Challenges and Progress in Dataset Search

Zhiyu Chen
Lehigh University
113 Research Drive, Bethlehem, PA
US
zhc415@lehigh.edu

Nowadays, data becomes an indispensable part of our life for different purposes, while current search engines do not support specialized functions for dataset search. A dataset usually consists of metadata and data content. Applying existing IR models to datasets will miss semantic information inside datasets and therefore cause low recall. Data content in tabular format usually contains schema labels which could be considered as metadata of data content. Matching schema labels with user queries is different from normal text matching task due to its functionality and heterogeneity. To address the problem, we propose the task of schema label generation which can facilitate dataset search by generating more common and understandable schema labels. Besides, a lot of datasets contain numerical columns which could have quantity names. When people search for data, quantity names commonly appear in the queries. Therefore, we propose the task of quantity name recognition which could help answer queries asking for numerical data.

dataset search;data linking; text normalization

1. INTRODUCTION

A lot of people nowadays rely on datasets for their work: data journalists need datasets to tell a good story, researchers use datasets for their research experiments. With increasing volume of datasets available on the Web, finding desired datasets becomes a non-trivial task for modern search engines.

A dataset usually is comprised of data content and metadata. In general, data content could be in diverse forms such as text, audio, images and video. Among all types of data content, text data in tabular format is one of the most important. In the rest of this paper, we only consider data content which is tabular data. Though some data portals (e.g., data.gov[1], datahub[2]), provide search functions to users, their functions are limited to matching text of user queries with text of metadata and data content, which means the metadata of datasets and the data content are treated in the same way.

However, different pieces of a dataset inherently take different roles and represent different information about a dataset. Metadata either summarizes the content of a dataset in a high level or illustrates the dataset by introducing the background, while data content contains the information that could satisfy the user need. Moreover, data content could have its internal structures and have its own metadata and content. Tabular data represents relational data in a compact way which has a header row, consisting of schema labels (attribute names), followed by data rows storing actual data values of corresponding attributes.

Equally dealing with metadata and data content is inappropriate since structure information will be missed. Besides, schema labels, served as metadata of data content, cannot be treated as normal text. Due to the existence of non-dictionary words (NDWs), text matching between schema labels with normal text will cause low recall of a traditional search engine. Also, it is possible that data publishers do not provide schema labels in the data content and require data users to infer schema labels from metadata which is very inconvenient.

To solve those challenges, we propose several research questions that could be combined with traditional IR problems to help dataset search in the following sections.

[1] https://www.data.gov/
[2] http://datahub.io/

2. SCHEMA LABEL GENERATION FOR DATASETS

For tabular data, each schema label has corresponding schema content. Due to their heterogeneity, schema labels in different datasets could have different identifiers even when they refer to the same concept. For example, a lot of datasets have a column describing the zip codes of addresses, but the schema label could be different such as "zipcode", "zip_code", "postal code" and "zip". It is common that NDWs are used in schema labels due to various naming conventions of different data publishers. Directly applying traditional retrieval methods which usually rely on text matching between query terms and document terms could cause low recall, because the out-of-vocabulary problem will be overly amplified. Through schema label generation, more common and thus understandable schema labels can be provided to datasets and therefore increase the recall of dataset retrieval system.

The closest work is Web table annotation which aims to annotate columns in a Web table with semantic concepts. Zhang (2014) uses features from context inside and outside of the table to help annotate columns containing entity mentions. Venetis et al. (2011) leverage a database to attach a class label to a column if a sufficient number of the values in the column are identified with the corresponding label in the database. Wang et al. (2012) use Probase to annotate a Table with related concepts. Similarly, Mulwad et al. (2013) also make use of knowledge bases to interpret Web Tables.

Different from Web tables, real-world datasets usually have not enough context such as surrounding paragraphs or semantic markups inserted in the Webpage. Moreover, there are few entities that can be linked to a knowledge base since the concepts contained in a dataset are usually too narrow (e.g., street names on a map) or too broad. Our method only uses generic features extracted from the datasets and therefore only annotates columns with labels from the datasets rather than concepts from other resources.

Chen et al. (2018) treat it as a multiclass classification problem where each schema label is considered as a single class. We assume that useful features could be extracted from data content that effectively characterize schema labels. By analyzing the content of dataset, we develop a variety of features to enable machine learning methods to recommend useful labels. Experiments on two real-world datasets show that the proposed method is able to outperform the baseline. We also find the generated "wrong" schema labels is highly semantic related to original labels such as shown in Table 1. For example, the "wrong" prediction is "Position" which is different from original label "Pos", while it represents the same meaning and could be used as an alternative label.

Table 1: *Examples of "wrong" predictions*

Original labels	**Predictions**
Year	Season
Opponent	Team
Pos	Position
Score in the final	Score

One limitation of the method proposed by Chen et al. (2018) is that features from one column are considered as independent from co-occurring columns. We believe that by considering the relationship of different columns, the models could be more robust and less ambiguous. The task could go further by finding the relationship of different schema labels. For example, one schema label may be a hypernym of another schema label. Identifying such relationships could help schema label generation when multiple labels are available.

3. RECOGNIZING QUANTITY NAMES OF TABULAR DATA

Numerical data accounts for a large proportion of all the datasets. When people search for data, quantity names (also called quantity kinds in QUDT[3] reference ontologies), which correspond to one or more units, commonly appear in the queries. Therefore inferring quantity names is an important task for IR systems to improve the ability to answer queries demanding numerical information.

Yi et al. (2018) propose a method to recognize and recommend quantity names for numeric columns. In the experiment, five popular quantities are selected: length, time, percent, currency and weight. Similar to the schema label generation method proposed in the previous section, features are extracted from column content. Hence the schema label also has great impact on quantity name detection, features extracted from schema labels are also used. Besides, the presence of quantity-specific terms within schema labels is used to create indicator features since schema labels belong to different quantity kind have different preferred terms. After applying those features, the random forest model with 200 trees and max depth 200 produces the accuracy of 89.5% on data.gov datasets.

An obvious weakness of the method proposed by Yi et al. (2018) is that the quantity-specific terms are manually curated which makes the method difficult

[3] http://www.qudt.org/

to be generalized to the detection of more quantity names. Making use of external resources such as QUDT reference ontologies could be useful. QUDT not only defines quantity kinds, but also units and data type ontology for science and engineering. In a lot of scientific articles, units often appear in the tables showing experimental results. So it is possible to extract the units and find the corresponding quantity types based on QUDT ontology, which means more features such as context words can be extracted for quantity kinds.

4. DATASET SEARCH

Both the schema label prediction and quantity name detection could facilitate dataset search. Kacprzak et al. (2018) shows that search queries from data portals have more mentions of geospatial and temporal information. It indicates that considering the matching of schema labels and quantity names with query terms is very important for dataset search, and is a better choice than treating text in queries, metadata and data content as normal text which misses semantic information in datasets. Gao and Callan (2017) propose a framework for scientific tables in research articles. Complementary to their method which expands user queries, we augment datasets by generating semantic information which could not only help dataset search, but also help other tasks such as data profiling and data linking.

In the future, we plan to integrate the schema label generation process and quantity name recognition to a real dataset search engine. To evaluate the performance, we first create a set of tasks. For each task, we will ask workers on Amazon Mechanical Turk (AMT) to provide queries that could find related datasets to complete the task. We do not require them to really feed the queries into a search engine and only ask them to provide their queries. After collecting queries, we will feed those queries to our several retrieval models and get top results from those systems. Then we will ask AMT workers to do relevance judgment on (query, dataset) pairs selected by different systems (pooling method). Then we can evaluate the performance using traditional metrics on this set.

5. CONCLUSION

In this paper, we discuss the task of dataset search. After analyzing its discrepancy with traditional information retrieval task, we propose the task of schema label generation and the task of quantity name recognition. In the end, we discuss how the tasks could help dataset search.

REFERENCES

Chen, Z., H. Jia, J. Heflin, and B. D. Davison (2018). Generating schema labels through dataset content analysis. In *Companion of the The Web Conference 2018 on The Web Conference 2018*, pp. 1515–1522. International World Wide Web Conferences Steering Committee.

Gao, K. Y. and J. Callan (2017). Scientific table search using keyword queries. *CoRR abs/1707.03423*.

Kacprzak, E., L. Koesten, J. Tennison, and E. Simperl (2018). Characterising dataset search queries. In *Companion of the The Web Conference 2018 on The Web Conference 2018*, pp. 1485–1488. International World Wide Web Conferences Steering Committee.

Mulwad, V., T. Finin, and A. Joshi (2013). Semantic message passing for generating linked data from tables. In *The Semantic Web – ISWC 2013*, pp. 363–378. Springer.

Venetis, P., A. Halevy, J. Madhavan, M. Paşca, W. Shen, F. Wu, G. Miao, and C. Wu (2011, June). Recovering semantics of tables on the web. *Proc. VLDB Endow. 4*(9), 528–538.

Wang, J., H. Wang, Z. Wang, and K. Q. Zhu (2012). Understanding tables on the web. In *Conceptual Modeling*, pp. 141–155. Springer.

Yi, Y., Z. Chen, J. Heflin, and B. D. Davison (2018). Recognizing quantity names for tabular data. In *ProfS/KG4IR/Data:Search@SIGIR*.

Zhang, Z. (2014). Towards efficient and effective semantic table interpretation. In *International Semantic Web Conference*, pp. 487–502. Springer.