FACTA UNIVERSITATIS
Series: **Automatic Control and Robotics** Vol. 20, Nº 1, 2021, pp. 21 - 31
https://doi.org/10.22190/FUACR201225002B

Regular Paper

GENERATING KNOWLEDGE STRUCTURES FROM OPEN DATASETS' TAGS - AN APPROACH BASED ON FORMAL CONCEPT ANALYSIS

UDC (004.6:81'37)

Miloš Bogdanović, Milena Frtunić Gligorijević, Nataša Veljković, Leonid Stoimenov

University of Niš, Faculty of Electronic Engineering, Department of Computer Science, Republic of Serbia

Abstract. *Under influence of data transparency initiatives, a variety of institutions have published a significant number of datasets. In most cases, data publishers take advantage of open data portals (ODPs) for making their datasets publicly available. To improve the datasets' discoverability, open data portals (ODPs) group open datasets into categories using various criteria like publishers, institutions, formats, and descriptions. For these purposes, portals take advantage of metadata accompanying datasets. However, a part of metadata may be missing, or may be incomplete or redundant. Each of these situations makes it difficult for users to find appropriate datasets and obtain the desired information. As the number of available datasets grows, this problem becomes easy to notice. This paper is focused on the first step towards decreasing this problem by implementing knowledge structures to be used in situations where a part of datasets' metadata is missing. In particular, we focus on developing knowledge structures capable of suggesting the best match for the category where an uncategorized dataset should belong to. Our approach relies on dataset descriptions provided by users within dataset tags. We take advantage of a formal concept analysis to reveal the shared conceptualization originating from the tags' usage by developing a concept lattice per each category of open datasets. Since tags represent free text metadata entered by users, in this paper we will present a method of optimizing their usage through means of semantic similarity measures based on natural language processing mechanisms. Finally, we will demonstrate the advantage of our proposal by comparing concept lattices generated using formal the concept analysis before and after the optimization process. The main experimental research results will show that our approach is capable of reducing the number of nodes within a lattice more than 40%.*

Key words: *Open data, formal concept analysis, semantic similarity, natural language processing*

Received December 25, 2020 / Accepted January 20, 2021

Corresponding author: Miloš Bogdanović
University of Niš, Faculty of Electronic Engineering, Aleksandra Medvedeva 14, 18000 Niš, Republic of Serbia
E-mail: milos.bogdanovic@elfak.ni.ac.rs

1. INTRODUCTION

In the past decade, there is a growing data transparency initiative which strives towards data openness of public and private institutions [1]. Open data initiative, as an idea to make public data available to anyone to use and republish, without restrictions from copyrights, patents or other controlling mechanisms [2], has influenced the governments' data transparency worldwide [3]. As one of the results, a significant number of Open Data Portals (ODPs) has been created. These portals offer anyone an ability to exploit the data and generate added value out of the data [4]. Due to its scale and a variety of subjects, government data is particularly interesting to both research and development communities [5].

Government data covers diverse areas, like statistics, transportation, environment, permits, licenses, budget, geography, elections, and etc. The scale of available data published by different countries and Open Data Portal (ODP) instances is growing every year. For example, a research regarding various data quality dimensions and end-user preferences conducted in 2018 by Neumaier, Umbrich and Polleres took into consideration 259 open data portals (ODPs) originating from 43 different countries, which all together hold more than 10TB of datasets [4]. Published data is organized into datasets which aggregate different data fields used to describe a particular dataset. These fields represent datasets' metadata - descriptive pieces of information, obtainable via Application Programming Interface (API), presented in a structured format that eases usage and data discovery. Metadata is organized as key-value pairs (meta-keys and values), where the key represents the property label while the value holds a numerical or textual representation [4]. Different ODPs organize metadata differently but every ODP controls the metadata usage through schemas consisting of pre-defined fields for specific information. In most cases, schema elements vary [6], but there are some common elements, like: title, description, groups, publisher, tags, resources and etc.

Majority of open data portals (ODPs) offer users an ability to further describe the content and structure of datasets by defining dataset tags. Tags present expressions that describe a specific open dataset. Each dataset's metadata contains a meta-key labeled as "tags" and values associated with this key that represents a particular tag(s). Thus, tag values can also be considered as a means to categorize open datasets or describe them. Although tags represent free text metadata, their usage in some cases becomes the basis for datasets search and discovery. Furthermore, if tag usage can be used to distinguish categories of datasets, tag sets can be observed as folksonomies of different open data portals (ODPs). Although folksonomy existence introduces various opportunities, the tag value folksonomy will not explicitly state a conceptualization shared among users. To do so, a separate data structure should be created with a purpose of revealing shared conceptualization originating from tags' usage. The optimal development of this structure is the focus of our research. In this paper we will take advantage of Formal Concept Analysis (FCA) to reveal the shared conceptualization originating from the tags' usage by developing a concept lattice per each category of open datasets.

2. RELATED WORK

2.1. Formal Concept Analysis

Formal Concept Analysis (FCA) is a mathematical method with a growing popularity across various fields of research and development. Although used in various fields, FCA is often considered a clustering technique. In the epicenter of its definition, FCA holds the term “concept”: a unit of thought constituted of two parts, its extent and its intent.

Formal Concept Analysis was introduced as a mathematical theory for the formalization of concepts. The following definitions constitute the foundation of FCA formal description:

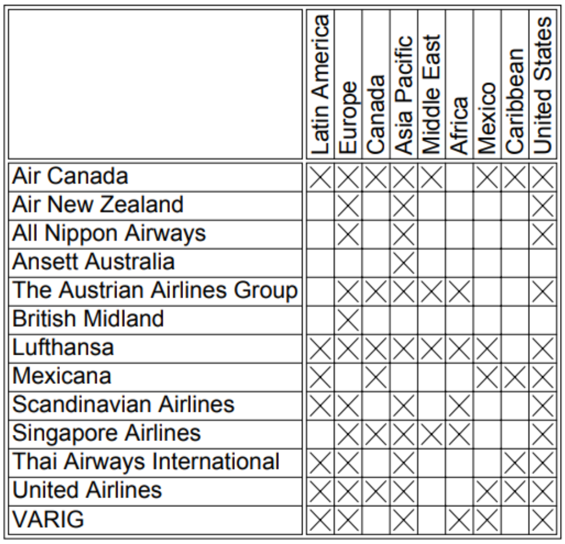
Definition 1 [14] A formal context is a triple $K := (G, M, I)$ which consists of a set G of objects, a set M of attributes, and a binary relation $I \subseteq G \times M$. $(g, m) \epsilon I$ is read as “object g has attribute m”.

Definition 2 [14] For $A \subseteq G$, let $A^I := \{m \epsilon M \mid \forall g \epsilon A: (g, m) \epsilon I \}$, and dually, for B $\subseteq M$, let $B^I := \{ g \epsilon G \mid \forall m \epsilon B: (g, m) \epsilon I \}$.

If the following conditions are met: $A \subseteq G$, $B \subseteq M$, $A^I = B$, $B^I = A$, then a pair (A, B) is a formal context. Set A is named concept extent while set B is named concept intent.

Definition 3 [14] The set S(C) of all concepts of a formal context C together with a partial order $(A_1, B_1) \leq (A_2, B_2) \leftrightarrow A_1 \subseteq A_2$ (which is equivalent to $B_1 \supseteq B_2$) is a complete lattice of C.

FCA derives concepts from incidence matrix, as shown in Figure 1, which uses the relationship between a particular set of objects and a particular set of attributes.



	Latin America	Europe	Canada	Asia Pacific	Middle East	Africa	Mexico	Caribbean	United States
Air Canada	☒	☒	☒	☒	☒		☒	☒	☒
Air New Zealand		☒		☒					☒
All Nippon Airways		☒		☒					☒
Ansett Australia				☒					
The Austrian Airlines Group		☒	☒	☒	☒	☒			☒
British Midland		☒							
Lufthansa	☒	☒	☒	☒	☒	☒	☒		☒
Mexicana	☒		☒				☒	☒	☒
Scandinavian Airlines	☒	☒		☒		☒			☒
Singapore Airlines		☒	☒	☒	☒	☒			☒
Thai Airways International	☒	☒		☒				☒	☒
United Airlines	☒	☒	☒	☒			☒	☒	☒
VARIG	☒	☒		☒		☒	☒		☒

Fig. 1 An example of formal context [25]

The main output of FCA is a concept lattice. The concept lattice is a collection of formal concepts logically organized into a hierarchy of concepts interconnected using subconcept-superconcept relations. Thus, the concept lattice reflects the generalization and specialization between formal concepts. In Figure 2 we can see a concept lattice generated for a formal context depicted in Figure 1. It is easy to notice that concept lattice reflects the generalization and specialization between formal concepts within a single formal context [16].

FCA has proven to be applicable in various domains and for various purposes. Over the last 20 years, literature reports extensive usage of FCA in knowledge discovery, software engineering and information retrieval. One of the best statements regarding the

FCA applicability is stated in [17]: "FCA enables the discovery and reasoning with concepts in data, discovery and reasoning with dependencies in data, and visualization of data, concepts, and dependencies".

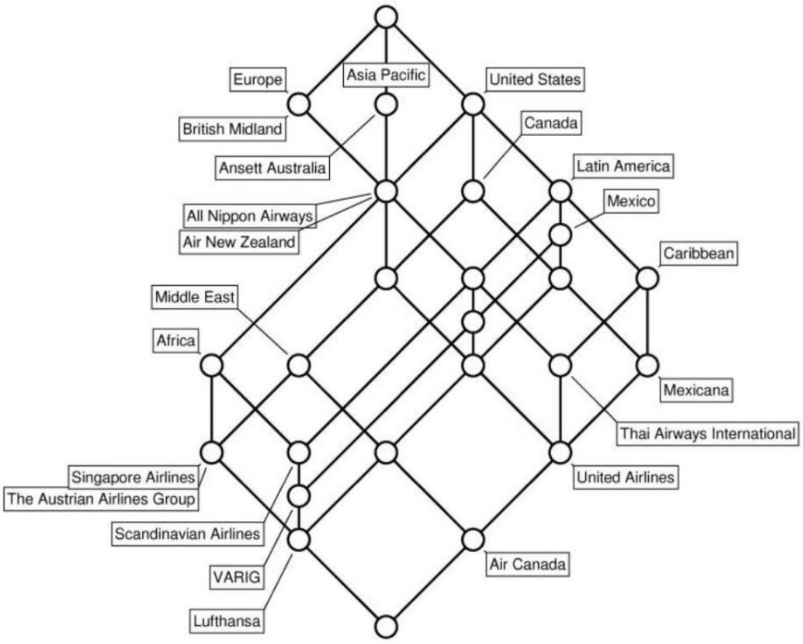


Fig. 2 The concept lattice of the formal context in Figure 1 [16]

2.2. Open dataset classification

From the users' perspective, open dataset should be visible and easily discoverable. Metadata can have a crucial role in satisfying both demands. The more metakeys have corresponding values, the more the dataset visibility increases and the more frequently it appears in search results.

To improve discoverability, OPDs commonly organize datasets into categories, which eases finding, discovery and combining data across different open data platforms. As a consequence, common use-case is that users will browse datasets from a certain category or perform the bulk download for the whole category and analyze the data. Depending on the ODP characteristics, categories may be static or dynamically determined based on the value of meta-key used to define dataset's category. In case a more precise and narrower collection of search results is necessary, majority of open data platforms rely on the tag based navigation. Thus, tag values can also be considered as a mean to categorize open datasets or describe them.

Although open data portals (ODPs) use both categories and tags to categorize datasets, their values originate from completely different sources - categories are mostly chosen from a controlled vocabulary while tags are entered at will (user defined). Since their values are predefined and controlled, categories enable the intuitive browsing of datasets [7]. The same assertion cannot be directly applied to tags - they express low consistency due to their origin. Nevertheless, tag values represent free text expression of the way users understand the available data already assigned to a particular category. Thus, their meaning is tightly coupled with dataset category.

While browsing datasets, it can be easily noted that many of the available open data portals (ODPs) contain significant amount of datasets with missing values for meta-keys. Furthermore, this is the case with meta-keys describing the dataset’s category. Since dataset’s category plays the crucial role in data discovery and usage, datasets become difficult to use if these pieces of data are missing. A number of questions arises from such situation: How to position an uncategorized dataset, into an existing set of categories? If a categorization is to be performed, what data is most appropriate to be used for these purposes? Can it be done automatically or at least semi-automatically?

Previous research reported various methods applicable to data categorization based on the relevant text attributes, text content, metadata analysis and metadata usage. Currently, machine-learning algorithms, including decision tree, nearest neighbor, Bayesian and neural networks are dominant when a text categorization should be performed [8]. Aside from previously mentioned approaches, dealing with categorization problems involves preprocessing (tokenization of a document), indexing (transformation into a vector model), feature selection (labeling important words or features in the document) and classification (determining a category using a-priori knowledge of already categorized data) [9]. Feature selection is particularly interesting since it influences the dimensionality of the feature space. If features are properly selected and reduced, there is a possibility to perform less computation and achieve a higher classification accuracy [10].

As for previous research using FCA, FCA-based approaches proved their potential as excellent classification methods. Prokasheva et al. present an excellent overview of simple classification methods based on FCA. According to [18], classification methods based on FCA are grouped into the following categories: hypothesis-based classification, concept lattice-based classification, classification based on Galois sub-hierarchies and cover-based classification. Definitions used for constructing concept-based hypotheses can be found in [19] and [20], while method examples can also be found in [21] and [22]. Various implementations have shown that a significant drawback of concept lattices is that the number of concepts may be exponential in the size of the relation. Thus, systems like CIBLe [23], CLNN&CLNB [24] and LEGAL [25] use a polynomial-size representation of the lattice while preserving the most relevant information. These systems build a Galois sub-hierarchy to reduce the exponential algorithmic complexity of generating a lattice. Another path already taken is the construction of concept cover - a part of lattice constructed using only pertinent concepts. IPR (Induction of Product Rules) [26] and the AdaBoost.M2 algorithm [27] benefit from using a local optimization of a measure function that defines pertinent concepts.

As previously mentioned, implementations dealing with the concept lattice construction and representation suffer from the lattice magnitude, in terms of a number of generated concepts and relations. Many of these algorithms and systems focus on reducing a search-space scale to achieve a higher efficiency. Research results we present in this paper make a step in the same direction - optimizing the formal context extent by optimizing the volume of relevant data.

3. Methodology

To propose a solution to the problem of uncategorized datasets, we focus on determining similarity between datasets. Datasets' metadata can be a useful source for similarity matching between datasets [11]. If dataset categorization should conform to users expectations, we find tags meta-key to be particularly significant. Descriptive knowledge of dataset's content and structure, as comprehended by a user, is contained within tags meta-key. Whether it holds simple or semantically rich terms, tags are the basis for datasets discovery [12]. The process of revealing the way tags meta-key is used can reveal the conceptualization shared among users. If conceptualization is determined, it becomes a powerful tool for categorization. For these reasons, we have decided to use tags and apply Formal Concept Analysis (FCA) on uncategorized datasets to suggest a match for the category. Further, we decided to apply semantic similarity measures based on natural language processing mechanisms with a purpose of reducing tag value space used for conceptualization determination.

FCA has received a significant attention from research and development communities of various fields. This method can be perceived as a clustering technique [13] that builds upon the definition of the term "concept": a unit of thought constituted of two parts, its extent and its intent [14]. Rudolf Wille introduced Formal Concept Analysis as a mathematical theory [14] for formalization of concepts. FCA provided us with the ability to extract a collection of formal concepts logically organized into a hierarchy of concepts (a concept lattice) starting from a set of objects and a set of attributes. Together, these two sets constitute a formal context. In our case, a set of object consists of datasets gathered from open data portals (ODPs), while a set of attributes contains a group of tags' values. The concept hierarchy generated starting from such formal context represents categories of datasets logically interconnected using generalization and specialization relationships according to tags usage. Our expectations are rather straightforward - users with similar interests are expected to use tags with similar meaning and this usage will in turn converge to a shared vocabulary of tags. By applying FCA to a shared vocabulary of tags, we have transformed it into shared tags category hierarchy. Generated hierarchy, represented via concept lattice, is capable of categorizing uncategorized datasets by examining their tags originating from datasets' metadata.

Due to the nature of the tags [15] and the fact that tags used for describing a dataset can be assigned an arbitrary value, the number of distinct tag values used across ODPs can be very large. Being so, within this research special attention was devoted to the fact tag values express low consistency due to their origin. This characteristic affects both computational time needed for generating shared tags category hierarchy and its structure, in terms of the scale of generated concepts. It is our opinion that the number of distinct tag values can be reduced by determining the same or very similar tag value meanings. To do so, we apply semantic similarity measures based on natural language processing mechanisms. As a result, we reduce the complexity of the generated hierarchy, represented by concept lattice. As the same time, our proposal will retain all necessary information regarding the meaning of distinct tag values thus retaining categorization capabilities of the generated knowledge structure, e.g. concept lattice.

Within this research we create relations between the tags and hierarchical order among them using Formal Concept Analysis on the combination of tags appearing in the datasets within the same category. Our approach gathers dataset metadata and per category formal contexts, thus generating concept lattices which contains the hierarchical order of all tags appearing as a part of datasets for each category. Before FCA is performed, we reduce formal contexts using semantic similarity measures. The overall process is shown in Figure 3. In the current state of implementation, we are using a custom tool we previously developed for generating concept lattices. This tool implements the Next Closure algorithm and was developed using the Microsoft .NET framework.

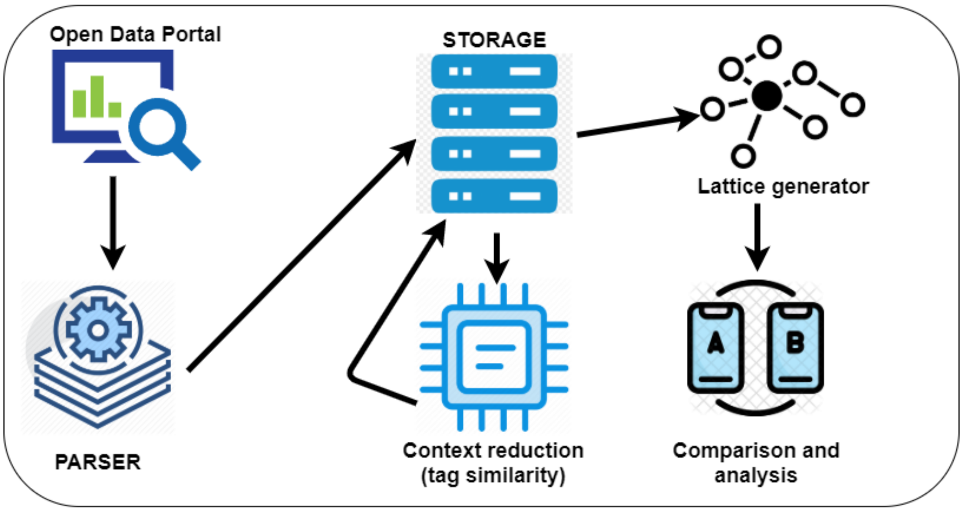


Fig. 3 Context reduction and lattice comparison

Since the concept lattice scale is proportional to the number of distinct tag values used to generate the formal context, open dataset categorization optimization is possible if an analysis of tag values meaning is performed. For this purpose, our approach relies on determining the tag values semantic similarity. Since the tag value is not necessarily a single word, each tag value is considered to be a sentence. Therefore, our approach uses GloVe (Global Vectors for Word Representation) model [28], trained on Common Crawl data with 840 billion words, to determine the similarity between sentences e.g. between tag values. GloVe model is a global log-bilinear regression model that combines the advantages of the two major model groups: global matrix factorization, such as latent semantic analysis (LSA) [29] and local context window methods, such as the skip-gram model of Mikolov et al. [30]. The model we use contains 2.2 million words, each represented via 300-dimension vector. We have chosen this model since it contains a large number of words gathered from different contexts and we consider it appropriate for our research because tag values contain a small number of words. Each tag value is first divided into words, whereby each word is defined by a vector contained within previously described GloVe model. When comparing two tag values, we have used the cosine similarity between each two vector pair corresponding to words generated by dividing tags being compared. Such measure can be applied for any pair of words or even syntagmas in majority of cases. Tag similarity is defined as the largest similarity between any two vectors of words belonging to tags being compared.

Table 1 Input reduction – per category comparison

Category	Original		Replaced	
	DSN	DCN	DCNRPL	PCT (%)
agriculture	622	601	436	27.45
arts_music_literature	18	80	76	5
economics_and_industry	66101	2756	1973	28.41
education_and_training	232	381	260	31.76
form_descriptors	67864	967	825	14.68
government_and_politics	64248	1973	1400	29.04
health_and_safety	1235	1578	1234	21.8
history_and_archaeology	98	155	136	12.26
information_and_communications	442	651	504	22.58
labour	602	604	404	33.11
language_and_linguistics	38	109	87	20.18
law	406	303	218	28.05
military	39	134	120	10.45
persons	2360	610	437	28.36
processes	76	201	161	19.9
science_and_technology	5699	1686	1312	22.18
society_and_culture	1463	1513	1154	23.73
transport	668	625	508	18.72

DSN – number of datasets per category
DCN – number of distinct tags per category before reduction
DCNRPL – number of distinct tags per category after reduction
PCT – reduction of distinct tags in percents

Two tags are considered similar if the similarity value exceeds the pre-defined similarity threshold. In our case, the similarity threshold was set to 0.8. In this way, a group of similar tags is defined for each tag in the observed set of tags. The next step is a comparison of generated groups. The content of each group, in terms of tags belonging to it, is checked against the content of every other group. If an intersection of groups contains more than one tag, each tag contained within intersection becomes a replacement candidate. Replacement candidates are further replaced in the group they originate from using a tag whose group of tags is considered similar (intersection contains more than one tag) to a group of tags replacement candidate belongs to. Thus, the size of the formal context used to generate concept lattice is effectively reduced. On the basis of the reduced context, a reduced concept lattice is generated while retaining dataset classification capabilities.

We have performed a comparison of the concept lattices generated before and after the reduction of the formal context is performed. A set of data used for reduction and comparison purposes was collected from the https://open.canada.ca/en open data portal (ODP), whereas the analysis was performed on a sample of tag values extracted from more than 80000 datasets. Overall results include per category comparison of inputs (distinct tags) and generated concept lattices, as shown in Tables 1 and 2, respectively. By analyzing the results, it can be determined that the approach we presented reduces complexity of both input and lattice structure. Also, reduction is not limited to categories containing large number of distinct tags, although reduction results for these dataset categories outperform the result generated for categories containing smaller number of distinct tags.

Table 2 Lattice reduction – per category comparison

Category	Original		Replaced		Reduction (%)	
	NLVL	NNO	NLVL	NNO	NLVL	NNO
agriculture	10	435	8	347	20	20.23
arts_music_literature	7	26	6	28	14.29	-7.69
economics_and_industry	14	2557	9	1777	35.71	30.5
education_and_training	7	200	7	129	0	35.5
form_descriptors	11	1036	10	956	9.09	7.72
government_and_politics	14	1484	10	1288	28.57	13.21
health_and_safety	11	978	8	753	27.27	23.01
history_and_archaeology	6	82	6	85	0	-3.66
information_and_communications	9	396	9	331	0	16.41
labour	13	435	8	249	38.46	42.76
language_and_linguistics	6	49	6	50	0	-2.04
law	9	216	7	177	22.22	18.06
military	4	50	4	53	0	-6
persons	15	739	13	455	13.33	38.43
processes	8	88	7	86	12.5	2.27
science_and_technology	10	1277	10	1025	0	19.73
society_and_culture	15	1271	13	1107	13.33	12.9
transport	10	313	9	298	10	4.79

NLVL – number of levels in the concept lattice
NNO – number of nodes in the concept lattice

Concept lattice complexity reduction (reduced number of levels and nodes) is more noticeable in lattices with originally higher number of levels. In smaller lattices, the number of levels remained unchanged in some cases. At the same time, the number of nodes in the concept lattices was reduced in almost all categories. This was not the case in few very small lattices. The overall reduction of the number of nodes for category *economics_and_industry* is 30%, for category *persons* 38% and goes up to 43% for category *labour*. The average reduction of the number of nodes is 14.6% for all categories.

4. Conclusion and Outlook

Governments around the world are adopting transparency and efficiency strategies. Publishing open datasets is a part of these strategies and results in a large amount of open datasets becoming publicly available. As the amount of open data grows, search and discovery capabilities become essential for end users. These capabilities often rely on metadata to offer users the ability to discover data and generate new knowledge out of it. For that reason, it is very important for an open dataset to be adequately described through metadata coupled with it and this process usually starts with defining open dataset category within a particular open data portal (ODP).

In this paper, our aim was to start migrating research focus from metadata content to metadata meaning by analyzing data used to categorize open datasets. To do so, we have analyzed tag values used to describe the content and meaning of a particular dataset. Since tags represent free text metadata entered by users, we have focused on using the meaning of tags to reduce the heterogeneity of tag values while retaining the meaning of tags for a particular open dataset category. Although presented approach can be improved in terms of semantic similarity measures used to reduce the number of distinct tag values, our approach shows promising result and effectively reduces computational time needed for developing an auxiliary structure used for categorizing open datasets.

Regarding further improvements, it is also necessary to implement a mechanism to extend the existing concept lattice in terms of including new tag values. This would enhance created knowledge and discovery data structure with learning capabilities. During the analysis we presented in this paper, concept lattice creation proved to be long-running task - in some cases couple of weeks. The variable that influences time the most is the number of datasets, the number of tags and the number of different combinations of tags per category in a single open data portal (ODP). There are various proposals regarding algorithmically improved concept lattice creation, for example the Parallel Recursive Algorithm for FCA and the In-Close algorithm. It is our aim to use some of them instead of the Next Closure algorithm we are currently using for lattice development. Further, in cases a new tag value appears, we plan to develop a separate service for this situation. The new service would implement the ability to perform incremental lattice construction and remove a necessity to reconstruct the whole lattice for this case. Up to our knowledge, parts of existing algorithms, such as the FastAddExtent algorithm, can be used for these purposes.

REFERENCES

[1] S. Kubler, J. Robert, S. Neumaier, J. Umbrich, Y. Le Traon, “Comparison of metadata quality in open data portals using the Analytic Hierarchy Process,” Government Information Quarterly, vol. 35, no.1, pp.13-29, 2018.

[2] S.R. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, ”DBpedia: A Nucleus for a Web of Open Data,” The Semantic Web, Lecture Notes in Computer Science, pp. 722-735, 2007.

[3] N. Veljković, S. Bogdanović-Dinić, L. Stoimenov, “eGovernment openness index,” Proceedings of the 11th European Conference on eGovernment, Ljubljana, pp. 571–577, 2011.

[4] S. Neumaier, J. Umbrich, A. Polleres, “Automated quality assessment of metadata across open data portals,” Journal of Data and Information quality, vol. 8, no.1, pp. 2:1-2:29, 2016.

[5] S. van der Waal, K. Węcel, L. Ermilov, V. Janev, U. Milošević, M. Wainwright, “Lifting open data portals to the data web,” In Linked Open Data--Creating Knowledge Out of Interlinked Data, Springer, Cham, pp. 175-195, 2014.

[6] P. Milic, N. Veljkovic, L. Stoimenov, “Comparative analysis of metadata models on e-government open data platforms,“ IEEE Transactions on Emerging Topics in Computing, 2018.

[7] F. Maali, R. Cyganiak, V. Peristeras, “Enabling Interoperability of Government Data Catalogues,” In Proceedings of EGOV 2010, pp. 339-350, 2010.

[8] M. El Kourdi, A. Bensaid, T.E. Rachidi, “Automatic Arabic document categorization based on the Naïve Bayes algorithm,” Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, pp. 51-58, 2004.

[9] V. Korde, C.N. Mahender, “Text classification and classifiers: A survey,” International Journal of Artificial Intelligence & Applications, vol. 3, no. 2, pp. 85-99, 2012.

[10] A.K. Uysal, S. Gunal, “A novel probabilistic feature selection method for text classification,” Knowledge-Based Systems, vol. 36, pp. 226-235, 2012.

[11] V. Korde, C.N. Mahender, “Text classification and classifiers: A survey,” International Journal of Artificial Intelligence & Applications, vol. 3, no. 2, pp. 85-99, 2012.

[12] A.K. Uysal, S. Gunal, “A novel probabilistic feature selection method for text classification,” Knowledge-Based Systems, vol. 36, pp. 226-235, 2012.

[13] R. Jaschke, Formal Concept Analysis and Tag Recommendations in Collaborative Tagging Systems, Dissertations in Artificial Intelligence, 2011.

[14] R. Wille, “Restructuring lattice theory: An approach based on hierarchies of concepts,” Ordered Sets, Springer, Dordrecht, pp. 445–470, 1982.

[15] D.D. Lewis, M. Ringuette, “A comparison of two learning algorithms for text categorization,” Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, 1994.

[16] B. Ganter, G. Stumme, “Formal concept analysis: Methods and applications in computer science,” Technical Report Otto – von – Guericke – Universitat Magdeburg

[17] A. M. Boutari, C. Carpineto, R. Nicolussi, R., “Evaluating term concept association measures for short text expansion: two case studies of clas-sification and clustering, ” In CLA 2010, pp. 163–174, 2010.

[18] O. Prokasheva, A. Onishchenko, S. Gurov, Classification methods based on formal concept analysis, FCAIR 2012 – Formal Concept Analysis Meets Information Retrieval, p. 95, 2012.

[19] S.O. Kuznetsov, Mathematical aspects of concept analysis, Journal of Mathematical Science, Vol. 80, Issue 2, pp. 1654–1698, 1996.

[20] S.O. Kuznetsov, Complexity of Learning in Concept Lattices from Positive and Negative Examples, Discrete Applied Mathematics, No. 142(1–3), pp. 111-125, 2004.

[21] V.K. Finn, The Synthesis of Cognitive Procedures and the Problem of Induction, Autom. Doc. Math. Linguist., 43, pp.149-195, 2009.

[22] V.K. Finn, On machine-oriented formalization of plausible reasoning in the style of F. Bacon and D.S. Mill [in Russian], Semiotika i Informatika, 20, pp.35–101, 1983.

[23] P. Njiwoua, Mephu Nguifo E, Améliorer l'apprentissage à partir d'instances grâce à l'induction de concepts: Le système CIBLe, Revue d'Intelligence Artificielle (RIA), vol. 13, 2, pp. 413–440, Hermes Science, 1999.

[24] Z. Xie, W. Hsu, Z. Liu, M. L. Lee: Concept Lattice based Composite Classifiers for high Predictability, Artificial Intelligence, vol. 139, pp.253–267, Wollongong, Australia, 2002.

[25] P. Njiwoua, E. M. Nguifo, Forwarding the choice of bias LEGAL-F Using Feature Selection to Reduce the complexity of LEGAL, In Proceedings of BENELEARN-97,ILK and INFOLAB, Tilburg University, the Netherlands, pp. 89–98, 1997.

[26] M. Maddouri, Towards a machine learning approach based on incremental concept formation, Intelligent Data Analysis, Volume 8, Issue 3, pp. 267–280, 2004.

[27] Y. Freund, R. E. Schapire, Experiments with a new boosting algorithm, International Conference on Machine Learning, pp. 148-156. Morgan Kaufmann Publications, Bari, 1996.

[28] J. Pennington, R. Socher, C. Manning, “Glove: Global vectors for word representation, ” In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543, 2014.

[29] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman, "Indexing by latent semantic analysis". Journal of the American Society for Information Science, 41(6): 391–407, 1990.

[30] T. Mikolov, W. T. Yih, G. Zweig, "Linguistic regularities in continuous space word representations", In Proceedings of NAACL-HLT, pages 746–751, 2013