USING THE MUSICBRAINZ DATABASE IN THE CLASSROOM

Cédric Mesnage
Southampton Solent University
United Kingdom

Abstract

Musicbrainz is a crowd-sourced database of music metadata. The level 6 class of Data Science teaches students about *big data*, using a large database, interacting with a database from a programming language, exporting data as XML, JSON or CSV, applying data mining algorithms and visualising data. This paper reports on using Musicbrainz in this class as well as the Orange educational tool. After a description of the class, this paper will give an evaluation based on students' questionnaires and a discussion on the experience and how to improve the class.

Introduction

Teaching and attending a class involving databases can be dull. Databases, data analysis and data science are important fields of computer science and often suffer from misconceptions as people feel they are boring, hard and not interesting when they are exciting, fun and essential. In the scope of the computing curriculum at Southampton Solent University is a level 6 class of Enterprise Data Modeling in which I chose to teach about data science and to use the Musicbrainz database to engage students better with the topic. Data science is a revolutionary field, which purpose is to make scientific discoveries out of data, and in which applications in business lead to improve existing businesses or find new strands to exploit. Through the class, students get to use the Orange educational data-mining tool, Postgres Musicbrainz database, migrate data to a MongoDB database (Docs.mongodb.com, 2017) and the Python programming language. The class is inspired by the work of Mesnage and Jazayeri (2008) in the context of the project based learning curriculum set in Jazayeri, (2004). This paper is structured as follows, a presentation of the Musicbrainz database, the Data Science class, a reflection on engagement with the class, challenges and technical settings.

Musicbrainz

Figure 1. MusicBrainz logo.

The Musicbrainz database is an open source music metadata database available at Musicbrainz.org. It is crowd sourced and is used by major digital music companies such as the BBC, iTunes and Amazon to identify artists, albums and tracks. Originally people submitted metadata when purchasing a new CD or other formats, and, playing it on their computers, they would enter metadata about the album, the artist, the list of tracks, the label and who published it. This data would be sent to the Musicbrainz server and related to previously entered data about the same elements. The database now stores and makes available metadata about more than a million artists and 16 million tracks from all other the world. It also contains links to music services and some information about genres as tags or concert and events. Musicbrainz is run by the Metabrainz foundation and supported by remote developers.

I have contacted the developers through their IRC chat and apart from the fact that they are very pleased for us to use their database, they also mentioned it has not been done before in a classroom.

Data Science Class

In the scope of the level 6 of the computing curriculum at Southampton Solent University, students can take the Data Science class as an option if they study the Web Development, Networking or Software Engineering course or have it as a core course if they follow the computing curriculum.

The class goal is to study data science, i.e., how to make scientific discoveries out of data. For this purpose the class starts with the current context, the world of big data. In fact it is predicted that by the year 2020 we will generate 40 Zettabytes of data per year, which converted in high definition video is equivalent to 4.5 billion years of video, the age of Earth to watch each year seems unfeasible. We study the history of database management systems going from mainframes to the more recent non-relational databases through the relational model as shown in Figure 2.

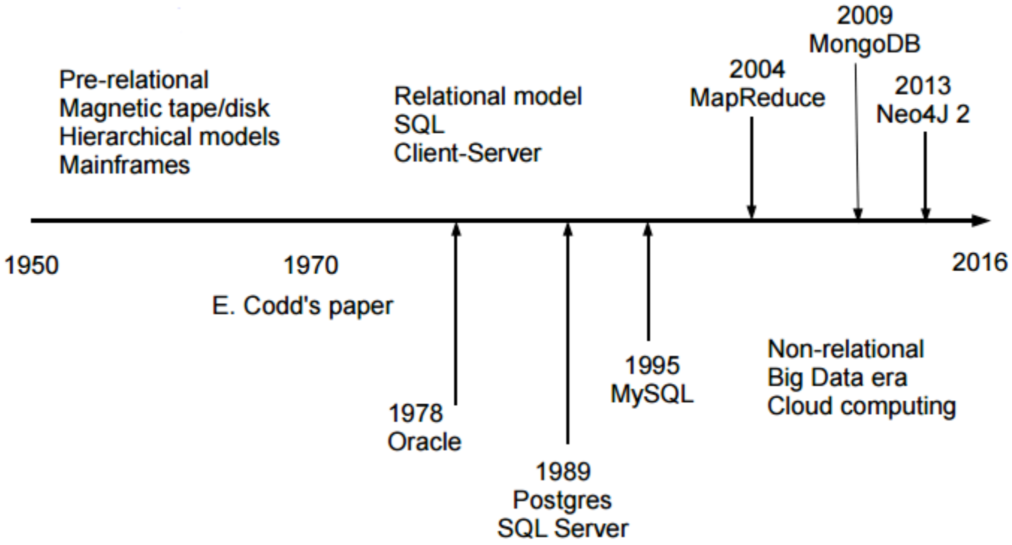

Figure 2. Evolution of database management systems.

As an example of a large database we use Musicbrainz; students get in touch with the database by first interacting with the website and searching for their favourite artist and albums. We then move on to actually connecting to the Postgres database (Postgresql.org, 2017) and students are set to write SQL queries about what they are interested in by looking at the schema in Figure 3.

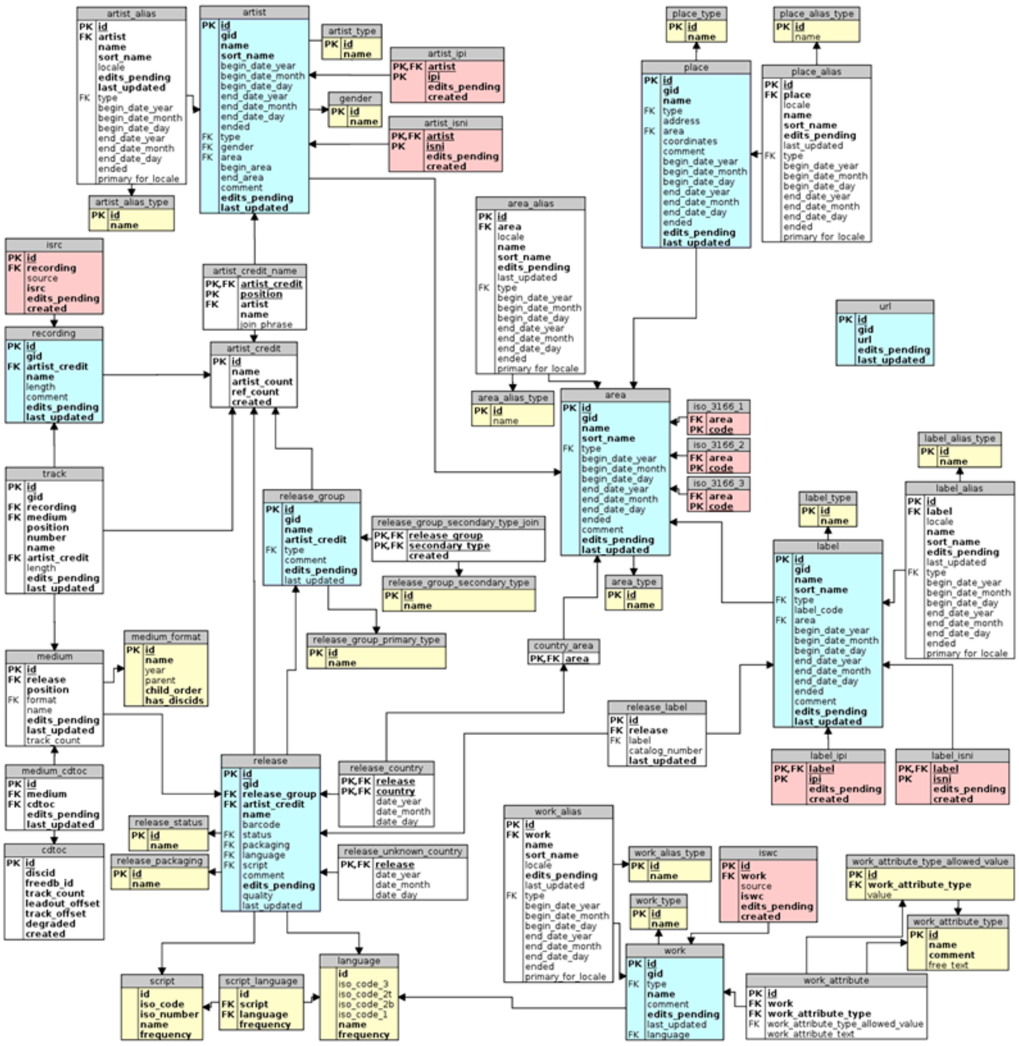

Figure 3. The MusicBrainz database schema.

The first assignment is to migrate parts of MusicBrainz to a MongoDB database. This is a case study of modernising a system as MongoDB is a recent database management system that enables easy distribution over multiple servers and facilitates retrieving data in now conventional formats such as JSON (JavaScript Object Notation). To perform the migration, students write Python programs.

The second part of the class focuses on the data science process as shown if Figure 4. Collecting data, pre-processing it, applying data mining algorithms and visualisations and interpreting them. The process is simplified, as it does not include hypothesis formulation nor hypothesis testing, verification or falsification. We study multiple data mining algorithms for clustering, prediction, classification, association rules, time series and basic statistics such as variables distributions. Since it is not the purpose of the paper we will not go into many details of this part of the class.

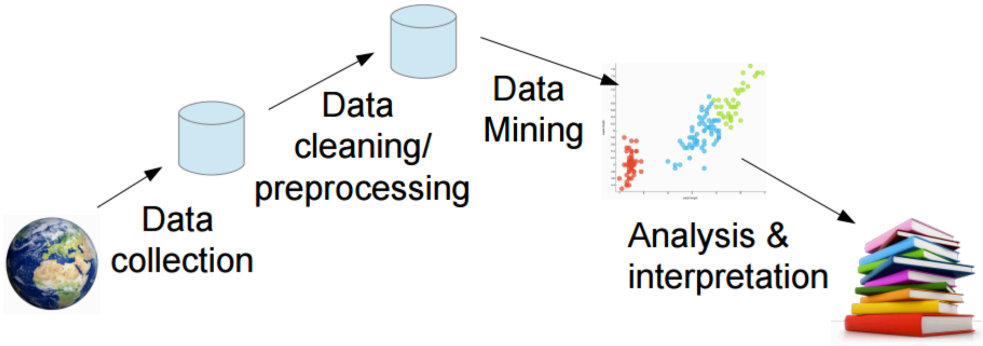

Figure 4. The data science process.

The second assignment of the class takes the students in extracting data from the Musicbrainz database, processing it and analysing it using the Orange open source data mining tool built by the bio lab of the University of Ljubljana (Bioinformatics Laboratory, 2017). It is an educational tool programmed in Python and on top of the Scipy, Numpy and Scikit learn scientific libraries. Figure 5 shows a scatterplot produced in Orange, which was part of the assignment.

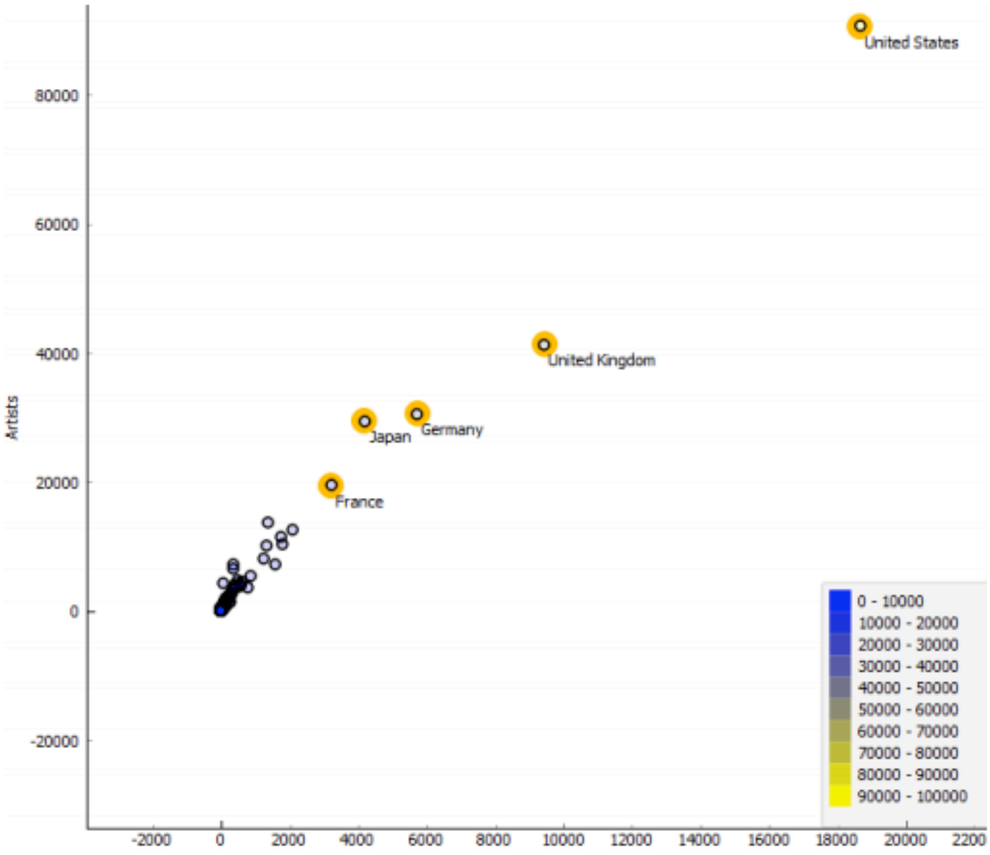

Figure 5. Example of a scatterplot produced by a student of the number of artists and music labels per country.

Engagement

I have not measured the student engagement with the class nor can I compare with previous instances since this was the first time I taught the class. I can transcribe my observations. Students engage better with Musicbrainz as the domain is music and most students are interested in music and have knowledge about it. Top students are engaged by the fact that this is a real database in use by popular services as shows the following student feedback.

Example of student feedback related to Musicbrainz:

> Going through the MusicBrainz and learning about that, as it was a real database used for professional applications so it was very interesting to look into. To sum it up, everything is recent and industry focused and that doesn't feel the same with the other units sometimes.

Students of a lower level are engaged as well, as they want to find data about their favourite artist, and I have observed a stronger interest from them in the class once we started using MusicBrainz.

I ran two surveys to get student feedback, one midway and one at the end of term. The only question from the survey that relates to engagement was “is the class interesting?” as shown in Figure 9. Out of the students who answered, 57% found the class interesting, and 89% did not find it not interesting. I believe this shows students engaged well with the class. The exit survey is not significant as only 4 students answered it.

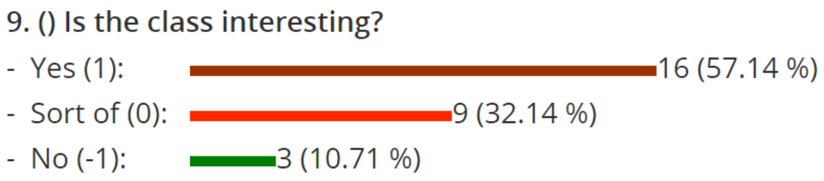

Figure 6. Result about engagement from the midway survey.

Challenges

It is challenging for students to work on such a large database schema as shown in Figure 3, grasping the semantics of more than 50 tables. The size of the data is a challenge as well; the database is 30GB, which forces the SQL queries to be not only correct but also efficient.

Another challenge in the class was to let students write queries about anything they want to look for, requiring them to be creative, which seems to be difficult for many of them who are used to having examples and more straightforward exercises to perform.

Most students passed the class, out of 37, 35 passed, and 2 are currently working on their retake. This is another measure of engagement, as to carry out the work of the assignment was challenging.

Technical settings

The Musicbrainz database is installed on an Ubuntu Linux virtual machine called Alexandria, which runs on the university servers. Students have a personal account on the machine and since using SSH (Secure Shell, which enables user to connect to a remote machine and run commands and programs on it) is new to most of them, this became part of the class as well.

One issue is that Alexandria is not accessible from outside the university for security reasons. The data-mining tool Orange is installed on the machines in the classroom where the class is held but not in the work areas of the university, which is a problem when students want to work outside of class.

Conclusion

In this paper we have seen the importance of using a database that engages students better with such a critical topic as data science. The Musicbrainz database is a large open database, which can be installed in any university on a Linux virtual machine, and students can connect to it with PostgreSQL and programming languages. Students found the class interesting in a feedback survey conducted within the class. Musicbrainz has gaps: it lacks data about events and venues, the genre information is very sparse, and the schema overly complex. We are developing a music database in house and might use it in the classroom as a replacement or together with Musicbrainz. The fact is music engaged students with a difficult topic, it would be interesting to experiment with other topics such as films (IMDB), video games (STEAM) or ultimately on anything they like.

Acknowledgements

I would like to thank the students who attended the class of Enterprise Data Modelling at Southampton Solent University the year of 2017, the previous unit leader for this class, Dr Jing Lu, for her input, the Musicbrainz developers, the Orange team and the Computing group of Southampton Solent University for giving me the opportunity to develop this class.

References

Bioinformatics Laboratory, University of Ljubljana. (2017). *Orange – Documentation* [online]. Docs.orange.biolab.si. Retrieved from http://docs.orange.biolab.si/

Docs.mongodb.com. (2017). *MongoDB Documentation* [online]. Retrieved from https://docs.mongodb.com/

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques* (3rd edition). Waltham, MA: Elsevier.

Harrison, G. (2015) *Next generation databases: NoSQL and big data*. 2015. New York, NY: APress.

Jazayeri, M. (2004). The education of a software engineer. *Proceedings of the 19th IEEE international conference on automated software engineering* (pp. 18-xxvii). Washington, DC: IEEE Computer Society.

Mesnage, C., & Jazayeri, M. (2008). Social thinking to design social software: A course experience report. *ASE'08 Proceedings of the 2008 23rd IEEE/ACM International Conference on Automated Software Engineering* (pp. 19-24). Washington, DC: IEEE Computer Society.

MusicBrainz. (2017). *MusicBrainz - The Open Music Encyclopedia* [online]. MetaBrainz Foundation. Retrieved from https://musicbrainz.org/

The PostgreSQL Global Development Group. (2017). *PostgreSQL: The world's most advanced open source database* [online]. Retrieved from http://www.postgresql.org/

Author Details

Cédric Mesnage
cedric.mesnage@solent.ac.uk