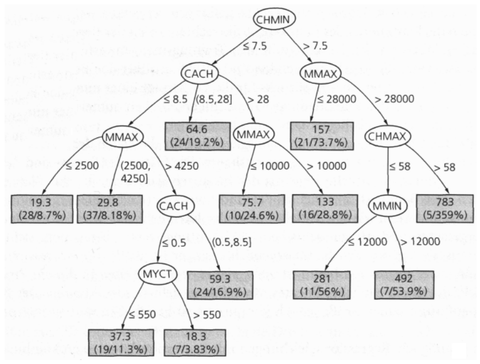
Fachhochschule Brandenburg
Fachbereich Informatik und Medien

Quelle: [1]

Analyse von Matching-Verfahren und Konzeption für eine auf Angebot und Nachfrage basierende Plattform – Prototypische Implementierung am Beispiel einer Lehrstellenbörse

Bachelorarbeit, vorgelegt von Helge Scheel



Quelle: [2]

Aufgabenstellung

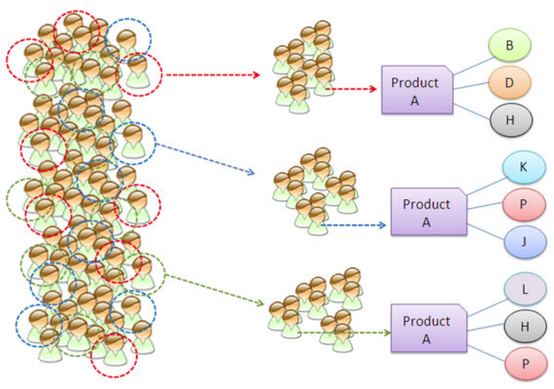
Ziel dieser Arbeit ist, Verfahren zur automatischen Auswahl von passenden Angeboten zu Nutzern zu untersuchen und eine Lösung für eine Lehrstellenbörse als Beispielanwendung zu konzipieren. Dafür werden verschiedene Lösungsansätze für diese Problematik analysiert und die Verfahren hinsichtlich ihrer Eignung beurteilt. Auf Basis dieser Evaluierung wird ein Konzept erarbeitet, welches die Lösung des Problems für eine konkrete Lehrstellenbörse ermöglicht und die spezifischen Anforderungen berücksichtigt. Abschließend wird das entstandene Konzept als Prototyp implementiert, der die Grundlage einer im produktiven Betrieb eingesetzten Lösung bilden kann.

Beispielanwendung

Das Zielsystem der Lehrstellenbörse ist eine Webapplikation, die von der]init[Aktiengesellschaft für Digitale Kommunikation für den Deutschen Industrie- und Handelskammertag entwickelt wird. Ziel der Plattform ist die deutschlandweit zentrale Erfassung offener Lehrstellen. Die Angebote einzelner regionaler Industrie- und Handelskammern werden zusammengestellt und auf einer zentralen Plattform veröffentlicht. Die automatisch empfohlenen Angebote müssen eine hohe Eignung für den Nutzer besitzen und ihm erklärt werden können. Einschränkungen ergeben sich dadurch, dass keine Funktionen vorgesehen sind, mit denen die Nutzer Lehrstellenangebote oder automatische Vorschläge bewerten können. Weiterhin liegen zur Entwicklungszeit keine Trainingsdaten oder Daten zu realen Nutzern vor.

Empfehlungssysteme

Empfehlungssysteme stellen eine häufig eingesetzte Möglichkeit dar, eine automatische, personalisierte Auswahl von Angeboten für eine Person anhand ihrer Bewertungen zu treffen. Inhaltsbasierte Filterung ermittelt die Ähnlichkeit von Objekten basierend auf deren Eigenschaften, um Vorschläge zu generieren. Dabei werden die vom Nutzer bereits bewerteten Objekte mit den unbewerteten Objekten der Datenbasis verglichen. Gemeinschaftsbasierte Filterung basiert auf den Ähnlichkeiten von Nutzern, dabei werden die Bewertungen eines Benutzers mit denen aller anderen Nutzer verglichen, um Nutzer mit einem ähnlichen Profil zu finden. Objekte, die von diesen Nutzern positiv bewertet wurden, können dem Benutzer empfohlen werden. Hybride Ansätze kombinieren Elemente beider Verfahren.



Zuordnung von Produkten zu Nutzern anhand ihrer Profile Quelle: [3]

Clusteranalyse

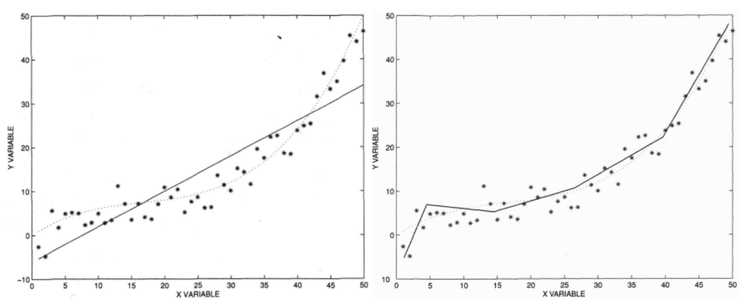
Clusterverfahren teilen Objekte anhand ihrer Ähnlichkeit in Gruppen ein und legen die Gruppen dabei selbstständig fest. Ziel sind Cluster, deren Objekte untereinander möglichst ähnlich sind, während die Cluster möglichst verschieden sind. Eine hohe Eignung wird Objekten aus Clustern unterstellt, aus denen bisher positiv bewertete Objekte stammen. Partitionierende Algorithmen teilen die Objekte auf eine festgelegte Anzahl an Clustern auf. In den Gruppen wird eine Aufteilung gesucht, die hinsichtlich des Gütekriteriums des jeweiligen Verfahrens optimal ist. Hierarchische Verfahren bilden mehrere Ebenen mit unterschiedlichen Anzahlen an Clustern, die sukzessiv zusammengesetzt oder aufgeteilt werden.

Klassifizierung und Regression

Klassifizierungs- und Regressionsverfahren bestimmen einen unbekannten Wert für ein Objekt anhand seiner Eigenschaften. Dieser Wert ermöglicht die Einteilung des Objekts in Klassen oder die Beurteilung hinsichtlich einer bestimmten Problemstellung. Im Anwendungsfall kann die automatische Auswahl von Objekten aufgrund des berechneten Wertes erfolgen, der Aussage über die Eignung der Objekte trifft. Für jede Kombination von Nutzer und Angebot kann die Eignung bestimmt und durch den Vorhersagewert repräsentiert werden.

Regressionsmodelle

Regressionsmodelle stellen den vorherzusagenden Wert als abhängig von anderen, zur Vorhersage ausgewerteten Variablen dar. Die abhängigen Variablen können Attribute eines Objekts repräsentieren, für das ein Vorhersagewert berechnet werden soll. Zur Bestimmung dieses Wertes werden die abhängigen Variablen gewichtet und aufsummiert. Die Gewichte können durch Trainingsdaten in einem Optimierungsprozess bestimmt werden.

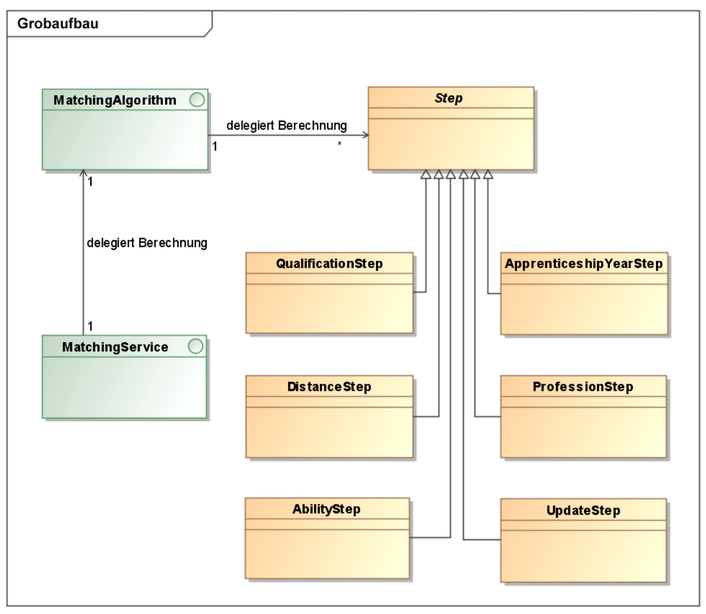


Lineare Funktion und lineare Teilstücke zur Annäherung an Trainingsdaten Quelle: [4]

Bei der linearen Regression wird eine lineare Abhängigkeit zwischen den ausgewerten Variablen angenommen. Das Prinzip kann durch Aufsummieren gewichteter Funktionswerte generalisiert werden. Die Funktionen werten jeweils ein Attribut aus, wobei sie nicht-linear sein können. Dadurch wird das Modell flexibler und kann nicht-lineare Beziehungen abbilden. Erweitert werden kann das Prinzip der Regression ebenfalls, indem das Modell statt einer Hyperebene eine Menge von Hyperebenen nutzt, die stückweise zusammengesetzt werden.

Konzept für Lehrstellenbörse

Das konzipierte Verfahren setzt ein Regressionsmodell um, dessen Terme durch die Übereinstimmung jeweils eines Attributs des Angebots und des Nutzers bestimmt sind. Dazu wird für jedes Attribut eine Funktion angewendet, die die Übereinstimmung ermittelt. Durch Auswertung aller Regressionsterme ergibt sich die vorhergesagte Eignung des Angebots für den Nutzer, ohne auf eine Bewertungsfunktion angewiesen zu sein. Regressionsmodelle weisen eine flexible und erweiterbare Struktur auf, die unabhängig von Traingsdaten erstellt werden kann. Parameter können dynamisch verändert werden, z. B. durch Ergebnisse von Lernverfahren, die auf später erhobenen Daten arbeiten. Die Vorschläge können dem Nutzer mit der Übereinstimmung einzelner Attribute durch geringe Unterschiede zwischen ihnen verständlich erklärt werden.



Grobe Übersicht über die Architektur des konzipierten Verfahrens

Das Interface MatchingService bildet die Schnittstelle zum Gesamtsystem um das Matching anzustoßen und zur Datenzugriffsschicht, um benötigte Datensätze für die Berechnung zusammenzustellen und die Resultate zu speichern. Die Berechnung der Matching-Ergebnisse wird an das Interface MatchingAlgorithm delegiert. Implementierungen der Schnittstelle iterieren über Teilschritte, die jeweils einen Regressionsterm repräsentieren, und aggregieren die Teilergebnisse zu dem Gesamtresultat. Die Auswertung der einzelnen Attribute erfolgt unabhängig voneinander in separaten Klassen.

Auswertung

Das implementierte Verfahren erfüllt die gestellten Anforderungen, berücksichtigt die spezifischen Einschränkungen und ist einfach anpassbar. Eine genaue Auswertung der Güte der Matching-Ergebnisse ist nicht möglich, da keine Testdaten vorliegen. Die Auswahl der auszuwertenden Attribute und der Parameter muss manuell erfolgen, Lernverfahren können aufgrund der fehlenden Daten nicht eingesetzt werden. Bessere Ergebnisse sind durch den Einsatz einer Bewertungsfunktion für die Angebote oder die Vorschläge und Daten realer Nutzer möglich.

Bildquellen: [1] http://www.pressrelations.de/new/standard/result_main.cfm?n_firmanr_=123990&sid=&aktion=jour_pm&pfach=1&sektor=pics&popup_vorschau=0 [2] Witten, Ian H.; Frank, Eibe. Data Mining Praktische Werkzeuge und Techniken für das maschinelle Lernen, S. 75 [3] http://www.bridgewell.com/ec%20portal.html [4] Hand, David; Mannila, Heikki; Smyth, Padhraic. Principles of Data Mining, S. 171, 173

Betreuer: Dipl.-Inform. Ingo Boersch
Dipl.-Inform. Dirk Wiezcorek (]init[AG)

Abgabe: 23.09.2011