Ajose-Ismail, B. M. & Osanyin, Q. A., Book of Proceedings of 4th National Development Conference of The School of Pure and Applied Science, The Federal Polytechnic Ilaro, Ogun State, 2nd – 5th December, 2019 249-256

A SEMANTIC APPROACH FOR FACILITATING SEARCH AND DISCOVERY OF OPEN GOVERNMENT DATA ON OPEN DATA PORTALS

[1]BABATUNDE MICHEAL, AJOSE-ISMAIL and [2]AYODEJI QUADRI, OSANYIN

[1]Department of Computer Science, School of Applied Science,
The Federal Polytechnic Ilaro, Ogun State. Nigeria.
babatunde.ajose29@federalpolyilaro.edu.ng
Phone no: 08061614570

[2]Department of Computer Science, School of Applied Science,
The Federal Polytechnic Ilaro, Ogun State. Nigeria.
quadri.osanyin@federalpolyilaro.edu.ng
Phone no: 070624048582

Abstract

The Open government initiative has seen a vast amount of open data portals being developed around the globe to promote the accessibility of government datasets which can yield high economic value if reused by researchers and developers. Their data needs may be met by submitting queries to a dataset search engine of an open data portal to retrieve relevant datasets. However, existing open data portals provide mostly keyword-based search without the ability to understand the user's intent and the contextual meaning of the datasets. Data search systems on open data portals tend to rely on the text contained in metadata and dataset descriptions to facilitate keyword search. A cursory review of the literature indicates that poor discovery of datasets is a critical problem on open data portals**.** Semantic search has been well explored in semantic web as an attempt to improve the quality of search for relevant documents and web pages. In this work, we present an approach for semantic search of open government datasets, a relatively underexplored domain using machine learning and natural language processing techniques that help match a user data need against a collection of datasets

This paper aims to consider the adoption of a semantic approach to open government dataset search that will help novice users search for open government datasets and improve the process of open data discovery.

***Keywords*-Open government, Open government data, Open data portal, Dataset, semantic search, Machine Learning, Natural Language Processing**

1. INTRODUCTION

Open government is the governing doctrine that supports the right of citizens to access the documents and proceedings of the government for an effective public oversight. Much related to open government is *Open data* which generally refers to the idea that some data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control (Porreca, Leotta, Mecella. Vassos, & Catarci, 2017). Opening up official information can support technological innovation and economic growth by enabling third parties to develop new kinds of digital applications and services. Generating value from data requires the ability to find, access and make sense of datasets. The statistical data and metadata exchange initiative (SDMX) (SDMX, 2018) defines a dataset as 'a collection of related observations, organized according to a predefined structure'. This definition is shared by the DataCube working group at the World Wide Web Consortium (W3C), which adds the notion of a 'common dimensional structure' (W3C, 2014).

Meanwhile, the Organization for Economic Co-operation and Development (OECD), citing the US bureau and census, defined open data as 'any permanently stored collection of information usually containing either case level data, aggregation of case level data, or statistical manipulations of either the case level or aggregated survey data, for multiple survey instances (SDMX, 2018)

Datasets are increasingly being exposed for trade within data markets (Balazinska, Howe, Koutris, Suciu & Upadhyaya, 2013; Grubenmann, Bernstein, Moor, & Seuken, 2018), open data portals, scientific repositories (Altman et al., 2015; Elsevier scientific repository, 2018) and other websites as required by EU Directives and national regulations. Communities such as Wikidata or the Linked Open Data Cloud (Linked open data cloud, 2018) come together to create and maintain vast, general-purpose data resources, which can be used by developers in applications as diverse as intelligent assistants, recommender systems and search engine optimization. The common intent is to broaden the use and impact of the millions of datasets that are being made available and shared across organizations. Despite all these initiatives, the available published open data descriptions are often unprecise or incomplete and identifying the relevant datasets is time-consuming. On the other hand, data portals or websites often provide only keyword-based search depending on the presence of the user entered keywords in the dataset description. The challenge is that the user does not necessarily know what to search and how to phrase the search keywords. Synonyms may exist matching user's intention, but the user does not know the terms used by the publishers to describe their data. Dataset search poses unique challenges and little attention has been given to the queries in dataset search, which may differ considerably from general Web search queries. Semantic search copes with the above challenges and attempts to improve search accuracy by understanding the searchers intent (user query) and the contextual meaning of terms in the searchable dataspaces (datasets). In other words, semantic search engine searches for concepts instead of keywords.

In this paper, we present an approach for open data search exploiting semantic search based on natural language processing techniques. Natural Language Processing (NLP) techniques are applied to transform user's query to concepts (search terms) and datasets descriptions to concepts

(searchable dataset terms). The semantic search is then based on the semantic similarity between the search terms and the concepts related to the datasets.

2. RELATED WORK

2.1 Barriers Related to Usage of Open Data

Considerable effort has been made by several researchers to identify the barriers related to usage of open government data on open data portals.

Crusoe & Melin (2018) also examined the existing barriers, myths, challenges or impediments inherent in the different phases of the open government data process. The authors discovered that despite the creation of open data portals to ease the discovery of open datasets by users, findability and accessibility issues still persist

Osagie et al. (2017) investigated the reason why open data platforms are currently being underutilized and discover the factors preventing the exploitation of available open data. Findings revealed that Existing portals fail to meet the simplicity and understandability required by most intended users due to complexity of the platform thereby making the findability, accessibility and exploitation of such data difficult.

Attard et al. (2015) attempted to explore the state of open government data initiatives, as well as existing tools and approaches using a systematic survey of current as well as past literatures. The authors also discovered that portals support only simple search functions which do not return only relevant data but related policies and document such as research papers resulting in information overload for the user

Zuiderwijk & Janseen (2014) in their work identified identify the socio-technical barriers of the six phases of the open data process and define development directions for each of the barriers. Findability issues related to open data by users still persists due to the fact that open data portals lack advanced search component and clear navigation. It also poses a challenge to data providers

2.2 Open Government Dataset Discovery

Chen et al. (2019) explored a novel method to analyze both metadata and data content in line with user submitted queries. In their methodology they performed a semantic annotation of queries using a novel fine grained scheme (Query Analysis) and made use of a query-biased and illustrative snippet generation technique to enhance dataset search. Results obtained revealed that illustrative snippet generation outperformed the query-biased in terms of data content but has limitations in that it often failed to match the keywords appearing in the query.

Jiang et al. (2019) investigated the application of semantic search methodologies on open dataset search. Their methodology comprised of automatic linking of ontology concept to dataset using NLP based techniques and computation of semantic similarity between concepts, concept labels and datasets using the Wu and Palmer algorithm. Findings show that the automatic linking of

ontology concepts to dataset overcomes the time consuming constraint of the manual process and thus enhances the online search of open data on open data portals

Lafia et al. (2019) focused on exploring the current capabilities of voice assistants in the discovery of geospatial open government data. The authors employed the Cook and Daniels software design methodology consisting of an essential, formal and system model was adopted to develop the proposed conceptual framework for facilitating geospatial data discovery

Swamiraj & Freund (2015) investigated the usefulness of an improved data search interface to help novice users discover relevant open government dataset. The authors utilized an exploratory search paradigm to develop the search interface and a network based visualization technique to display search results

3. METHODOLOGY

This work investigates the feasibility of semantic search on open data portals. One of the challenges of traditional search or keyword –based search is that the search result hardly return the relevant datasets the user is interested in. This is due to the fact that in keyword-based search, the terms in the user query are simply matched with the terms in the dataset description (Metadata) and then results are returned based on similarity matching. In most cases, datasets relevant to the user query exists but poor quality metadata or descriptions about these datasets usually renders the search process ineffective.

We take a semantic approach based on natural language processing techniques and show the feasibility of discovering relevant datasets starting from when a query is issued on the search engine of a data portal to when the datasets are retrieved. Our approach consists of five modules which is discussed below.

3.1 Dataset

In this phase, datasets regarding a particular domain of interest can be gathered or crawled using a web crawler and stored in the repository of the open data portal. Next, preprocessing is performed on the datasets and indexing done with Latent Semantic Indexing (LSI). This yields a Dataset-Concept-Similarity-Vector (DCSV)

3.2 Query Preprocessing

Queries are often posed in form of questions and in several cases need refinement in order to generate the right answers. In most cases, the queries are often short, may contain spelling errors and might be incomplete. Based on the known facts about queries, we suggest a query reformulation strategy specifically query expansion to expand the query by adding terms that could facilitate the search and retrieve the relevant dataset. A probabilistic pre-trained word embedding model called n-gram is used to achieve this expansion. Spelling error correction can be achieved using edit distance, which calculates the distance of the spelt word by the user to the dictionary version of that word. Subsequently, terms in the query can be extracted using NLP techniques and query indexing performed with Latent Semantic Indexing (LSI) to map the query to a concept

yielding a Query-Concept-Similarity-Vector (QCSV) where each vector element contains a score reflecting the similarity of the query to a concept. Finally the query is stored for later use

3.3 Search

The search step can also utilize LSI to perform similarity computation between the query concept vectors (QCSV) with all dataset concept vectors (DCSVs) to produce an unranked list of datasets.

3.4 Ranking

Ranking is done using ranking algorithms like BM25, PageRank or any other state-of-the art algorithm that has shown good performance and has stood the test of time. This is performed to sort the retrieved datasets using their relevance scores i.e. making the document with the highest similarity to the entered query the topmost viewed document (Only the top-k result is viewed with k representing the dataset limit usually ranging from 10-50).

3.5 Visualization Interface

This displays the ranked list of datasets. Several methods can be used to view this on a browser e.g. Graph-Based and list method.

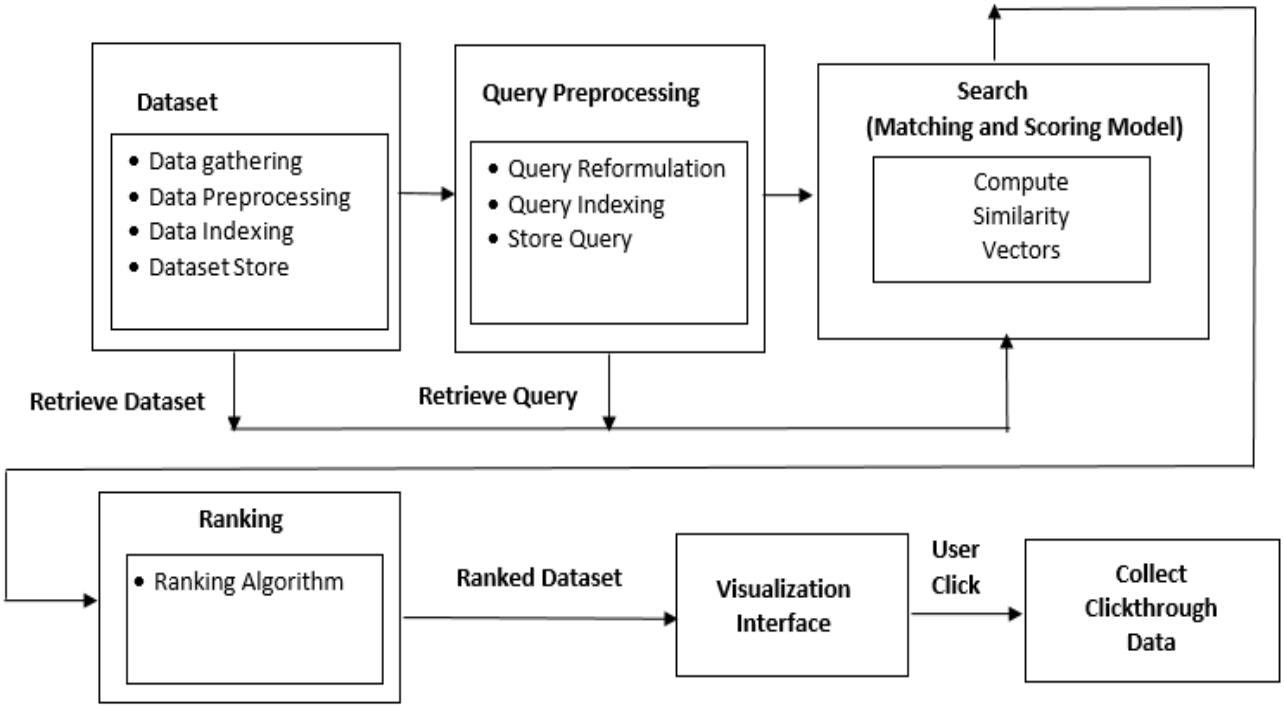

Figure 1: General Workflow of Semantic Dataset Search Approach

4. POSITIVE IMPLICATIONS OF PROPOSED SEMANTIC SEARCH TO NATIONAL DEVELOPMENT

Described below are some of the positive implications of semantic search to national development but not limited to:

4.1. Wealth creation through the development and sales of applications: The discovery of relevant dataset for a specific closed domain will facilitate the development and sale of web and mobile applications for different enterprise which would lead to an increase in the per capita income of the country.

4.2. Employment Opportunities: The development of such applications will require trained personnel and this would lead to the creation of more employment opportunities.

4.3. Cost Efficient and Resource Optimization: The cost incurred by developers in gathering data on the field would be greatly reduced because they would only need to surf several open data portals to get the relevant data they are interested in. Also, much of the tools needed to develop these applications are open source software tools meaning buying tools to build these needed applications might not be necessary and this leads to resource optimization.

5. CONCLUSION AND FUTURE WORK

In this paper, we present an approach for semantic search of open government datasets on open data portals, a relatively underexplored domain using machine learning and natural language processing techniques that help match a user data need against a collection of datasets and improve the process of open data discovery. Semantic approach to searching and discovering data is very crucial and could have huge economic impact on a nation's economy. Applications could be developed by programmers using this datasets which could render convenient services to the populace and if sold, generate revenue for the stakeholders involved.

However, semantic search to open government dataset search is not limited to our approach. As future work, after the dataset has been retrieved and shown on the visualization interface, click-through data can be obtained. A deep learning model can then be trained on learning query and document similarities from a click-through bipartite graph. The click-through bipartite represents the click relations between queries and documents, gotten from the query logs of a search engine. Query Suggestion which aims to recommend or suggest full semantically relevant queries that have been formulated by previous users can then be used in place of query expansion as proposed in our approach.

REFERENCES

Altman, M., Castro, E., Crosas, M., Durbin, P., Garnett, A., & Whitney, J. (2015). Open Journal Systems and Dataverse Integration–Helping Journals to Upgrade Data Publication for Reusable Research. Code4Lib Journal, (30).

Attard, J., Orlandi, F., Scerri, S., & Auer, S. (2015). A systematic review of open government data initiatives. Government Information Quarterly, 32(4), 399-418.

Balazinska, M., Howe, B., Koutris, P., Suciu, D., & Upadhyaya, P. (2013). A discussion on pricing relational data. In In Search of Elegance in the Theory and Practice of Computation (pp. 167-173). Springer, Berlin, Heidelberg.

Chen, J., Wang, X., Cheng, G., Kharlamov, E., & Qu, Y. (2019, November). Towards More Usable Dataset Search: From Query Characterization to Snippet Generation. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (pp. 2445-2448). ACM.

Crusoe, J., & Melin, U. (2018, September). Investigating open government data barriers. In International Conference on Electronic Government (pp. 169-183). Springer, Cham.

Elsevier scientific repository (2018). Retrieved fromhttps://datasearch.elsevier.com/

Grubenmann, T., Bernstein, A., Moor, D., & Seuken, S. (2018, April). Financing the web of data with delayed-answer auctions. In Proceedings of the 2018 World Wide Web Conference (pp. 1033-1042). International World Wide Web Conferences Steering Committee.

Jiang, S., Hagelien, T. F., Natvig, M., & Li, J. (2019). Ontology-Based Semantic Search for Open Government Data. In 2019 IEEE 13th International Conference on Semantic Computing (ICSC) (pp. 7-15). IEEE.

Lafia, S., Xiao, J., Hervey, T., & Kuhn, W. (2019). Talk of the Town: Discovering Open Public Data via Voice Assistants (Short Paper). In 14th International Conference on Spatial Information Theory (COSIT 2019). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Linked open data cloud (2018). https://www.lod-cloud.net/

Osagie, E., Waqar, M., Adebayo, S., Stasiewicz, A., Porwol, L., & Ojo, A. (2017, June). Usability evaluation of an open data platform. In Proceedings of the 18th Annual International Conference on Digital Government Research (pp. 495-504). ACM.

Porreca, S., Leotta, F., Mecella, M., Vassos, S., & Catarci, T. (2017, June). Accessing Government Open Data Through Chatbots. In International Conference on Web Engineering (pp. 156-165). Springer, Cham.

SDMX. (2018). Sdmx glossary. Technical Report, SDMX Statistical Working Group

Swamiraj, M., & Freund, L. (2015). Facilitating the discovery of open government datasets through an exploratory data search interface.

W3C. (2014). The rdf data cube vocabulary. Retrieved from https://www.w3.org/

Zuiderwijk, A., & Janssen, M. (2014). Barriers and development directions for the publication and usage of open data: A socio-technical view. In Open government (pp. 115-135). Springer, New York, NY.