JWEBMINER: A WEB-BASED FEATURE EXTRACTOR

Cory McKay **Ichiro Fujinaga**
Music Technology Area and CIRMMT, Schulich School of Music, McGill University
Montreal, Quebec, Canada
cory.mckay@mail.mcgill.ca, ich@music.mcgill.ca

ABSTRACT

jWebMiner is a software package for extracting cultural features from the web. It is designed to be used for arbitrary types of MIR research, either as a stand-alone application or as part of the jMIR suite. It emphasizes extensibility, generality and an easy-to-use interface.

At its most basic level, the software operates by using web services to extract hit counts from search engines. Functionality is available for calculating a variety of statistical features based on these counts, for variably weighting web sites or limiting searches only to particular sites, for excluding hits that do not contain particular filter terms, for defining synonym relationships between certain search strings, and for applying a number of additional search configurations.

1. JMIR AND GENERAL-PURPOSE MIR TOOLS

Many MIR research areas are strongly dependent upon the central tasks of extracting features and applying classification algorithms to them. Examples include music recommendation, playlist generation, performer or composer identification, genre classification, instrument identification, and many others. The jMIR software suite has been developed as an integrated set of tools for general-purpose MIR research in these areas. jMIR is designed to provide an open framework that encourages collaborative research and sharing of algorithms. Ease of use for researchers with varying technical backgrounds is a central priority, and all components include well-documented GUIs. The jMIR components are all open source and implemented in Java, with a plugin-based architecture that emphasizes extensibility.

One of the key goals of jMIR is the facilitation of research that combines low-level features (based on basic signal processing and human physiology), high-level features (based on musical abstractions) and cultural features (based on sociocultural information outside the scope of musical content itself). Each of these feature types encompasses potentially significantly different information, with the implication that combining them could improve the performance of many areas of MIR research. This supposition has been supported by experimental gains in performance when low-level and cultural features have been combined in the past [8].

jMIR therefore includes, among other components, a low-level feature extractor for processing audio files (jAudio [4]), a high-level feature extractor for processing MIDI files (jSymbolic [5]), and a web-based cultural feature extractor (jWebMiner, the subject of this paper).

2. CULTURAL FEATURES AND THE WEB

There is psychological and musicological reason to believe that cultural factors beyond the content of music itself play an essential role in how humans interpret and organize music. North and Hargreaves, for example, found experimentally that the style of a piece can influence listeners' liking for it more than the piece itself [6], and Fabbri has argued that content-based technical and formal aspects of music represent only one of five ways in which musical genres can be characterized [2].

The web offers a valuable source of information from which cultural features can be extracted. Data mined from the web can also be useful in acquiring ground truth for use in training and evaluating MIR systems.

A number of important MIR studies have been published experimentally investigating co-occurrence analysis of web data (e.g., [1], [3], [7], [8], [9]). jWebMiner, however, is the first out-of-the-box cultural feature extractor designed for general-purpose MIR research.

3. OVERVIEW OF JWEBMINER

jWebMiner is a software package for extracting cultural features from the web using web services for use in a variety of MIR research areas. At its most basic level, jWebMiner operates by accessing search engines to acquire hit counts for various search strings. For example, calculations involving how often the names of different musicians co-occur on the same web pages (compared to how often they occur individually) can provide insights on the relative similarity of the musicians to each other. Similarly, the cross tabulation of song names and musical genres can be used to classify music by genre. Such basic hit counts can result in noisy results, however, so it is necessary to include additional functionality.

jWebMiner begins by parsing either iTunes XML, ACE XML, Weka ARFF or text files in order to acquire strings to use in searches. Users may also manually enter search strings in the GUI (see Figure 1). The software then accesses the web to either measure the co-occurrence of each value in one field with other values in the same field, or to measure the cross tabulation of values in different fields.

Research has indicated (e.g., [3], [7]) that the best choice of statistical procedure for processing hit counts can vary depending on the task at hand. For example, one must consider not only the accuracy of an approach, but also its search complexity, as web services typically involve daily limits on queries. jWebMiner therefore allows users to choose between a variety of metrics and scoring systems to base features upon.

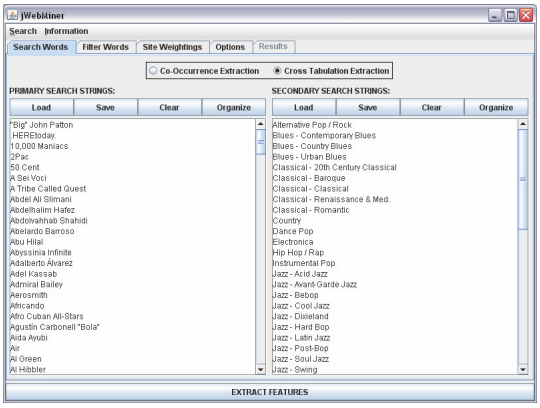

Figure 1. jWebMiner's GUI.

Users can specify string synonyms so that hit counts will be combined for linked synonyms. This could be useful, for example, in a genre classification task where the class names "R&B" and "RnB" are equivalent.

jWebMiner also allows user-definable filter strings. The software can be set to ignore all web pages that do not contain general filter terms such as "music," for example, or application-specific terms such as "genre" or "mood." This can be useful in avoiding irrelevant and noisy hit counts. For instance, a feature extraction should not count co-occurrences of "The Doors" with "Metal" or "Rap" unless they refer to music rather than the building industry or door knockers.

It is also possible to set jWebMiner to limit searches to particular sites, such as the All Music Guide, Pitchfork, etc. in order to emphasize musically relevant and reliable sites. jWebMiner also allows users to assign varying weights to particular sites as well as to the web as a whole when feature values are calculated.

jWebMiner outputs feature values as ACM XML, Weka ARFF or delimited text files. Feature values may also be accessed directly via the GUI.

4. JWEBMINER AND WEB SERVICES

jWebMiner utilizes web services to extract features from the web using, currently, either Yahoo! (via REST) or Google (via SOAP). The included Yahoo! license allows 5000 queries per day per IP address. Users must provide their own Google SOAP API License, however, as Google ceased releasing new licenses in 2006.

jWebMiner's API includes an extensible plugin interface that facilitates the addition of further web service resources in the future. A highly configurable search dialog box is also available for use in debugging new implementations and comparing specific results from different web services side by side.

5. CONCLUSIONS AND FUTURE RESEARCH

jWebMiner is a general-purpose tool for easily extracting cultural features from the web. The flexibility of the interface and the extensibility of the API encourage experimentation with various techniques. Extracted features can be used directly or combined with other feature types extracted with software such as jAudio and jSymbolic. Future research will focus on analyzing the actual content of web sites as well as utilizing additional web service resources, such as Audioscrobbler and Amazon.

jWebMiner and the other components of jMIR are available at http://jmir.sourceforge.net.

6. ACKNOWLEDGEMENTS

We would like to thank Mark Zadel for making the details of his research available. We would also like to thank the SSHRC and the Centre for Interdisciplinary Research in Music Media and Technology for their generous financial support.

7. REFERENCES

[1] Ellis, D. P. W., B. Whitman, A. Berenzweig, and S. Lawrence. 2002. The quest for ground truth in musical artist similarity. *Proceedings of the International Conference on Music Information Retrieval.* 170–7.

[2] Fabbri, F. 1981. A theory of musical genres: Two applications. In *Popular Music Perspectives,* D. Horn and P. Tagg, eds. Göteborg: IASPM.

[3] Geleijnse, G., and J. Korst. 2006. Web-based artist categorization. *Proceedings of the International Conference on Music Information Retrieval.* 266–71.

[4] McEnnis, D., C. McKay, and I. Fujinaga. 2006. jAudio: Additions and improvements. *Proceedings of the International Conference on Music Information Retrieval.* 385–6.

[5] McKay, C., and I. Fujinaga. 2006. jSymbolic: A feature extractor for MIDI files. *Proceedings of the International Computer Music Conference.* 302–5.

[6] North, A. C., and D. J. Hargreaves. 1997. Liking for musical styles. *Music Scientae* 1: 109–28.

[7] Schedl, M., T. Pohle, P. Knees, and G. Widmer. 2006. Assigning and visualizing music genres by web-based co-occurrence analysis. *Proceedings of the International Conference on Music Information Retrieval.* 260–5.

[8] Whitman, B., and P. Smaragdis. 2002. Combining musical and cultural features for intelligent style detection. *Proceedings of the International Conference on Music Information Retrieval.* 47–52.

[9] Zadel, M., and I. Fujinaga. 2004. Web services for music information retrieval. *Proceedings of the International Conference on Music Information Retrieval.* 478–83.