jWebMiner: A Web-Based Feature Extractor

Cory McKay and Ichiro Fujinaga

Social Sciences and Humanities Research Council of Canada | McGill | Schulich School of Music / École de musique Schulich | CIRMMT Centre for Interdisciplinary Research in Music Media and Technology | jMIR

Overview

jWebMiner is a software package for extracting cultural features from the web. At its most basic level, it operates by using web services to extract hit counts for sets of query strings from search engines. It then processes these hit counts to calculate feature values denoting classifications or similarity measurements. jWebMiner emphasizes extensibility, generality and an easy-to-use interface.

Cross-Tabulation Extraction

One of the two main types of feature extraction that can be performed by jWebMiner is "cross tabulation extraction." This involves finding relatively how often each query string in one set of query strings appears on the same web page as each string in another set of query strings. This is useful for classifying instances from one list into classes denoted in the other list, as shown in Figures 1 and 2:

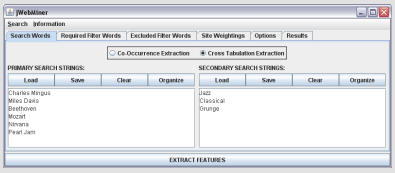
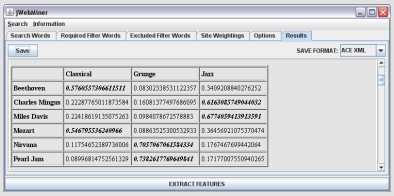


Figure 1: Query terms used for a sample cross tabulation extraction that will classify the six musical artists amongst the three musical genres.



	Classical	Grunge	Jazz
Beethoven	**0.576055730661511**	0.083023385311223 57	0.340920884027625 2
Charles Mingus	0.222877450118735 84	0.160813774976860 95	**0.616398574944032**
Miles Davis	0.224186191359752 63	0.098407867257888 3	**0.677405943913591**
Mozart	**0.546795536249966**	0.088635253005329 33	0.364569210753704 74
Nirvana	0.117546523897360 86	**0.785706706258433 4**	0.176746769442064
Pearl Jam	0.089968147525613 28	**0.758261776960451 1**	0.171770075509402 65

Figure 2: The results for the queries shown in Figure 1. The figures represent the normalized weighted relative frequencies with which each artist appears on the same web page with each genre. The highest value for each artist is in bold.

Synonyms

Multiple terms are often used to refer to the same entity. For example, "Charles Mingus" and "Charlie Mingus" are often used interchangeably. jWebMiner therefore makes it possible to define synonyms so that hits for corresponding synonyms are combined.

Co-Occurrence Extraction

The other main type of feature extraction that can be performed by jWebMiner is "co-occurrence extraction." This involves finding relatively how often each query string in a set of query strings appears on the same web page as every other query string in the same set. This is useful for similarity measurement, as demonstrated in Figures 3 and 4:

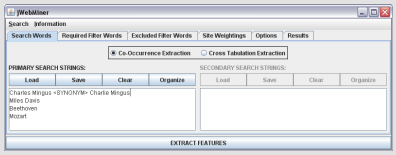
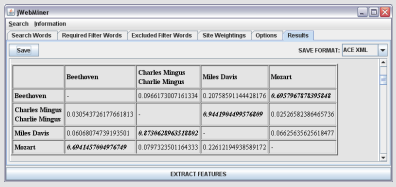


Figure 3: Query terms used for a sample co-occurrence feature extraction that will measure the similarity between each of the four musical artists. A synonym for "Charles Mingus" is included.



	Beethoven	Charles Mingus Charlie Mingus	Miles Davis	Mozart
Beethoven	-	0.0966173007161334	0.207585914442817 6	**0.495796787835284 8**
Charles Mingus Charlie Mingus	0.0305437261776618 13	-	**0.944394449957609 9**	0.025265813864657 36
Miles Davis	0.0606807473919350 1	**0.573662956331889 2**	-	0.066256356256184 77
Mozart	**0.494165790497674 9**	0.0797323501164333	0.226121949385891 72	-

Figure 4: The results for the queries shown in Figure 3. The figures represent the normalized weighted relative frequencies with which each artist appears on the same web page with each other artist. The highest value for each artist's row is in bold.

Required String Filters

There is usually some feature noisiness due to hits not related to the topic being studied. There are many Hinduism-related sites that contain the string "Nirvana," for example, but that have nothing to do with the band. jWebMiner allows users to define required string filters (e.g., "music") that must be found on a site in addition to the query strings for it to be counted as a hit.

Excluded String Filters

Excluded string filters offer another way for jWebMiner users to reduce feature noisiness. There are user-definable strings that may not appear on a web site if it is to be included in feature counts. For example, one might wish to exclude sites containing the words "paradise" or "zen" in order to avoid false hits relating to the band "Nirvana."

Site Weightings

Another way to reduce feature noisiness is to limit searches to sites that are known to be relevant to music, such as the All Music Guide, Pitchfork, etc. jWebMiner therefore allows users to assign relative weightings to both the Internet as a whole and to individual sites. These weightings control the effect that hits from certain sites have relative to hits from other sites on the final feature calculations.

Additional Options

jWebMiner offers a variety of additional ways that users can customize their feature extractions. Users may choose among a variety of formulas and normalization schemes for calculating feature values. Users may also choose to use either Google, Yahoo! or both to perform searches, and may configure these engines using a variety of options, as shown in Figure 5:

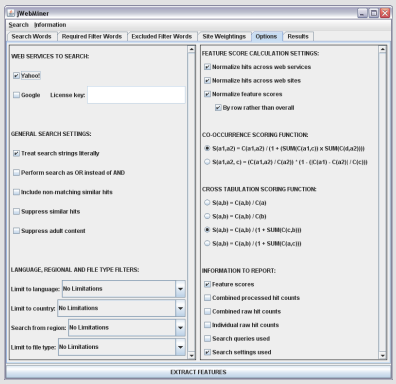


Figure 5: Additional ways of customizing jWebMiner feature extractions.

Usability & Extensibility

jWebMiner includes a simple and intuitive GUI as well as a full user manual. The software is open source and the Java code is well-documented and implemented using a modular plug-in design to facilitate the addition of new functionality such as additional web services and feature calculation schemes. jWebMiner can parse query terms from text, iTunes XML, ACE XML, or Weka ARFF files. Feature values can be saved as text, ACE XML, Weka ARFF or HTML files. jWebMiner is one component of the jMIR software suite, which includes audio and symbolic feature extractors and machine learning software.