ECU178 Computer Science:
220CT Data and Information Retrieval
Case Study
Time to dig deeper!

Due on March 20th 2015

Robert Rigler : 4939377

Contents

Data Mining Definitions

Forecasting

Definition:

Forecasting is the act of predicting a future activity by the analysis of past and present data trends. This is done via the visualisation of relevant data-set, so to derive patterns and

Regression

Definition:

Regression is the application of certain equations to a data-set in order to better model the data.

For example applying the equation of a straight line ($y = mx + c$) to a data-set and seeing if it forms a correlation to a straight line, then future values of y and x can be predicted.

Time Series

Definition:

A time series is a collection of data that has been collected at regular, sequential intervals over a specific period of time. Time-Series analysis is the use of statistical techniques to model and derive patterns from that data set.
Time-series can be applied in many different fields, but most notably:

- Financial:
  Forecasting inflation and stock prices.
- Scientific:
  Forecasting results from experiments.

Associtaiton

Definition:

Defining a set of rules, that imply a relationship(an association) between data items.
By analysis of a dataset, certain relationships will be discovered. Using these relations, rules can be made which can help the forecasting of future data.

Sequencing

Definition:

The act of analysing a dataset for sequences of action, and then from this data find the set of most frequent sequences.

Data Mining and Music

Data-Mining and music are not regularly linked together when studying either of the topics. But with digital music distribution (via download and streaming) being the most lucrative method of distribution.

Digital Music is a very popular business, a lot of money is made via advertising and downloads. To keep customers' loyalty , a recommendation system is often used to notify the users of other music that they might be interested in. The more music that the user is aware of, the greater potential profit increases. Currently, data-mining in music is most used in music streaming, where users can create their own 'Radio Station'. The user chooses an artist or a genre and the service will then play songs that are similar to the song or match the given criteria, the user can then 'like' or 'dislike' the current song, and the service will improve the station based on that information.

How?

Firstly, the data is gathered. This comes in many formats:
Meta data: Name, Genre, Label, Sales.
Associations: The various relationships between artists, genres and labels.
User Data: Although this is less relevant, because individual tastes can not be necessarily applied to others especially in music.

After all the relevant data has been gathered the data is then grouped into similarities and then analysed. The data can be made into a graph and then similarity relationships can be put into place. The relationships point to other nodes on the graph (called neighbours). The graph can then be traversed so similar music can be found.

The individual data items could also be represented as a matrix which could be populated by how many times two song were recommended, or played sequentially or appeared together. From the matrix a distance function (Regression) could be used to analyse how similar those two items are.
A similar approach would be to model the user as a vector. The music played by this user influence the direction and magnitude of this vector, which can then be used to predict similarities.