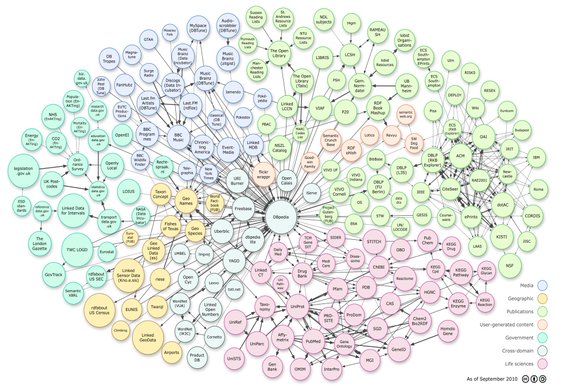
FACHHOCHSCHULE BRANDENBURG
FACHBEREICH INFORMATIK UND MEDIEN

Kontext-spezifische Analyse von Benutzerpräferenzen mittels Clustering für Musikempfehlungen auf Grundlage von Semantic-Web-Metadaten

Masterarbeit, vorgelegt von Christian Freye



Aufgabenstellung

Ziel dieser Arbeit ist die Entwicklung, Implementierung und Evaluation eines Verfahrens zur Extraktion von Nutzerprofilen aus einer Menge von Künstlern, die ein Nutzer zuvor gehört hat.

Hierbei sollen die Künstler mit Metadaten aus dem Semantic-Web angereichert werden, um so das Clustering zu ermöglichen. Anschließend werden Merkmale aus den Clustern extrahiert und als beschreibende Labels genutzt. Es soll dann überprüft werden, ob diese Labels spezifisch genug sind, um die Künstler in den Clustern zu beschreiben und ob es mit ihnen möglich ist, die unterschiedlichen Musikrichtungen eines Nutzers zu identifizieren.

Umfeld

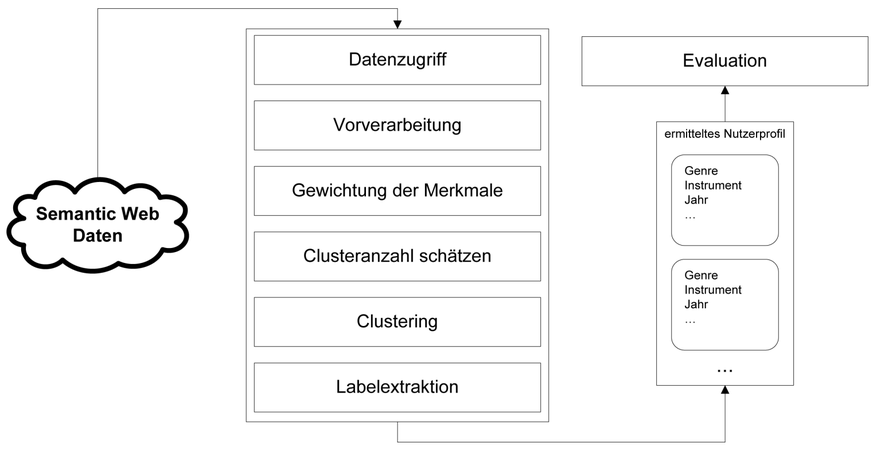
Zu dem ersten Punkt in dem Umfeld der Arbeit zählen Empfehlungssysteme. Diese lassen sich in drei unterschiedliche Kategorien unterteilen:

- **Inhaltsbasiert** - Dem Nutzer werden Dinge empfohlen, die ähnlich zu denen sind, die er bereits in der Vergangenheit bevorzugt hat.
- **Kollaborativ** - Dem Nutzer werden Dinge empfohlen, die andere (ihm ähnliche) Nutzer in der Vergangenheit favorisiert haben.
- **Hybrid** - Diese Methoden kombinieren die Eigenschaften inhaltsbasierter und kollaborativer Empfehlungssysteme.

Der zweite Punkt in dem Umfeld der Arbeit ist das Semantic-Web. Der Kerngedanke dabei besteht darin, dass das heutige WWW um Strukturen erweitert werden soll, mit denen es möglich ist Informationen mit maschinenlesbaren Bedeutungen anzureichern.

Von besonderem Interesse ist dabei das Prinzip der Linked Data, bei dem Daten miteinander verknüpft und öffentlich zur Verfügung gestellt werden.

Zwei wichtige Semantic-Web-Dienste, die zur Gewinnung von Metadaten der verschiedenen Künstler genutzt werden, sind *Freebase* und *MusicBrainz*. Es handelt sich dabei um Online-Datenbanken, die Informationen von verschiedenen Quellen sammeln und aufbereitet zur Verfügung stellen.



Schematische Darstellung des Data-Mining-Prozesses.

Durchführung

Im ersten Schritt werden Metadaten zu einer großen Anzahl an Künstlern aus dem Semantic-Web gesammelt. Anschließend werden die Künstler ermittelt, die ein Nutzer gehört hat, welche dann mit den Metadaten verknüpft werden.

Es folgt die Vorverarbeitung der Daten, wobei Attribute entfernt werden, die nur selten vorkommen. Die übrigen Daten werden mittels TF-IDF-Maß gewichtet.

Die Künstler werden dann mit Hilfe des k-Means Algorithmus geclustert, wobei k im Intervall [2,20] festgelegt wird. In diesem Intervall wird die Anzahl der präferierten Musikrichtungen eines Nutzers vermutet. Die endgültige Auswahl der Clusteranzahl richtet sich nach dem Verlauf der RSS-Werte (*residual sum of squares*), die innerhalb der Cluster bestimmt werden.

Nach der Bestimmung der Clusteranzahl werden aus den Clustern Merkmale extrahiert. Es werden dabei zwei Ansätze verglichen, mit denen die Relevanzwerte der Attribute bestimmt werden können. Zum Einen wird ein häufigkeitsbasierter Ansatz verwendet und zum Anderen erfolgt die Merkmalsauswahl mittels Chi-Quadrat-Test. Die Attribute werden dann abhängig von der Relevanz des Top-Attributs ausgewählt. Es erfolgt die Unterteilung in die Relevanzklassen 100%, 80%, 60%, 40% und 20%, d.h. es werden nur Attribute zu dem Label hinzugefügt die einen Mindestwert an Relevanz für das Cluster besitzen. Diese Labels für die unterschiedlichen Cluster bilden somit das Nutzerprofil, welches im weiteren Verlauf evaluiert werden soll.

Evaluation

Für die Evaluation der Nutzerprofile werden die Messwerte Precision, Recall und F-Maß der Labels bestimmt. Diese Werte werden pro Cluster errechnet und dann durch das arithmetische Mittel pro Benutzer und Relevanzklasse zusammengefasst.

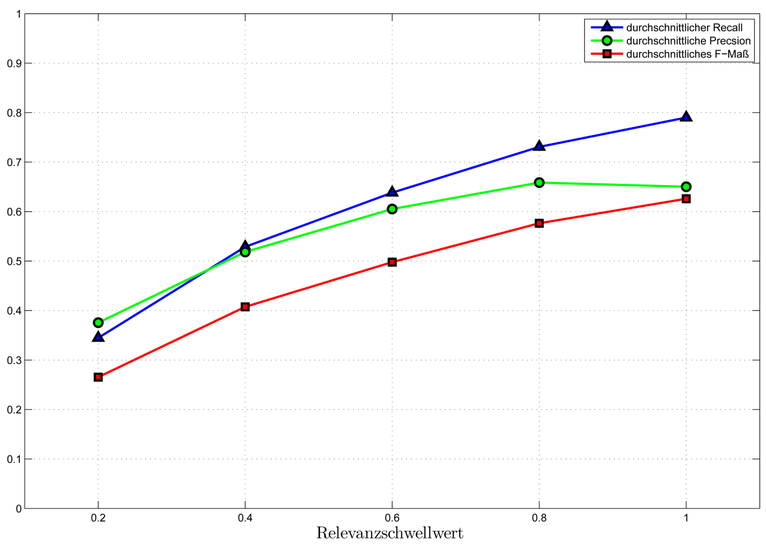
Mit Hilfe der Precision kann die Genauigkeit der extrahierten Labels bestimmt werden, d.h. Wie viel Prozent der gefundenen Künstler korrekte Ergebnisse sind. Durch den Recall wird die Vollständigkeit der gefundenen Künstler überprüft, d.h. wie viele Künstler aus einem Cluster können mit dem Label gefunden werden. Das F-Maß bildet eine Kombination der beiden Messwerte, da sich Precision und Recall gegenseitig beeinflussen.

Des Weiteren werden die Messwerte in Form von Lorenz-Kurven dargestellt. Mit ihnen ist es möglich relative Konzentrationsverteilungen von Merkmalen darzustellen. Somit kann überprüft werden für wie viele Nutzer sinnvolle Cluster gefunden werden können.

Die besten Ergebnisse ließen sich bei dem Experiment mit der häufigkeitsbasierten Merkmalsauswahl erreichen. Das Maximum, welches für das F-Maß erreicht werden konnte, befindet sich dabei bei Labels, die nur die relevantesten Attribute ihrer Cluster enthalten. Anhand der Lorenz-Kurven ist sichtbar, dass mit diesen Labels ein F-Maß von über 0,5 bei ca. 80% der Nutzer erreicht werden kann.

Fazit

In dieser Arbeit konnte gezeigt werden, dass Nutzer unterschiedliche Musikrichtungen hören und sich diese mit dem hier vorgestellten Verfahren identifizieren lassen. Es konnten Labels extrahiert werden, die spezifisch für diese Musikrichtungen sind. Somit konnte gezeigt werden, dass Semantic-Web-Metadaten ein großes Potential für zukünftige Empfehlungssysteme besitzen. Bei dem Vergleich der beiden Methoden zur Merkmalsauswahl konnte ermittelt werden, dass sich mit dem häufigkeitsbasierten Ansatz bessere Ergebnisse erzielen lassen, als es mit dem Chi-Quadrat-Test möglich ist. Es konnte ebenfalls gezeigt werden, dass sich mit diesem Verfahren für einen Großteil der Nutzer gute Ergebnisse erzielen lassen.



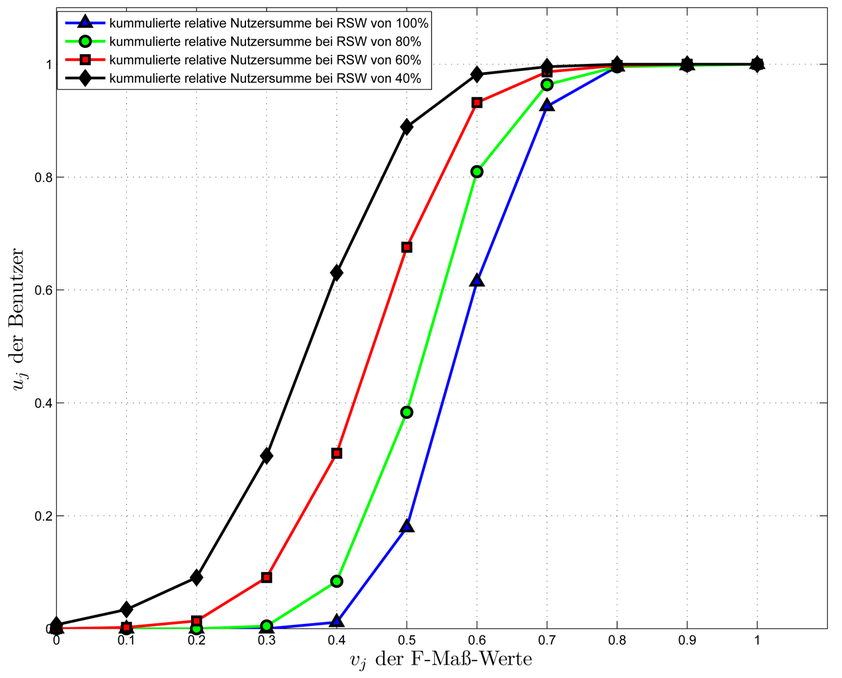
Die durchschnittlichen Ergebnisse pro Benutzer, die mit den Labels der häufigkeitsbasierten Merkmalsauswahl erreicht werden konnten.

Nutzer-ID	Labels
232	{electronic music} {hip hop} {indie rock}
402	{indie rock} {smooth jazz} {salsa music} {pop music}
511	{electronic music} {pop music} {string instrument} {indie rock} {indie pop} {film score} {garage rock} {hard rock} {dream pop} {power pop} {britpop} {dance-punk}
740	{rock music} {speed metal} {punk rock} {hard rock} {grindcore} {melodic death metal} {trash metal} {death metal} {progressive metal}
865	{rock music} {string instrument} {punk rock} {indie rock} {heavy metal} {electronic music}

Extrahierte Nutzerprofile für fünf zufällig gewählte Nutzer.

Publikation

Rafael Schirru, Stephan Baumann, Christian Freye, and Andreas Dengel. Towards Context-Sensitive Music Recommendations Using Multifaceted User Profiles. *The Proceedings of the AES 42nd International Conference,* 2011, 54-59.



Lorenz-Kurven für die F-Maß-Werte bei der häufigkeitsbasierten Merkmalsauswahl.

Betreuer: Prof. Dr.-Ing. Jochen Heinsohn
Dipl.-Inf. Rafael Schirru (DFKI Projektbüro Berlin)
Dipl.-Inf. Ingo Boersch

Abgabe: 10.08.2011