PEACHNOTE: MUSIC SCORE SEARCH AND ANALYSIS PLATFORM

Vladimir Viro
Ludwig-Maximilians-University Munich

ABSTRACT

Hundreds of thousands of music scores are being digitized by libraries all over the world. In contrast to books, they generally remain inaccessible for content-based retrieval and algorithmic analysis. There is no analogue to Google Books for music scores, and there exist no large corpora of symbolic music data that would empower musicology in the way large text corpora are empowering computational linguistics, sociology, history, and other humanities that have printed word as their major source of evidence about their research subjects. We want to help change that. In this paper we present the first result of our work in this direction - the Music Ngram Viewer and search engine, an analog of Google Books Ngram Viewer and Google Books search for music scores.

1. INTRODUCTION

This project seeks to do for music scores what Google Books Search does for books. We are aiming at indexing all scanned music scores and making their content available for querying and algorithmic analysis. We would like to help build up the foundation needed for computational musicology research by assembling a large corpus of symbolic music data.

We have developed a search engine and processing pipeline for scores from the Petrucci Music Library (IMSLP, http://imslp.org), the largest music score library on the Internet. Our system takes the scores in PDF format, runs optical music recognition (OMR) software over them, indexes the data and makes them accessible for querying and data mining. The search engine is built upon Hadoop and HBase and runs on a cluster. Our system has already recognized more than 250 million notes from about 650 thousand sheets, or 45 thousand scores.

We chose the Petrucci library as our first data source because of the low entry barrier: both the scores and their scans at the IMSLP are free from copyright, and so we were free to use them without asking for permission. Therefore at the beginning of the development it was the easiest collection to work with. But the Petrucci Library contains only a small part of all scores digitized by the libraries worldwide. We would like to help libraries not only make their score collections searchable, but also to present them in novel ways. In this paper we present one such interface - the Music Ngram Viewer and search engine.

The paper is structured as follows. First, we provide a short review the related work in the areas of symbolic music corpora and music search engines. Then we introduce our search engine and analysis platform, describe its architecture and talk about the data collected so far. The next section presents the application built on top of the platform, the Music Ngram Viewer and search engine. We provide some statistics collected during the first three months after the public launch of the Ngram Viewer. This section is followed by a short conclusion.

2. RELATED WORK

2.1 Music data collections

Existing corpora of symbolic music data vary in size and quality. Probably the largest collection is the Kunst der Fuge collection with about 18,000 MIDI files (mostly piano works or reductions) contributed by the Internet users. A comparably large collection can be accessed via the search engine at Musipedia.com, although the data set is not available for download or purchase. A collection from the Center for Computer Assisted Research in the Humanities at Stanford University is of excellent quality, containing complete orchestral scores in MusicXML format, but is comparably small with 880 manually encoded compositions in 4116 movements. It also provides a search interface for the collected data, the Themefinder. The online version of Barlow and Morgenstern's Dictionary of Musical Themes contains 9,825 monophonic melodies of a few measures length.

2.2 Search engines and interfaces

Two existing systems are most relevant for our work: the Musipedia search engine and the Probado project.

Musipedia offers multiple querying interfaces: query by humming, virtual keyboard, search by rhythm and by typed

in melody. The database behind Musipedia is assembled from different MIDI and MusicXML collections. Most music is either composed or transcribed for piano, and there are few orchestral scores in the system.

The Probado project offers a very advanced interface for simultaneously browsing the scores and the audio recordings aligned to them (cf. [2], [3]). The scores have been recognized using the SharpEye OMR software.

3. SEARCH ENGINE AND ANALYSIS PLATFORM

3.1 System architecture

Our system consists of two major components: the frontend and the backend.

The backend is responsible for importing, processing and indexing the scores and the metadata. For importing and preprocessing the scores we use a cluster of Linux machines. The workflow relies on Amazon's Simple Queue Service for passing tasks between different processing steps.

We have implemented wrappers for various optical music recognition systems: an open source Java-based Audiveris, and the proprietary Windows-based and GUI-only SharpEye, CapellaScan and Smartscore. For the GUI-only OMR systems we implemented wrapper scripts that allow us to integrate these systems into the recognition workflow while running inside the VMWare virtual machines. After evaluating these OMR systems in our environment we came to the conclusion that Smartscore currently offers the best recognition rates among the four systems we tested, and so the majority of the scores in our database are recognized using Smartscore 10.3.2.

The workflow components responsible for indexing and metadata processing are running in the Hadoop and HBase environments [8]. The frontend presenting the processed data is hosted on Google's App Engine.

Using HBase for data storage offers the advantage of built-in redundancy and compression. Currently, the inverse index of the ngram viewer and the search engine, which are described in the next section, uses 50 Gigabytes. Without compression, this number would be an order of magnitude higher. Another advantage of using Hadoop in the processing backend is the ability to scale it easily with various providers, like Amazon EC2 or supercomputing centers, which is beneficial for a research project.

Using Google App Engine for the frontend has the benefit of reliability, security and ease of development and deployment. In our setup we use the App Engine also as a caching layer for the Hadoop backend, where the bulk of the data is stored.

3.2 Data

Currently the search engine contains the data from the Petrucci Music Library. The system has already recognized more than 1,000,000 sheets from more than 65,000 scores. Here are some occurrence counts of musical symbols recognized by the system. The database contains 264M notes, 45M measures, 3.7M keys, 2.8M parts, 630K staves, 52K trill marks and 23 fffff signs. The following figure contains the occurrence counts of piano signs:

```
p  1808243
pp  403366
ppp  20945
pppp  1024
ppppp  10
pppppp  2
```

Figure 1. Counts of piano signs in the IMSLP scores recognized so far.

4. MUSIC NGRAM VIEWER AND SEARCH ENGINE

Inspired by the Google Books Ngram Viewer [1], we implemented a similar application for music scores on top of our platform. We extracted the score metadata provided by the users of Petrucci Music Library from the web site. For all scores with available date of composition or at least of first publication (about two thirds of all scores), for all voices we extracted all melodies of up to fifteen notes length. Chords were represented as rising note sequences. Then, for each year we stored the occurrence counts of melodies that occurred three or more times in scores published or composed during that year. We published our system at www.peachnote.com. We also provided the dataset behind the Ngram Viewer under the Creative Commons Attribution license. As far as we know, these are the first publicly available system and dataset of the kind.

4.1 User Input

Users can use virtual piano keyboard implemented in Flash to enter their queries. In the current version query are sequences of pitches. The note duration is not considered.

4.2 Ngram Viewer

Currently the database contains ngrams up to the length 15, or melodies of up to sixteen notes. If a melody occurs in some year more than two times, it is stored in the database. This results in approximately 200 million ngram-year records in the database.

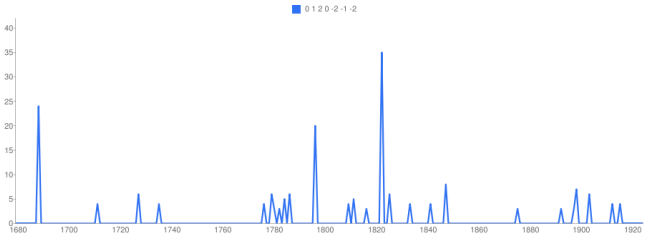
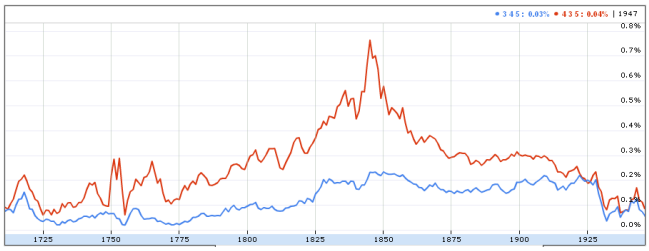


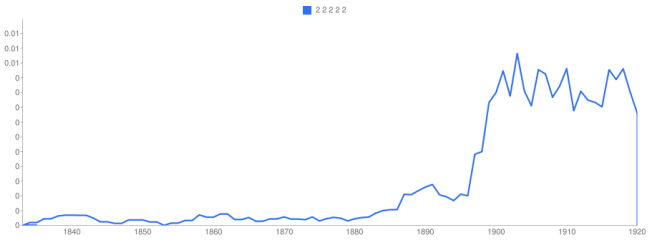
Figure 2. Occurrences of the Ode to Joy motif

The above chart shows the occurrences of the Ode to Joy motif from Beethoven's Ninth Symphony, encoded differentially (the numbers represent differences between consequent notes) - a 7-gram, "0 1 2 0 -2 -1 -2". What the y-axis shows is this: of all the 7-grams contained in the OMR'ed scores from IMSLP, the Petrucci Music Library, how many are identical with the first 8 notes of Ode to Joy up to a patch shift? Here, you can observe a peak around 1822 - the year of the Ninth's composition. Apparently, the score of the Ninth symphony contains most occurrences of this pattern. It is interesting to learn what the other peaks are. Our search engine described in the next section provides an answer to this question.

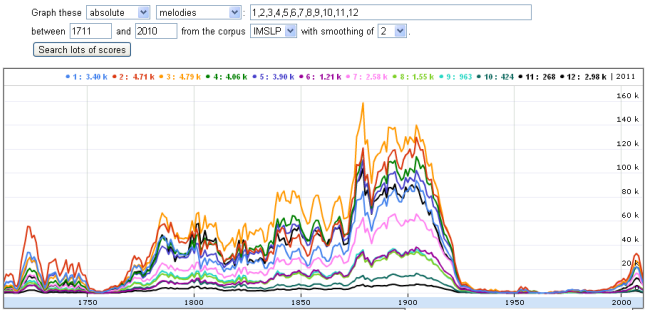
The next graph shows the frequency of occurrence of major and minor chords:



The following graph shows the emergence of the whole-tone scale at the turn of the 20th century.

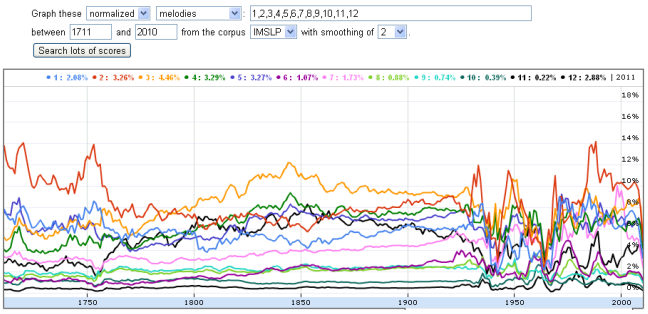


The graph below depicts the number of occurrences of twelve intervals from the minor second to the octave in our database, by year:



The gap between 1925 and 2000 is due to scores still being under copyright protection and hence unavailable on IMSLP. Modern composers, however, are free to upload their own compositions, and indeed they do so, as the bump on the right tells.

The next figure shows the data for the same time frame and same intervals, but this time it is normalized by the total number of notes published in a given year and stored in our database.



The more scores we have for any given year, the more reliable are the statistics.

4.3 Search Engine

For each ngram which is stored in the Ngram Viewer dataset, we also provide the information about the scores containing the given sequence. Using the dynamic ngram chart users can select the time range and get the list of scores composed during this time which contain the given note sequence. The list of compositions is paginated and sorted by the number of occurrences of the query in the scores. For each score we provide a list of pages containing the query. In future releases we will display the score sheets and highlight the locations of the queried note sequences. Also, for queries returning less than 10,000 scores we provide users the ability to filter the search results by text, using corresponding

tags provided by users of the IMSLP website. This way users can select pieces of particular genre (for example symphonies or quartets), participating instrument or instrument group (harp, winds), or composer.

4.4 Usage data

The system has been launched on May 5-th of this year, when the Petrucci library added the "Search by Melody" link on its home page. There has been a short announcement on the IMSLP Journal, but apart from that we have not promoted the search engine in any way, since we wanted to test it and improve its quality first. We installed Google Analytics to gain insights into our users' behavior. In the following we present a few data points we collected using Google Analytics.

In the first three months the system has been used by more than 50,000 people from over 160 countries. On average the search engine processed a search query every 5 seconds.

To see how the system has been used by people who are really interested in the insights it provides and to separate them from casual users, we looked at the statistics for visits with duration longer than 20 minutes. There have been 1385 such visits, and the average time on site was 60 minutes, which gives a total of 1385 hours of intensive research using the database. We also looked at the number of users who visited the website often. More than 1500 people used the system more than 10 times, 426 users visited the site more than 50 times, and 177 of them visited more than 100 times.

The files from the Ngram dataset have been downloaded more than 800 times.

5. CONCLUSION

In this paper we have presented a new music score search engine and analysis platform. The system opens new ways to explore notated music. The users can easily obtain insights that were hard to come by in the past. We also provide a large data set that can be used in computational musicology research. We continue digitizing score collections and will build additional search indexes that will allow more precise and musically meaningful queries.

6. REFERENCES

[1] Jean-Baptiste Michel, Yuan Kui Shen, Aviva P. Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden: Quantitative Analysis of Culture Using Millions of Digitized Books. Science 1199644 Published online 16 December 2010.

[2] Juergen Diet, Christian Goehlert: Innovative Erschliessung und Bereitstellung von Musikdokumenten im Probado-Projekt. Zeitschrift Forum Musikbibliothek, Vol. 30 No 3/2009.

[3] F. Kurth, D. Damm, C. Fremerey, M. Mueller and M. Clausen: A Framework for Managing Multimodal Digitized Music Collections. Proceedings of 12th European Conference on Research und Advanced Technology for Digital Libraries (ECDL 2008), Aarhus, Denmark, September 14-19, 2008.

[4] Akira Maezawa, Hiroshi G. Okuno, Tetsuya Ogata, Masataka Goto: Polyphonic Audio-to-Score Alignment Based on Baysian Latent Harmonic Allocation Hidden Markov Model. ICASSP 2011.

[5] M. Szwoch: Using MusicXML to evaluate accuracy of OMR systems. In Diagrammatic Representation and Inference: Proc. Diagrams 2008, volume 5223 of Lecture Notes in Computer Science, pages 419-422. Springer Verlag, Berlin. Herrsching, Germany, September 19-21, 2008.

[6] M. Droettboom, I. Fujinaga: Symbol-level groundtruthing environment for OMR. International Symposium on Music Information Retrieval. Barcelona, Spain. 2004.

[7] D. Byrd, W. Guerin, M. Schindele, I. Knopke, OMR Evaluation and Prospects for Improved OMR via Multiple Recognizers. 2009.

[8] Jeffrey Dean and Sanjay Ghemawat: MapReduce: simplified data processing on large clusters. Commun. ACM 51, 1 (January 2008), 107-113.

[9] Shyamala Doraisam: Polyphonic Music Retrieval: The N-gram Approach, PhD Thesis, Imperial College London, 2004.

[10] Shyamala Doraisamy and Stefan Rüger: Robust polyphonic music retrieval with n-grams. Journal of Intelligent Information Systems 21 (1): 5370. 2003.

[11] J. Stephen Downie: Evaluating a Simple Approach to Music Information Retrieval: Conceiving Melodic N-Grams as Text. Ph.D. Thesis, The University of Western Ontario. 1999.