MythMiner

Ein Empfehlungssystem für Fernsehprogramme auf Basis von RapidMiner

Balázs Bárány

Predictive-Analytics-Konferenz 2012

Inhalt

Motivation

- Lern- und Übungsprojekt für Text Mining
- Eigene Nutzung
- Open-Source-Veröffentlichung

Geschichte

- Beginn der Entwicklung: November 2010

Geschichte

- Beginn der Entwicklung: November 2010
- Veröffentlichung am 30. 1. 2011
  - Ankündigung in RapidMiner- und MythTV-Foren
  - MyExperiment.org
  - Projekt auf Freecode.com (ehemals Freshmeat.net)
  - MythTV-Wiki

Geschichte

- Beginn der Entwicklung: November 2010
- Veröffentlichung am 30. 1. 2011
  - Ankündigung in RapidMiner- und MythTV-Foren
  - MyExperiment.org
  - Projekt auf Freecode.com (ehemals Freshmeat.net)
  - MythTV-Wiki
- Vortrag bei den Linuxwochen in Wien am 7. 5. 2011

Rezeption

- Im RapidMiner-Forum positive Postings mit Bitte um weitere Informationen
  - Im Mai im offiziellen Rapid-I-Blog erwähnt
  - Testinstallation bei Rapid-I

Rezeption

- Im RapidMiner-Forum positive Postings mit Bitte um weitere Informationen
  - Im Mai im offiziellen Rapid-I-Blog erwähnt
  - Testinstallation bei Rapid-I
- Auf der MythTV-Mailingliste keine Reaktionen

Rezeption

- Im RapidMiner-Forum positive Postings mit Bitte um weitere Informationen
  - Im Mai im offiziellen Rapid-I-Blog erwähnt
  - Testinstallation bei Rapid-I
- Auf der MythTV-Mailingliste keine Reaktionen
- Beim Linuxwochen-Vortrag etwa 10 Zuhörer

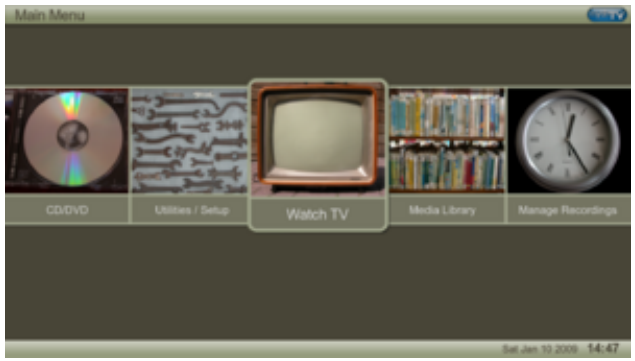
Nutzungsstatistiken

- 41 Downloads auf MyExperiment.org
- Ca. 30 Downloads von der Homepage seit Anfang 2012
- Drei Abonnenten auf Freecode.com
- Mail-Kontakt mit zwei verschiedenen Usern

MythTV

Open Source Digital Video Recorder

- Komplettlösung für „Unterhaltungscomputer”
  - Fernsehen, Video, Audio, Bilder, Wetter, Nachrichten, ...
- Verteilte Architektur
- Web-Frontend
- Theme-Support



Installation und Nutzung von MythTV

Voraussetzungen

- Angepaßte Hardware empfehlenswert
  - Wohnzimmer-PC
  - Fernsehkarte, Fernbedienung
  - HDMI

Installation und Nutzung von MythTV

Voraussetzungen

- Angepaßte Hardware empfehlenswert
  - Wohnzimmer-PC
  - Fernsehkarte, Fernbedienung
  - HDMI
- Fertiges Paket in Linux-Distribution *oder*
- Eigene Distribution: Mythbuntu, MythDora, KnoppMyth

Installation und Nutzung von MythTV

Voraussetzungen

- Angepaßte Hardware empfehlenswert
  - Wohnzimmer-PC
  - Fernsehkarte, Fernbedienung
  - HDMI
- Fertiges Paket in Linux-Distribution *oder*
- Eigene Distribution: Mythbuntu, MythDora, KnoppMyth
- Programminformationen für EPG notwendig
  - Bei DVB-T automatisch dabei
  - kommerzielle und Community-Anbieter

RapidMiner

Open-Source-System für Data Mining

- Komplette Umgebung für Data Mining:
  - Grafische Oberfläche, visuelle Modellierung
  - Explorative Datenanalyse
  - Datenintegration

RapidMiner

Open-Source-System für Data Mining

- Komplette Umgebung für Data Mining:
  - Grafische Oberfläche, visuelle Modellierung
  - Explorative Datenanalyse
  - Datenintegration
  - Data-Mining-Verfahren: hunderte eingebaut
    - zusätzlich Weka-, R- und Octave/Matlab-Plugins

RapidMiner

Open-Source-System für Data Mining

- Komplette Umgebung für Data Mining:
  - Grafische Oberfläche, visuelle Modellierung
  - Explorative Datenanalyse
  - Datenintegration
  - Data-Mining-Verfahren: hunderte eingebaut
    - zusätzlich Weka-, R- und Octave/Matlab-Plugins
  - Berichtsfunktionen: Tabellen und Diagramme in HTML, PDF
  - Plugins für diverse Aufgaben: Web-Mining, automatische Prozesserstellung, Empfehlungssysteme, Image Mining, ...
  - Große Community und Plugin-Ökosystem

RapidAnalytics

- Serverversion von RapidMiner
  - Community- vs. Enterprise-Version
  - Repository für Teams
  - Prozesserstellung in RapidMiner, Verwaltung der Prozesse in der Web-Oberfläche

MythMiner installieren

- **MythTV installieren und einige Wochen lang benutzen!**

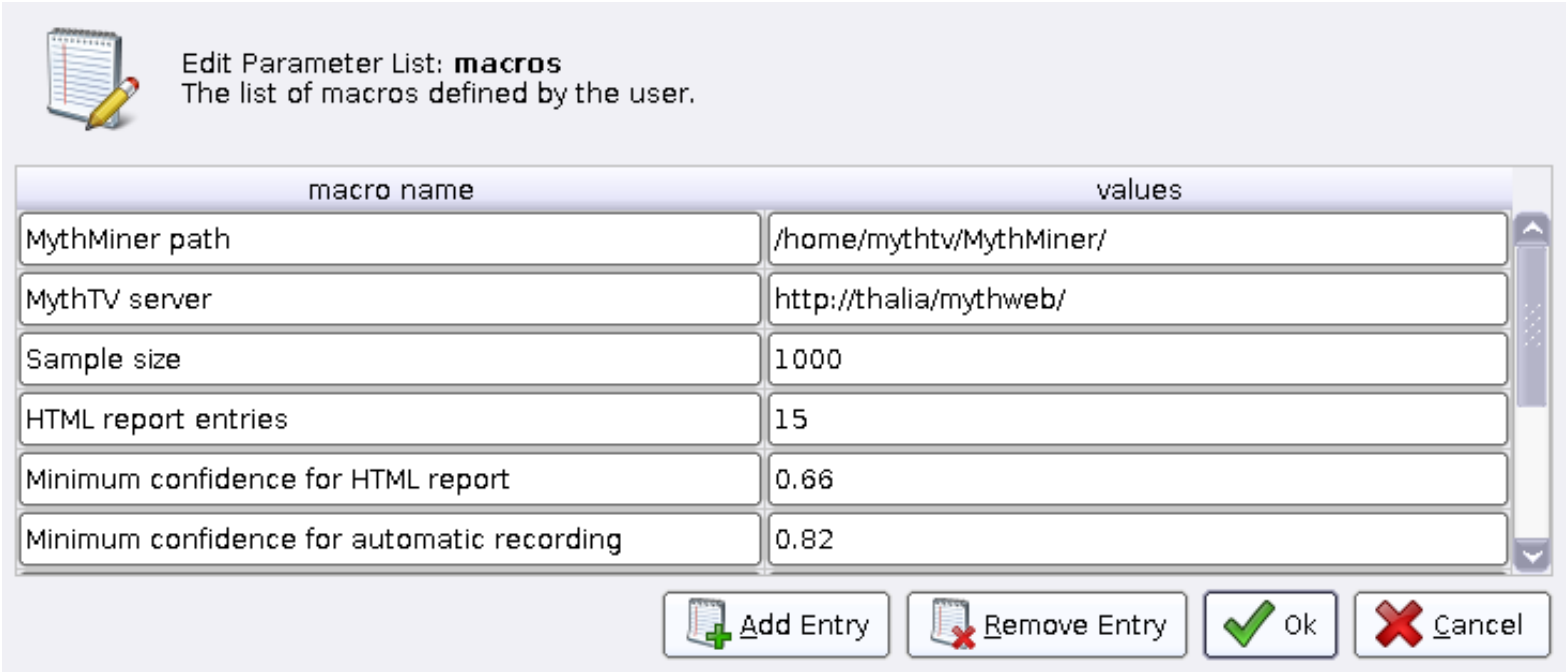
MythMiner installieren

- **MythTV installieren und einige Wochen lang benutzen!**
- RapidMiner installieren

MythMiner installieren

- **MythTV installieren und einige Wochen lang benutzen!**
- RapidMiner installieren
- MythMiner herunterladen:
  `http://tud.at/programm/mythminer/`
  - entpacken
- In RapidMiner öffnen und konfigurieren
  - Datenbankverbindung zu MythTV
  - „Configure process"
- (optional RapidAnalytics)

MythMiner-Konfiguration



Ergebnis der Ausführung

- HTML-Datei
- Optional: Umwandlung mit dem mitgelieferten Shellskript
- Optional: tägliche E-Mail

Ergebnis der Ausführung

- HTML-Datei
- Optional: Umwandlung mit dem mitgelieferten Shellskript
- Optional: tägliche E-Mail

- Ergebnisse nicht immer optimal

Ergebnis der Ausführung

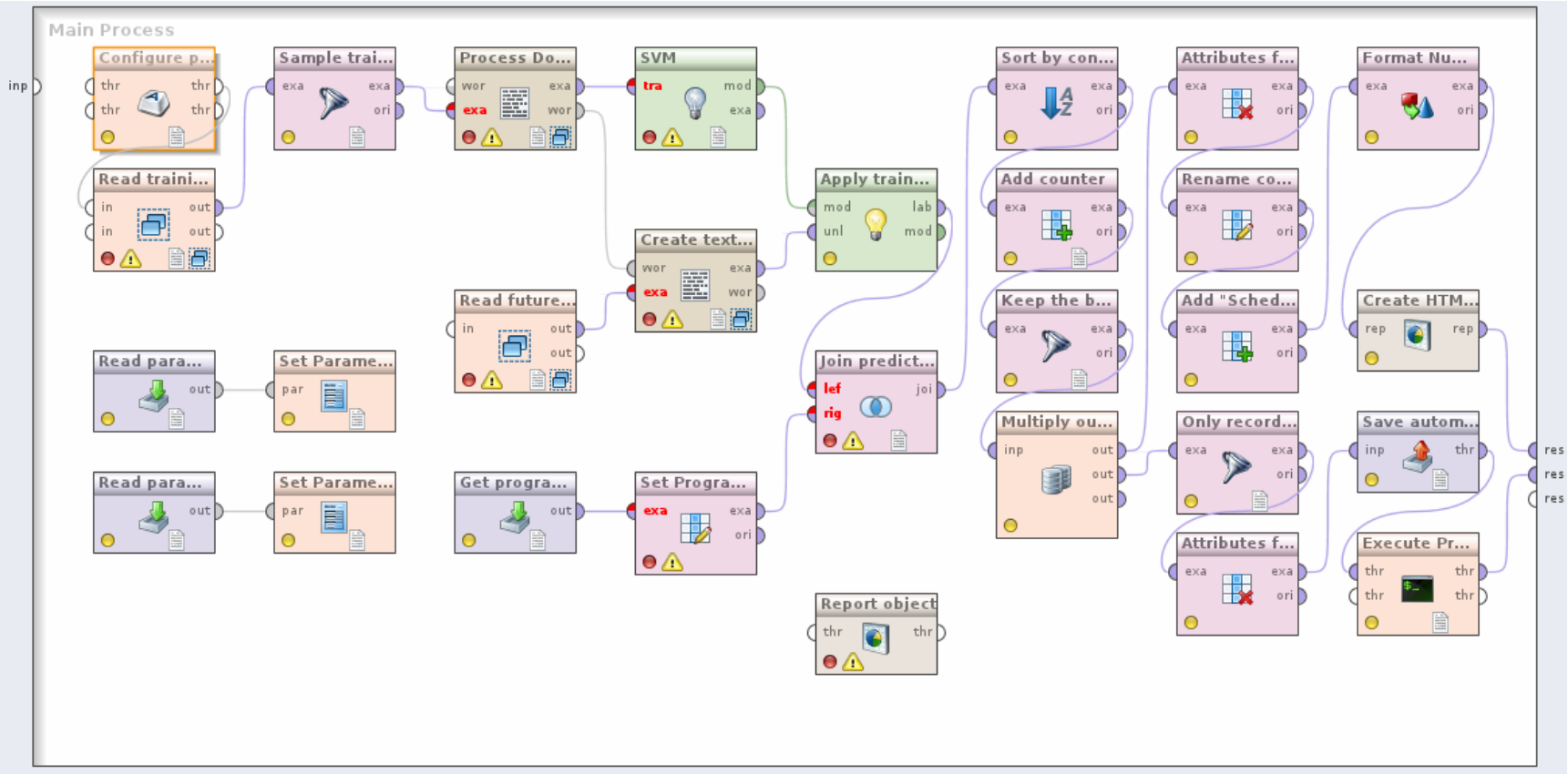
- HTML-Datei
- Optional: Umwandlung mit dem mitgelieferten Shellskript
- Optional: tägliche E-Mail
- Ergebnisse nicht immer optimal
- Trotzdem nützlich

Beispiel für Ergebnis-Mail

Program list

Program info	Category	Title	Description	Scheduled for recording	Confidence
SW/BW, 08:30 - 09:15	geschichte	Rom - Marmor, Macht und Märtyrer: Vom Hüttendorf zur Metropole	Die erste Folge "Vom Hüttendorf zur Metropole" untersucht den Gründungsmythos der Stadt am Tiber, die Sage von Romulus, Remus und der Wölfin, und sie beleuchtet die Rolle der Nachbarvölker, den Einfluss der Etrusker und Griechen auf die frühe Siedlung. Sie fragt nach der politischen Organisation und den sozialen Problemen der römischen Republik. Wie funktionierte dieses antike Gemeinwesen? Was waren die entscheidenden Gründe, die Roms Größe und Macht bestimmten. Wie muss man sich die städtische Infrastruktur vorstellen?	yes	83.0 %
SW/BW, 09:15 - 10:00	geschichte	Rom - Marmor, Macht und Märtyrer: Der Aufstieg zur Kaiserstadt	Die zweite Folge spannt den Bogen vom Charisma Cäsars über die Expansion des römischen Weltreichs bis zur Friedensherrschaft des Kaiser Augustus. Er war es, der die Stadt aus Ziegeln zu einer Weltstadt aus Marmor machte. Von Rom aus erstreckte sich strahlenförmig ein gut ausgebautes Straßennetz bis in die letzten Winkel des Reiches. Dies erlaubte rasche Truppenverschiebungen, förderte die Handelsströme und den kulturellen Austausch zwischen den Provinzen und der		76.2 %

Überblick


Main Process
inp
Configure p...
Sample trai...
Process Do...
SVM
Read traini...
Apply train...
Sort by con...
Attributes f...
Format Nu...
Add counter
Rename co...
Create text...
Keep the b...
Add "Sched...
Create HTM...
Read future...
Join predict...
Read para...
Set Parame...
Multiply ou...
Only record...
Save autom...
Read para...
Set Parame...
Get progra...
Set Progra...
Attributes f...
Execute Pr...
Report object
res
res
res

Herausforderungen

- Aus schlechten Trainingsdaten schlechte Ergebnisse
  - Aber es kann schwer sein, gute Trainingsdaten zu bekommen!

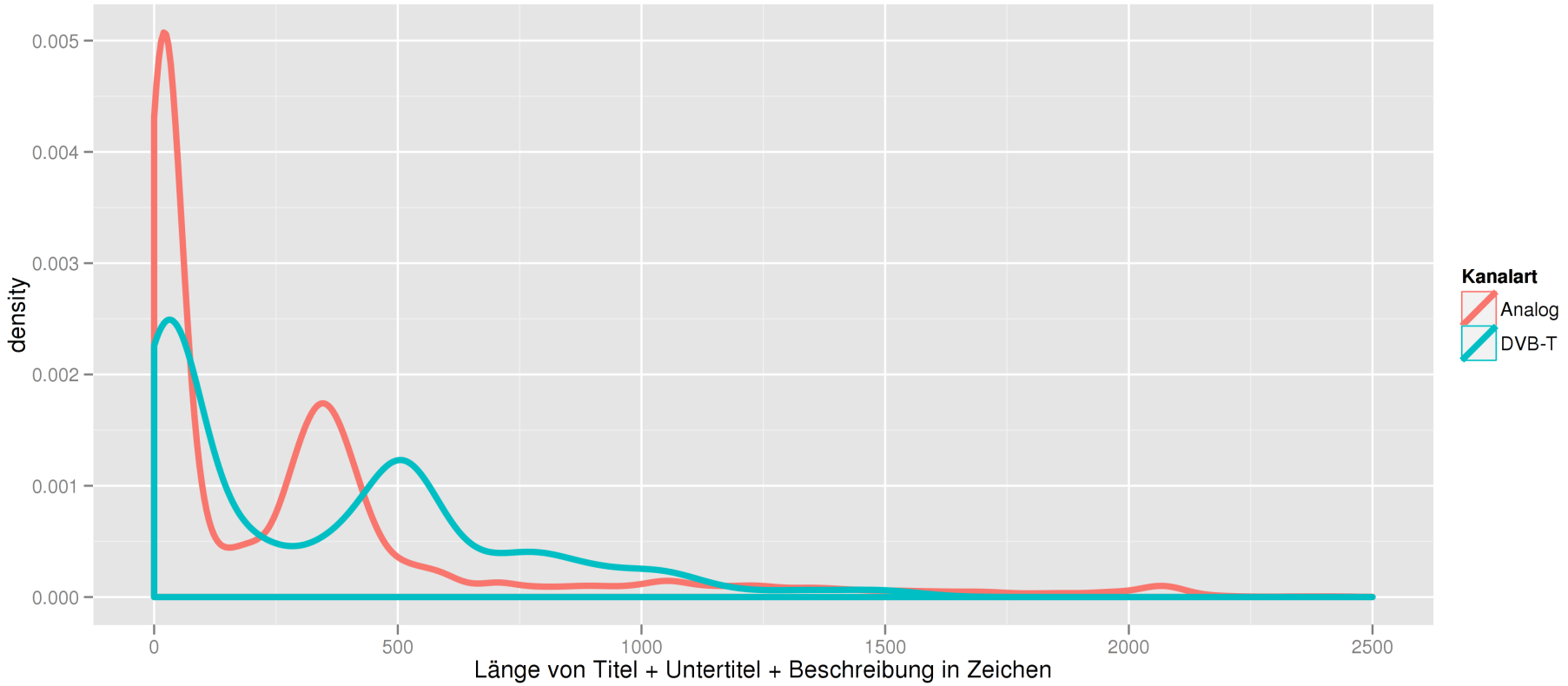
Herausforderungen

- Aus schlechten Trainingsdaten schlechte Ergebnisse
  - Aber es kann schwer sein, gute Trainingsdaten zu bekommen!
- Wörter in den Beschreibungen nicht das einzige Kriterium
  - Eine Krankenhausserie holt andere nach
  - Serien mit allgemeinem Inhalt verwässern die Ergebnisse

Herausforderungen

- Aus schlechten Trainingsdaten schlechte Ergebnisse
  - Aber es kann schwer sein, gute Trainingsdaten zu bekommen!
- Wörter in den Beschreibungen nicht das einzige Kriterium
  - Eine Krankenhausserie holt andere nach
  - Serien mit allgemeinem Inhalt verwässern die Ergebnisse
- Unterschiedliche Daten bei verschiedenen EPG-Quellen

Beschreibungslänge



(X-Achse auf 2.500 Zeichen eingeschränkt; einige Beschreibungen haben bis zu 3.600 Zeichen)

Möglichkeiten für Weiterentwicklung

- Verbesserung der Ergebnisse
  - Ungenutzte Sender automatisch ignorieren?

Möglichkeiten für Weiterentwicklung

- Verbesserung der Ergebnisse
  - Ungenutzte Sender automatisch ignorieren?
  - Bessere Erkennung von Wiederholungen und Mehrfach-Ausstrahlungen

Möglichkeiten für Weiterentwicklung

- Verbesserung der Ergebnisse
  - Ungenutzte Sender automatisch ignorieren?
  - Bessere Erkennung von Wiederholungen und Mehrfach-Ausstrahlungen
- Modularisierung
  - Benutzerkonfiguration von der Prozessdatei getrennt

Möglichkeiten für Weiterentwicklung

- Verbesserung der Ergebnisse
  - Ungenutzte Sender automatisch ignorieren?
  - Bessere Erkennung von Wiederholungen und Mehrfach-Ausstrahlungen
- Modularisierung
  - Benutzerkonfiguration von der Prozessdatei getrennt
  - Anwendung auf die Daten anderer Unterhaltungssysteme (Windows Media Center, Dreambox, VDR)

Möglichkeiten für Weiterentwicklung

- Verbesserung der Ergebnisse
  - Ungenutzte Sender automatisch ignorieren?
  - Bessere Erkennung von Wiederholungen und Mehrfach-Ausstrahlungen
- Modularisierung
  - Benutzerkonfiguration von der Prozessdatei getrennt
  - Anwendung auf die Daten anderer Unterhaltungssysteme (Windows Media Center, Dreambox, VDR)
  - Parameteroptimierung beim Enduser

Möglichkeiten für Weiterentwicklung

- Verbesserung der Ergebnisse
  - Ungenutzte Sender automatisch ignorieren?
  - Bessere Erkennung von Wiederholungen und Mehrfach-Ausstrahlungen
- Modularisierung
  - Benutzerkonfiguration von der Prozessdatei getrennt
  - Anwendung auf die Daten anderer Unterhaltungssysteme (Windows Media Center, Dreambox, VDR)
  - Parameteroptimierung beim Enduser
- Mehr Information
  - „Top keywords"

Schluß

- Fragen?
- http://tud.at/programm/mythminer/
- mailto:balazs@tud.at