Mining a Bird Observatory Dataset

Mikko Koho

Technical report
UNIVERSITY OF HELSINKI
Department of Computer Science

Helsinki, May 18, 2015

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Laitos — Institution — Department
Faculty of Science	Department of Computer Science

Tekijä — Författare — Author

Mikko Koho

Työn nimi — Arbetets titel — Title

Mining a Bird Observatory Dataset

Oppiaine — Läroämne — Subject

Computer Science

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Technical report	May 18, 2015	6

Tiivistelmä — Referat — Abstract

Avainsanat — Nyckelord — Keywords

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Contents

1 Introduction

This technical report examines the use of data mining techniques [TSK+06] to find interesting patterns from a bird observatory dataset.

Bird species names are written written with finnish, english and scientific names, to be accessible to different audiences. The format used is <finnish name, *scientific name*, english name>.

2 Dataset

Hanko Bird Observatory, *Halias*, is located at the southernmost tip of Finland in Hanko, Finland. Halias has been gathering bird observation data from 1979 onward and it has been used intensively in research [Han14]. However data mining or machine learning methods have been never applied to it, so these methods could perhaps uncover some interesting patterns in the data.

Gathering the data at Halias is highly standardized and main focus is on counting the daily migration of each bird species. Manning the station is based on volunteers, which causes some gaps in the data when no observers are present. The dataset is not publicly available, but is available for research projects at request.

The data in original digital format consists of two Excel files, containing half a million rows of distinct daily counts per species for local and migrating birds from 1979 to 2009. Some rows contain a *taxon* (plural *taxa*) that is higher than species-level, for example a genus or a species pair.

However, in this case study we will be using a linked data publication made from the original dataset. The dataset has been transformed [KHL14, Koh15] to RDF data model and linked with weather data from nearby weather station in Hanko, Russarö and a bird taxon ontology. The linked dataset is structured using the RDF Data Cube Vocabulary [CR14], containing distinct data cubes for both daily bird observations and daily weather observations. An example of the data in this format is given in figure 1.

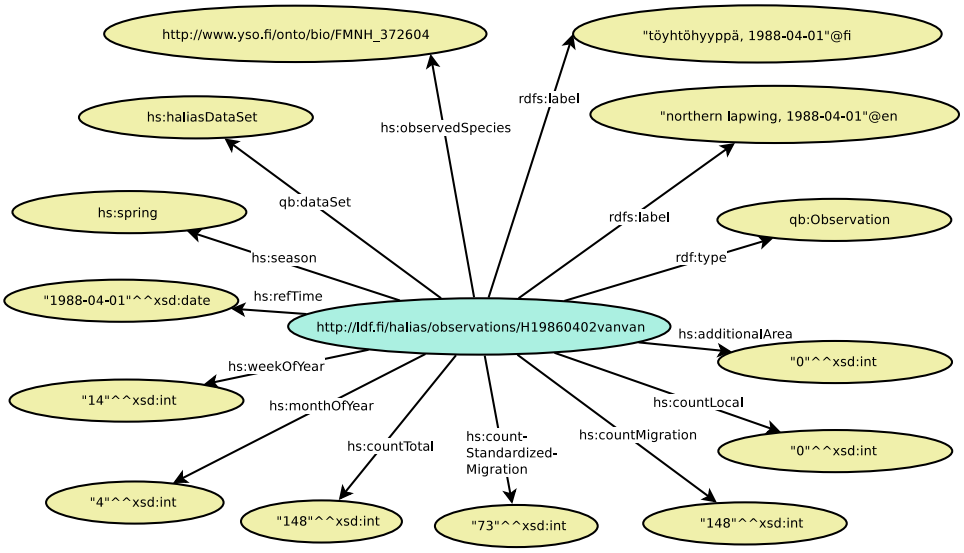

Figure 1: Daily observation data of one species from one day in RDF format [Koh15].

3 Methodology

For association analysis we take a subset of the dimensions of the data cubes and transform these to market basket transactions or sequence database. In this transformation it is easy to combine bird observations, weather observations and information from the used ontologies. The transformed data is stored in *JSON* or text files, which is then read in when doing analysis.

For simplicity, we first try to find interesting patterns using just the bird observations. We will do frequent pattern mining to find the most frequently observed species and species combinations, without using the daily counts. We could also use the observed counts and generate quantitative association rules, which might reveal some interesting patterns, but is probably very slow to calculate.

Next, we will try sequential pattern mining, using the timestamps already present in the data. This would probably reveal something interesting, but is again very slow to calculate.

The analysis is done using Python 3 and various modules. Plotting and

examining the data was done mainly with *IPython notebook*[1] and *pandas* (Python Data Analysis Library)[2]. Conversion from RDF uses *RDFLib*[3].

Pattern mining implementations are self-made[4]. They include algorithms for finding frequent itemsets, frequent sequential patterns and rule generation. The algorithms are based on the *Apriori algorithm* [TSK+06].

The pattern mining algorithms are somewhat memory efficient, but are computationally slow. They only use a single CPU core for all calculations and they use standard Python data types, which are flexible, but computationally inefficient. An attempt was made to optimize an Apriori based algorithm for speed, by first profiling the algorithm at execution time and finding the bottlenecks. The pruning step seems to be clearly the most computationally intensive part of the algorithm.

We attempted to optimize the pruning by using multiple processes, running in different CPU cores, using Python module Joblib[5], which provides an improved interface to Python standard library *multiprocessing* module. However, the processes running in parallel cannot access shared resources in memory, leading to extensive copying of large data structures and longer execution times. This could be overcome by switching to use data structures from NumPy[6], which are optimized for speed and allow shared access from multiple processes.

Orange Data Mining Toolbox[7] was also tried for analysis, but it wasn't able to cope with even simple market basket transaction data of all the daily observed taxa in the dataset (7419 days, 378 taxa).

4 Analysis results

The dataset is quite large for many association analysis tasks, which is why the focus has been on using simple methods.

[1] `http://ipython.org/notebook.html`
[2] `http://pandas.pydata.org/`
[3] `https://github.com/RDFLib/rdflib`
[4] `https://github.com/razz0/DataMiningProject`
[5] `https://pythonhosted.org/joblib/`
[6] `http://www.numpy.org/`
[7] `http://orange.biolab.si/`

4.1 Visualization

Visualizations of the data can easily convey important patterns in the data. 2

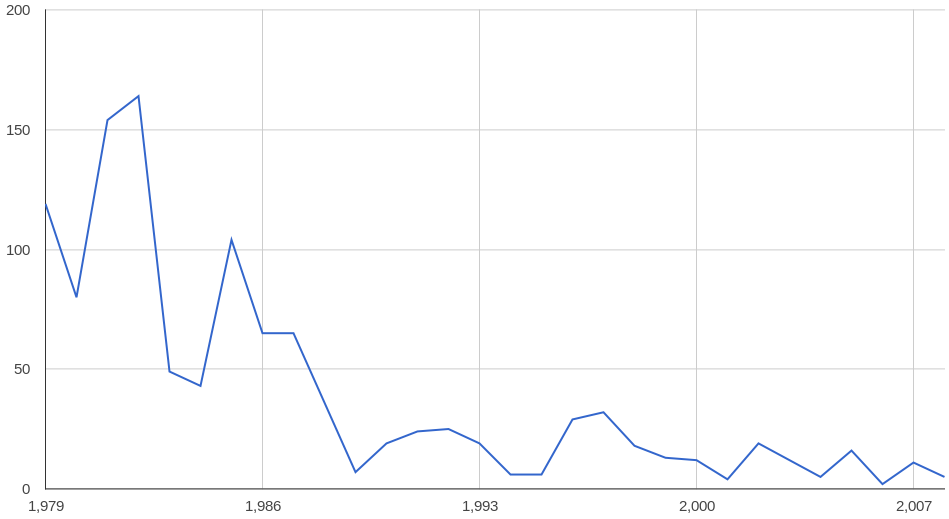

Figure 2: embhor years

Foo3.

Bar4.

Baz5.

4.2 Frequent itemsets

Analysing daily observed species as market basket transactions, we can examine the most commonly observed species and species combinations.

The most common species is <varis, *Corvus corone*, carrion crow> with support 0.951. For support 0.5, the largest frequent itemset consists of the following species:

<harmaalokki, *Larus argentatus*, herring gull>,

<isokoskelo, *Mergus merganser*, goosander/common merganser>,

<kalalokki, *Larus canus*, mew gull>,

<merilokki, *Larus marinus*, great black-backed gull>,

<sinisorsa, *Anas platyrhynchos*, mallard>,

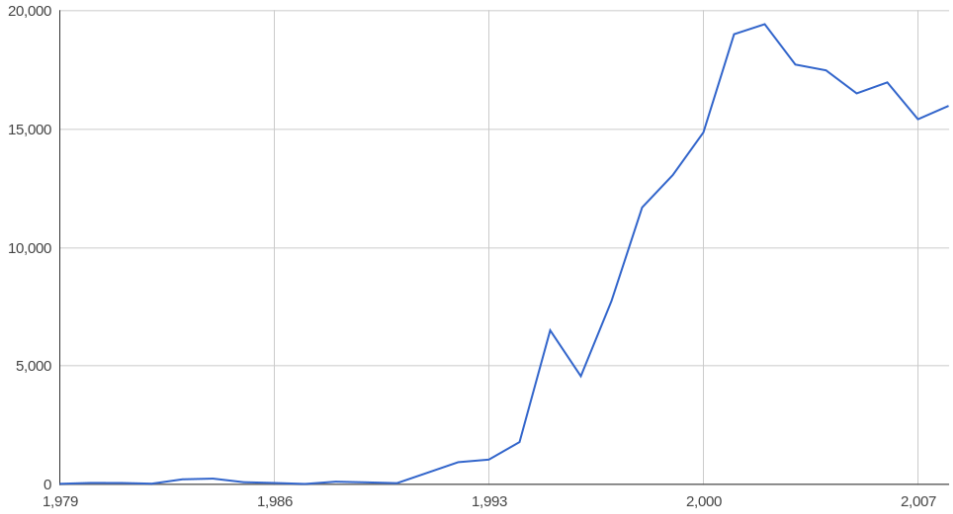

Figure 3: phacar years

<sinitiainen, *Parus caeruleus*, blue tit>,
<talitiainen, *Parus major*, great tit>,
<telkkä, *Bucephala clangula*, common goldeneye>,
<varis, *Corvus corone*, carrion crow>,
<viherpeippo, *Carduelis chloris*, european greenfinch>

This is the largest combination of species that is observed in half of the observation days. These species are very common and observable almost all year round. Individually all of the listed species have support over 0.75.

4.3 Frequent taxon sequences

4.4 Rule generation

We generated some rules from frequent itemsets...

5 Conclusions

6 Discussion

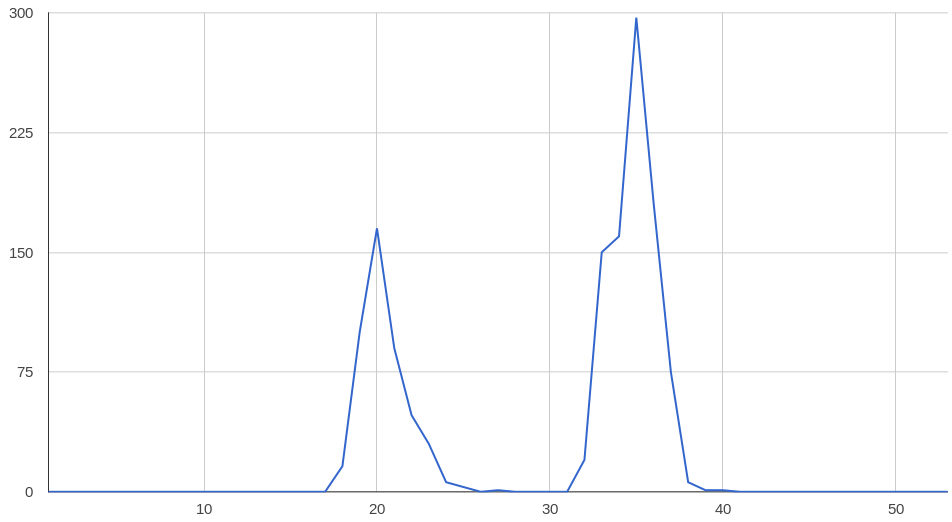

Figure 4: embhor weeks

References

[CR14] Cyganiak, R. and Reynolds, D.: *The RDF data cube vocabulary*. W3C recommendation, 2014. `http://www.w3.org/TR/2014/REC-vocab-data-cube-20140116/` [28.02.2014].

[Han14] *Hangon lintuaseman julkaisuluettelo 1980-*, 2014. `http://www.tringa.fi/hangon-lintuasema/julkaisut/` [5.05.2015].

[KHL14] Koho, M., Hyvönen, E., and Lehikoinen, A.: *Ornithology based on linking bird observations with weather data*. In *Proceedings of the 4th Workshop on Semantic Publishing (SePublica), ESWC 2014*. CEUR Workshop Proceedings, Heraklion, Kreikka, toukokuu 2014.

[Koh15] Koho, M.: *Linked data -palvelu luontohavaintoaineistoille*. Master's thesis, University of Helsinki, 2015.

[TSK+06] Tan, Pang Ning, Steinbach, Michael, Kumar, Vipin, *et al.*: *Introduction to data mining*, volume 1. Pearson Addison Wesley Boston, 2006.

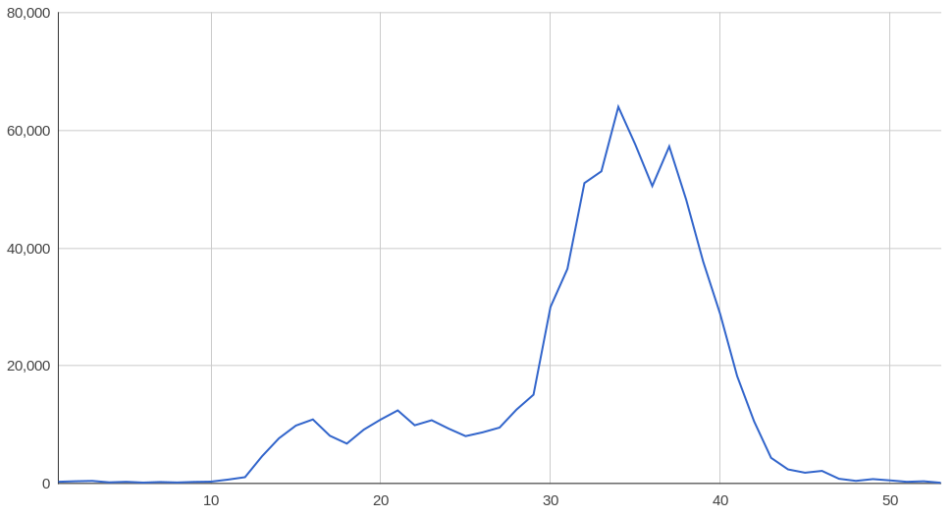

Figure 5: phacar weeks