A Methodology for Retrieving Datasets from Open Government Data Portals Using Information Retrieval and Question and Answering Techniques

Raissa Barcellos, Flavia Bernardini, Jose Viterbo

A methodology for retrieving datasets from open government data portals using information retrieval and question and answering techniques

Raissa Barcellos[1], Flavia Bernardini[1], and Jose Viterbo[1]

Fluminense Federal University, Institute of Computing, Rio de Janeiro, Brazil
raissabarcellos@id.uff.br,{fcbernardini,viterbo}@ic.uff.br

Abstract. Public administration is one of the largest producers and collectors of data in several domains. Bearing in mind that one of the main pillars of public transparency is the release of social value through government data, and that this right remains a challenge for citizens, in this work we implemented a methodology that uses techniques and Answers and Information Retrieval to retrieve data sets and respond to questions from citizens on a Brazilian open government data portal. Using metrics of information retrieval evaluation, such as accuracy and recall, we obtained satisfactory results in applying the proposed methodology.

Keywords: Open Government Data · Information Retrieval · Question and Answering

1 Introduction

Around the world, governments have already realized the importance of higher data openness and data transparency [3]. Currently, there are open data portals in whole world, which provide access to large volumes of them in increasing numbers around the globe. Nowadays, more than a million datasets have been made available by governments around the world [14]. This fact is a consequence of a large amount of public information generated continuously, which can refer to finance, health, human development, among other topics [9]. Open government data allows easy access to large amounts of data without the need for repeated data requests, transcription of data from printed formats to electronics, and other tasks that would limit the user's interest and usefulness of the data [22]. Through this openness, the government has the potential to promote transparency, increase citizen participation, and stimulate innovation. Also, open data initiatives can help citizens learn about government activities [22].

Government data geared to the citizen's objective is a useful contribution to decision making. However, open government data portals are complex and multifunctional, due to a large amount of data, usually fragmented and hidden within the portal. An earlier study, conducted by [15], with data professionals across a wide range of domains and skillsets states that, in most cases, data search shows characteristics of a complicated task, involving multiple queries, iterations

and improvement of the need for original information, in addition to complex cognitive processing. It becomes a challenge for citizens to access, retrieve, and understand the data made available [14]. So, the goal of this work is to present a methodology that uses Question and Answering and Information Retrieval techniques to retrieve datasets and answer citizen' questions, using technologies like natural language processing and the CKAN API for data retrieving.

The work is organized as follows: In Section 2, we present a background on open government data, information retrieval, and Question-Answering. In Section 3, we present a literature review. In Section 4, we present our methodology. In Section 5, we present our experimental analysis. Finally, in Section 6, we present our conclusions and future work.

2 Background

Open Government Data: According to Jetzek *et al.* [11], the open data term refers primarily to data that has been created or collected by government agencies for one purpose, but which is now available to the public for other purposes. Opening government data leads to several different communities using it simultaneously for different and productive purposes, and can be used as a means to achieve many ends. According to Jetzek *et al.* [11], the following three criteria are arguments in favor of the potential benefit of opening government data: (i) the data can be shared and used by many at the same time at no additional cost; (ii) the data have productive value, often used as a resource for the production of something of interest and are rarely consumed directly; and (iii) the data is versatile and can be used as an input for a wide range of goods and services.

Also, according to Alzamil *et al.* [2], the availability of government data can play a crucial role in decision-making by government agencies. By making government data available to the public, citizens, professionals, and other interested groups can access the data to help monitor public spending and increase overall participation. In addition to better decision making, making data available can help prevent or minimize the abuse of government resources. When government officials know that their work is being monitored, they are less likely to make mistakes [2].

In this scenario, one important issue in Open Government Data is finding and recovering information from different sources. However, there is a lack of patterns for providing data in data portals, as well as data models interpretable by machines that could be useful for this task. In this work, we propose the use of Information Retrieval (IR) and Question and Answering (QA) techniques for this end.

Information Retrieval: According to [18], Information Retrieval (IR) is related to finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored in computers). So, IR is frequently used for helping people to recovery documents, typically by librarians, paralegals, and similar professional searchers.

In order to assess the quality of IR systems, we should measure retrieved data sets as relevant and non-relevant, through formal and appropriate IR assessment metrics. Two metrics widely applied for evaluating IR systems are Precision and Recall. For calculating these metrics, for each search result, all retrievable items fit one and only one of the four cells in a confusion matrix, shown in Table 1: (i) Retrieved or Not Retrieved, and (ii) Relevant or Not Relevant [5]. In this same Table, N_{rec} is the number of retrieved data and N_{rel} is the number of relevant data sets.

	Relevant	**Not Relevant**	**Total**
Retrieved	$N_{rec \bigcap rel}$	$N_{rec \bigcap \overline{rel}}$	N_{ret}
Not Retrieved	$N_{\overline{rec} \bigcap rel}$	$N_{\overline{rec} \bigcap \overline{rel}}$	$N_{\overline{ret}}$
Total	N_{rel}	$N_{\overline{rel}}$	N_{tot}

Table 1: Confusion Matrix

For a retrieved dataset, with each item labeled as relevant or irrelevant, and its respective constructed confusion matrix, Recall R is the proportion of the number of relevant retrieved items regarding to all relevant items, i.e., $R = N_{\mathrm{rec} \bigcap \mathrm{rel}} \div N_{rel}$.Precision P is the proportion of the number of relevant items regarding to all retrieved items, i.e., $P = N_{rel} \div N_{rec}$.

Question-Answering: Question-Answering Systems (QAS) are an extension of search engines using IR techniques, as they aim to automatically provide users accurate answers to questions asked in natural language, rather than only returning a list from relevant sources based on a set of keywords [6]. The architecture of this type of system depends on the type of explored sources. A common approach is to use traditional IR techniques combined with Machine Learning and Natural Language Processing methods to extract answers from simple questions [6]. According to [10], the QA systems present an architecture composed of three basic modules: (i) processing of the question, (ii) processing of the document, and (iii) processing of the response, as shown in Figure 1.

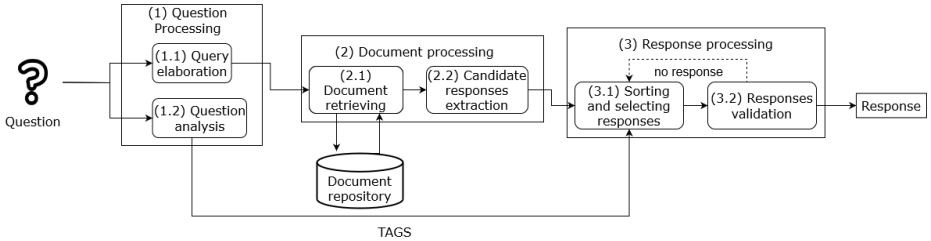

Fig. 1: Basic architecture of QA systems.
Fonte: [1]

The first step "(1) Question Processing" consists of transforming the question asked by the user — from natural language to computational language — and understanding the question in order to extract information that supports the selection of answers. In the second stage, "(2) Document Processing", the query already generated is used to retrieve relevant documents or datasets, and to extract candidate responses. Finally, in the third step, "(3) Response Processing", the candidate responses are ordered based on the analysis of the similarity between the question and the candidate responses and, thus, presented to the user. QAS commonly present a single answer to the user's question. However, they can also present several answers ordered by relevance, whether these responses are in text or visual format [4].

3 Literature Review

There are some works in literature that tries to enrich user experience when leading to open government data, turning easier to manipulate the available data. Lahti et al (2017) [17] introduce a package in R programming language, called Eurostat, for providing a rich collection of data through the open data service, including thousands of demographic datasets, economy, health, infrastructure, traffic, and other European topics. Statistics are generally available with optimal geographic resolution and include time-series that span several years or decades. The implemented package provides customized tools for accessing Eurostat's open data. The main features, such as cache, date formatting, organized data principles, and table data format, support seamless integration with other tools for data manipulation and analysis.

Kuo and Chou (2019) [16] propose an evaluation mechanism based on metadata for analysis of metadata quality, spatial similarity and temporal similarity between user requests and open data. When evaluating metadata for each dataset, the metadata score in the engine contributes two scores, mandatory and optional scores. From the required fields and a combination of recommended and optional metadata fields, they calculate two scores. By extracting keywords from the user's input or addressing the location on a map, they perform the spatial similarity analysis. Also, they conducted a temporal similarity analysis comparing a query and the open dataset. A developed platform presents data related to punctuation, similarity, and classification that help users to obtain the expected data.

A significant problem that Smart Cities' information and data management systems are currently facing is the heterogeneity, not only of data flow but also of external data sources, such as the Linked Data web, whose use is inevitable in making decision making on the scale of a city. Kettouch et al (2017) [12] present a theoretical framework to manage data flow in Smart Cities and integrate it immediately with Linked Data for data retrieving and web semantics. The structure consists of a modular architecture that receives, as input, semi-structured, and heterogeneous linked data retrieved from various sensors and the data network, respectively. The general objective of the system is to offer an interface to

facilitate automatic access to data from Smart Cities and publish it on Linked Datacloud to allow for future reuse.

Oliveira et al (2016) [19] propose a platform capable of enriching police reports with public data available on the Web. The authors present an evaluation tool, which consumes the linked data provided by a Web API, expands it with data extracted from other sources, and provides an interface to assess the relevance of the expansions. The tool is capable of retrieving police reports to store data in a structured format. The main component of the proposed platform is a REST Web API that offers police report data extracted from the Public Security Secretariat of the state of São Paulo system website. The Web API provides report information in a structured and semantically annotated manner with vocabulary terms and ontologies used in linked open data.

As we can observe in these works, none of them present a general proposal that gather data in order to smartly responding open government data portal users. As far as we know, there is not any work in literature that explores IR and QA techniques in this scenario. In what follows, we present our methodology.

4 Methodology

Usually, open government data portals offer to users at maximum a basic search interface, similar to the basic internet search engines. The goal of our methodology is to allow smart data retrieval by users in open government data portals. Is is basically devided into two steps: (i) Processing User Questions and (ii) Retrieving and Delivering Data. Figure 2 illustrates the methodology used in this work.

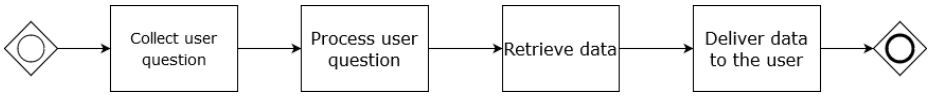

Fig. 2: Methodology illustration

Processing User Questions: Our methodology firstly identify the citizen's question. We used Natural Language Processing (NLP) techniques, like tokenization and removal of stopwords [18], for question processing, in order to point out the real purpose of the citizen when typing his search in the search bar of the application. For this, we used nlpnet library [8], a Python library for NLP tasks, based on neural networks. Currently, it performs identification of semantic functions. It can be used as a Python library or through its independent scripts. Most of the architecture is independent of the language, but some functions have been specially adapted to work with Portuguese. The nlpnet uses word vector representations as input resources for a conventional neural network architecture [7]. Based on the result of the semantic interpretation of the user's question, we retrieve the title and description of the datasets. In work [21], the authors also take

a similar approach, implementing a semantic analysis based on open government data, but the analysis is performed to minimize misunderstandings with the the semantics of the dataset categories used.

Retrieving and Delivering Data: From the results of the semantic interpretation of the user's question, a tool for collecting or havesting data is necessary in this step. We use the CKAN API or Comprehensive Knowledge Archive Network for data retrieval task. CKAN [1] is a web-based management system, developed by Open Knowledge Foundation [2] and is being used by more than 192 governments, institutions, and other organizations worldwide to manage open government data. CKAN is written in Python and it is useful for developers who want to write code that interacts with CKAN sites and their data. The API is also extensively documented and provides a comprehensive way to retrieve the metadata from the [13] data catalog. For performing data retrieval, it is necessary to perform an API request, where the answer is a JSON file in a data dictionary format. In order to demonstrate the results, we developed two usage scenarios for the experiment.

5 Experimental Analysis

We implemented a tool that follows the steps of our methodology, available at `https://github.com/RaissaBarcellos/openIR`, using Python. For conducting our experiment, we retrived data from a catalog of open government data in Brazil called *dados.gov* [3]. It aggregates data from the 26 Brazilian federal states and their municipalities. In order to evaluate the results obtaines in both scenarios, we must label each answer for a question from the retrieved answers as relevant and not relevant, in order to calculate P and R metrics. For calculating Recall, it should be necessary to verify each of the 39269 datasets if they are relevant or not for each question. So, for calculating R (Recall), we used Gauss curve equation to sample data fromt he portal [20]. Eq. 1 defines the number n of samples should be selected from the entire population. In this equation, n is the sample size to be calculated; N is the population size (in the case of the experiment, 39269 datasets); Z is the standard normalized variable associated with the confidence level (in this case, a confidence level of 95%); p is the proportion we want to find (in this case, $\mathrm{p} = 0.5$); e is the sampling error (in this case, $\mathrm{e} = 0.05$). Using this equation, we sampled 380 documents at random from the entire data.gov portal.

$$\mathrm{n} = \frac{\mathrm{N} \cdot \mathrm{Z}^2 \cdot \mathrm{p} \cdot (\mathrm{n} - \mathrm{p})}{\mathrm{Z}^2 \cdot \mathrm{p} \cdot (1 - \mathrm{p}) + \mathrm{e}^2 \cdot (\mathrm{N} - 1)} \tag{1}$$

Based on this, in order to validate our methodology, we conducted two experiments in two different usage scenarios.

[1] https://ckan.org/
[2] https://okfn.org/
[3] https://dados.gov.br/

Usage Scenario 1 (UC1): Maria wants to promote a debate about the sports context of Brazil, in her class of the 1st year of high school. With the platform, she can efficiently introduce students to the reality of the situation regarding initiatives to support sports in the country, using open government data. For this, using our methodology, she follows the following steps: (i) She connect to the platform; (ii) She writes "I want to know about sports in Brazil"; and (iii) She gets data related to the requested, shown in Table 2. The first column shows the title of the retrieved dataset, and the second column shows if the dataset is relevant or not for the question, labeled by a researcher in our university. We can observe that all of the datasets are relevant in this case. For matters of comparison, we asked the same question:"I want to know about sports in Brazil", in the search box on *dados.gov* , the results are shown in the Table 3. We can observe that none of the datasets returned by *dados.gov* are relevant to the question.

Retrieved Dataset	**Relevant/Not Relevant**
UFMS Sport Projects 2019	Relevant
Number of projects that raised funds under the Sports Incentive Law	Relevant
High Performance Sports Equipment	Relevant
Number of municipalities involved in the Sport and Leisure program of the City	Relevant
Sport and recreational craft by DN	Relevant
Diagnosis of Youth with Sport and Leisure Theme - Year 2008	Relevant
Sport Initiation Center	Relevant
Total funds raised by sports projects linked to the Sports Incentive Law	Relevant
Number of centers implemented by the Sport and Leisure program of the City	Relevant
High Performance Sports Equipment in the INDE viewer	Relevant

Table 2: 10 first datasets retrieved in our platform for UC1

Retrieved Dataset	**Relevant/Not Relevant**
[2013 to 2016] Stricto Sensu Graduate Teachers in Brazil	Not Relevant
Methodological adjustment on internal debt - Balances (R$)	Not Relevant

Methodological adjustment on external debt - Balances (R$)	Not Relevant
Demographic Census	Not Relevant
Union Properties	Not Relevant
School-age distortion rates in Basic Education	Not Relevant
Marcantonio Vilaça Plastic Arts Award	Not Relevant
School Census Microdata	Not Relevant
Primary income - monthly - expenditure	Not Relevant
Primary income - monthly - revenue	Not Relevant

Table 3: 10 first datasets retrieved in *dados.gov* for UC1

Usage Scenario 2 (UC2): Sebastiao wants to understand more about Science in Brazil. With the platform, he can get to know the real situation about the country's Science, using open government data. For this, he follows the following steps: (i) He connect to the platform; (ii) He writes "Science in Brazil"; (iii) He gets data related to the requested, shown in Table 4. We can observe that all the datasets are relevant in this case. For matters of comparison, we asked the same question: "Science in Brazil", in the search box on *dados.gov* , the results are shown in the Table 5. We can observe that only one of the datasets returned by *dados.gov* are relevant to the question.

Retrieved Dataset	**Relevant/Not Relevant**
Number of Selected in the Research Support Notice of the Women and Science Program	Relevant
Scholarships awarded in the Science Without Borders Program	Relevant
2.1.1 Brazil: National expenditure on science and technology (S&T) (1), in current values, by activity, 2000-2016	Relevant
Expenditure on science and technology	Relevant
2.1.2 Brazil: National expenditure on science and technology (S&T) (1), in current values, in relation to total S&T and gross domestic product (GDP), by institutional sector, 2000-2016	Relevant
Computer Science - 2011	Relevant
2.2.1 Brazil: Federal government expenditure on science and technology (S&T) (1) by activity, 2000-2016	Relevant
2.2.2 Brazil: Federal government expenditures on science and technology (S&T) (1) (2) by agency, 2000-2016	Relevant

2.2.3 Brazil: Federal government expenditure on science and technology (S&T) (1) (2), applied by the Ministry of Science, Technology and Innovation (MCTI), by budgetary unit and activity, 2000-2016	Relevant
Number of science and technology dissemination and popularization projects supported	Relevant

Table 4: 10 first datasets retrieved in our platform for UC2

Retrieved Dataset	**Relevant/Not Relevant**
Innovation Research	Relevant
National Science, Technology and Innovation Indicators	Relevant
Brazil-Argentina: A Strategic Relationship	Not Relevant
Strategic Alliances for Brazil: China and India	Not Relevant
List of Species of Flora do Brasil 2015 - Brazilian Flora Checklist	Not Relevant
VII National Meeting of Strategic Studies - Volume II	Not Relevant
VII National Meeting of Strategic Studies - Volume I - 2007	Not Relevant
Perspectives for the Boundary Strip	Not Relevant
Study Meeting - Indigenous Question	Not Relevant
Study Meeting - Climate Change	Not Relevant

Table 5: 10 first datasets retrieved in *dados.gov* for UC2

In addition to the natural language processing work, our approach performs a search engine where we thoroughly analyze all the titles and descriptions of the datasets, so that there is a more effective result for the user. For each usage scenarios, in our platform, we obtained the following P values (Table 6); R values (Table 7).

	Total retrieved items	**Total relevant items**	**Precision**
UC1	21	21	100%
UC2	77	74	96%

Table 6: Precision Results

	Total relevant items in the sample	Total retrieved items	Recall
UC1	1	1	100%
UC2	3	3	100%

Table 7: Recall results

From these results, we observed high values of precision and recall. The high precision obtained means that our experimental platform returned substantially more relevant than irrelevant results. While the high recall value obtained means that the experimental platform returned most of the relevant results present on the portal.

6 Conclusions

We present in this work a methodology for retrieving data using Information Retrieval and Question and Answering techniques and methods. Our methodology involves using Natural Language Processing techniques to extract semantic meaning of the user's question and then perform the IR task. We implemented our methodology in a tool, available at `https://github.com/RaissaBarcellos/openIR`. We carried out an experiment performing two task of recovering open government data taking into consideration the user's question. We used the Brazilian open government data portal *dados.gov*. We compared ur mechanism for retrieving data with the simple mechanism available by *dados.gov*. The results were expressive, considering that high values of precision and recall indicate a high power of data recovery of the platform. As future work, we intend to conduct a discussion of how semantic information retrieval mechanisms are essential in open data portals As a limitation of this work, we experimented with just one open government data portal, we can still perform the integration with other portals.

References

1. Almansa, L.F.: Uma arquitetura de question-answering instanciada no domínio de doenças crônicas. Ph.D. thesis, Universidade de São Paulo (2016)
2. Alzamil, Z.S., Vasarhelyi, M.A.: A new model for effective and efficient open government data. International Journal of Disclosure and Governance pp. 1–14 (2019)
3. Anastasiu, I., Foth, M., Schroeter, R., Rittenbruch, M.: From repositories to switchboards: Local governments as open data facilitators. In: Open Cities| Open Data, pp. 331–358. Springer (2020)
4. Athenikos, S.J., Han, H.: Biomedical question answering: A survey. Computer methods and programs in biomedicine **99**(1), 1–24 (2010)
5. Buckland, M., Gey, F.: The relationship between recall and precision. Journal of the American society for information science **45**(1), 12–19 (1994)

6. Dimitrakis, E., Sgontzos, K., Tzitzikas, Y.: A survey on question answering systems over linked data and documents. Journal of Intelligent Information Systems pp. 1–27 (2019)
7. Falci, D.H.M.: An automatic semantic role labeler for the portuguese language. Ph.D. thesis, Mestrado em Sistemas de Informação e Gestão do Conhecimento (2018)
8. Fonseca, E.R., Rosa, J.L.G.: A two-step convolutional neural network approach for semantic role labeling. In: The 2013 International Joint Conference on Neural Networks (IJCNN). pp. 1–7. IEEE (2013)
9. Gascó-Hernández, M., Martin, E.G., Reggi, L., Pyo, S., Luna-Reyes, L.F.: Promoting the use of open government data: Cases of training and engagement. Government Information Quarterly **35**(2), 233–242 (2018)
10. Hirschman, L., Gaizauskas, R.: Natural language question answering: the view from here. natural language engineering **7**(4), 275–300 (2001)
11. Jetzek, T., Avital, M., Bjorn-Andersen, N.: The sustainable value of open government data. Journal of the Association for Information Systems **20**(6), 6 (2019)
12. Kettouch, M., Luca, C., Khorief, O., Wu, R., Dascalu, S.: Semantic data management in smart cities. In: 2017 International Conference on Optimization of Electrical and Electronic Equipment (OPTIM) & 2017 Intl Aegean Conference on Electrical Machines and Power Electronics (ACEMP). pp. 1126–1131. IEEE (2017)
13. Kirstein, F., Dittwald, B., Dutkowski, S., Glikman, Y., Schimmler, S., Hauswirth, M.: Linked data in the european data portal: A comprehensive platform for applying dcat-ap. In: International Conference on Electronic Government. pp. 192–204. Springer (2019)
14. Koesten, L., Singh, J.: Searching data portals-more complex than we thought? In: SCST@ CHIIR. pp. 25–28 (2017)
15. Koesten, L.M., Kacprzak, E., Tennison, J.F., Simperl, E.: The trials and tribulations of working with structured data: -a study on information seeking behaviour. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. pp. 1277–1289 (2017)
16. Kuo, C.L., Chou, H.C.: Metadata assessment for efficient open data retrieval. In: Accepted Short Papers and Posters from the 22nd AGILE Conference on Geoinformation Science (AGILE 2019). Cyprus, Greece (2019)
17. Lahti, L., Huovari, J., Kainu, M., Biecek, P.: Retrieval and analysis of eurostat open data with the eurostat package. The R Journal **9**(1), 385–392 (2017)
18. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval. Cambridge university press (2008)
19. Oliveira, B.C., Salvadori, I., Huf, A., Siqueira, F.: A platform to enrich, expand and publish linked data of police reports. In: Proceedings of the 15th International Conference WWW/Internet. pp. 111–118 (2016)
20. Pasquali, L.: A curva normal. _____. Matemática Discreta. Rio de Janeiro (2006)
21. Pinto, H.d.S., Bernardini, F., Viterbo, J.: How cities categorize datasets in their open data portals: an exploratory analysis. In: Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age. pp. 1–9 (2018)
22. Whitmore, A.: Using open government data to predict war: A case study of data and systems challenges. Government Information Quarterly **31**(4), 622–630 (2014)