Search Sounds: An audio crawler focused on weblogs

Òscar Celma
Music Technology Group
Universitat Pompeu Fabra
Pg. Circumval.lació 8, 08003
Barcelona, SPAIN
ocelma@iua.upf.edu

Pedro Cano
Music Technology Group
Universitat Pompeu Fabra
Pg. Circumval.lació 8, 08003
Barcelona, SPAIN
pcano@iua.upf.edu

Perfecto Herrera
Music Technology Group
Universitat Pompeu Fabra
Pg. Circumval.lació 8, 08003
Barcelona, SPAIN
pherrera@iua.upf.edu

Abstract

In this paper we present a focused audio crawler that mines audio weblogs (MP3 blogs). This source of semi-structured information contains links to audio files, plus some textual information that is referring to the media file. A retrieval system —that exploits the mined data— fetches relevant audio files related to user's text query. Based on these results, the user can navigate and discover new music by means of content-based audio similarity. The system is available at: **http://www.searchsounds.net**.

Keywords: focused audio crawler, weblogs, music recommendation, content-based similarity.

1. Introduction

In recent years the typical music consumption behaviour has changed dramatically. Personal music collections have grown favoured by technological improvements in networks, storage, portability of devices and Internet services. In the context of the World Wide Web, the increasing amount of available music makes very difficult, to the user, to find music he/she would like to listen to. To overcome this problem, there are some audio search engines [1] that can fit the user's needs [1]. Some of these search engines are nevertheless not fully exploited because their companies would have to deal with copyright infringing material.

2. Syndication of Web Content

During the last years, syndication of web content —a section of a website made available for other sites to use— has become a common practice for websites. This originated with news and weblog sites, but nowadays is increasingly used to syndicate any kind of information. Since the beginning of 2003, a special type of weblog, named audio weblogs (or MP3 blogs), has become very popular. These blogs make music titles available for download. The music posted is explained by the blog author, and usually it has links that allow to buy the complete album or work. Sometimes, the music posted is hard to find or has not been issued in many years, and many MP3 blogs link strictly to music that is authorized for free distribution. In other cases, MP3 blogs include a disclaimer stating that they are willing to remove music if the copyright owner objects. Anyway, this source of semi-structured information is a jewel for web crawlers, as it contains the user's object of desire —e.g. an audio file—, and some textual information that is referring to the object.

2.1. File formats

The file format used to syndicate web content is XML. The XML description to use is defined in the RSS (and Atom) family. The RSS abbreviation is variously to refer to the following standards: Really Simple Syndication (RSS 2.0), Rich Site Summary (RSS 0.91 and 1.0) or RDF Site Summary (1.0).

2.2. Multimedia syndication

Of special interest are the feeds that syndicate multimedia content. These feeds publish audiovisual information that is available on the net. An interesting example is the Media RSS (mRSS) specification [2], lead by Yahoo! and the multimedia RSS community. mRSS allows to syndicating multimedia files (audio, video, image) in RSS feeds, and adds several enhancements to RSS enclosures. Although mRSS is not yet widely used on the net, there are some websites that syndicates their multimedia content following the specification [3].

3. Crawling RSS feeds

The implemented audio crawler starts the process from a manually selected list of RSS links (MP3 blogs). Each RSS file contains a list of entries (or *items*). The crawler seeks for new incoming items —using the $pubDate$ item value

[1] To mention a few (accessed on June, 1st, 2006):
http://search.singingfish.com/,
http://audio.search.yahoo.com/,
http://www.audio crawler.com/,
http://www.alltheweb.com/?cat=mp3 and
http://www.altavista.com/audio/

[2] http://search.yahoo.com/mrss/

[3] One of the most important ones is http://www.ourmedia.org

and comparing with the latest entry in the database— and stores the new information into the database. Thus, the audio crawler system has an historic information of all the items that appeared in a feed.

4. Audio Retrieval System

The logical view of a feed item can be described by the bag-of-words approach: a document is represented as a number of unique words, with a weight assigned to each word [2]. Special weights are assigned to the music related terms, as well as the metadata (e.g ID3 tags) extracted from the audio file. Similar to our approach, [3] presents a proposal of modifying the weights of the terms pertaining to the musical domain. A more sophisticated method based on unsupervised learning of text profiles for music (from unstructured data crawled from the web) is explained in [4]

Moreover, basic natural language processing methods are applied to reduce the size of the document (elimination of stopwords, and apply Porter's stemming algorithm [5]). The information retrieval (IR) model used is the classic vector model approach, where a given document is represented as a vector in a multidimensional space of words (each word of the vocabulary is a coordinate in the space).

4.1. Full text search

The similarity function, $sim(d_j, q)$, between a query (q) and a document (d_j) is based on the TF/IDF weightning function, where TF (term frequency) denotes the frequency of the word t in the document d_j, and IDF (inverse document frequency) measures the general importance of t in the overall collection [2]:

$$sim(d_j, q) \sim \sum_{t \in q} TF_{t,j}/|\vec{d_j}| \cdot IDF_t \qquad (1)$$

This IR model is well suited not only for querying via artists' or songs' names, but for more complex text queries such as: "funky guitar riffs" or "traditional Irish tunes".

The retrieval system outputs the documents (i.e. feed entries) that are relevant to the user's query, ranked by the similarity function. All the audio links posted in the entry are displayed too, so the user can listen to the audio files associated to that entry.

4.2. Content based similarity

Based on the results obtained from the user's textual query, the system allows to find similar audio files by means of content-based audio similarity. Each link to an audio file has a "Similar sounds" button that retrieves the most similar audio files, based on a set of mid-level audio descriptors. These descriptors are extracted from the audio and represent properties such as: rhythm, harmony, timbre and instrumentation, intensity, structure and complexity [6].

This exploration (or browsing) mode allows to the user to discover new music, related to her original (text-based) query, that would be more difficult to discover by using textual queries only. To our knowledge, nowadays, this is the only web-based audio search engine that allows this type of content-based navigation. There is an a analogy between this type of navigation and, for example, Google's "find web pages that are similar to a given HTML page". In our case, similarity among items are based on audio similarity, whereas Google approach is based on the content of the HTML page. Both browsing approaches are, then, based on the *content* analysis of the retrieved object.

5. Conclusions and Pending Work

We have presented an audio crawler focused on weblogs that publish music related information. Out of the crawling process, each feed item is represented as a text document, containing the item content, as well as the links to the audio files. Then, a classic text retrieval system outputs relevant items related to the user's query. Moreover, a content-based navigation allows to browse among the retrieved items and discover new music and artists by means of audio similarity.

An interesting remaining task could be a set of experiments addressing the interaction between content-based and text-based querying, and how users differently employ them. Finally, a relevance feedback method to tune the system and get more accurate results (specially for the content-based navigation) should be taken into account.

6. Acknowledgements

We are very greatful for the help of Markus *Koppi* Koppenberger, Nicolas Falquet and Xavier Oliver in the design and implementation of the system.

References

[1] I. Knopke. *AROOOGA: An Audio Search Engine for the World Wide Web*, in Proceedings of the International Computer Music Conference, 2004, Barcelona, Spain.

[2] R. Baeza-Yates, *Modern Information Retrieval*, Addison-Wesley, 1999.

[3] S. Vembu and S. Baumann, *A Self-Organizing Map Based Knowledge Discovery for Music Recommendation Systems*, Computer Music Modeling and Retrieval, 2004, Esbjerg, Denmark.

[4] B. Whitman and S. Lawrence. *Inferring Descriptions and Similarity for Music from Community Metadata*, in Proceedings of the 2002 International Computer Music Conference, 2002, Goteborg, Sweden.

[5] M. F. Porter, (1980). *An Algorithm for Suffix Stripping*. Program 14, 130–137.

[6] P. Herrera et al. *SIMAC: Semantic Interaction with Music Audio Content*, in Proceedings of the 2nd European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies, 2005, Savoy Place, London.