DataHunter: A System for Finding Datasets Based on Scientific Problem Descriptions

MICHAEL FÄRBER, Karlsruhe Institute of Technology (KIT), Germany
ANN-KATHRIN LEISINGER, Karlsruhe Institute of Technology (KIT), Germany

The number of datasets is steadily rising, making it increasingly difficult for researchers and practitioners in the various scientific disciplines to be aware of all datasets, particularly of the most relevant datasets for a given research problem. To this end, dataset search engines have been proposed. However, they are based on the users' keywords and thus have difficulties in determining precisely fitting datasets for complex research problems. In this paper, we propose the system at http://data-hunter.io that recommends suitable datasets to users based on given research problem descriptions. It is based on fastText for the text representation and text classification, the Data Set Knowledge Graph (DSKG) with metadata about almost 1,700 unique datasets, as well as 88,000 paper abstracts as research problem descriptions for training the model. Overall, our system demonstrates that recommending datasets facilitates data provisioning and reuse according to the FAIR principles and that dataset recommendation is a promising future research direction.

Additional Key Words and Phrases: datasets, recommendation, machine learning, text classification, FAIR principles

ACM Reference Format:
Michael Färber and Ann-Kathrin Leisinger. 2021. *DataHunter*: A System for Finding Datasets Based on Scientific Problem Descriptions. In *Fifteenth ACM Conference on Recommender Systems (RecSys '21), September 27-October 1, 2021, Amsterdam, Netherlands.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3460231.3478882

1 INTRODUCTION

The number of available datasets in the various scientific fields has grown vastly and is continuously on the rise [10]. For instance, OpenAIRE [12] contains the metadata of more than 23,000 datasets. Several hundred datasets are listed for machine learning tasks on Wikipedia.[1] In recent years, large national and international initiatives, such as the German national research data infrastructure (NFDI) and the initiatives around the FAIR data principles [17], have been established to foster the reuse of datasets [10]. In the process of accessing and reusing datasets from repositories, one of the most challenging tasks is to identify the most relevant datasets [13]. Dataset search engines, such as Google Dataset Search [3] and Zenodo search,[2] are prominent assistance tools in this regard. They help users retrieve the most relevant datasets for their research problem. However, existing dataset search engines using the datasets' metadata are limited in their applicability [5]. Apart from the fact that search engines relying on the metadata depend on the accuracy and maintenance of the metadata [4], existing dataset search engines are not suitable for the user's very specific and comprehensive information needs [5] (see the example queries in Table 1). Chen et al. [5] found that real data needs are most often formulated as phrases and not as keywords. The latter case constitutes only 32 % of the investigated queries.

[1] https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research
[2] https://zenodo.org/

Table 1. Example dataset recommendations.

Input text (e.g., paper abstract without dataset mentions)	**Rec. datasets**
This paper presents an algorithms for tagging words whose part-of-speech properties are unkown. Unlike previous work, the algorithm categorizes word tokens in context instead of word types.	Brown Corpus
Given a set of images with related captions, our goal is to show how visual features can improve the accuracy of unsupervised word sense disambiguation when the textual context is very small, as this sort of data is common in news and social media. We extend previous work in unsupervised text-only disambiguation with methods that integrate text and images. [...]	ImageNet, SemCor

Overall, and to the best of our knowledge, systems beyond keyword search for retrieving relevant datasets for a given research problem are still missing.

In this paper, we propose a publicly available system for dataset search that is not based on keywords or faceted search; instead, the recommendation algorithm relies on a text classification model that predicts relevant datasets for a user's input. The user input is a text that describes the research that the user plans to conduct (see Table 1). A classifier predicts all relevant datasets indexed in a given repository based on the entered text. The hypothesis is that the quality of the dataset search can be considerably improved when using a rich formulation of the research problem in natural language, rather than relying purely on isolated keywords or attributes. Our system is available online at **http://data-hunter.io**. The source code of the frontend and backend and a video showing the main system features is available online at https://github.com/michaelfaerber/datahunter.

2 APPROACH

We model the dataset recommendation as a supervised multi-class, multi-label text classification, as for each input text, one or more datasets might be relevant. We differentiate between the following steps.

Text Preprocessing. We consider several text representation methods: (1) *tfidf*; (2) *doc2vec* [16], a way of representing an entire document; (3) *fastText embeddings* [2] as high quality pretrained word and phrase representations achieving or even outperforming state-of-the-art results on various tasks [15]; (4) *SciBERT embeddings* [1] as a widely used language model based on BERT and trained on scientific texts; and (5) *Transformer-XL embeddings* [6], based on the self-attention model Transformer-XL that models longer-term dependency. In experiments [8], we identified fastText as text representation method that achieves a high effectiveness and efficiency in user studies and that results in more recommended datasets than other representation methods. Thus, we use it in our online demonstration system.

Text Classification. In preliminary research [8], we implemented several text classification methods combined with text representation methods: (1) *classification based on tfidf similarity*; (2) *classification based on the BM25 score*; (3) *random forest*; (4) *logistic regression*; (5) *Gaussian, multinomial and complement naïve Bayes*; (6) *convolutional neural network (CNN)*; (7) *recurrent neural networks (Simple RNN, LSTM and BiLSTM)*; (8) *CNN-LSTM* [18]; (9) *fastText classification* [9]; and (10) *finetuned SciBERT with subsequent classification layer* [1, 11]. Based on our evaluation [8] and our findings concerning text representation, we choose fastText classification for our running system.

Ranking. The recommended datasets are ranked by their confidence scores.

Table 2. Dataset attributes with filling degree in the Data Set Knowledge Graph (DSKG) [14].

Dataset attribute	Filling degree
Title	100.0%
Source	100.0%
Description	99.9%
Link to dataset	78.7%
Topic	76.5%
Date of last modification	71.2%
Size (in byte)	28.1%
Date of issue	20.7%
Identifier	18.2%
Language	17.8%
Data format	13.2%
Format	13.2%
Access rights	12.9%

3 SYSTEM

3.1 Data

Dataset collection. As a database for datasets, we use the Data Set Knowledge Graph (DSKG) [14]. This up-to-date collection, containing the dataset attributes listed in Table 2, is based on dataset entries in Wikidata and OpenAIRE. It is characterized by rich and highly accurate metadata. Specifically, it contains dataset attribute information, such as the title, description, topic assignment, date of issue, date of last modification, language, access rights, data format, and size. Detailed information about the DSKG is given in [14]. One of the peculiarities of the DSKG is that all modeled datasets are linked to scientific publications in which they are mentioned. Specifically, the datasets are linked to the Microsoft Academic Knowledge Graph [7], containing the metadata of 240 million publications. We can use this fact to display detailed information about publications that mention or even use the recommended datasets. This might help users to choose appropriate datasets. We focus on computer science, resulting in a set of 1,691 unique datasets.

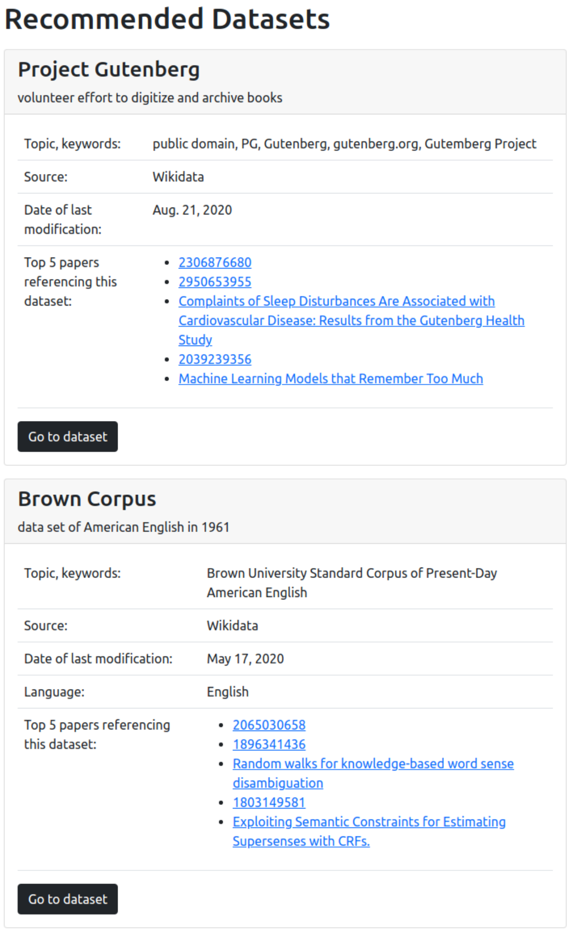
Scientific problem descriptions. To train and evaluate our recommendation model, the datasets need to be linked to textual problem descriptions. Since pure research problem descriptions are not available to a large extent, we use paper abstracts from given scientific papers that contain the given datasets as in-text mentions.[3] Paper abstracts are very similar to problem descriptions as they both summarize in a few sentences the examined task for which a dataset has been used or will be used. Specifically, we use the 88,047 paper abstracts from the Microsoft Academic Graph that reference the datasets in our collection. Most of the paper abstracts (85 %) reference only one dataset in our collection. Each abstract references at most 20 different datasets.

3.2 Implementation and User Interface

The backend is implemented in Python and uses SQLite. For the frontend, we use Django.

Figure 1 shows the user interface of our system with dataset summaries of the recommended datasets as output. A dataset summary consists of the dataset's title as well as the dataset's attributes (e.g., source, date of last modification, language, download link), as far as available. Given the links to papers modeled in the Microsoft Academic Knowledge Graph, in which the datasets are mentioned, we provide the links to the five papers with the highest citation count

[3] The dataset mentions were removed for the training for better generalization.



Recommended Datasets

Project Gutenberg

volunteer effort to digitize and archive books

Topic, keywords:	public domain, PG, Gutenberg, gutenberg.org, Gutemberg Project
Source:	Wikidata
Date of last modification:	Aug. 21, 2020
Top 5 papers referencing this dataset:	• 2306876680 • 2950653955 • Complaints of Sleep Disturbances Are Associated with Cardiovascular Disease: Results from the Gutenberg Health Study • 2039239356 • Machine Learning Models that Remember Too Much

Brown Corpus

data set of American English in 1961

Topic, keywords:	Brown University Standard Corpus of Present-Day American English
Source:	Wikidata
Date of last modification:	May 17, 2020
Language:	English
Top 5 papers referencing this dataset:	• 2065030658 • 1896341436 • Random walks for knowledge-based word sense disambiguation • 1803149581 • Exploiting Semantic Constraints for Estimating Supersenses with CRFs.

Fig. 1. Screenshot of *DataHunter*.

according to the Microsoft Academic Knowledge Graph. In case the Microsoft Academic Knowledge Graph does not contain the paper's title, we show the paper's Microsoft Academic Knowledge Graph ID on the result page. The full paper title is then given on the linked Microsoft Academic Graph web page.

4 CONCLUSION

In this paper, we developed a dataset recommender system that uses text classification to predict relevant datasets for given scientific problem descriptions. We considered several state-of-the-art classification models combined with text representation methods and use fastText for the final text representation and text classification. Overall, this paper shows how researchers and practitioners can be assisted when searching for data sets, facilitating data provisioning and reuse according to the FAIR principles.

REFERENCES

[1] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (Hong Kong, China) *(EMNLP-IJCNLP’19)*. 3613–3618.

[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomás Mikolov. 2017. Enriching Word Vectors with Subword Information. *Trans. Assoc. Comput. Linguistics* 5 (2017), 135–146.

[3] Dan Brickley, Matthew Burgess, and Natasha F. Noy. 2019. Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. In *Proceedings of the 28th International World Wide Web Conference* (San Francisco, CA, USA) *(WWW’19)*. 1365–1375.

[4] Adriane Chapman, Elena Simperl, Laura Koesten, George Konstantinidis, Luis-Daniel Ibáñez, Emilia Kacprzak, and Paul Groth. 2020. Dataset search: a survey. *VLDB J.* 29, 1 (2020), 251–272.

[5] Jinchi Chen, Xiaxia Wang, Gong Cheng, Evgeny Kharlamov, and Yuzhong Qu. 2019. Towards More Usable Dataset Search: From Query Characterization to Snippet Generation. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (Beijing, China) *(CIKM’19)*. 2445–2448.

[6] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc Viet Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In *Proceedings of the 57th Conference of the Association for Computational Linguistics* (Florence, Italy) *(ACL’19)*. 2978–2988.

[7] Michael Färber. 2019. The Microsoft Academic Knowledge Graph: A Linked Data Source with 8 Billion Triples of Scholarly Data. In *Proceedings of the 18th International Semantic Web Conference* (Auckland, New Zealand) *(ISWC’19)*. 113–129.

[8] Michael Färber and Ann-Kathrin Leisinger. 2021. Recommending Datasets Based on Scientific Problem Descriptions. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management* (Virtual Event) *(CIKM’21)*.

[9] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomás Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* (Valencia, Spain) *(EACL’17)*. 427–431.

[10] Laura Koesten, Elena Simperl, Tom Blount, Emilia Kacprzak, and Jeni Tennison. 2020. Everything you always wanted to know about a dataset: Studies in data summarisation. *Int. J. Hum. Comput. Stud.* 135 (2020), 1–21.

[11] Xiaofei Ma, Zhiguo Wang, Patrick Ng, Ramesh Nallapati, and Bing Xiang. 2019. Universal Text Representation from BERT: An Empirical Study. *CoRR* abs/1910.07973 (2019).

[12] Paolo Manghi, Claudio Atzori, others, and Friedrich Summann. 2019. OpenAIRE Research Graph Dump. https://doi.org/10.5281/zenodo.3516918

[13] Yasmmin Cortes Martins, Fábio Faria da Mota, and Maria Cláudia Cavalcanti. 2016. DSCrank: A Method for Selection and Ranking of Datasets. In *Proceedings of the 10th International Conference on Metadata and Semantics Research* (Göttingen, Germany) *(MTSR’16)*. 333–344.

[14] Michael Färber and David Lamprecht. 2021. Creating a Knowledge Graph for Data Sets. *Quantitative Science Studies* (2021). Minor revision, available at http://dskg.org/publications/DSKG_QSS2021.pdf.

[15] Tomás Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in Pre-Training Distributed Word Representations. In *Proceedings of the 11th International Conference on Language Resources and Evaluation* (Miyazaki, Japan) *(LREC’18)*.

[16] Maria Mihaela Truşcă. 2019. Efficiency of SVM classifier with Word2Vec and Doc2Vec models. In *Proceedings of the International Conference on Applied Statistics*, Vol. 1. Sciendo, 496–503.

[17] Mark D. Wilkinson, Michel Dumontier, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* 3, 1 (2016), 1–9.

[18] Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis C. M. Lau. 2015. A C-LSTM Neural Network for Text Classification. *CoRR* abs/1511.08630 (2015).