Towards Using Semantic Data Mining for Predicting Community Assembly

Dina Sharafeldeen
Heinz-Nixdorf Chair for Distributed Information Systems
Friedrich-Schiller-Universität Jena
dina.sharafeldeen@uni-jena.de

Michael Owonibi
Heinz-Nixdorf Chair for Distributed Information Systems
Friedrich-Schiller-Universität Jena
michael.owonibi@uni-jena.de

Birgitta König-Ries[1]
Heinz-Nixdorf Chair for Distributed Information Systems
Friedrich-Schiller-Universität Jena
birgitta.koenig-ries@uni-jena.de

ABSTRACT

While data mining techniques are extensively applied in different domains, the potential of semantic techniques to enhance the data mining process is not yet fully used. This is true in general and applies especially to association problems. In this paper, we propose a framework to explore the benefits of semantic data mining to a particularly challenging association problem, namely the prediction of ecological community assembly. After introducing our framework, we present first results of applying association rule mining on species abundance data to extract species co-occurrence patterns which will help in community assembly prediction to show the general feasibility of using data mining approaches to this class of problems. We then explain how we will extend our system with semantic techniques to further enhance result quality.

Categories and Subject Descriptors

H.2.8 [**Information Systems**]: Database Management- Database Applications- Data mining; I.2.6 [**Computing Methodologies**]: Artificial Intelligence- Learning- Knowledge acquisition

Keywords

Data mining, association rules, semantic, community assembly, biodiversity, species co-occurrence.

1. INTRODUCTION

In recent years, the amount of available data in different domains has grown exponentially creating the need for new methods to access, manage and analyze these data [29]. Thanks to great advances in data mining techniques on the one hand and knowledge engineering on the other hand, semantic data mining approaches have started to appear. Semantic data mining refers to the process of involving semantics into the tasks of the data mining process. Previous research efforts have shown the advantage of incorporating such domain knowledge into data mining process. It provides the tasks in the data mining process with additional knowledge which enhances the output of these tasks, and provides a formal way for representing the data mining flow, from data preprocessing to mining results [3].

Biodiversity research is one of the disciplines that has experienced a tremendous increase in available data. It aims to study genetic diversity, species diversity, and ecosystem diversity. Genetic diversity refers to the genetic variation and heritable traits within organisms. Species diversity refers to the variety of living organisms within an ecosystem, a habitat or a region. It is evaluated by considering two factors: species richness and species evenness. Ecosystem diversity refers to the variety of ecosystems in each region of the world. An ecosystem is a combination of communities - associations of species - of living organisms with the physical environment in which they live (e.g., air, water, mineral soil, topography, and climate) [9]. The biologist Edward O. Wilson, known as the "father of biodiversity" said: "It is reckless to suppose that biodiversity can be diminished indefinitely without threatening humanity itself". Considering the rapid loss in biodiversity we evidence right now [1], understanding the mechanisms behind biodiversity is crucial. Associations between species (like complex food webs) are key factors for maintaining ecosystem stability. However, environmental changes can produce adverse impacts on species interactions to the threatening of ecosystem stability. Data intensive approaches seem promising in capturing these complex relationships [11]. One such example is the prediction of community assembly. Recent work in this area studies how species interactions could generate predictive patterns of species co-occurrence in communities (i.e., assembly rules) [8]. We believe, that the relatively new techniques of data mining offer promising ways to extract knowledge and patterns from large, multidimensional and complex data sets [12]. These extracted patterns provide new insights to scientists to answer biodiversity questions. Therefore, a wider adoption of data mining techniques in ecology and earth sciences can potentially improve the quality of science. However, the full potential of data mining techniques in ecology and earth sciences has not been fully achieved yet [24].

Our research work will focus on how to utilize different semantic data mining techniques. We will use different biodiversity and environmental data sources to support finding answers to important research questions in that field. As our research work is in an early stage, this paper will focus on our general framework and the initial results.

29th GI-Workshop on Foundations of Databases (Grundlagen von Datenbanken), 30.05.2017 - 02.06.2017, Blankenburg/Harz, Germany.

[1] Birgitta König-Ries was on sabbatical at the German Center for Integrative Biodiversity Research (iDiv) while working on this paper.

This paper is organized as follows: the state of the art is presented in the next section. Then our objectives and our initial proposed architecture are presented. Finally, preliminary results are illustrated and discussed.

2. STATE OF THE ART

Applying data mining techniques in biodiversity research integrates several research areas. However, in this paper, we focus on two important aspects only: semantic data mining and to what extent biodiversity research already benefits from these techniques.

2.1 Semantic Data Mining

Data mining, also known as knowledge discovery from databases (KDD), is the process of nontrivial extraction of implicit, and potentially useful knowledge from data [6] or "the analysis of (often large) observational datasets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner'' [7]. The tasks performed in the data mining process are knowledge intensive and can often benefit from using domain knowledge from various sources. Therefore, data mining may integrate techniques from machine learning, knowledge engineering, artificial intelligence, statistics, and other algorithms to analyze large and complex data sets on the one hand and to extract the domain knowledge on the other hand. Semantic data mining refers to data mining approaches that systematically incorporate domain knowledge into all tasks of the data mining process [3] as shown in Figure 1. Recently, semantic data mining research has emphasized the positive influence of domain knowledge on data mining. For example, the *preprocessing* can benefit from domain knowledge that can help filter out redundant or inconsistent data. During the *searching and pattern generating* process, domain knowledge can work as a set of prior constraints to help in reducing the search space and to guide the search path. Furthermore, in the *post processing*, the discovered knowledge/patterns can be cleaned [15,16] or made more visible by encoding them formally [28].

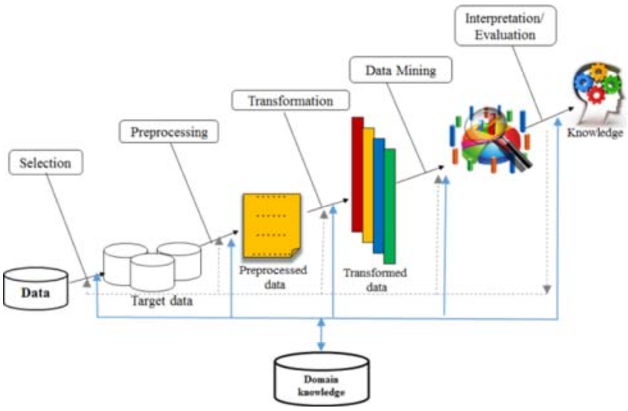

Figure 1:overview of incorperating domain knowledge in data mining

Based on our literature review, semantic data mining can be classified based on how the domain knowledge (semantics) is represented. Semantics can be represented by concept hierarchies [10], knowledge bases [31], ontologies [19], graphs [13], Linked Opened Data (LOD) [12], or meta-paths [23]. Meta-paths are a new representation that has been designed for semantic data mining tasks. The meta-path is a path that defines a composition of relations between the set of terms on the path [23].

2.2 Data Mining in Biodiversity

Recently, Ristoski and Paulheim [21] made a comprehensive survey on about 100 publications in different domains. These publications use semantic techniques with data mining and knowledge discovery in different stages. As an example, they showed how LOD can be used in different stages for building content-based recommender systems. Their survey showed that, even though there are numerous interesting research works performed, the full potential of using semantics for data mining and KDD is still to be achieved especially in association problems.

Biodiversity research is an interdisciplinary research field that is complex and very dynamic where new data from different sources are being observed and created all the time. Especially, in ecology, scientists study complex interactions between biotic and abiotic systems to understand these interactions and make predictions for biodiversity preservation and to answer biodiversity questions [24]. Based on our literature review, while data mining and KDD techniques have been applied in various domains, they are not yet fully utilized in biodiversity research. The authors in [9] provide examples of data mining applications for biodiversity and environmental studies. They showed that data mining can successfully discover new results and information to help environmental scientists to explain phenomena and to get new insights. These results can be improved by semantically integrating data and knowledge from different related application domains into the data mining process. To that extent, our research work focuses on developing innovative techniques utilizing the development in knowledge engineering to solve association problems. We will apply these techniques trying to answer biodiversity questions for community assembly by integrating semantic data mining approaches.

As far as we know, there is no research that automatically extracts co-occurrence patterns in an ecosystem where different types of species exist and uses these patterns to predict the community assembly. Recently, Silvaet al. [22] developed a data mining method for the analysis of multi-species in multi-scale form to assist ecologists in the assessment of patterns of occurrences of species in plant communities only. They allowed analysis of pairs and groups of species, identifying species that indicate a positive and negative co-occurrence in the same analysis. However, they need more studies to determine how to use their method to prove existing hypotheses in the field of plant communities. They plan to work on the selection of another existing metric to evaluate the generated co-occurrence patterns. Also, they recommend the integration of domain knowledge in the process, considering the reduction of the number of rules.

3. OBJECTIVES AND PROPOSED APPROACH

In an ecosystem, each species has a fundamental role in the circle of life. Hence, all species interact and depend on each other based on what each supply, e.g., food, oxygen, shelter, and soil enrichment. These associations can be positive [17] or negative [26]. These interactions produce one of the critical ecological associations between any species which is co-occurrence [18]. Based on our literature review, and based on our discussions with ecological scientists, we believe that it is quite difficult for scientists to detect all these co-occurrence patterns or to predict the existence of certain species manually.

Despite all the progress achieved in the data mining area and after considering that ecologists have long been researching effective new methods [25] to understand the mechanisms of species co-occurrence, competition and distribution of species [27], we found no research work using data mining techniques to automatically extract co-occurrence patterns and using these patterns to predict the existence of species then community assembly prediction.

To achieve this, our research objectives are as follows:

- Taking advantage of the progress in data mining techniques to apply complex analysis to extract hidden knowledge about species co-occurrence.
- Finding a (automatic or semiautomatic) methodology to evaluate the extracted domain knowledge.
- Incorporating the extracted domain knowledge into the analysis process trying to answer scientists' questions.
- Applying the complete cycle of machine learning techniques (training and testing) in our proposed framework.

Figure 2 shows the initial proposed framework to achieve our research objectives. The core challenge in biodiversity data is the diversity of data sources like abundance data, trait data, taxa, images, publications, phylogenies, species interactions, and even knowledge hidden in domain experts' minds. In the domain knowledge extraction process, hidden domain knowledge is extracted and collected from different biodiversity data sources. Then, this knowledge will be evaluated automatically or by a domain expert. Biodiversity scientists will use the analysis framework to answer their questions. This framework applies semantic analysis techniques.

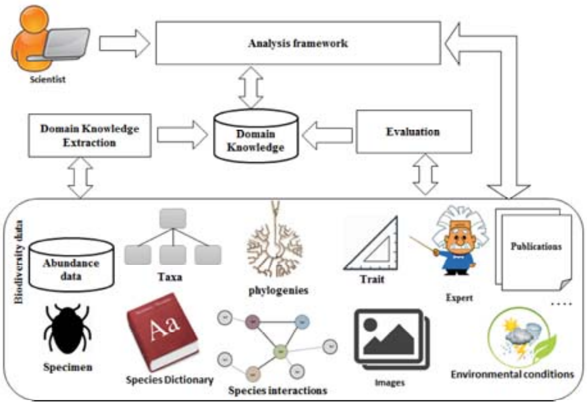

Figure 2: Our proposed framework

Talking about our current stage, our initial focus is community assembly prediction. Therefore, all domain knowledge that affect the prediction process should be extracted. As a beginning step, species co-occurrences patterns were extracted in the domain knowledge extraction process, then these patterns should be evaluated. At that point, the main goal for the analysis framework will be the prediction of different species exitance using semantic data mining. Then, this will lead to the community assembly prediction. Therefore, the process of semantic data mining will go through the whole steps as shown in figure 1 taking into consideration the extracted species co-occurrences patterns.

4. PRELIMINARY RESULTS

Our research work is in an early stage and we started working on the domain knowledge extraction process. One type of domain knowledge that is relevant in our sample domain is information about species interaction. This can be considered as semantic graph. Once known, these interactions can be used for community assembly prediction. Based on our proposed framework, we started to extract patterns of co-occurrences of pairs and groups of species in different communities. To obtain valid results, this analysis should be done on different data sets on different taxa and levels, land use, and ecosystem processes. Such a data set is available in the information system of the Biodiversity Exploratories (BE) [4]. The Exploratories are a long term, large scale project funded by Deutsche Forschungsgemeinschaft (DFG). The Exploratories use the BExIS platform [14] for central data management. The large collection of data sets in the BE BExIS is the result of research activities by many scientists in different disciplines involved in biodiversity science over the last ten years. We work on the publicly available data, which is 117474 records containing 4692 different species (plantae, animalia, and fungi). Each of these records contains information about a research plot and a specimen observed on that plot.

Association rule mining aims to discover frequent items from a set of transactions, deriving rules from associations among the items involved in each transaction [30], and these associations could be positive or negative. A *transaction* refers to the set of items in an operation, such as products purchased by a customer for market basket analysis [2]. For species data, the species that exist in the same plot form a transaction. Table 1 shows examples of species transactions. A *rule* can be written as a logical statement between two items, *A* (antecedent) and *B* (consequent). For example, sp-*1→sp-3* can be explained as a positive association or co-occurrence pattern where *sp-1* and *sp-3* exist together. On the other hand, *sp-1→(NOT)sp-4* can be explained as negative association or co-occurrence where *sp-1* and *sp-4* do not exist together.

Table 1. Examples of species transaction

Transaction	species
T1	sp-1, sp-2, sp-3
T2	sp-1, sp-3
T3	sp-2, sp-4

We apply association rule mining to extract species co-occurrences patterns. Our method was implemented with R [20] and Apriori with package Arules [5]. This method follows the data mining process tasks. During the *selection and preprocessing* task*,* species occurrence data is selected. In the *transformation* task, the data is transformed to fit the Apriori algorithm input format. Based on the values of the parameters used in the algorithm, 6351 co-occurrence patterns are extracted. There are two types of co-occurrence, positive and negative.

The quality of the extracted association rules can be evaluated by several metrics. We used the following set of measures: support, confidence, and lift [6]. For example, considering two species sp-1 and sp- 2, the support is the probability P of transactions with both species *support (sp-1 → sp-2) = P (sp-1 ∪sp-2).* The confidence is defined as the proportion with which item sp-1 is found in transactions containing sp-2 and is defined as the conditional probability *conf (sp-1 → sp-2) = P (sp-1| sp-2).* The lift is the measure of importance of a rule and can be defined by *P (sp-1 ∪sp-2) / (P (sp-1) *P (sp-2)).* The lift measure is used to differentiate between positive and negative co-occurrence. We set the lift threshold to be >= 2 for positive co-occurrence.

Table 2 shows a sample of the extracted (positive and negative) co-occurrence patterns. These patterns can be interpreted as a network of species co-occurrence. Figure 3 shows a part of the derived species interactions network.

Table 2. Sample of the extracted species co-occurrence patterns

Positive co-occurrence			
Antecedent→ Consequent	***Sup***	***Conf***	***Lift***
Incana→ Cupressiforme	0.15	0.95	2.80
Cupressiforme→ Heterophylla	0.22	0.65	2.78
Aenea,Rutabulum→ Cupressiforme	0.14	0.98	2.88
Rutabulum, Sylvatica→ Incana	0.14	0.47	2.91
Rutabulum,Sylvatica, Heterophylla → Cupressiforme	0.19	0.96	2.82
Excelsior, Odoratum, Perennis → Sylvatica	0.17	0.98	2.06
Negative co-occurrence			
Incana → (NOT) Sylvatica	0.15	0.93	1.94
Pineti → (NOT) Rutabulum	0.12	0.94	1.92
Incana, Cupressiforme →(NOT) Sylvatica	0.14	0.94	1.96
Carthusiana,Cupressiforme→(NOT) Sylvatica	0.14	0.90	1.90
Incana, Cupressiforme, Heterophylla → (NOT) Rutabulum	0.12	0.93	1.91
Incana, Aenea, Sylvatica →(NOT) Rutabulum	0.11	0.93	1.90
Plantae: Cupressiforme, Heterophylla, Rutabulum, Sylvatica, Excelsior, Odoratum, Perennis, Carthusiana **Fungi**: Incana, Aenea, Pineti			

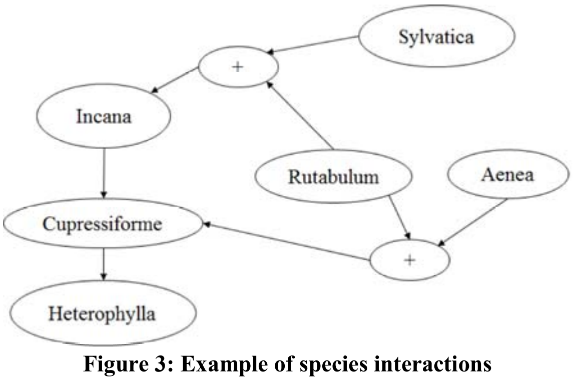

Figure 3: Example of species interactions

As a next step, we will work on finding a better metric to evaluate the extracted co-occurrence patterns taking domain knowledge into consideration. Also, further studies will be made on integrating domain knowledge on the evaluation of these extracted patterns. On the other hand, we believe that these co-occurrence patterns could be interpreted as not only pair associations, but also as a chain of associations model like food consumption chains. These interpretations help scientists to gain a better understanding and to prove existing hypotheses.

5. CONCLUSIONS

The evolution of data mining and knowledge engineering help in proposing new innovative methodologies to analyze large data sets. However, in biodiversity research, the full potential of these methodologies has not been fully achieved. Association problems in data mining are considered one of the problems that require more semantic involvement in extraction, and evaluation. On the other hand, the prediction process will benefit from the semantic involvement as well. Our research in its early stage, this paper focuses on species co-occurrence patterns extraction. We believe that using semantic data mining techniques in ecology will help to answer biodiversity questions. We will work on finding better methods to evaluate species co-occurrence patterns. Also, using these patterns to predict community assembly. On the other hand, we will work on extracting other hidden knowledge to enhance the community assembly prediction process. Finally, an ontology that defines the whole process of association extraction and prediction could be another fruitful output. This ontology would guide users in community assembly prediction.

6. ACKNOWLEDGMENTS

We would like to thank Andreas Ostrowski and Pete Manning for their valuable comments and suggestions. We thank the Biodiversity Exploratories (BE) project which is funded by the DFG for publicly providing the species data we work on. This work was partly funded by the DFG in the framework of the PP Biodiversity Exploratories as well as the DFG Research Centre iDiv.

7. REFERENCES

[1] Barnosky, A.D., Matzke, N., Tomiya, S., Wogan, G.O., Swartz, B., Quental, T.B., Marshall, C., McGuire, J.L., Lindsey, E.L., Maguire, K.C. and Mersey, B., 2011. Has the Earth/'s sixth mass extinction already arrived?. *Nature*, *471*(7336), pp.51-57.

[2] Brin, S., Motwani, R. and Silverstein, C., 1997, June. Beyond market baskets: Generalizing association rules to correlations. In *Acm Sigmod Record* (Vol. 26, No. 2, pp. 265-276). ACM.

[3] Dou, D., Wang, H. and Liu, H., 2015, February. Semantic data mining: A survey of ontology-based approaches. In *Semantic Computing (ICSC), 2015 IEEE International Conference on* (pp. 244-251). IEEE.

[4] Fischer, M., Bossdorf, O., Gockel, S., Hänsel, F., Hemp, A., Hessenmöller, D., Korte, G., Nieschulze, J., Pfeiffer, S., Prati, D. and Renner, S., 2010. Implementing large-scale and long-term functional biodiversity research: The Biodiversity Exploratories. Basic and Applied Ecology, 11(6), pp.473-485.

[5] Hahsler, M., Buchta, C., Grün, B. and Hornik, K., arules: Mining Association Rules and Frequent Itemsets, 2010. URL http://cran. r-project. org/package= arules. R package version 0.6-8.

[6] Han, J., Pei, J. and Kamber, M., 2011. *Data mining: concepts and techniques*. Elsevier.

[7] Hand, D.J., Mannila, H. and Smyth, P., 2001. *Principles of data mining*. MIT press.

[8] HilleRisLambers, J., Adler, P.B., Harpole, W.S., Levine, J.M. and Mayfield, M.M., 2012. Rethinking community assembly through the lens of coexistence theory. *Annual Review of Ecology, Evolution, and Systematics*, *43*, pp.227-248.

[9] Inthasone, S., Pasquier, N., Tettamanzi, A.G. and Pereira, C.D.C., 2015. Biodiversity and Environment Data Mining. *Scientific Journal of National University of Laos*, *9*, pp.116-128.

[10] Kamber, M., Winstone, L., Gong, W., Cheng, S. and Han, J., 1997, April. Generalization and decision tree induction: efficient classification in data mining. In *Research Issues in Data Engineering, 1997. Proceedings. Seventh International Workshop on* (pp. 111-120). IEEE.

[11] La Salle, J., Williams, K.J. and Moritz, C., 2016. Biodiversity analysis in the digital era. *Phil. Trans. R. Soc. B*, *371*(1702), p.20150337.

[12] Lausch, A., Schmidt, A. and Tischendorf, L., 2015. Data mining and linked open data–New perspectives for data analysis in environmental research. *Ecological Modelling*, *295*, pp.5-17.

[13] Liu, H., Dou, D., Jin, R., LePendu, P. and Shah, N., 2013, December. Mining biomedical ontologies and data using rdf hypergraphs. In *Machine Learning and Applications (ICMLA), 2013 12th International Conference on* (Vol. 1, pp. 141-146). IEEE.

[14] Lotz, T., Nieschulze, J., Bendix, J., Dobbermann, M. and König-Ries, B., 2012. Diverse or uniform? — Intercomparison of two major German project databases for interdisciplinary collaborative functional biodiversity research. Ecological Informatics, 8, pp.10-19.

[15] Mansingh, G., Osei-Bryson, K.M. and Reichgelt, H., 2011. Using ontologies to facilitate post-processing of association rules by domain experts. *Information Sciences*, *181*(3), pp.419-434.

[16] Marinica, C. and Guillet, F., 2010. Knowledge-based interactive postmining of association rules using ontologies. *IEEE Transactions on Knowledge and Data Engineering*, *22*(6), pp.784-797.

[17] Monge, J.A. and Gornish, E.S., 2014. Positive species interactions as drivers of vegetation change on a barrier island. *Journal of Coastal Research*, *31*(1), pp.17-24.

[18] Neeson, T.M. and Mandelik, Y., 2014. Pairwise measures of species co-occurrence for choosing indicator species and quantifying overlap. *Ecological Indicators*, *45*, pp.721-727.

[19] Phan, N., Dou, D., Wang, H., Kil, D. and Piniewski, B., 2015, September. Ontology-based deep learning for human behavior prediction in health social networks. In *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics* (pp. 433-442). ACM.

[20] Team, R.C., 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013.

[21] Ristoski, P. and Paulheim, H., 2016. Semantic Web in data mining and knowledge discovery: A comprehensive survey. *Web semantics: science, services and agents on the World Wide Web*, *36*, pp.1-22.

[22] Silva, L.A.E., Siqueira, M.F., dos Santos Pinto, F., Barros, F.S.M., Zimbrão, G. and Souza, J.M., 2016. Applying data mining techniques for spatial distribution analysis of plant species co-occurrences. *Expert Systems with Applications*, *43*, pp.250-260.

[23] Sun, Y., Norick, B., Han, J., Yan, X., Yu, P.S. and Yu, X., 2013. Pathselclus: Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *7*(3), p.11.

[24] Thessen, A., 2016. Adoption of machine learning techniques in Ecology and Earth Science. *One Ecosystem*, *1*, p.e8621.

[25] Veech, J.A., 2013. A probabilistic model for analysing species co-occurrence. *Global Ecology and Biogeography*, *22*(2), pp.252-260.

[26] Veech, J.A., 2014. The pairwise approach to analysing species co-occurrence. *Journal of Biogeography*, *41*(6), pp.1029-1035.

[27] Wiegand, T., Huth, A., Getzin, S., Wang, X., Hao, Z., Gunatilleke, C.S. and Gunatilleke, I.N., 2012, August. Testing the independent species' arrangement assertion made by theories of stochastic geometry of biodiversity. In *Proc. R. Soc. B* (Vol. 279, No. 1741, pp. 3312-3320). The Royal Society.

[28] Wimalasuriya, D.C. and Dou, D., 2010, October. Components for information extraction: ontology-based information extractors and generic platforms. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 9-18). ACM.

[29] Wu, X., Zhu, X., Wu, G.Q. and Ding, W., 2014. Data mining with big data. *ieee transactions on knowledge and data engineering*, *26*(1), pp.97-107.

[30] Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Philip, S.Y. and Zhou, Z.H., 2008. Top 10 algorithms in data mining. *Knowledge and information systems*, *14*(1), pp.1-37.

[31] Yu, X., Ma, H., Hsu, B.J.P. and Han, J., 2014, February. On building entity recommender systems using user click log and freebase knowledge. In *Proceedings of the 7th ACM international conference on Web search and data mining* (pp. 263-272). ACM.