Mining a Bird Observatory Dataset

Mikko Koho

Technical report
UNIVERSITY OF HELSINKI
Department of Computer Science

Helsinki, May 24, 2015

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty Faculty of Science	Laitos — Institution — Department 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 Technical report	Aika — Datum — Month and year May 24, 2015	Sivumäärä — Sidoantal — Number of pages 12

Tiivistelmä — Referat — Abstract

This report examines the use of data mining techniques to find interesting patterns from a bird observatory dataset.

The bird observations are modeled as an RDF Data Cube dataset and linked with weather data from nearby weather station in Hanko, Russarö and a bird taxon ontology. This RDF data is transformed to simpler formats for association analysis, by grouping together relevant pieces of information from the RDF datasets and related ontologies. This allows to analyse the information directly with pattern mining algorithms and association rule generation. Using these methods, we were able to find some interesting patterns from the dataset.

Data mining approach can be applied to nature observation datasets to find interesting patterns. Organising the data for pattern mining is however not trivial as there are numerous possible ways to do this.

Avainsanat — Nyckelord — Keywords

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Contents

1 Introduction

This report examines the use of data mining techniques [TSK+06] to find interesting patterns from a bird observatory dataset. Section 2 shows some related work. Section 3 explains the structure and content of the dataset. Section 4 explains the used analysis methods and section 5 lays out results obtained using those methods. Sections 6 and 7 provide conclusions and discussion.

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

2 Related work

Kelling et al. have used exploratory analysis to find patterns in observational records of wintering birds in North America [KHF+09]. They also lay out the importance of data-driven approaches in biodiversity research. Caruana et al. [CEM+06] and Hochaka et al. [HCF+07] apply data mining and machine learning to bird observation datasets. Also other articles use classification models to predict occurence of bird species[GSZ+14, MH14].

The dataset used in this report has already been presented in earlier work [KHL14, Koh15], including some exploratory analysis with visualizations.

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].

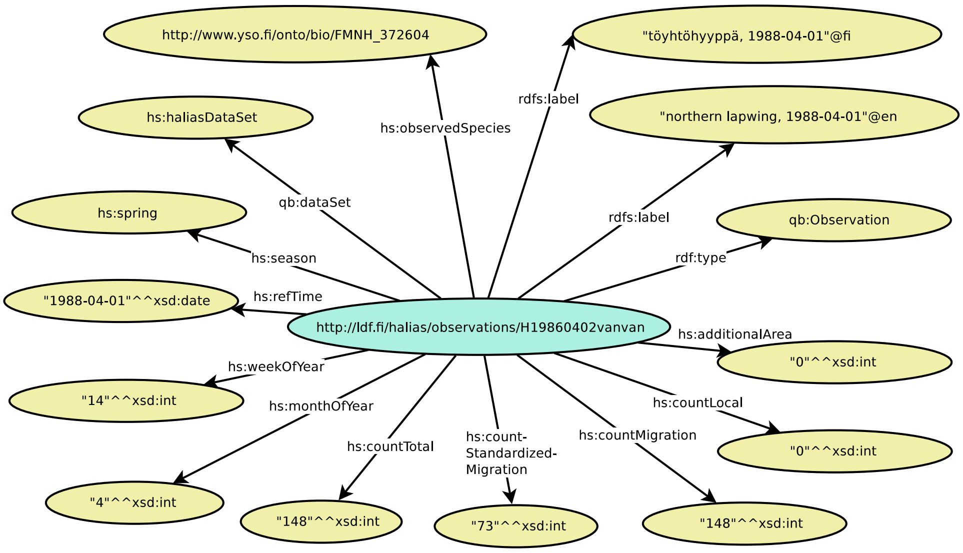

Figure 1: Daily observation data of one species from one day in RDF format [Koh15].

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

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 [TSK$^+$06]. 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.

We first analysed only the bird observations without any related information. Simple visualizations provide some insight into the occurence of different species. Frequent pattern mining was used to find the most frequently observed species and species combinations, without using the daily counts.

A sequential pattern discovery [TSK$^+$06] attempt was also made using daily observed species as sequential patterns, where each year was used as a distinct sequence. The data set as sequence data was too large to get any good results in feasible time, and as the initial results did not seem to provide any interesting information, this attempt was not pursued further.

The analysis is done using Python 3 and various modules. Initial examining the data was done mostly 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 implemented 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 bottleneck, which is the pruning step.

We tried to optimize the pruning by using multiple processes, running

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

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

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

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

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 for multiple processes.

Orange Data Mining Toolbox[8] was also tried for analysis, but it wasn't able to cope with large market basket transaction datasets.

5 Results

In this section we look at results of applying data mining methods to the dataset.

5.1 Visualization

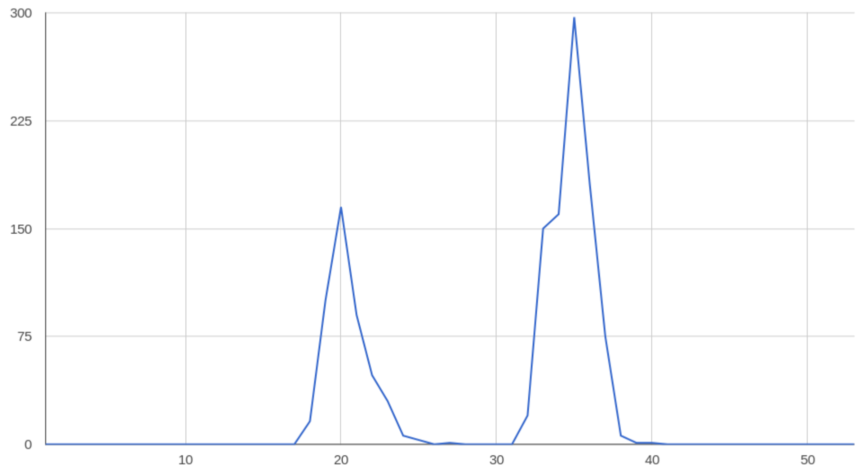

Figure 2: Yearly occurence of ortolan bunting. X-axis contains week number, y-axis contains total observed individuals for week number.

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

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

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

Visualizations of the data can easily convey important patterns in the data. Yearly occurence of species <peltosirkku, *Emberiza hortulana*, ortolan bunting> if given in figure 2. This shows a common yearly occurence of a migratory species not found at the bird observatory during breeding season. Most species are more frequent during autumn migration than during spring migration [LV00].

5.2 Frequent itemsets

By using an algorithmic analysis of the dataset using daily observed species as market basket transactions, we can examine the most commonly observed species and species combinations. An implementation of Apriori algorithm was used to get itemsets that occur frequently in the transactions.

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.

5.3 Association rules

We can create some association rules [TSK+06] that better explain the occurence of species at same dates. Examples of association rules with species

<uuttukyyhky, *Columba oenas*, stock dove> and <sepelkyyhky, *Columba palumbus*, woodpigeon>, are shown in table 1, which contains some rules with their measures [TSK+06] for confidence, support, lift and IS measure. The direction of the rule influences some of the given measures, but not all. Confidence is a frequentistic probability of the consequent happening if the antecedent has happened. So if a stock dove is observed it's very likely that a woodpigeon is also observed during the day, but not the other way around. Support gives the ratio of itemsets that contain both the antecedent and the consequent. Lift measures if the antecedent and consequent are dependent of each other, where value 1 implies independence. So woodpigeon and stock dove are strongly dependent according to the observations. The IS measure is large when lift and support are large.

Rule	**Confidence**	**Support**	**Lift**	**IS measure**
{woodpigeon} → {stock dove}	0.49	0.22	1.86	0.63
{stock dove} → {woodpigeon}	0.82	0.22	1.86	0.63

Table 1: Some association rules of species occurence at same dates.

5.4 Analysing species itemsets

For another type of analysis we created itemsets of all observed species. This creates 302 itemsets which consist of species name, its' characteristics, conservation status and commonness measure at Halias. All of the information are represented as strings. Many species lack characteristics annotations and for those species the itemsets are rather small.

All used species information is present in the taxon ontology and we transform the information into the itemsets as strings. 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 2 shows two examples of a frequent generated rule with the rules' descriptive measures. The "Late spring – early summer" and "Late summer – autumn" are part of the characteristics ontology, describing seasonal occurence of the species in Finland.

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 2: Two examples of frequent association rules generated from species itemsets.

5.5 Including weather variables

The species itemsets were further processed by leaving out all taxa that don't have characteristics annotations, and enriching the itemsets with the names of upper levels of taxon hierarchy. After this process we have a total of 239 itemsets, which are all created of species, as other taxonomic ranks lack characteristics annotations.

For each species we also add an average of the weather variables measured each day from sunrise to sunset and weight them with the total daily count of the species. So we get an average 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 3.

The itemset sizes vary between each species, consisting of 36 to 66 items. The size depends on the amount of characteristics annotations, how many upper taxa it has in the taxon ontology and whether it has a conservation status or not.

We generated association rules from these species itemsets using an Apriori-based algorithm. The rule generation included using frequent itemsets with minimum support 0.3 and using minimum confidence of 0.5, which

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

resulted in $105,949$ association rules. Some examples of the generated rules with their measurements are given in table 4. The last rule in the table is one that has very high lift, indicating a strong dependency between species of the order <varpuslinnut, *Passeriformes*, passerine> and some characteristics.

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
{Common at Halias, walks} → {Wind speed average, underside atleast two-coloured}	0.80	0.38	1.02	0.62
{Beak dark, iris dark} → {Upperside atleast two-coloured, common at Halias, underside atleast two-coloured}	0.63	0.45	1.09	0.70
{Passerine} → {Head multicoloured, legs short, late summer – autumn, beak sharp, iris dark, upperside atleast two-coloured, late spring – early summer}	0.84	0.31	2.09	0.81

Table 4: Some association rules generated from species itemsets.

Table 5 shows some relationships between bird sizes and weather variables. From the confidence and lift we can see that the they correlate to air pressure with different sign depending on the bird size. Larger birds are more dependent on clear weather when migrating, and higher air pressure indicates clearer weather. Small birds are not very much dependent on the weather when migrating, and also local rainfall, which is associated with low air pressure, may cause migrants to pause the migration, at which time they may be feeding locally at Halias and thus be easily observed. The rules seem to support these assumptions. Also all bird of prey of different sizes migrate mostly during clear, sunny weather and this can be stronly seen in association rules including the order <petolinnut, *Falconiformes*, birds of prey>. No species of the order has an average observation air pressure in the ”low” category, although some species of the order are not included in the analysis as they are lacking characteristics annotations and thus are not included in the itemsets.

Rule	**Confidence**	**Support**	**Lift**	**IS measure**
{Size huge} → {Air pressure low}	0.15	0.01	0.45	0.06
{Size huge} → {Air pressure average}	0.38	0.02	1.16	0.16
{Size huge} → {Air pressure high}	0.46	0.03	1.40	0.19
{Size small} → {Air pressure low}	0.38	0.13	1.11	0.39
{Size small} → {Air pressure average}	0.40	0.14	1.21	0.41
{Size small} → {Air pressure high}	0.22	0.08	0.67	0.23
{Bird of prey} → {Air pressure low}	0.0	0.0	0.0	0.0
{Bird of prey} → {Air pressure average}	0.39	0.03	1.18	0.19
{Bird of prey} → {Air pressure high}	0.61	0.05	1.85	0.29

Table 5: Some association rules between bird sizes and weather variables.

6 Conclusions

Data mining approach can be applied to biodiversity data to find interesting patterns. There are numerous possibilities how to organize the data into itemsets for pattern mining. By grouping relevant information of each species to a single itemset, we can efficiently analyse them for frequent patterns. Semantically linked datasets and related ontologies can be used to easily build the itemsets.

7 Discussion

It could be interesting for frequent rule generation to provide also monthly occurence at the bird observatory to species itemsets. The occurence information could use some quantitative categorization. This would allow for analysing which species characteristics are frequent during different months.

Also other kind of information, like a trend of yearly observation counts could be added to itemsets, to analyse what kind of trends do endangered species have. Also there could possibly be found unexpected association rules for species with clear negative or positive trends.

Directly including quantitative occurence data of each taxon in the daily observation itemsets could provide some interesting patterns.

References

[CEM+06] Caruana, R., Elhawary, M., Munson, A., Riedewald, M., Sorokina, D., Fink, D., Hochachka, W. M, and Kelling, S.: *Mining citizen science data to predict orevalence of wild bird species.* In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 909–915. ACM, 2006.

[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].

[GSZ+14] Goetz, S. J., Sun, M., Zolkos, S., Hansen, A., and Dubayah, R.: *The relative importance of climate and vegetation properties on patterns of north american breeding bird species richness.* Environmental Research Letters, 9(3):034013, 2014.

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

[HCF+07] Hochachka, W. M., Caruana, R., Fink, D., Munson, A. R. T., Riedewald, M., Sorokina, D., and Kelling, S.: *Data-mining discovery of pattern and process in ecological systems.* The Journal of Wildlife Management, 71(7):2427–2437, 2007.

[KHF+09] Kelling, S., Hochachka, W. M., Fink, D., Riedewald, M., Caruana, R., Ballard, G., and Hooker, G.: *Data-intensive science: a new paradigm for biodiversity studies.* BioScience, 59(7):613–620, 2009.

[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.

[LV00] Lehikoinen, A and Vähätalo, A: *Phenology of bird migration at the hanko bird observatory, finland, in 1979–1999.* Tringa, 27:150–224, 2000.

[MH14] Mentch, L. and Hooker, G.: *Ensemble trees and clts: Statistical inference for supervised learning.* arXiv preprint arXiv:1404.6473, 2014.

[TSK^{+}06] Tan, P., Steinbach, M., Kumar, V., *et al.*: *Introduction to data mining*, volume 1. Pearson Addison Wesley Boston, 2006.