

Spezielle Kapitel der Informatik:
Music Information Retrieval
KV SS2009

Protokoll der Übungsaufgabe:

Informationsextraktion mit LastFM
im Vergleich mit Google

Datum der Präsentation: 24. Juni 2009

Matr. Nr.	Kennzahl	Name
		Jakob Doppler
		Doris Zachhuber
0856711	086	Matthias Husinsky

Inhaltsverzeichnis

1	Einleitung	3
1.1	Aufgabenstellung	3
1.2	Zielsetzungen	3
2	LastFM	4
2.1	Hintergrund	4
2.2	Geschichtlicher Überblick	4
2.3	Die LastFM-API	4
2.3.1	Verfügbare Information	4
2.3.2	Technischer Überblick	5
3	Systemarchitektur	6
3.1	Tools und Frameworks	6
4	Auswertung der Daten	7
4.1	Datenquellen	7
4.1.1	Textbasiertes Genre-Artist-Mapping	7
4.1.2	LastFM	7
4.1.3	Google	7
4.2	Feature Extraction	7
4.2.1	Albumbasierte Epoche	7
4.2.2	TagClouds	7
4.3	Ähnlichkeitsmessungen	7
4.3.1	Differenz von skalaren Werten	7
4.3.2	Cosinus-Similarity	7
4.4	Visualisierung	7
5	Ergebnisse	8
5.1	Epochen von Artists	8
5.2	LastFM- zu Google-Genreähnlichkeit	8
5.3	Genreklassifizierung basierend auf Tags der Top-Artists	8
6	Referenzen	9

1 Einleitung

1.1 Aufgabenstellung

Was in der Angabe steht in aller Kürze

1.2 Zielsetzungen

Contentbasierte Features (TagClouds, Epochen), Datasourcen (LastFM, Google, Textdateien - nur ein grober Umriss!), Visualisierung, Machine Learning LastFM TagClouds gewichtet mit LastFMRanking und mit Termfrequency von Google. Performanceunterschied (Ground Truth ist LastFM) Epochen anhand von AlbumReleasedates

2 LastFM

2.1 Hintergrund

LastFM ist heute ein personalisiertes Online-Radio - abspielbar mit einer eigenen, proprietären Software. Dieses ermöglicht es den Usern einen auf ihren Geschmack zugeschnittenen Musikstream abzuspielen. Angegeben kann der momentan gewünschte Musikstil via Tags - entweder auf einen Künstler bezogen oder auf einen Musikstil bzw. Genre. Ebenso ist es möglich über Plug-Ins das Abspielverhalten von Usern in anderen Audio-Playern zu beobachten um damit die Erstellung des Nutzerprofils zu unterstützen. Weiters ist *LastFM* auch ein großes soziales Netzwerk, dass vor allem auf das zusammenbringen von Menschen mit ähnlichem Musikgeschmack ausgelegt ist. Die User der Community tragen zur Klassifikation des Musikbestandes durch tagging, Wiki-Beträge und einfach ihr Hörverhalten bei. Mehr dazu in Abschnitt 2.3.

2.2 Geschichtlicher Überblick

LastFM wurde in den späten 90er Jahren als Online-Musiklabel gegründet und bot als Feature die Möglichkeit, sich durch die Art der konsumierten Musikstücke ein Profil über seinen Musikgeschmack zu erstellen. Die Firma *Audioscrobbler*, hervorgegangen aus einem Informatikprojekt, hatte sehr ähnliche Ideen, worauf hin beide Unternehmen sehr eng zusammenarbeiteten, bis sie schließlich im Jahr 2005 fusionierten und unter dem Namen *LastFM* die Funktionen beider Technologien zur Verfügung stellen. 2007 wurde *LastFM* um den Preis von 280 Millionen Dollar an das US-amerikanische Medienunternehmen CBS verkauft. Diese Übernahme gehört damit zu den größten dotcom Aquisitionen bisher.

Seit April 2009 ist die Benutzung des Radiodienstes nur mehr in den USA, Großbritannien und Deutschland kostenlos möglich. In anderen Ländern muss ein Entgelt von 3 Euro für die monatliche Nutzung erbracht werden.

2.3 Die LastFM-API

2.3.1 Verfügbare Information

Das große Motto von LastFM lautet Finde die Musik, die du magst und zeige mir die Leute, die einen ähnlichen Musikgeschmack haben. Dazu ist es natürlich notwendig, dass der Dienst eine Menge Daten über die User aquiriert und ein genaues Nutzerprofil erstellt. Ebenso muss der Dienst so gut wie möglich über die Bands und deren Eigenschaften bescheid wissen, damit eine sinnvolle Einteilung möglich wird.

Hauptquelle für Informationen über Musikstücke und KünstlerInnen sind die User der Community. Das Profil über das Hörverhalten eines Users wird durch sein/ihr Hörverhalten bestimmt. Der LastFM-Player übermittelt dabei an LastFM, welche Musikstücke konsumiert werden (das sogenannte *scrobblen*). Auf der Ebene der KünstlerInnen werden Informationen aquiriert, indem die User diese taggen können oder detaillierte Informationen in einem Band-Wiki eintragen. Ob auch Band-Infos von Plattenlabels oder anderen professionellen Marketingagenturen

hier mit einbezogen werden ist nicht bekannt, jedenfalls kann man über das Online-Portal auch eine Menge Hintergrundinformationen und aktuelle Auftritte (künstlerInnen- oder regionsbezogen) erfahren.

Ein Problem, das sich mit dieser Art der Informationssammlung ergibt ist, dass die Information nur auf KünstlerInnenebene, aber nicht auf Track-Ebene vorhanden ist. Dadurch wird die eventuell gegebene Vielseitigkeit oder musikalische Entwicklung einer Band verschwommen und wird nicht so behandelt, wie es notwendig wäre.

Ein weiteres Problem ist die fehlende Unterscheidung von unterschiedlichen Bands, die aber denselben Namen haben. Diese werden zwangsläufig in ein Profil zusammengeführt. Umgekehrt kann es auch in seltenen Fällen passieren, dass ein Künstler in mehreren Profilen vorhanden ist (für Mozart existieren etwa die Profile: "Wolfgang Amadeus Mozart und MMozart").

2.3.2 Technischer Überblick

Die meisten dieser Infos lassen sich direkt über die Webseite, oder eine Webservice-API abrufen. Die API ist zu denselben Bedingungen verfügbar wie die Nutzung des Radiodienstes, d.h. außerhalb US, UK und DE muss man ein geringes monatliches Entgelt entrichten. Nur wenn man die API in großem Stil nutzen will sind eigene Nutzungsbedingungen mit den Betreibern auszuverhandeln.

Technisch basiert die Web-API auf dem REST-Standard (XML), Abfragen sind somit unkompliziert via HTTP möglich. Unter anderem lassen sich Künstlerdaten (Bio, Alben, Events, Top-Songs, Tags, ...), Geodaten (Events in der Umgebung, beliebteste Künstler, ...), Userdaten (Top-Artists, Friends, Neighbours, ...) und viele mehr. Für eine Dokumentation mit Anwendungsbeispielen sei auf <http://www.lastfm.de/api/intro> verwiesen.

Es existieren eine Reihe von Wrapper-APIs von Dritten für unterschiedliche Technologien, damit man die Web-API auch unkompliziert in andere Anwendungen einbinden kann. Die Qualität dieser Implementierungen ist unterschiedlich und es kann durchaus sein dass diese nicht am aktuellen Stand der Web-API sind.

Es existieren bereits eine Reihe von Audioplayern und anderen Applikationen, die diese API auch direkt nutzen. Unter <http://build.last.fm> gibt es einen Showcase.

3 Systemarchitektur

3.1 Tools und Frameworks

4 Auswertung der Daten

4.1 Datenquellen

4.1.1 Textbasiertes Genre-Artist-Mapping

Verwendet für...

4.1.2 LastFM

Verwendet für Albeninfos für Epochenberechnung und für TagClouds, auch als Ground-Truth für die Klassifizierung.

4.1.3 Google

TagClouds durch Termfrequency

4.2 Feature Extraction

4.2.1 Alumbasierte Epoche

4.2.2 TagClouds

4.3 Ähnlichkeitsmessungen

4.3.1 Differenz von skalaren Werten

4.3.2 Cosinus-Similarity

Ein Maß zur Bestimmung der Ähnlichkeit der Künstler in der LastFM Datenbank ist die Auswertung der Tags und ein Vergleich auf deren Übereinstimmung. Dazu wurden die Tags der Künstler mit der API ausgelesen und anschließend die Ähnlichkeit mit der aus der LV bekannten Formel der Cosine Similarity bestimmt:

$$sim(a,b) = \frac{a * b}{|a| * |b|} \quad (1)$$

Die API gibt auf Anfrage die Tags eines Künstlers/Band mit einer Gewichtung zwischen 100 und 1 zurück. Zwei kleine Probleme traten hier aber auf: Zum werden auf jeden Fall eine Menge Tags zurückgegeben. Nach ein paar Versuchen wurde aber klar, dass die Tags mit Gewicht 1 nur zufälliger Natur sind und in die Berechnung nicht mit einzubeziehen sind (Mozart hat die Attribute *psychodelic* bzw. *Heavy-Metal* eher nicht verdient). Weiters lieferte unsere Wrapper-API für Java leider nur die Tags, aber ohne Gewichtung. Die API musste daher geringfügig modifiziert werden.

4.4 Visualisierung

5 Ergebnisse

5.1 Epochen von Artists

5.2 LastFM- zu Google-Genreähnlichkeit

5.3 Genreklassifizierung basierend auf Tags der Top-Artists

6 Referenzen