This project involves the development of a Music Information Storage & Retrieval system that can provide the user with a robust and informative experience when listening to her digital music. Imagine an interface to one’s digital music collection that provides rich and detailed metadata content, album reviews, interviews, discographies, and biographies on the band members. All of this data is available for free on the Internet. Our challenge is to collect this data, generate metadata about it, and use this metadata to power a user interface that enables a user to explore her collection in novel ways. A major milestone in this project is to provide a graph or network representation of a user’s music collection and related artists based on data culled from the web and matched against a collection of ID3 tags extracted from the user’s digital music collection. We would like to provide the user with a dynamic graph visualization, where the vertices are artists and the edges represent relations among artists. We also intend to provide a similar visualization based on one’s social network. That is, if a user has identified other users of the system as ‘friends’, the application will be able to create graphs of the interrelation of artists across music collections.

A relation may include links like ‘played in,’ ‘toured with,’ ‘collaborated with,’ ‘sounds similar to,’ ‘produced by,’ etc. We hope to build these relations by mining and manipulating several web-based sources. We plan to make some simple inferences on relations using NLP machine learning techniques, and we also plan to mine more structured sources of data. In a nutshell, we seek to build a stand alone MP3 player that provides a rich and dynamic interface for exploring a large music space. Our ultimate goal is to build recommendations into this system, so that music that might be of interest to the user can be highlighted within the graph visualization. Do to scope issues however, our primary focus is on collecting the data, and building an application that makes use of this information to create meaningful and helpful visualizations to the user.

This project has several components, described below.

Web Crawling and Building a Named Entity Dictionary

We are presently building a set of tuned web crawlers to collect data from websites such as epitonic.com, pitchforkmedia.com and tinymixtapes.org. The former two sites have HTML formatting that allows us to extract named entities with ease and a high degree of confidence. Having collected a list of named entities, in this case artists and albums, we then feed each entry back into a web crawler that selectively queries various sites for the text of record reviews, biographies and interviews. Currently we
have a dictionary of roughly 8,000 albums and 5,000 artists. The process of collecting the text about these entities is presently underway.

Natural Language Processing and Building Relations

Once we retrieve the text, we will pass the text through a set of NLP modules that will first try to match terms in the text against our database of named entities, and then attempt make some inferences about how the named entity is related to the artist being discussed in the text. We recognize that this is a difficult task, and we do not attempt to build a complex machine learning system to identify all of possible relations among named entities. Instead, we plan to build a set of fairly simple NLP modules capable of assigning default categories to relations. We can then improve upon manually and by mining more structured data – such as discographies and band information pages that explicitly describe relations among artists. Natural Language Processing will also help automate or speed up many of the tasks related to data collection and organization. NLP is an important portion of this project, but the focus is not to build an application solely on the quality of our NLP algorithms. Instead we seek to use some NLP techniques to automate some of the metadata generation necessary for this application to work.

System Architecture, Data Modeling and Storage

We have designed this system to be a client/server architecture, with a central repository of information that can communicate with multiple clients to provide content. The data model for the central repository needs to support large volumes of structured and unstructured text, possibly complex numbers, file links to media assets, and the ability to retrieve information from multiple tables in a very efficient manner. We will have data from the following sources: online music magazines, user information including preferences and ID3 tags from their digital assets, user playlists, digital radio playlists and music blogs. Much of the actual text will be stored in a file system, but all metadata we generate will be stored in normalized relational format. The data are generally organized according to the source with relations and overlaps where necessary to minimize redundancy.

We are prototyping the application in mySQL, although a later iteration will likely require a more robust database that can support sub-queries and stored procedures. A detailed data model is currently being designed, and we have tables to support record reviews, playlists, user profiles, and metadata annotation.

The client application will be written in Java, and will communicate with the server over HTTP. Another data-modeling task is to determine a messaging format for this communication.
User Interface Design and Information Visualization

There are interface design and information visualization challenges inherent in this project. The quantity of information we would like to make available to the user necessitates a smart and efficient design. The variation in content, including photographs, categories of text, liner notes, etc., requires careful planning so as to not overwhelm the user. Building the networks, and managing all of the interactions between the graph and the underlying text is a complicated task, and much of the project time is dedicated to getting this component working.

A Novel Recommendation System

A feature we hope to include in this project is a customizable recommendation system that will use our knowledge of relations among artists and friends to offer new or overlooked music, using the same visualization scheme described above. However, the prerequisite to offering this functionality is having the data and the framework for building visualizations in the client application. As such, we plan to build the application with a novel recommendation system in mind, and evaluate the feasibility of developing this component once we are satisfied that all of the other components are under control.

Motivation

This work is motivated by a strong interest in music, and the challenges of information retrieval in the music domain. The research area of Music Information Retrieval is an emerging field, and it brings together many of the challenges we have encountered in the course of the M.I.M.S curriculum. Namely, the need for a semantic and contextual understanding of content, the need to index and retrieve this information in a meaningful way, and the development of an application that can present this information to the user in a meaningful way.

By leveraging community metadata to describe digital content, and by developing algorithms to generate metadata automatically, as well as interactively, we will be applying many of the concepts and approaches we have learned in the last 1 1/2 years. Additionally, we seek to present this research to the MIR community and ultimately publish our work, thus contributing to the community of researchers interested in this topic.

Roles of the Project Members

The project members, Brooke Maury, Vijay Viswanathan and Jeannie Yang will work closely on all elements of this project. However, each member brings their own skills to the tasks at hand. Brooke will provide database development expertise and programming skills. His primary foci will be system architecture, backend system development, database development and, reluctantly, project management.
Vijay’s primary focus will be on the design and development of the user interface. Jeannie will focus on the backend system code, the crawlers, and the communication b/w server and client.

Our advisor, Marti Hearst, brings expansive expertise and guidance to this project, in the areas of Information Visualization, Natural Language Processing, User Interface design and well-informed approaches to application development and research.

Because this project necessarily involves interaction with the MIR community, we also hope to seek guidance from two researchers in the field of MIR research: Jonathan Foote and Brian Whitman.