Mining a Bird Observatory Dataset

Mikko Koho

Tiivistelmä — Referat — Abstract

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 Related work

3 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 bird observation data in this format is given in figure 1.

The used bird taxon ontology [Koh15] contains finnish conservation

statuses for endangered species, a coarse measure of commonness at Halias and characteristics of many finnish species, such as size, beak color and plumage coloring. The taxon ontology is open and available[1].
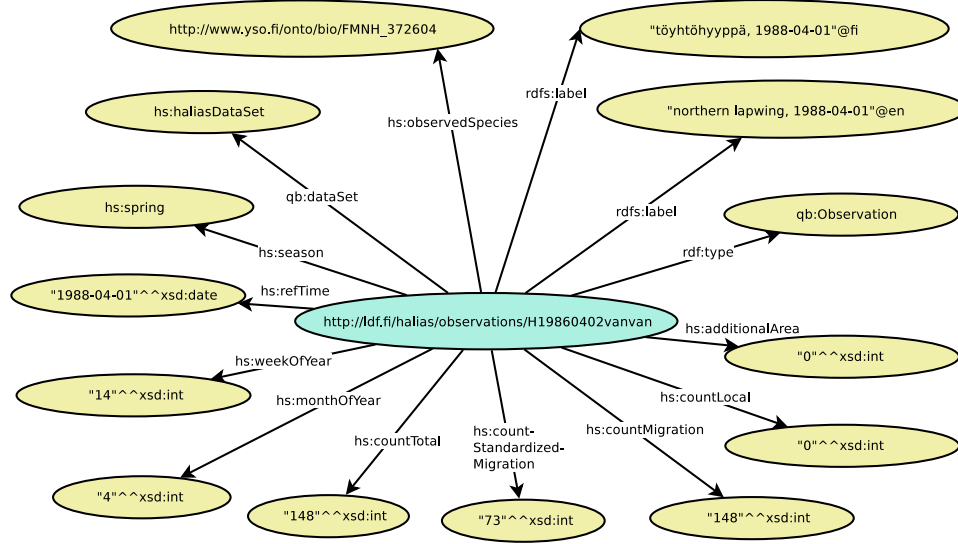

Figure 1: Daily observation data of one species from one day in RDF format [Koh15].

4 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

---

[1] http://www.ldf.fi/dataset/halias/rdf/halias_taxon_ontology.zip

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*[2] and *pandas* (Python Data Analysis Library)[3]. Conversion from RDF uses *RDFLib*[4].

Pattern mining implementations are self-made and available online[5]. 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[6], 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[7], which are optimized for speed and allow shared access from multiple processes.

Orange Data Mining Toolbox[8] was also tried for analysis, but it wasn't

---

[2]http://ipython.org/notebook.html

[3]http://pandas.pydata.org/

[4]https://github.com/RDFLib/rdflib

[5]https://github.com/razz0/DataMiningProject

[6]https://pythonhosted.org/joblib/

[7]http://www.numpy.org/

[8]http://orange.biolab.si/

able to cope with even simple market basket transaction data of all the daily observed taxa in the dataset (7419 days, 378 taxa).

5 Analysis results

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

5.1 Visualization

Visualizations of the data can easily convey important patterns in the data. 2
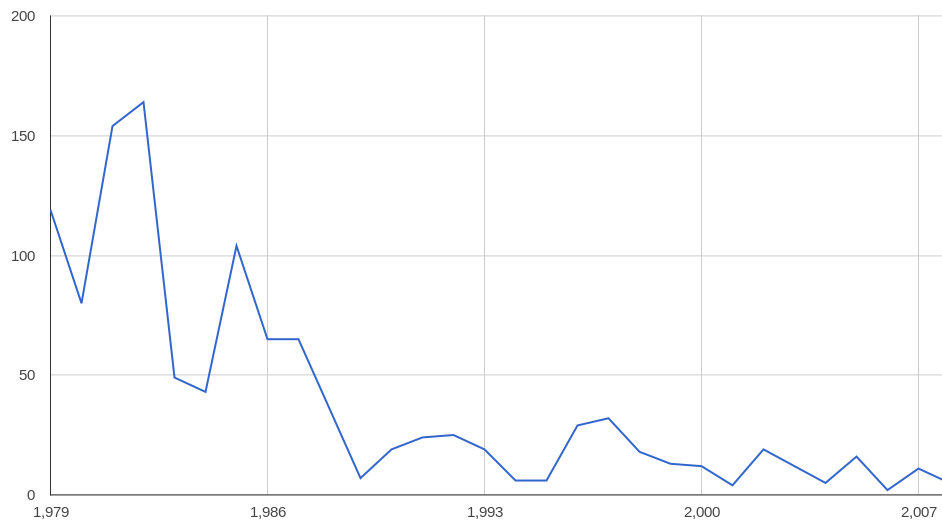
Figure 2: embhor years

Foo3.
Bar4.
Baz5.

5.2 Frequent itemsets

Analysing daily observed species as market basket transactions, we can examine the most commonly observed species and species combinations.
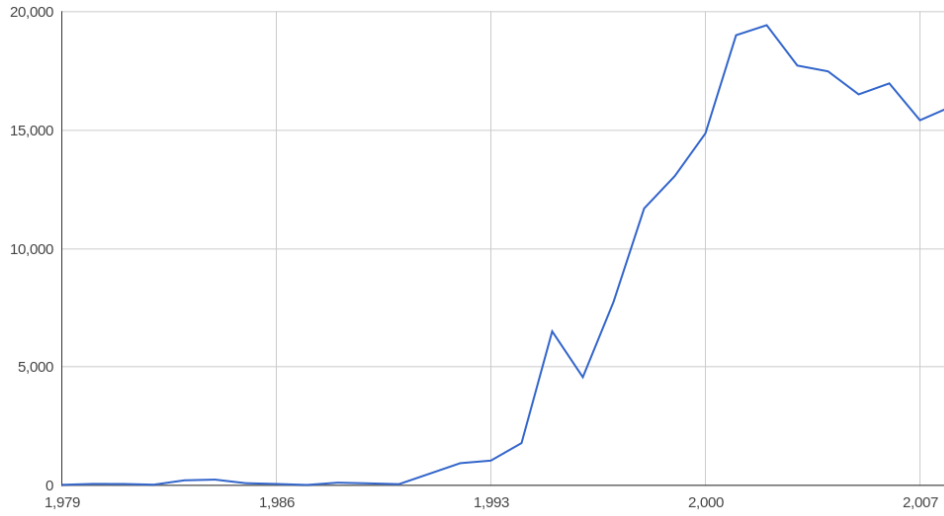
Figure 3: phacar years

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>,

<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.
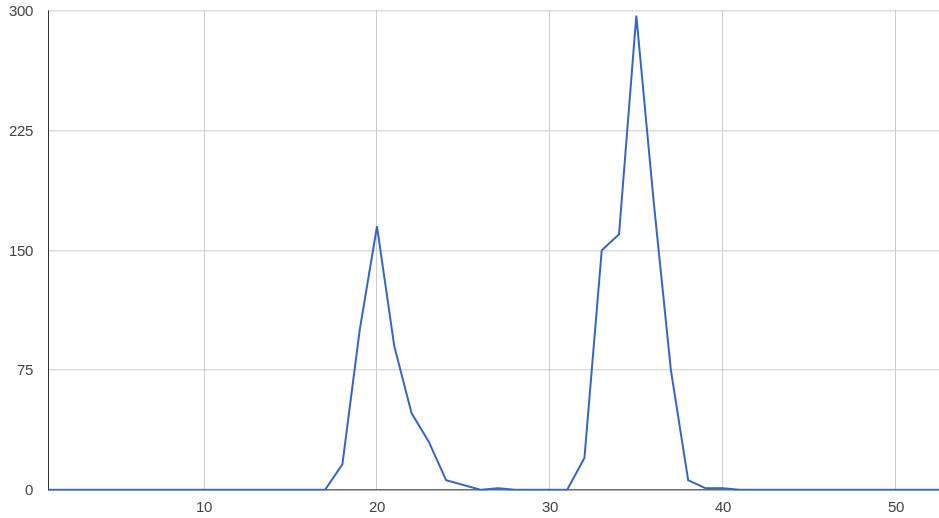
Figure 4: embhor weeks

5.3 Analysing species itemsets

We created itemsets of all observed taxa, consisting of species name, its' characteristics, conservation status and commonness measure at Halias. All of this species information is already present in the taxon ontology, but here we transform the information from ontologies to a format that allows us to use common data mining algorithms. Analysing these itemsets as market basket transactions we can infer frequent rules, but these mostly tell us about the used characteristics ontology and its' annotations. Table 1 shows two examples of a frequent generated rule with the rules' measures [TSK+06] for confidence, support, lift and IS measure. The "Late spring – early summer" and "Late summer – autumn" are part of the characteristics ontology, describing seasonal occurence of the species in Finland.

5.4 Species itemsets with weather variables

The species itemsets were further processed by leaving out all taxa that don't have characteristics annotations, and enriching the itemsets with weather averaged variables. For each species we add an average of the weather variables measured each day from sunrise to sunset and weight them with the count of birds that have been observed that day. So we get an average
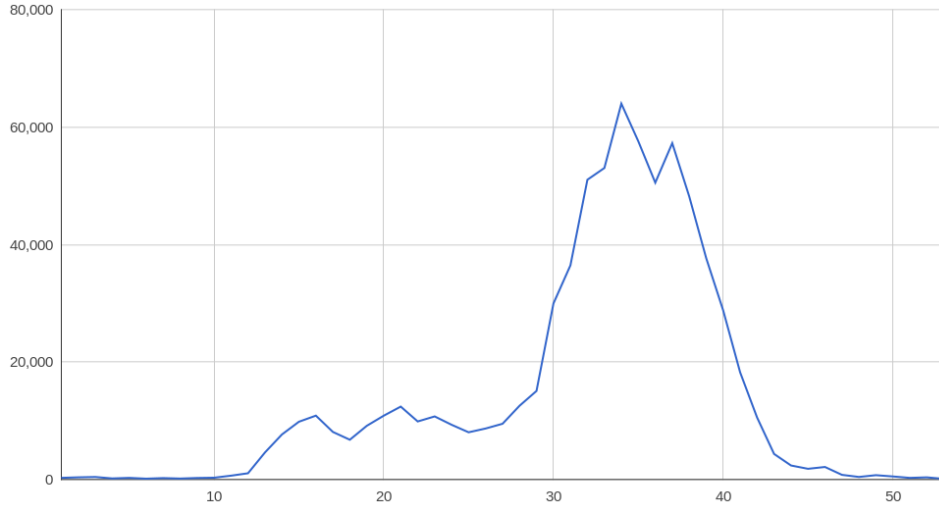
Figure 5: phacar weeks

Rule	Confidence	Support	Lift	IS measure
{Late spring – early summer, beak dark, head multicoloured, iris dark, upperside atleast 2-coloured} → {Late summer – autumn}	0.99	0.51	1.26	0.80
{Common species at Halias} → {Late spring – early summer, Late summer – autumn}	0.93	0.51	1.19	0.78

Table 1: Two examples of frequent rules generated from species itemsets.

weather condition in which the species has been observed. All of the averaged weather variables were categorized in three categories: "low", "average" and "high". The categories contain respectively the lowest, middle and highest third of all numbers. The limits of the categories are shown in table 2.

We generated association rules from these species itemsets. Some examples of these rules with their measurements are given in table 3.

6 Conclusions

7 Discussion

Variable	Low	Average	High
Pressure	[994.3, 1013.0]	(1013.0, 1014.5]	(1014.5, 1039.0]
Cloud cover	[0.0, 3.9]	(3.9, 4.5]	4.5, 8.0]
Humidity	[56.0, 79.7]	(79.7, 81.8]	(81.8, 93.7]
Rainfall	[0.0, 1.0]	(1.0, 1.4]	(1.4, 18.0]
Temperature	[-2.4, 8.4]	(8.4, 12.1]	(12.1, 20.0]
Wind speed	[0.1, 8.0]	(8.0, 18.8]	(18.8, 2109.0]

Table 2: Weather variable category limits derived from averaged weather variables by dividing them into three equally sized chunks.

Rule	Confidence	Support	Lift	IS measure
{Rare at Halias} → {Head multicoloured, Late summer – autumn, Upperside atleast two-coloured, Late spring – early summer}	0.89	0.30	0.99	0.55
{a} → {b}	0	0	0	0

Table 3: Some association rules generated from species itemsets.

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.