Music Data Mining Based on MusicXML

Gen-Fang Chen
Department of Music Engineering, Zhejiang Conservatory of Music, Hangzhou 310024, China
genfangchen@163.com

Abstract. In this paper, we use the music resources of the Internet to analyze and mine the data of MusicXML file. First, the structure of MusicXML file is introduced briefly, then the data source is explained, then the database is established by MySQL and PHP, and finally the data is analyzed by MATLAB. The experimental results show that the music group is centered on the group of small characters in which the center C is located, showing normal distribution, while the music group is centered on the group of small characters in G, showing normal distribution.

1. Introduction
MusicXML is a digital sheet music interchange and distribution format. The goal is to create a universal format for common Western music notation, similar to the role that the MP3 format serves for recorded music. The musical information is designed to be usable by notation programs, sequencers and other performance programs, music education programs, and music databases[1].

Data mining is the process of knowledge discovery. Through the analysis of large-scale data, people can use a certain pattern model to find out the correlation among many data, to show the distribution of data and reflect the statistical characteristics of data. In the MusicXML file, there are a lot of music data, which represent different music knowledge, such as music score, beat, tonality, melody, passage, harmony. This paper will use data mining technology to search music tags in musicxml files, and then use visualization technology to reflect the statistical characteristics and distribution of music data.

2. Music Data Source --Finale's MusicXML File
The data source for the MusicXML file is MakeMusic's website (www.finalemusic.com)[2] Finale demo software (version No.: 25.3.0.276) is a free trial software provided. In this trial software, there are some music score examples for users to use. They are scattered in the directories "tutorials" and "worksheets and". In repertoire, there are 1236 music score examples in the two directories.

The sub-directories under these two directories have a six layer structure. The deepest directory is located in the "worksheets and repertoire- > repertoire- > holiday patriotic - > instrumental - > suite from nuttracker - > nuttracker solos - > nuttracker bass". Here, 610 music score examples are selected from "tutorials" and "worksheets and repertoire". The file names of these 610 music score examples can be downloaded from reference [3]. They are distributed in different directories. The number of files in different directories is shown in table 4-6. The number of files selected under different directories is different, among which the classical piano music is the most, with 57 pieces.

These files are music score examples of finale demo software. Their storage format is finale software's proprietary digital music format. The suffix of the file is ". Musx". The ". Musx" file format
needs to be converted into a general "XML" file format, so as to become MusicXML music file. The conversion tool is to use the functions provided by finale software.

In the process of data mining, the initial preparation of data often takes up about 70% of the workload of the whole data mining project. After preparing the MusicXML digital music file according to the above process, you can enter the next step of mining modeling and visualization process. Because the data mining itself integrates the comprehensive knowledge of statistics, database, machine learning and other disciplines, here we use MySQL database technology, statistical methods, PHP programming language and visual tools provided by MATLAB to mine the music information and graphical analysis results of MusicXML digital music files.

3. Music Information Mining Based on Tags

3.1. MusicXML Tags
MusicXML file is an XML format file based on user tags, the number of tags reaches 682 in MusicXML 3.0. To facilitate data analysis, we use these 682 tags to mine the information of the 610 MusicXML digital music files mentioned above. In order to store the mining results, we use SQL statements to create MySQL data table, which is used to store the content of each tag mined from these files and the number of tags.

3.2. Some Results
We use MySQL management tools to run SQL statements for statistical analysis of data table. Some of the results are as follows:
(a) There are 56897 records in the data table, and the label of "name" field of each record appears at least once in the file of this record.
(b) The total number of time value tags in all 610 MusicXML files is 227528.
(c) There are only 235 different tags in 682 tags of MusicXML3.0. The total number of tags is 2185606, and the average number of tags is 9300.4511.
(d) In all 610 MusicXML files, the file "Beethoven Polonaise.xml" The total number of occurrences of all tags contained in is the highest, 17760.
(e) In the MusicXML file "Mozart Eine Kleine Nachtmusik.xml" The number of tags "duration" contained in is 4183, which is the most frequent tag in all 610 MusicXML files.

Through observation, it is found that some tags appear the same times in 610 MusicXML files, indicating that they appear in combination form in MusicXML files. These tags and their total times are shown in Table 1. Obviously, it is more likely to appear in combination form with more tags with the same total times, and vice versa.

| Total | Tag name | Total | Tag name |
|-------|----------|-------|----------|
| 227528| duration | 85468 | staff |
| 190320| note     | 54181 | notations |
| 177539| octave   | 49944 | alter |
| 169451| pitch    | 36170 | measure |
| 167322| voice    | 27473 | chord |
| 162894| type     | 22361 | slur |
| 154724| step     | 21897 | articulations |
| 154052| stem     | 19480 | rest |
| 140918| beam     | 15360 | dot |
Table 2 Statistics of common music symbols in 610 MusicXML music files

| music notation | Number of files | Total | music notation | Number of files | Total |
|----------------|----------------|-------|----------------|----------------|-------|
| clef           | 610            | 1317  | pitch          | 572            | 169451|
| key            | 605            | 1052  | rest           | 535            | 19480 |
| mode           | 603            | 1030  | alter          | 535            | 49944 |
| time           | 542            | 2118  | voice          | 585            | 167322|

Some common music symbols will be frequently used in MusicXML files. Here, SQL statements are designed on MySQL platform to count the number of tags such as clef, key, tonality, beat, pitch note, rest, up and down sign and strength.

According to table 2, the music score label appears in every MusicXML music file, the pitch note label has the most times in all the tags, and the mode label has the least times.

Table 3 total number of occurrence of groups and names in 610 MusicXML music files

| Pitch name | Total  | Octave number | Total |
|------------|--------|---------------|-------|
| C          | 25619  | c8–b8         | 4     |
| D          | 24625  | c7–b7         | 57    |
| E          | 22416  | c6–b6         | 2739  |
| F          | 22814  | c5–b5         | 42643 |
| G          | 26650  | c4–b4         | 71849 |
| A          | 23402  | c3–b3         | 37848 |
| B          | 23932  | c2–b2         | 13622 |
|            |        | c1–b1         | 695   |
|            |        | c0–b0         | 1     |

3.3. Visualization of Analysis Results

Visualization is a graphics and image technology that displays the results of data mining. It transforms data into graphics or images and displays them on the screen, and can process them interactively. It is a comprehensive technology to study a series of problems such as data representation, data processing, decision analysis, etc. Visualization can be used to show the change trend of data when the data is clearly understood by users, which is convenient for decision-making. Sometimes, when there are
many data, tables, etc., which cannot show the change trend of data, they can be represented by graphics.

In Figure 1, the abscissa is the serial number of 235 labels, and the ordinate is the total number of times of each label. It can be seen from the figure that the change trend of the number of times is obvious. When the serial number is less than 20 or so, the curve declines rapidly. When the serial number is greater than 20, the curve decreases steadily.

![Figure 1](image1)

**Fig.2 The ratio of the frequency of each voice group to the total frequency and the standard normal distribution curve**

Figure 2 is divided into upper and lower subgraphs. The upper subgraph shows the occurrence probability of 9 groups and the corresponding normal distribution curve. The lower graph shows the occurrence probability of 63 international standard tones and the corresponding normal distribution curve. The deviation of the normal distribution curve of the upper subgraph is 5, and the standard deviation is 1, indicating that it is the standard normal distribution curve. The deviation of the normal distribution curve of the lower graph is 32, indicating that it is 63 musical tones. The middle note.

Observe the subgraph in Figure 2, the one-lined octave, two-lined octave and three-lined octave are very close to the function value corresponding to the standard normal distribution curve, other octaves also fluctuates not far up and down the curve. It can be concluded that the probability distribution of the groups in all music scores is in accordance with the standard normal distribution. Looking at the subgraph, the most frequent of all tones is “g” tone of one-lined octave, and the ratio of “g” tone and all music tones higher than “g” tone is very close to the normal distribution curve.

4. Conclusion

In this paper, we use the music resources of the Internet to analyze and mine the data of MusicXML file. First, the structure of MusicXML file is introduced briefly, then the data source is explained, then the database is established by MySQL and PHP, and finally the data is analyzed by MATLAB. The experimental results show that the music group is centered on the group of small characters in which the center C is located, showing normal distribution, while the music group is centered on the group of small characters in G, showing normal distribution.

Acknowledgments

This work was financially supported by the Humanities and Social Sciences Foundation of the Ministry of Education in China (19YJAZH004) and the cultural innovation team of Zhejiang Province.
References
[1] https://www.MusicXML.com/
[2] www.finalemusic.com.
[3] http://www.musicdata.cn/htdocs/data/FinaleMusicXMLFiles/finaleFileListNoZero.csv