Mining a Bird Observatory Dataset

Mikko Koho

Technical report
UNIVERSITY OF HELSINKI
Department of Computer Science

Helsinki, May 14, 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 14, 2015	4

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.

2 Dataset

Halias bird observatory 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.

Data gathering at the bird observatory is highly standardized and main focus is on counting the daily migration of each bird species. The data unfortunately is not publicly available, but is available for research projects at request.

The data in digital form 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.

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.

3 Methodology

For association analysis we take a subset of the values in 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 dataset is very large for many association analysis tasks, and therefore

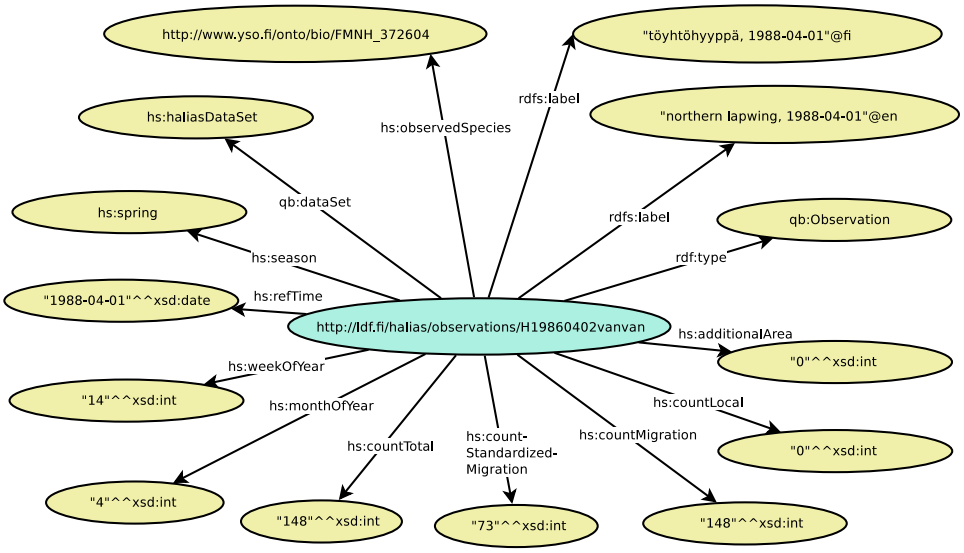

Figure 1: Daily observation data of one species from one day in RDF format [Koh15].

it may be neccessary to aggregate some variables.

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 2.7 and Orange Data Mining Toolbox [DCE+13].

4 Analysis results

4.1 Frequent itemsets

Analysing daily observed species as market basket transactions, we can see what are the most commonly observed species. For support 0.5, the largest

frequent itemset consists of the following species. Finnish names are used here for clarity.

> telkkä, sinisorsa, sinitiainen, isokoskelo, viherpeippo, varis, kalalokki, harmaalokki, merilokki, talitiainen

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.

4.2 Rule generation

This dataset was too large for Orange Data Mining Toolbox...

5 Conclusions

6 Discussion

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].

[DCE+13] Demšar, Janez, Curk, Tomaž, Erjavec, Aleš, Gorup Črt, Hočevar, Tomaž, Milutinovič, Mitar, Možina, Martin, Polajnar, Matija, Toplak, Marko, Starič, Anže, Štajdohar, Miha, Umek, Lan, Žagar, Lan, Žbontar, Jure, Žitnik, Marinka, and Zupan, Blaž: *Orange: Data mining toolbox in python.* Journal of Machine Learning Research, 14:2349–2353, 2013. `http://jmlr.org/papers/v14/demsar13a.html`.

[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.