UNIVERSITY OF TARTU
Institute of Computer Science
Software Engineering Curriculum

Ramil Huseynov

A recommender system for improved data findability in open government data portals

Master's Thesis (30 EAP)

Supervisor(s): Anastasija Nikiforova
Dimitrios Symeonidis

Tartu 2025

A recommender system for improved data findability in open government data portals

Abstract:

Despite the large amount of data available through OGD (Open Government Data) portals, most of it remains “dark data” which means it is not being used. A significant factor contributing to it can be the usability challenges such poor data findability and discoverability associated with these portals. One of the ways to contribute to the solution of these challenges is a recommendation system that can suggest related datasets. Unlike other domains, the recommendation system in the OGD portals is special as it can’t rely on user profile as most OGD portals don’t require authentication. Moreover, this recommendation method should be adaptable to the diverse structures of these portals. Finally, existing recommendation systems for OGD portals mostly focus on tags/category recommendations not datasets recommendations or fail to capture the semantic meaning of dataset’s metadata when making recommendations. This study focuses on these challenges by proposing a new datasets recommendation method based on dataset’s metadata that can capture its semantic meaning without relying on user’s profile and compatible with wider range of OGD portals. To capture the semantic relations between dataset’s metadata the proposed recommendation system relies on pretrained Word2Vec model. Additionally, the prototype of the proposed recommendation system was implemented for the usability testing and feedback was collected and analyzed.

Keywords:

Open data, open government data, recommendation system

CERCS:

P170 Computer science, numerical analysis, systems, control

Soovitussüsteem andmete parema leitavuse parandamiseks avatud valitsuse andmeportaalides

Lühikokkuvõte:

Vaatamata OGD (Open Government Data) portaalide kaudu saadaolevale suurele hulgale andmetele jääb suurem osa neist "tumedateks andmeteks", mis tähendab, et neid ei kasutata. Selle oluliseks teguriks võivad olla nende portaalidega seotud kasutatavuse probleemid, näiteks halb andmete leitavus ja leitavus. Üks võimalus nende väljakutsete lahendamisele kaasa aidata on soovitussüsteem, mis võib soovitada seotud andmekogumeid. Erinevalt teistest domeenidest on OGD-portaalide soovitussüsteem eriline, kuna see ei saa tugineda kasutajaprofiilile, kuna enamik OGD-portaale ei vaja autentimist. Lisaks peaks see soovitusmeetod olema kohandatav nende portaalide erinevatele struktuuridele. Lõpuks keskenduvad OGD-portaalide olemasolevad soovitussüsteemid enamasti siltidele/kategooriasoovitustele, mitte andmekogumisoovitustele või ei suuda soovituste tegemisel tabada andmestiku metaandmete semantilist tähendust. See uuring keskendub nendele väljakutsetele, pakkudes välja uue andmekogumite soovitusmeetodi, mis põhineb andmekogumi metaandmetel, mis suudab tabada selle semantilist tähendust ilma kasutaja profiilile tuginemata ja ühildub laiema hulga OGD-portaalidega. Andmestiku metaandmete vaheliste semantiliste seoste tabamiseks tugineb pakutud soovitussüsteem eelnevalt väljaõpetatud Word2Vec mudelile. Lisaks rakendati kasutatavuse testimiseks pakutud soovitussüsteemi prototüüpi ning koguti ja analüüsiti tagasisidet.

Võtmesõnad:

Avaandmed, avatud valitsuse andmed, soovitussüsteem

CERCS:

P170 Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine (automaatjuhtimisteooria)

Table of Contents

List of figures

1. Introduction

1.1 Background and Context

Open data is the data that is under Open License, which means that it can be accessed, read, and used by anyone for any purposes [18]. It is meant to remove the copyright and patent restrictions for accessing, using data and encourage innovation and transparency [18].

Open Government Data (OGD) is the type of Open Data published by the governmental bodies [18]. It includes data ranging from socio-economic metrics to environmental data. The availability of OGD is seen to be able to enhance transparency of governmental bodies, facilitate more informed decision-making and encourage innovation as these datasets can be used to create new applications, services or insights that benefit society, businesses and research [4] [18]. Repositories that host this data and provide interface for users to access and search for them and their metadata are called OGD portals [18].

Despite the wealth of data available in OGD portals, a significant portion becomes "dark data", data that remains unused or underutilized [19]. One of the main causes behind this is the difficulty in finding relevant datasets [1][2][4]. With the ever-increasing volume of data being provided by these portals, users often find it overwhelming to locate the specific data they need due to issues such as insufficient documentation and metadata around the dataset [15][20][22][23], weaknesses of search mechanisms [4][20][24].

This issue of data findability has a negative impact on the effectiveness of OGD portals. When users cannot easily find the data they need, potential benefits of OGD portals are significantly challenged. As such, there is a need for better findability mechanisms or tools to deal with this problem. One of the tools that can be used to challenge this problem is recommendation systems [20]. These systems are widely used in commerce and social media platforms [5]. Recommendation systems are tools that can recommend relevant content to the user. They are classified into 3 types based on their input data: content-based, collaborative and hybrid [20]. Content-based recommender systems rely only on the single user's profile and description, while collaborative recommender systems consider other similar users' behavior [20]. Hybrid recommender systems combine the results from both [20]. Implementing recommender systems

on the OGD portals may help users find the datasets that they are interested in, enhancing the effectiveness of OGD portals and OGD initiative [5].

The objective of this thesis is to develop a recommender system for OGD portals, aiming to bridge the gap between the wealth of available data and the specific needs of its diverse users. This is expected to enhance the findability, accessibility and usability of open government data in accordance with the FAIR (Findability, Accessibility, Interoperability and Reusability) principles [17], ultimately fostering a more informed and engaged quadruple helix (university, industry, government, public, environment) [1][4].

Although the recommendation systems are widely used throughout commerce and social media services, the research around their usage in the OGD portals is still limited. Most of the related research is around recommending tags or categories based on the dataset's metadata [4][6][11][22]. While some studies focus on recommending datasets in OGD portals [1][2][7][12], they are either recommending based on user's characteristics [1] or the selected dataset's metadata without considering semantics of the metadata[2][7][12]. Lack of authentication requirement by OGD portals requires a content-based recommendation system that is not based on user's characteristics.

Building a content-based recommendation system that can suggest related datasets based on user selected dataset's metadata requires calculation of relatedness between datasets. Since dataset's most metadata features - title, description, category – are typically provided in text format, this task falls under the natural language processing. In natural language processing (NLP), relatedness between texts can be lexical relatedness which can be calculated, for example, by using word embedding techniques such as TF-IDF for converting text to vectorial representation and by calculating a similarity between vectors using a similarity measure such as Cosine and Jaccard [26]. There are also other word embedding techniques that can also capture the latent semantic relations between words such as synonyms, antonyms, hypernyms [17][26]. This study proposes a recommendation system that uses one of these word embeddings, namely Word2Vec that can capture semantic relations.

1.2 Research Questions

The aim of this thesis is to develop a recommender system for OGD portals to enhance data findability in these portals. To attain the objective, the following research questions were defined:

1. **RQ1: What are the existing OGD portal recommender systems?**
   Through systematic literature review this question examines existing proposals for OGD portal recommender systems, identifying current research trends, gaps and opportunities that will inform the development of a new recommendation system. This question also investigates the types of recommendation techniques used in the current research of recommendation systems in OGD portals and evaluates their suitability for the problem.
2. **RQ2: How can a recommender system be developed for OGD portals based on the identified techniques and features?**
   This question aims to synthesize learnings from literature review and come up with systematic steps that can be used to recommend relevant datasets for the given dataset. This includes the definition of the OGD portal features, the recommendation techniques to be used and their integration.

By addressing these questions, this thesis seeks to create a foundation for the OGD recommendation systems to enhance the data findability in OGD portals.

The methodology for this research is designed to systematically address the research questions and achieve the outlined objectives. It consisted of reviewing literature, proposing a new recommendation method, implementation of a prototype for the method, evaluating the method through UAT, analyzing the results and defining refinements for the developed artifact based on this analysis.

The following tasks are put together to guide the development of a recommender system for OGD portals. First, a systematic literature review (SLR) to gather and assess existing proposals and implementations of recommender systems within OGD domain is conducted. Techniques used in these recommendation methods such as content-based, collaborative filtering, and hybrid are evaluated for their suitability and effectiveness in the context of OGD portals. Additionally, the features of the dataset that were used in these recommendation methods are cataloged and

evaluated for their suitability. Based on these insights, a recommendation system framework that integrates the most appropriate techniques and features is developed. To evaluate the effectiveness of the recommendation system, a prototype is developed with a sample dataset. Then, user acceptance testing (UAT) is conducted to assess the system's effectiveness in improving data findability.

By completing these tasks, the research aims to contribute to enhancing the usability and effectiveness of OGD portals, thereby supporting the broader goals of open government and data-driven decision-making.

1.3 Significance of Study

The significance of this research lies in improving the data findability in OGD portals. Improving data findability in OGD portals eliminate the dark data presence by enabling users to navigate to the data of their interest with further facilitation of FAIRness of these data. Better access to government data can lead to increased transparency in governmental processes [1][4]. With easier access to relevant data, citizens, policymakers, researchers, and businesses can make more informed decisions [4]. This step is important for creating data-driven policymaking, academic research, and business strategies. Improving the usability of OGD portals can encourage greater innovation among analysts, businesses, software engineers[1]. When users can easily find and understand government data, they are more likely to use this data in developing new services. Moreover, the study's findings and methodologies can set a precedent for further research in data findability and accessibility, particularly in the context of OGD. Additionally, this study focuses on proposing a recommendation method that is adaptable to diverse structures of OGD portals. Finally, it can also inspire similar initiatives in other domains where dark data is a challenge such as proprietary data stored and underutilized by private or public organizations.

2. Methodology

This chapter discusses the methodology used in this study which consists of 3 parts, namely Systematic Literature Review (SLR), implementation and evaluation of the proposed recommendation system prototype.

2.1 SLR

The SLR followed guidelines defined by Kitchenham (2004), which suggest SLR to be conducted in 5 stages: (1) study identification, (2) study selection, (3) study quality assessment, (4) data extraction, (5) data synthesis.

The first step of conducting SLR was to study identification, where research questions were considered to come up with the search query. As the objective of this research was to look for existing recommendations systems for OGD portal, the search query was deduced to be the as in Table 1.

Table 1. Search query

("open data" OR "open government data" OR "OGD") AND (("recommend*" AND " system") OR "data* recommend*").

As the second step of SLR, the results were filtered out to align with our research focus. Firstly, Scopus and the Web of Science were chosen as data sources because of their quality [27]. Then, we limited the scope of the search query to title, abstract and keywords to focus on the papers that deal with the topic as the main part rather than merely mentioning it. Secondly, the publication period was selected to 2017-2024 to cover the most recent literature as of October 2024 when the search conducted. We also limited the language to English and selected only journal articles, conference papers, and book chapters as the document types.

In the third phase we eliminated studies not related to the thesis topic by based on their title and abstract. After elimination of 364 studies, 13 studies remained and also categorized them based on their relevance as low, medium or high. Figure 1 presents Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) diagram of this process.

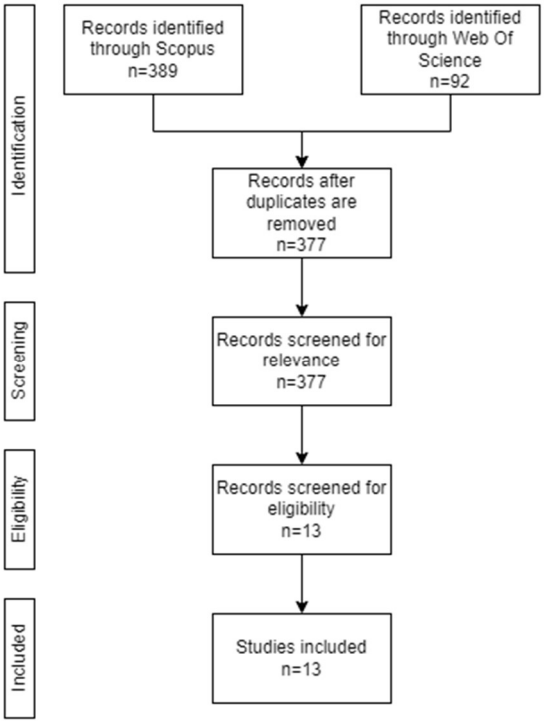

Figure 1. Flowchart Depicting the Study Selection Process

The fourth step, which is data extraction, was about gathering data from collected studies and aggregating them in an Excel spreadsheet. Based on the research question the protocol was designed to extract relevant information from selected studies (see Table 1) by adapting protocol suggested in [25].

Table 2. Protocol for Documenting Relevant Studies

Category	Metadata	Description
Descriptive Information	Article number	A study number, corresponding to the study number assigned in an Excel worksheet

	Complete reference	The complete source information to refer to the study (in APA style), including the author(s) of the study, the year in which it was published, the study's title and other source information.
	Year of publication	The year in which the study was published.
	Journal article / conference paper / book chapter	The type of the paper, i.e., journal article, conference paper, or book chapter.
	Journal / conference / book	Journal article, conference, where the paper is published.
	DOI / Website	A link to the website where the study can be found.
	Number of words	A number of words of the study.
	Number of citations in Scopus and WoS	The number of citations of the paper in Scopus and WoS digital libraries.
	Availability in Open Access	Availability of a study in the Open Access or Free / Full Access.
	Keywords	Keywords of the paper as indicated by the authors (in the paper).
	Relevance for our study (high / medium / low)	What is the relevance level of the paper for our study?

Approach- and research design-related information	Objective / Aim / Goal / Purpose & Research Questions	The research objective and established RQs.
	Study's contributions	The study's contribution as defined by the authors.
	Availability of the underlying research data	Whether the paper has a reference to the public availability of the underlying research data e.g., code repository etc., or explains why these data are not openly shared?
	Recommendation system techniques	Recommendation system techniques proposed in the study which can be content-based, collaborative, hybrid or other
	Features	Features of OGD Portals that were used as Input to Generate a Recommendation
	Environment in which tested	The environment the proposed recommendation system was tested

In the last step, data synthesis, collected data was analyzed and the insights of this analysis guided the implementation of the proposed recommendation method.

2.2 Implementation

In the second part of this study, a prototype for the proposed recommendation system is designed and implemented. The prototype is based on a sample of dataset's metadata used as an input, producing recommendations for each of the user selected dataset as an output. To evaluate the

effectiveness of the prototype, a survey page was designed to conduct UAT as alternatives such as Google Forms didn't fit our purposes due to lack of flexibility. This survey displayed the dataset metadata with the list of recommendations and enabled users to vote for the quality of the recommendations.

Functional requirements of the online survey were defined to understand respondents' usage of OGD portals and evaluate the effectiveness of the prototype while requirements of the recommendation system were defined

1. The online survey should show users selected dataset metadata alongside with their recommendations.
2. The online survey should allow users to rate the relatedness of recommendations to the chosen dataset.
3. The recommendation system should generate relevance ranking of recommended datasets for the main dataset from most relevant to least based on the metadata features

2.3 Evaluation

Evaluation of this recommendation method focused primarily on the accuracy. Accuracy evaluation is concerned with the quality of the suggested recommendation, while performance evaluation is concerned with time and memory resources required to generate recommendations. These two factors are important to understand the feasibility of applying this method in real life OGD portals.

To evaluate the accuracy of the recommendation, UAT was conducted. To this end, a survey page was designed providing users with 20 datasets with their calculated recommendations. It was hosted on Github Pages and disseminated through social networks of author and supervisors such Linkedin and Users were requested to rate the set of recommendations using 5-points Likert scale value from "Very irrelevant", "Irrelevant", "Neutral", "Relevant", "Very relevant".

3. Literature review results

This chapter covers the overview of the findings during the literature review, which includes analysis of descriptive information of reviewed studies, approach- and research-design related information about selected studies.

3.1 Descriptive analysis

This part discusses the descriptive analysis of the studies that are included in the SLR as this helps to identify trends in this field such as volume of studies over time and most active periods of research. In scope of descriptive analysis, we extracted data about the year of publication, the year in which the study was published, the type of paper, i.e., journal article, conference paper, or book chapter, journal article, conference, where the paper is published, a link to the website where the study can be found, a number of words of the study, the number of citations of the paper in Scopus and WoS digital libraries, availability of a study in the Open Access or Free / Full Access, keywords of the paper as indicated by the authors (in the paper) and finally, what is the relevance level (low/medium/high) of the paper for our study.

In terms of the year of publication, it seems like the popularity of the topic fluctuated through 2017-2023 (see Figure 2. Number of papers per year of the publication). Most of the papers were published in 2020, which was 4, while no papers were published in 2022. 3, 2, 1 papers were published in the years of 2017, 2018, 2019 respectively . Only 2 papers we published in 2023 while no papers were found from 2024.

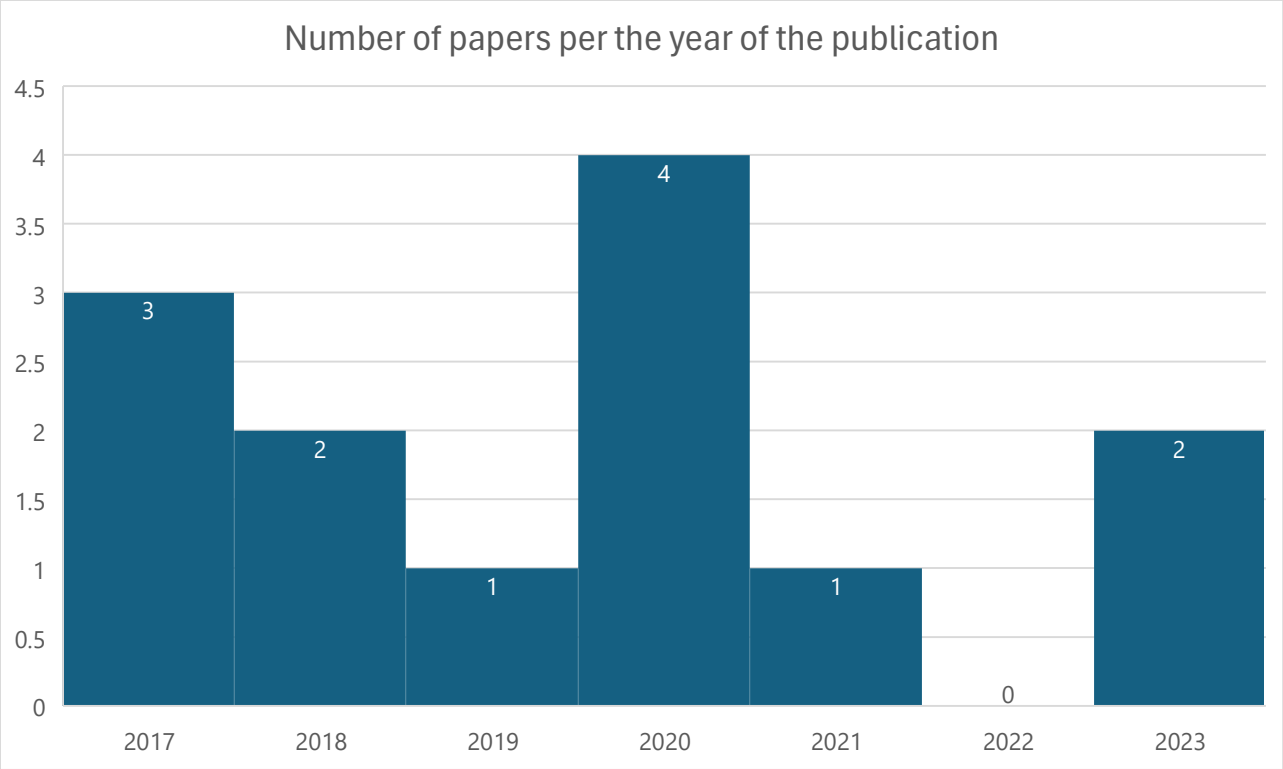

Figure 2. Number of papers per year of the publication

Regarding the type of papers, 9 of them were conference papers while only 4 were journal articles. The average number of words per paper was 5763 with the highest being 6058 and lowest being 1613. The average number of citations per paper according to Scopus in September 2024 was around 5, ranging from 0 to 23. Regarding keywords, the most frequent one was "data" with frequency of 20, followed by "open", "government" and "recommendation" with frequency of 15, 6 and 6 respectively. More detailed depiction of the keywords' frequency can be found in the Figure

3. Keywords cloud of the studies included in the literature review where the font size of the keyword is proportional to its frequency.

Figure 3. Keywords cloud of the studies included in the literature review

As for the relevance of the studies, 3 of them were highly relevant, while only 1 of them had low relevance. Studies that were about datasets recommendation systems were considered highly relevant while studies that focused tags or category recommendations were considered medium. Studies not falling under these 2 categories were accepted with low relevance.

3.2 Approach- and research- related information

This section is meant to cover information about the current approaches for recommending datasets, their effectiveness and the gaps in current literature.

Regarding the objectives, all selected studies [1-13] focused on improving the findability of OGD with a recommendation system except one [9] that focused on the assessment of the current state of open data portals focusing on the Latvia and recommended corrective actions. Some studies'

[4][6][11][22] objectives were to improve the findability of OGDs by suggesting a tag or category recommendation system which would automatically fill the category or tag field of dataset's metadata based on its content.

Another group of studies [4][8] was aiming to generate a knowledge graph over the OGD's which can be used to recommend datasets. One study [13] aimed to improve the process of making geographic models by automatically recommending open datasets that fit geographics model's input data requirement. Another paper [3] investigated the importance of user-driven feature weights in the open dataset recommendation. [2][7][8] focused on creating a recommendation

method that output related datasets based on chosen dataset's metadata using TF-IDF vectorization technique.

All proposed recommendation methods were tested on OGD portals as prototype is listed in Table 2.

Table 3. Studies and the portal used for testing recommendation method

Study	**Open Data portal**
Ahmed(2023).	European data portal
Yamada, Y. (2023).	Japanese data portal
Hsu, I., & Lin, Y. (2020).	Taiwanese data portal
Sornkongdang, N., Sanglerdsinlapachai, N., & Anutariya, C. (2021).	Taiwanese data portal
Devaraju, A., & Berkovsky, S. (2018).	CSIRO data access portal
Chen, I.-C., & Hsu, I.-C. (2019).	Taiwanese data portal
Ojo, A., & Sennaike, O. (2019)	DubLinked(CKAN)
Silva, G. O. M. D., Souza, P. R. D., & Durão, F. A. (2020).	DbPedia
Yamada, Y., & Nakatoh, T. (2018)	Data.gov (USA)
Sennaike, O. A., Waqar, M., Osagie, E., Hassan, I., Stasiewicz, A., Porwol, L., & Ojo, A. (2017)	DubLinked
Zhu, Y., Zhu, A.-X., Feng, M., Song, J., Zhao, H., Yang, J., Zhang, Q., Sun, K., Zhang, J., & Yao, L. (2017).	GeoData.CN

Next, the features of dataset that were used by proposed recommendation methods in studies were reviewed. Figure 4. Dataset metadata fields used for in the proposed recommendation

systemsshows most used features. “Title” was the most common feature, followed by “description” and “keywords”. Some studies [4][6][11][22] were focused on recommending the “category” for the dataset, despite that “category” was the third most common feature used as input in recommendations. Besides these, there were also other features that were used only by one recommendation method. For example, one study [1] was using Facebook posts liked by the user to recommend datasets based on them. Another study [13] about recommending datasets suitable for “geography research” was also utilizing temporal and spatial coverage of the data. In the scope of linked open data, one study [10] combined links with the RDF literals in the dataset metadata as an input for the recommendation system.

Dataset metadata fields	**Studies**
Title	[2][4][5][6][11][12][13]
Description	[2][4][5][6][11][13]
Keywords/tags	[2][5][11][12][13]
Category/theme/topic	[2][5][7][8][12][13]
Author	[2]
Spatial coverage	[2][12][13]
Other	[1][2][12][13]

Figure 4. Dataset metadata fields used for in the proposed recommendation systems

3.3 Results and their analysis

3.3.1 Research question 1: What are the existing proposals for OGD portal recommender systems?

Regarding existing proposals of OGD recommender systems, there were 2 types. The first type focused on recommending tags and categories; the others were recommending other datasets as in this study.

There were 4 studies [4][6][11][22] about tags recommendation systems. One study proposes two tag recommendation methods for OGDs [11]. The first used method is based on multi-label classification (support vector machine, random forests and multinomial naive Bayes methods) based on the already set tags of dataset, while the other is based on extracting frequent noun phrases from the title and description of the dataset [11]. Another study was continuation of this

focusing on the infrequent tags which are harder to recommend [6]. In contrast to these, the study [4] utilized pretrained transformer-based model WangchanBERTa instead of multi label classification to capture the semantic meaning of the dataset's textual metadata. Finally, the study [22] used Large Language Model (LLM) GPT-4 to automatically recommend tags.

As for the studies proposing systems recommending datasets, [2] combined content-based similarity with item-to-item co-occurrence to recommend datasets and conducted a user study to evaluate its effectiveness. Using item-to-item co-occurrence requires user's login to OGD portal, which is not common among OGD portals. It is also dependent on "title", "description", "keywords", "activity", "research fields", "creators", "contributors", "spatial", "search" and "download," which makes it less compatible with different OGD portals since not all of them are supported by the. For textual fields, TF-IDF term weighing and Cosine Similarity to compute the similarity score for each feature which fails to capture the semantic meaning. There were 2 studies -[8][12]- that focused on using Self Organized Maps (SOM) to generate knowledge graph of the datasets and calculate relatedness of datasets based on that. These also rely on TF-IDF term weighing technique for processing textual fields [8][12]. Additionally, there were 2 proposed recommendation systems that were based on linked open data [7][10]. One of them used a hybrid solution combining both links between datasets and literals in them to calculate their relatedness [10]. The other proposed a recommendation platform with a query interface that would recommend datasets based on a query [7]. One study proposed the recommendation system by integrating user's Facebook profile to OGD portal and using K-NN to recommend datasets based on the Facebook posts that user liked [1]. Another study - [20] - aimed to ease the process of making geographic models by proposing a recommendation system for open datasets that are fit to geographic model's input data requirement [20].

3.3.2 Research question 2: How can a recommender system be developed for OGD portals based on the identified techniques and features?

This question deals with the design of a recommender system which includes features and techniques to be used. Recommendation system to be utilized in the scope of the OGD should fit the following criteria to make sure it is adaptable to the diverse structures of OGD portals:

- It should not depend on the user's preferences or profile since most OGD portals don't require login for access
- It should rely on aspects of open dataset specification that are fulfilled in most OGD portals so that it is suitable for most.

Regarding features to be used, study [3] conducted a survey about importance of the metadata features in measuring similarity between datasets. In this survey, "title", "description" and "category" turned out to be most important while "creators" and "contributors" less important.

Moreover, study [15] conducted analysis of 41 open data portals where among open dataset specification aspects, average highest scores belonged to aspects of "Thematic categories and tags", "open data license" and "description of the dataset" in that order.

Therefore, it was concluded to propose recommendation system that is based on "title", "description" and "category". This makes recommendation method adaptable to a wider range of OGD portals.

As for techniques utilized in existing recommendation systems, almost all of them were content based except one which used hybrid combining both collaborative and content-based techniques. This is related to the fact that most OGD portals do not require login for data access. Going into more detail about techniques, to calculate the similarity measure between dataset's metadata, studies used different methods such as Cosine [2][10], Jaccard [2], Euclidean distance [2] [6] [8][12], Linked data semantic distance (LDSD) [10]. Cosine similarity was used to measure similarity between vector representation of text features for dataset metadata [1][10]. As for categorial fields such as "creators", "contributors", Jaccard similarity was preferred [1]. In scope of linked open data LDSD was also used to measure similarities between datasets based on direct

or indirect links between them [10]. [8] [12] utilized Euclidean distance as distance metric in the scope of SOM to calculate similarity between datasets. In the study [2] Euclidean distance was used to measure distance between spatial information by first converting it to a coordinate system. Since some of the dataset metadata is in text format, it is required to convert it machine understandable, word embeddings. Most studies [1][2][10] used TF-IDF for this which fails to capture the semantic meaning of the text.

As the proposed recommendation method will be dependent on textual features of dataset metadata: “title”, “description” and “category”, it was decided instead of TF-IDF to use pretrained Word2Vec model which is more advanced method of vectorization that can also capture the semantics of these features.

4. Implementation

This chapter covers the implementation of the prototype for the proposed OGD recommendation method.

The proposed method of providing recommendations for OGD datasets consisted of 4 stages: downloading a sample of dataset metadata, preprocessing, vectorization and calculating the similarities between datasets. It was implemented in the Kaggle Python notebook since it offers good performance for large computations. European data portal (https://data.europa.eu/) was chosen as data source since it aggregates OGD portals of 35 European countries as of December 2024.

As the first stage of implementation, "requests" python library was used to download a sample of 35308 dataset metadata from European data portal, which had 1 815 240 datasets as of December 2024. Every 50th dataset metadata including "id", "title", "description", "last modified date", "categories" from European data portal's API were downloaded and saved in the CSV file to be available for later access using "pandas" library.

In the next step, in the scope of preprocessing the data, the category label in the English language was extracted from the "categories" column that is from European data portal and added to newly created "category" column in scope of the code.

The third step was vectorization, which is the process of converting textual data to vector representation of it. The "title", "description" and "category" columns that contained text were converted to "title_vector", "description_vector", "category_vector" columns that contained word vectors using library Spacy. Spacy is an open-source library that offers trained pipelines for building natural language processing tools. Among these pipelines "**en_core_web_lg**" was chosen as it is the largest pipeline spacy offers for English language including word vectorization.

In the fourth stage, the word vectors were used to calculate similarity between title, description and categories of sampled datasets. Spacy's similarity function was utilized for this task which calculated the cosine similarity using an average of word vectors.

The square sum of calculated similarity scores for titles, description and categories between datasets was accepted as similarity score between datasets. These scores were used to create a similarity matrix with value in matrix at index (i, j) denoting similarity scores between i-th and j-th datasets. This similarity matrix was used to rank the most similar datasets to i-th dataset by sorting i-th row of the matrix.

5. Evaluation

To evaluate effectiveness of the proposed recommendation system prototype, UAT was conducted through an online survey which was hosted on Github Pages. It was disseminated through social networks such as Linkedin. The introduction part of the survey including consent collection can be found in the Appendix. It consisted of 3 parts. The first part collected data about responder's usage of OGD portals and European OGD portal such frequency of the use and satisfaction. The second part was intended to evaluate the quality of the recommendations through the acceptability task, where 20 dataset metadata with 5 recommendations produced by the prototype were provided for the users. 10 of the 20 set of recommendations were mandatory to rate while the rest was optional. The third part gathered general feedback about the recommender system to be transformed into the prototype improvement agenda, which, however, was optional.

In terms of response, 37 users responded to the survey and 671 datasets with recommendations were rated for their quality.

Regarding the first part, Figure 5 Figure 5. "How often do you use any open data portal?" responsesshows the responses for the first question "How often do you use any open data portal". Most respondents use an open data portal once a month, followed by once in a year and never. Only one respondent used any open data portal daily, the same as once a week.

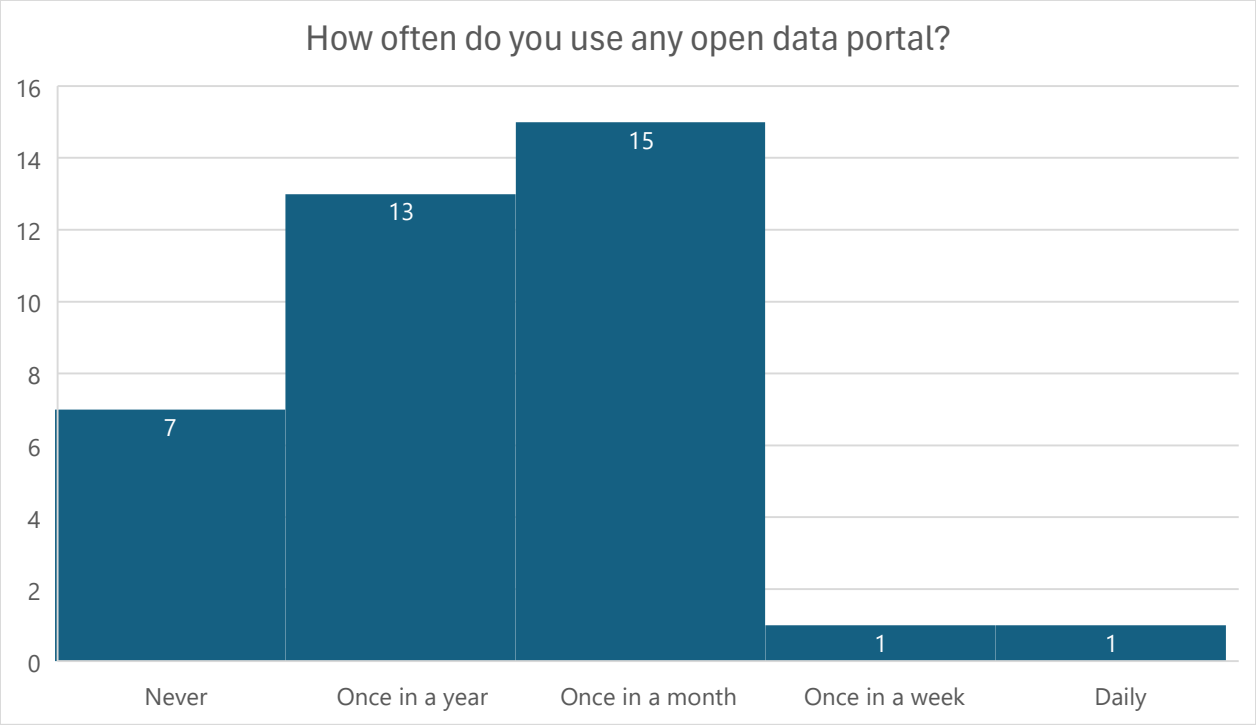

Figure 5. "How often do you use any open data portal?" responses

Compared to this, European open data portal usage is less frequent than any open data portal usage as can be seen in Figure 6. "How often do you use European open data platform?" response. The number of respondents who used European OGD portal once a year was the

highest followed by the number of “never” responses and the number is decreasing with the frequency, with none of the users using it daily.

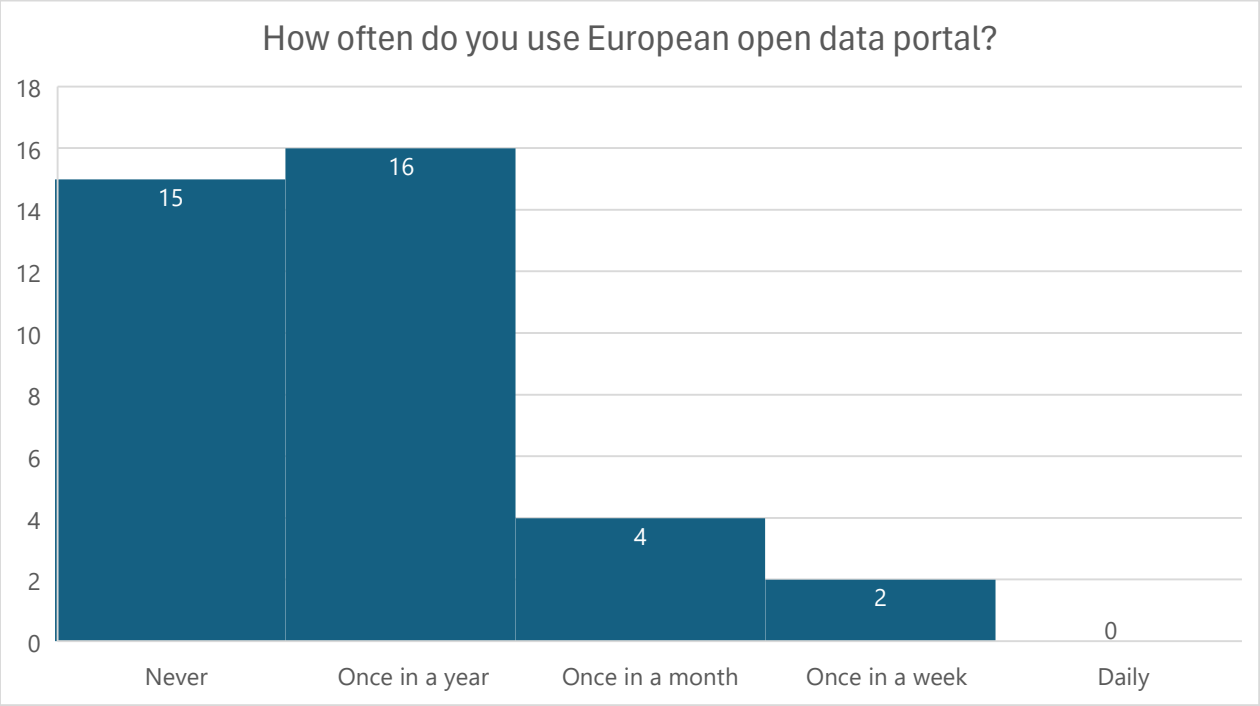

Figure 6. "How often do you use European open data platform?" response

Another question from the first part of the survey investigated how easy do users find navigating to the datasets of their interest. As can be seen from the Figure 7. "How easy do you find it to navigate datasets of your interest?" response, 6 of 37 users find it difficult, while 10 users find it easy or very easy. Most of the users responded neutral to the question.

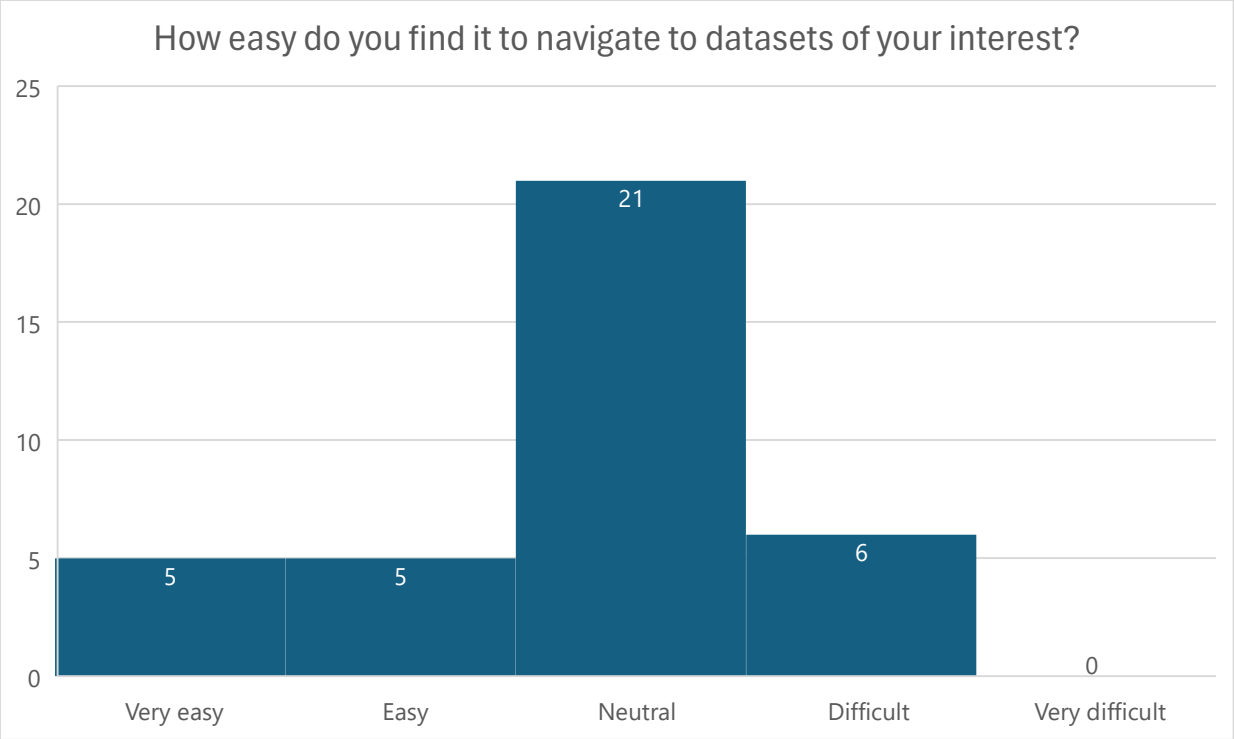

Figure 7. "How easy do you find it to navigate datasets of your interest?" response

Final question from the second part aimed to measure satisfaction of the respondents with the content of OGD portals on 5-points Likert scale. Figure 8. "To what extent do the contents of portals meet your expectations/needs?" responsesThe highest number of responses was 3 (see Figure 8), while the lowest, which was only 3 responses, was 5. Only 6, 7 and 6 of the responses

were 1, 2, 4 respectively. In general, it can be said that contents of portals tend to not meet expectations of respondents on average.

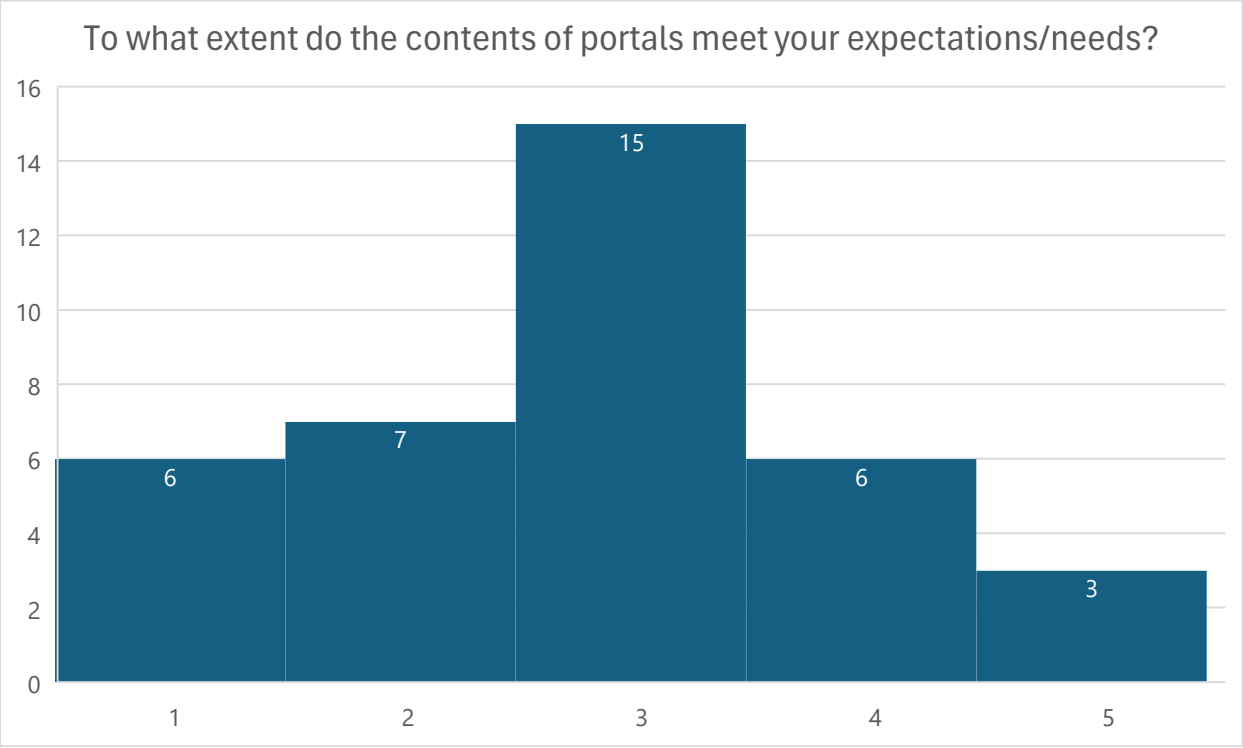

Figure 8. "To what extent do the contents of portals meet your expectations/needs?" responses

The second part of the survey aimed to evaluate the quality of prototype by asking respondents to rate the quality of recommendations per dataset. Every dataset was illustrated with 5 top recommended recommendations and with an option to rate the quality from “Very irrelevant” to

“Very relevant” as shown in Figure 9Figure 9. Survey - Example question to rate the quality of recommendations.

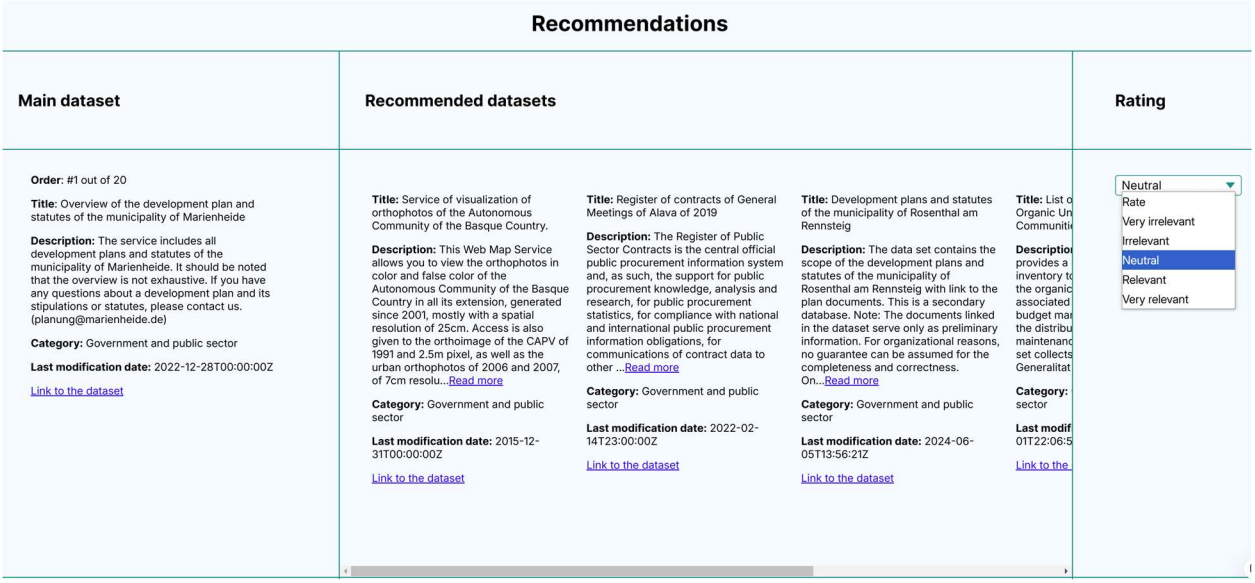

Figure 9. Survey - Example question to rate the quality of recommendations

As can be seen from Figure 10. Recommendation of quality responses, 63 percent of respondents found recommendations “Relevant” or “Very relevant,” while only 15 percent found it “Very

irrelevant" or "Irrelevant". The rest which comprised of 22 percent of respondents rated the quality of recommendations as "Neutral".

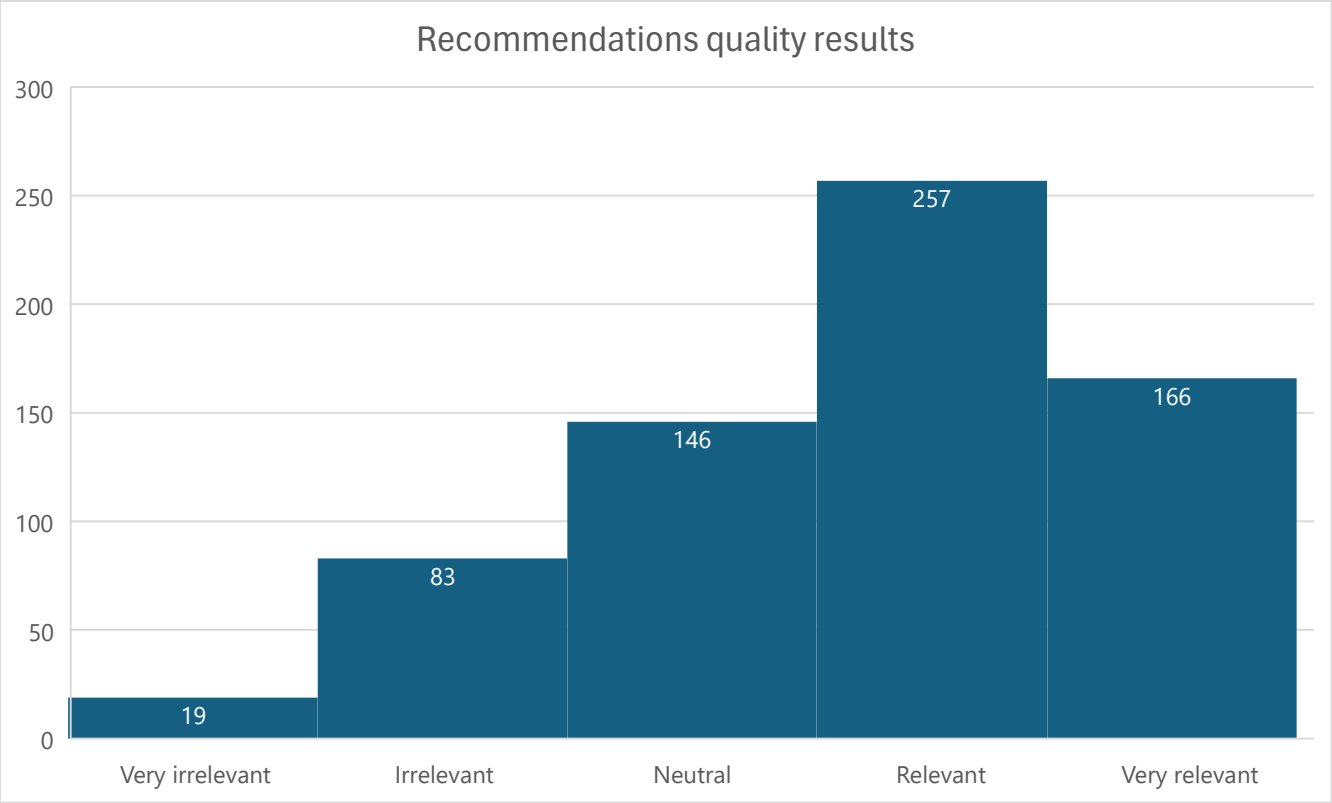

Figure 10. Recommendation of quality responses

In the last section, respondents were asked to optionally comment on the recommender system. There were 4 comments in total. There were 3 respondents that suggested that it was difficult to rate the quality of the recommendations for some datasets as the title and description of the datasets contained terminology and abbreviations that are hard to understand. The other comment suggested to include the approach for recommending datasets in the survey page.

6. Discussion

This chapter discusses the results of the study. The results suggest that majority of respondents use Open Data portals at most once a month and even more infrequently use European Open Data portal. They are also mostly neutral in terms of data findability and user satisfaction with OGD portals. As for the quality of the dataset recommendations by the prototype, they were found to be mostly relevant although the descriptions of the datasets were described as difficult to understand as they were of scientific nature.

These results illustrate that it is possible for the proposed recommendation to provide relevant recommendations just by relying on dataset's title, description and category.

This study has limitations, one of which is that this study didn't consider already integrated recommender systems for OGD portals, only the ones found in the literature. Moreover, the prototype was implemented and evaluated with only limited number of datasets from European data portals, namely only 35308 datasets out of 1815240 datasets present as of December 2024. Additionally, this study didn't evaluate the proposed recommendation method for its performance.

The main limitation of the survey is the number and the nature of respondents. Although there were 671 explicit judgements of the quality of recommendations, it came from 37 respondents, a significant number of which never uses OGD portals. Moreover, this prototype was built only on the sample datasets from European Open Data portal.

7. Conclusion

This study aims to contribute to improving data findability in OGD portals by proposing a recommendation method for datasets. In this study, SLR was conducted about existing OGD portal recommendation methods and techniques utilized in them. This review revealed that most recommendation systems focus on tags and category recommendations while the ones recommending datasets use vectorization methods that fail to capture semantics of dataset metadata. Based on these insights, a new recommendation method was proposed that recommend relevant datasets for a certain dataset by calculating relatedness between dataset's metadata such as title, description, and category. In contrast to existing recommendation methods, this method also captures the semantic relatedness between dataset's metadata by relying on Word2Vec method of word embeddings instead of TF-IDF that is used in existing methods. Additionally, by relying on only the most common dataset metadata i.e. title, description and category, this method is suitable for a larger number of OGD portals than existing methods. At last, UAT was conducted to collect feedback which proved the proposed recommendation method to be effective. As for future research, the effectiveness of this method can be evaluated on a larger sample along with the performance.

Acknowledgements

I am grateful to Anastasija Nikiforova and Dimitrios Symeonidis for their support and guidance during this research. I am also grateful to the University of Tartu for giving me the opportunity to pursue my studies with tuition waiver and Dora Plus scholarship.

Writing Assistance

ChatGPT 4-o is Transformer-based pre-trained model by Open AI which was used in some parts of this thesis to improve writing and correct grammatical mistakes.

References

1. Hsu, I. C., & Lin, Y. H. (2020). Integrated machine learning with semantic web for open government data recommendation based on cloud computing. Software: Practice and Experience, 50(12), 2293-2312.
2. Devaraju, A., & Berkovsky, S. (2018, July). A hybrid recommendation approach for open research datasets. In Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization (pp. 207-211).
3. Devaraju, A., & Berkovsky, S. (2017, August). Do Users Matter? The Contribution of User-Driven Feature Weights to Open Dataset Recommendations. In RecSys Posters.
4. Sornkongdang, N., Sanglerdsinlapachai, N., & Anutariya, C. (2021, December). DataCat: Attention-based Open Government Data (OGD) Category Recommendation Framework. In 2021 16th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP) (pp. 1-6). IEEE.
5. Ahmed, U. (2023). Reimagining open data ecosystems: a practical approach using AI, CI, and knowledge graphs. In BIR Workshops (pp. 235-249).
6. Yamada, Y. (2023). Tag Recommendation System for Data Catalog Site of Japanese Government. In KMIS (pp. 325-331).
7. Chen, I. C., & Hsu, I. C. (2019). Open Taiwan Government data recommendation platform using DBpedia and Semantic Web based on cloud computing. *International Journal of Web Information Systems*, *15*(2), 236-254.
8. Ojo, A., & Sennaike, O. (2019, December). Constructing knowledge graphs from data catalogues. In *International Conference on Distributed Computing and Internet Technology* (pp. 94-107). Cham: Springer International Publishing.
9. Nikiforova, A. (2020, July). Assessment of the usability of Latvia's open data portal or how close are we to gaining benefits from open data. In *IADIS 14th International Conference on Interfaces and Human Computer Interaction* (pp. 51-28).
10. Silva, G. O. M. D., Souza, P. R. D., & Durão, F. A. (2020). HSLD: a hybrid similarity measure for linked data resources. International Journal of Metadata, Semantics and Ontologies, 14(1), 16-25.
11. Yamada, Y., & Nakatoh, T. (2018). Tag Recommendation for Open Government Data by Multi-label Classification and Particular Noun Phrase Extraction. In KMIS (pp. 81-89).

12. Sennaike, O. A., Waqar, M., Osagie, E., Hassan, I., Stasiewicz, A., Porwol, L., & Ojo, A. (2017, September). Towards intelligent open data platforms: Discovering relatedness in datasets. In 2017 Intelligent Systems Conference (IntelliSys) (pp. 414-421). IEEE.
13. Zhu, Y., Zhu, A. X., Feng, M., Song, J., Zhao, H., Yang, J., ... & Yao, L. (2017). A similarity-based automatic data recommendation approach for geographic models. International Journal of Geographical Information Science, 31(7), 1403-1424.
14. Patel, B., Desai, P., & Panchal, U. (2017, March). Methods of recommender system: A review. In 2017 international conference on innovations in information, embedded and communication systems (ICIIECS) (pp. 1-4). IEEE..
15. Nikiforova, A. (2020, July). Comparative analysis of national open data portals or whether your portal is ready to bring benefits from open data. In *IADIS International Conference on ICT, Society and Human Beings* (pp. 21-23).
16. Arnaud, F., Pignol, C., Stéphan, P., Develle, A. L., Sabatier, P., Evrard, O., ... & Caillo, A. (2017). From core referencing to data re-use: two French national initiatives to reinforce paleodata stewardship (National Cyber Core Repository and LTER France Retro-Observatory). In *5th PAGES Open Science Meeting*.
17. Selva Birunda, S., & Kanniga Devi, R. (2021). A review on word embedding techniques for text classification. *Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2020*, 267-281.
18. Attard, J., Orlandi, F., Scerri, S., & Auer, S. (2015). A systematic review of open government data initiatives. *Government information quarterly*, *32*(4), 399-418.
19. Barcellos, R., Bernardini, F., & Viterbo, J. (2022). Towards defining data interpretability in open data portals: Challenges and research opportunities. Information systems, 106, 101961.
20. Molodtsov, F., & Nikiforova, A. (2024). From an Integrated Usability Framework to Lessons on Usability and Performance of Open Government Data Portals: A Comparative Study of European Union and Gulf Cooperation Council Countries. arXiv preprint arXiv:2406.08774.
21. Kitchenham, B. (2004). Procedures for performing systematic reviews. Keele, UK, Keele University, 33(2004), 1-26.

22. Kliimask, K., & Nikiforova, A. (2024, September). TAGIFY: LLM-powered Tagging Interface for Improved Data Findability on OGD portals. In *2024 Fifth International Conference on Intelligent Data Science Technologies and Applications (IDSTA)* (pp. 18-27). IEEE.
23. Ahmed, U., Alexopoulos, C., Piangerelli, M. and Polini, A., 2024. BRYT: Automated keyword extraction for open datasets. *Intelligent Systems with Applications*, *23*, p.200421.
24. Francey, A. (2023, August). Drivers of Dissatisfaction with an Open Government Data Portal: A Critical Incident Technique Approach. In *International Conference on Electronic Government* (pp. 279-294). Cham: Springer Nature Switzerland.
25. Zuiderwijk, A., Chen, Y.C. and Salem, F., 2021. Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. *Government information quarterly*, *38*(3), p.101577.
26. Tsatsaronis, G., Varlamis, I., & Vazirgiannis, M. (2010). Text relatedness based on a word thesaurus. *Journal of Artificial Intelligence Research*, *37*, 1-39.
27. Visser, M., Van Eck, N. J., & Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative science studies*, *2*(1), 20-41.

Appendix

I. Survey introduction

“Dear Participant,

While open government data (OGD) portals, such as the European Open Data Portal provide a plenty of datasets, it is often difficult to navigate to those of interest. To simplify this navigation, we propose a recommendation system, which, for every selected dataset, recommends potentially relevant other datasets (to be used instead or as complementary to the selected dataset). With the developed prototype, we seek your feedback on the effectiveness of this approach, asking you to evaluate the recommendations generated by our prototype through the below survey.

Please, note that this is a prototype, where we are testing the relevance of produced recommendations. Once developed, it will be integrated in a portal (not standalone tool).

Main dataset column contains metadata (title, description, category, modified date and the link to the dataset) of several datasets from the European Open Data Portal.

Recommended datasets column provides metadata (title, description, category, modified date and the link to the dataset) of datasets that are recommended based on the former dataset (based on similarity assessment). The number of recommended datasets per dataset of interest is 5, where to see all of them, we kindly ask you to use the slider. We ask you to rate the “quality”/relevance of recommendation under the Rating column. After rating individual datasets, you will be able to provide free-form feedback, where we encourage you to share your opinion and suggest improvements. Should you be willing to comment on an individual dataset and respective recommendations, please use ID of dataset (column 1).

While the survey contains 20 datasets, we ask you to assess at least 10 of them, which will take you 10 minutes, and, should you have a bit more time, we will be very grateful if you will assess the remaining 10 datasets, which will take you an additional 10 minutes. Filling this survey is expected to take 20 minutes.

The data collected in this study are completely anonymous. No personal data or identifiable information will be collected and the information you choose to provide in this study cannot be

connected back to you. Results from this study will be aggregated and may be used in scientific articles to be published in a journal or presented at scientific conferences. Your participation in this study is entirely voluntary and you can withdraw at any time.

For questions, please contact ramil.huseynov@ut.ee.

Given the above, you consent voluntarily to be a participant in this study, and understand that you can withdraw from the study at any time, without having to give a reason.

We appreciate your input very much!”

II. Code

Repository link: https://github.com/huseynovramilx/OGD-recommendation

III. License

Non-exclusive licence to reproduce the thesis and make the thesis public

I, _Ramil Huseynov_____,
(*author's name*)

1. grant the University of Tartu a free permit (non-exclusive licence) to

reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, my thesis

_A recommender system for improved data findability in open government data portals_____
_____,
(*title of thesis*)

supervised by _Anastasija Nikiforova, Dimitrios Symeonidis_____.
(*supervisor's name*)

2. I grant the University of Tartu a permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in points 1 and 2.

4. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Ramil Huseynov
10/01/2000