VU Research Portal

Ontology-Based Semantic Similarity Approach for Biomedical Dataset Retrieval

Wang, Xu; Huang, Zhisheng; van Harmelen, Frank

published in

Health Information Science
2020

DOI (link to publisher)

[10.1007/978-3-030-61951-0_5](https://doi.org/10.1007/978-3-030-61951-0_5)

document version

Publisher's PDF, also known as Version of record

document license

Article 25fa Dutch Copyright Act

[Link to publication in VU Research Portal](#)

citation for published version (APA)

Wang, X., Huang, Z., & van Harmelen, F. (2020). Ontology-Based Semantic Similarity Approach for Biomedical Dataset Retrieval. In Z. Huang, S. Siuly, H. Wang, Y. Zhang, & R. Zhou (Eds.), *Health Information Science: 9th International Conference, HIS 2020, Amsterdam, The Netherlands, October 20–23, 2020, Proceedings* (pp. 49–60). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 12435 LNCS). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/978-3-030-61951-0_5

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

E-mail address:

vuresearchportal.ub@vu.nl

Ontology-Based Semantic Similarity Approach for Biomedical Dataset Retrieval

Xu Wang^(✉), Zhisheng Huang, and Frank van Harmelen

Department of Computer Science, Vrije Universiteit Amsterdam,
Amsterdam, The Netherlands

{xu.wang,z.huang, Frank.van.Harmelen}@vu.nl

Abstract. Ontology-based semantic similarity approaches play an important role in text-similarity task, thanks to its ability of explanation. Ontology-based semantic similarity approaches can explain how two terms are similar with help of rich knowledge in ontology. Information retrieval aims to find relevant information for given user query. As a subareas of information retrieval, dataset retrieval is an activity to find dataset which are relevant to an information need, by using full-text indexing approach or content-based indexing approach. Ontology-based semantic similarity approaches can not only do some information retrieval tasks, such as full-text mapping, but also finding deeper similar information with the help of knowledge-richness in ontology. Because of the advantage of ontology-based similarity approaches, we are looking forwards to find the possibility to using ontology-based similarity for datasets retrieval. In this paper, we provide an ontology-based similarity approach for dataset retrieval. We run our novel approach on the bio-CADDIE 2016 Dataset Retrieval Challenge. After ruining experiments, we evaluate our results with several information retrieval evaluation measures. The evaluation results show that our approach could perform well.

Keywords: Dataset retrieval · Semantic similarity · Biomedical dataset

1 Introduction

Ontology-based semantic similarity approaches are popular semantic similarity measures, which are often used to calculate similarity between two ontology concepts or terms. Semantic similarity is always classified into two types for calculating topological similarity between ontological concepts: edge-based measure and node-based measure¹.

Dataset retrieval [6] is a specialization of information retrieval, which returns a list of relevant datasets instead of documents. Dataset retrieval always considers many aspects in metadata of datasets during retrieval process, such as topic

¹ https://en.wikipedia.org/wiki/Semantic_similarity.

(vocabularies, classes, properties and individuals), time, location (geo information) and so on. However, test-content (such as title and abstract/description) of one dataset also plays an important role. For instance, if the title of one dataset shares same meaning with given query, we can also say that this dataset should be retrieved for this query.

In this paper, we present a dataset retrieval approach, which is based on ontology-based similarity measure. We choose two different type of ontology-based similarity measures to work with this dataset retrieval approach, and apply them on bioCADDIE 2016 Dataset Retrieval Challenge. With the help of gold standard in bioCADDIE 2016 Dataset Retrieval Challenge, we use both average precision (AP) [14] and normalized Discounted Cumulative Gain (nDCG) [4] measure to evaluate our approaches. Finally, we find that ontology-based semantic similarity approaches could perform well on biomedical dataset retrieval.

The main contributions of this paper are: 1) we provide two novel ontology-based semantic similarity approaches for biomedical dataset retrieval; 2) we set up an experiment to evaluate the dataset retrieval approach with the help of bioCADDIE 2016 Dataset Retrieval Challenge; 3) we find that ontology-based semantic similarity approaches could perform well on biomedical dataset retrieval task.

2 Preliminaries

2.1 Ontology-Based Semantic Similarity

As we introduced above, semantic similarity are categorised into two type for calculating similarity between ontological concepts: edge-based measure and node-based measure. Edge-based measure often consider edge counting in ontology hierarchical structure. There are many existing edge-based measures, such as Wu-Palmer [13], Pekar [10], Cheng and Cline [1] and so on. Node-based measure will consider nodes and their properties in ontology, where nodes always means the concepts in ontology. Many existing and popular node-based measures could be used, such as Resnik [12], Lin [7], Align Disambiguate Walk (ADW) [11], Jiang and Conrath [5] and so on. In this section we will introduce Wu-Palmer measure and Resnik measure, which we used in our experiments.

Wu-Palmer. Wu-Palmer measure is a very popular edge-based semantic similarity measure. It has some advantage: 1) normalized measure; 2) similarity score never goes zero; 3) no distinction between similarity/relatedness. Only thing to notice is that Wu-Palmer heavily rely on quality of graph. So, with help of a good quality ontology, Wu-Palmer could work very well for similarity task. Least common subsumer is a very important terminology in ontology-based semantic similarity measure, which is used not only in Wu-Palmer measure but also in Resnik measure. In the hierarchical structure of ontology, least common subsumer of two nodes is the lowest node which has these two nodes as descendants. Given two concepts $C1$ and $C2$ in ontology, Wu-Palmer similarity between them is:

$$Sim_{WP} = \frac{2 * Dep(LCS(C1, C2))}{Dep(C1) + Dep(C2)} \quad (1)$$

where $LCS(C1, C2)$ is the least common subsumer of $C1$ and $C2$; $Dep(C1)$ is the least path from ROOT node to $C1$.

Resnik. Resnik measure is a semantic similarity measure based on the information content (IC) of the least common subsumer. The information content of a concept is the logarithm of the probability of finding the concept in a given corpus. We chose to use Resnik measure because many other node-based similarity measures, such as Lin and Jiang\$Conrath, are based on Resnik. Given two concepts $C1$ and $C2$ in ontology, Resnik similarity between them is:

$$Sim_{Resnik} = -\log P(LCS(C1, C2)) = -\log \frac{\sum_{c \in Words(LCS(C1, C2))} count(c)}{N} \quad (2)$$

where $Words(LCS(C1, C2))$ is the set of instances (i.e. terms) of $LCS(C1, C2)$; $count(c)$ is the number of counting all of c in $Words(LCS(C1, C2))$; N is the whole number of instances in ontology.

Similarity Between Sets of Concepts. After introducing two measure for calculating similarity between two concepts, we will introduce similarity measure for two sets of concepts. Given two sets $S1$ and $S2$ of concepts, the similarity between $S1$ and $S2$ is:

$$Sim(S1, S2) = \frac{sum\{Sim(C1, C2) | C1 \in S1, C2 \in S2\}}{|S1| * |S2|} \quad (3)$$

where $Sim(C1, C2)$ is the similarity between concepts $C1$ and $C2$ by Wu-Palmer or Resnik; $|S1|$ and $|S2|$ is the size of sets $S1$ and $S2$, respectively.

2.2 Evaluation Measure for Retrieval Approach

We use evaluation measures of information retrieval to evaluate our dataset retrieval approaches. In this paper, we use two popular evaluation measures, which are average precision (AP) and normalized discounted cumulative gain (nDCG), to evaluate our experiment results.

Average Precision. Average precision (AP) is a measure for ranked retrieval results [14]. For one information need, the average precision is the mean of the precision scores after each relevant document is retrieved. Average precision at position n is defiend as follow:

$$AP_n = \frac{\sum_{k=1}^n P@k}{R} \quad (4)$$

where $P@k$ is the precision of the top-k retrieved documents; R is the number of all retrieved datasets.

Normalized Discounted Cumulative Gain. Discounted cumulative gain (DCG) is a measure of ranking quality, which uses a graded relevance list of documents/datasets from the result to evaluate the gain of a document/dataset based on its position in the result ranking list [3]. Ideal discounted cumulative gain (IDCG) is the DCG score of ideal ranking list (list ranked by graded relevance). Normalized discounted cumulative gain (nDCG) is the normalized measure based on DCG [4]. DCG, IDCG and nDCG through top rank position p are defined as follow:

$$DCG_p = \sum_{i=1}^p \frac{2^{rel_i} - 1}{\log_2(i + 1)}, \quad (5)$$

$$IDCG_p = \sum_{i=1}^{|Rel_p|} \frac{2^{rel_i} - 1}{\log_2(i + 1)}, \quad (6)$$

$$nDCG_p = \frac{DCG_p}{IDCG_p} \quad (7)$$

where rel_i is the graded relevance score in position i .

3 Case Study: bioCADDIE Dataset

We use bioCADDIE Dataset 2016 as a case study of the proposed approach. The main objective of the 2016 bioCADDIE Dataset Retrieval Challenge is to create innovative ways for biomedical researchers to search and discover biomedical research data². For the challenge, a collection of metadata (structured and unstructured) from biomedical datasets were generated from a set of 20 individual repositories. It consist of about 800,000 datasets of metadata. The following is an example of a metadata description with the DATS model in JSON. In this example, we can see that the property “DOCNO” is used to define the local ID of a dataset. The “DOCNO” follows by the property “metadata” which states various values of the data, which includes the data origin such as “clinical_study”, data item such as data type, study group, grant, detailed study description, dataset identifier, detailed dataset description (title, description, and others).

```
{ "DOCNO": "10", "METADATA":
  { "dataResource":
    { "keywords": [], "altNames": [], "acronyms": [] },
    "citation": { "count": "0" },
    "organism": { "experiment": { "species": "Homo sapiens" } },
    "dataItem": { "description": "gene expression at 6h of
differentiation of Human endometrial stromal cell
expressing either or both of PRA and PRB Endogenous PGR
expression is silenced with siRNA mediated knockdown.
Then, cells are transduced with adenovirus expressing flag
```

² <https://biocaddie.org/biocaddie-2016-dataset-retrieval-challenge-registration>.

```

tagged PRA or flag tagged PRB. Human endometrial
stromal cell expressing one or both isoforms are treated
with differentiation cocktail for 6h.",
"title": "Expression data of PGR isoforms, PRA and PRB,
regulated genes in differentiating human endometrial
stromal cells",
"releaseDate": "2015-04-27", "lastUpdateDate": "2015-05-02",
"dataTypes": ["organism", "dataItem", "citation"],
"ID": "522705",
"experimentType": "transcription profiling by array"}},
"REPOSITORY": "arrayexpress_020916",
"TITLE": "Expression data of PGR isoforms, PRA and PRB,
regulated genes in differentiating human endometrial
stromal cells"}
, ...
    
```

The main advantage of using bioCADDIE Dataset 2016 is that the organizers provide a set of test queries and their expected answers, which can serve as a gold standard for the evaluation of our proposed approach.

Fig. 1. Pipeline of similarity-based biomedical dataset retrieval.

4 Similarity-Based Biomedical Dataset Retrieval

In this section we will introduce the pipeline of using ontology-based semantic similarity for biomedical dataset retrieval, which is shown in Fig. 1.

Because of the gold standard of 2016 bioCADDIE Dataset Retrieval Challenge, our biomedical dataset retrieval approach aims to find a ranking list of datasets for each query. So we will introduce a whole end2end process (from given queries and datasets to final ranking list) in this section.

4.1 Biomedical Concepts Extraction

We use ontology entity exact-matching approach to extract biomedical entities from datasets, and return concepts which include extracted entities in ontology. Given a dataset and an ontology, Exact-matching extracts words/phrases which are not only words/phrases in dataset but also entities in ontology.

Definition 1. Given dataset D and ontology O , exact-matching would extract a set of matched entities $EM_O(D) = \{e | e \in W(D) \cap E(O)\}$, where $W(D)$ is the set of all words/phrases of D and $E(O)$ is the set of all entities in O .

For exact-matching approach, we also have a proposition. Exact-matching would not cover any words/phrases which are not ontology entity, but can not cover all the ontology entities in datasets. This is because exact-matching approach is soundness but not completeness.

Proposition 1. Given dataset D , ontology D and a set of matched entities $EM_O(D)$. Then we have:

- $W(D) \models e^+, \forall e^+ \in EM_O(D)$
- $E(O) \not\models e^-, \forall e^- \notin EM_O(D)$.

Example 1. We show an example of bioCADDIE dataset and the results of exact-matching for this dataset. The example dataset is formatted as JSON as follow:

```
{
  "id": "793860",
  "abstract": "The goal of the study is to examine how cognitive behavioral therapy (CBT), a common treatment for adolescent depression, affects brain functioning. Depressed adolescents will complete an initial assessment consisting of an interview, questionnaires, computer tasks, and an EEG recording. EEG (electroencephalography) measures brain activity by recording the electrical activity along the scalp caused by the firing of neurons within the brain. They will then complete 12 sessions of cognitive behavioral therapy, which will be 50 minutes long and held once a week for 12 weeks. Before their third therapy session, participants will complete a computer task while EEG data are recorded. After completing the treatment, the participants will complete a final assessment, which will include questionnaires, computer tasks, and an EEG recording. They will also complete follow-up assessments over the phone 1, 3, and 6 months after completing the treatment. This study will also include healthy control participants. They will complete an initial assessment consisting of an interview, questionnaires, computer tasks, and an EEG recording. Three weeks later, they will return to complete a behavioral task while EEG data are recorded. Twelve weeks after the initial assessment, they will complete a final assessment, which will include questionnaires, computer tasks, and an EEG recording.",
  "title": "Examination of the Neural Components Underlying the Treatment of Adolescent Major Depression"
}
```

Then we use exact-matching approach to extract entities from title and abstract, and return concepts which include extracted entities in ontology. In MeSH ontology, we consider MeSH Descriptor as our concept in general. MeSH Descriptor consists of a set of MeSH Concepts and a set of MeSH Terms³. In one MeSH Descriptor, all the MeSH Concepts are synonymous with each other. All the MeSH Terms in same MeSH Descriptor are the entities of the MeSH Concepts in this MeSH Descriptor. The extracted concepts (matched entities and MeSH Descriptor IDs in brackets) are shown as follow:

```
{
  "Extracted_Concepts_from_Title": ["Depression (Depression, D003863)", "Adolescent (Adolescent, D000293)"],
  "Extracted_Concepts_from_Abstract": ["Goals (Goal, D006040)", "Cognitive Behavioral Therapy (Cognitive Behavioral Therapy, D015928)", "Depression (Depression, D003863)", "Brain (Brain, D001921)", "Adolescent (Adolescent, D000293)", "Surveys and Questionnaires (Questionnaires, D011795)", "Computers (Computers, D003201)", "Weights and Measures (Measures, D014894)", "Scalp (Scalp, D012535)", "Neurons (Neurons, D009474)", "Electroencephalography (EEG, D004569)"]
}
```

³ https://www.nlm.nih.gov/mesh/concept_structure.html.

4.2 Biomedical Concepts Similarity with MeSH

In MeSH ontology, each MeSH Descriptors have at least one tree number in MeSH Tree Structure⁴. With the given tree number, we can easily use both Wu-Pamler and Resnik similarity to calculate similarity between MeSH Descriptors. This is because we can find the least common subsumer of two MeSH Descriptors by using tree number. For instance, in MeSH ontology, Goals (D006040) has tree number “F01.658.500” and Depression (D003863) has tree number “F01.145.126.350”. Then we can easily see that Goals and Depression has least common subsumer with tree number “F01”, which is Behavior and Behavior Mechanisms (D001520). Also with the help of tree number in MeSH ontology, we can easily know the shortest path from ROOT to MeSH Descriptor. We continue with above example of Goals and Depression. Tree number “F01.658.500” means the shortest path from ROOT of MeSH to Goals is 3. Tree number “F01.145.126.350” means the shortest path from ROOT of MeSH to Depression is 4. Tree number “F01” states that Behavior and Behavior Mechanisms, the least common subsumer of Goals and Depression, only has shortest path 1 to ROOT of MeSH. Then we have Wu-Palmer similarity between Goals and Depression, which is $\frac{2*1}{3+4} = 0.2857$. For Resnik similarity, we try to find all the MeSH Descriptors, which have “F01” in their tree number. Then we count the number of all the MeSH terms in these MeSH Descriptors, which is 1735. Also, we count all the MeSH terms in MeSH ontology, which is 284832. Now we have Resnik similarity between Goals and Depression, which is $-\log \frac{1735}{284832} = 2, 2152$.

4.3 Similarity-Based Ranking List

In 2016 bioCADDIE Dataset Retrieval Challenge, the gold standard is a set of test queries and their expected answers (a list of datasets ranked by relevance score for each query). So our final results are lists of datasets, one list of datasets per query. After we calculate all the Wu-Palmer and Resnik similarity, we could rank all the dataset candidates based on similarity results for each query.

5 Experiments

In this section we will introduce our experiments of using ontology-based similarity approach for bioCADDIE Dataset Retrieval Challenge.

5.1 Biomedical Datasets Retrieval Experiment

As we mentioned in previous section, our biomedical datasets retrieval approach is based on Wu-Palmer and Resnik similarity measures. The output of experiments are ranked list of datasets based on similarity results.

⁴ https://www.nlm.nih.gov/mesh/intro_trees.html.

Datasets. The datasets we used are all from 2016 bioCADDIE Dataset Retrieval Challenge. We use these datasets because they are all biomedical datasets from bioCADDIE, a very popular biomedical dataset search engine. Also, all the datasets include text-content in their metadata, such as title and abstract. We can use our entity matching approach and ontology-based similarity measures on these text-content directly.

Query. 15 queries we used in our experiments are from 2016 bioCADDIE Dataset Retrieval Challenge:

1. Find protein sequencing data related to bacterial chemotaxis across all databases
2. Search for data of all types related to MIP-2 gene related to biliary atresia across all databases
3. Search for all data types related to gene TP53INP1 in relation to p53 activation across all databases
4. Find all data types related to inflammation during oxidative stress in human hepatic cells across all databases
5. Search for gene expression and genetic deletion data that mention CD69 in memory augmentation studies across all databases
6. Search for data of all types related to the LDLR gene related to cardiovascular disease across all databases
7. Search for gene expression datasets on photo transduction and regulation of calcium in blind *D. melanogaster*
8. Search for proteomic data related to regulation of calcium in blind *D. melanogaster*
9. Search for data of all types related to the ob gene in obese *M. musculus* across all databases
10. Search for data of all types related to energy metabolism in obese *M. musculus*
11. Search for all data for the HTT gene related to Huntington's disease across all databases
12. Search for data on neural brain tissue in transgenic mice related to Huntington's disease
13. Search for all data on the SNCA gene related to Parkinson's disease across all databases
14. Search for data on nerve cells in the substantia nigra in mice across all databases
15. Find data on the NF-kB signaling pathway in MG (Myasthenia gravis) patients

Gold Standard. As we mentioned in Sect. 3, we use given a set of queries and their expected answers as gold standard. For each query, its expected answers are a list of bioCADDIE datasets ranked by relevant score. The relevant score are ranged from 0 to 3:

- score 0: no relevance between dataset and query;
- score 1: possible no relevance between dataset and query;
- score 2: possible relevance between dataset and query;
- score 3: relevance between dataset and query.

Evaluation. As we mentioned above, we would use AP and nDCG measure to evaluate our retrieval approach. The relevant score (ranged from 0 to 3) in gold standard could be used in nDCG directly. This is because relevant score ranged from 0 to 3 is also used as standard in definition of nDCG [3]. However, for average precision (AP) measure, the relevant score is always ranged from 0 (not relevance) to 1 (relevance). So split our AP evaluation experiments into two scenarios:

- Brave scenario: We consider score 2 and 3 of relevant score ranged from 0 to 3, as score 1 of relevant score ranged from 0 to 1.
- Caution scenario: We consider only score 3 of relevant score ranged from 0 to 3, as score 1 of relevant score ranged from 0 to 1.

5.2 Results

Now we can go straight to results of evaluation for our dataset retrieval approach. In Table 1, we can see the evaluation results for dataset retrieval approach based

on Wu-Palmer. For all the scenarios of AP measure, the evaluation results are not good. In Chapter 8 of [8], the disadvantage of AP measure is introduced: AP is the least stable of the commonly used evaluation measures and that it does not average well, since the total number of relevant documents for a query has a strong influence on precision at k. So we can know the reason why the results of AP measure are not stable. For nDCG evaluation scores, we can find that nDCG scores are quite good while nDCG@10 scores are not good. This means that our dataset retrieval approach performs not so well if we just consider Top10 retrieved datasets. The highest nDCG score can reach almost 0.9 (0.8928), for Query 15. This is a very good result for nDCG evaluation. For this query, nDCG@10 score is 0.6976, a very good score in nDCG@k evaluation. We also achieve high nDCG score in Query 1,3,4,6,14. This means that in these queries, our approach can retrieve very relevant datasets by using edge-based similarity measure. A more clear and direct view can be found in Fig. 2.

Table 1. Evaluation results of dataset retrieval approach based on Wu-Palmer.

QueryID	Brave_AP	Brave_AP@10	Caution_AP	Caution_AP@10	nDCG	nDCG@10
1	0.2674	0.9091	0.0639	0.4826	0.6813	0.5566
2	0.0011	0	0	0	0.2963	0
3	0.2857	0.3754	0.0002	0	0.7867	0.2543
4	0.0438	0	0.0002	0	0.5321	0
5	0.0014	0	0	0	0.2914	0
6	0.119	0.1192	0.0006	0	0.6749	0.1004
7	0.0027	0	0	0	0.4668	0
8	0.0022	0.0227	0	0	0.3668	0.0489
9	0.0251	0	0.0029	0	0.4376	0
10	0.0551	0.0283	0.0054	0	0.4777	0.0433
11	0.0135	0	0.0083	0	0.4	0
12	0.0422	0.0909	0.0039	0	0.4433	0.0734
13	0.0134	0	0.0025	0	0.3945	0
14	0.0565	0.1111	0.0267	0.1111	0.4756	0.1909
15	0.2129	0.4527	0	0	0.8928	0.6976

From Table 2 and Fig. 3, we can state that dataset retrieval approach based Resnik has evaluation results similar to Wu-Palmer one. Resnik-based approach also has poor scores in all AP scenarios but have good scores in nDCG scenario. The highest nDCG score is 0.8888 (Query No. 15). For this query, the nDCG@10 score can also reach 0.6563. So we can know that our approach performs very well in this query, according to nDCG evaluation. We also achieve high nDCG score in Query 1, 3, 4, 6, 10, 12, 14. This means our approach can work well in some queries with both edge-based and node-based similarity measures.

Summary of Results. Through the evaluation results, we can have a brief summary for our approach’s performance:

Fig. 2. Evaluation results for dataset retrieval approach based on Wu-Palmer.

Table 2. Evaluation results of dataset retrieval approach based on Resnik.

QueryID	Brave_AP	Brave_AP@10	Caution_AP	Caution_AP@10	nDCG	nDCG@10
1	0.2619	0.8182	0.0618	0.5071	0.6813	0.5535
2	0.0013	0	0	0	0.3142	0
3	0.2857	0.3754	0.0002	0	0.7867	0.2543
4	0.0497	0	0.0003	0	0.5435	0
5	0.0014	0	0	0	0.2971	0
6	0.1412	0.6032	0.0004	0	0.7178	0.2628
7	0.0037	0.0455	0	0	0.5214	0.1389
8	0.0022	0	0.0001	0	0.3662	0
9	0.0251	0	0.0029	0	0.4376	0
10	0.0753	0.1662	0.0056	0	0.5127	0.1223
11	0.0214	0.053	0.0135	0.053	0.4534	0.1155
12	0.0873	0.2489	0.0137	0.1169	0.5697	0.2585
13	0.0171	0.0227	0.0026	0	0.4106	0.0316
14	0.0713	0.1545	0.0371	0.1545	0.5116	0.2459
15	0.216	0.4057	0	0	0.8888	0.6563

Fig. 3. Evaluation results for dataset retrieval approach based on Resnik.

- Our approach performs well in nDCG evaluation (very well in some queries, such as Query 15);
- Our approach cannot get a good result if we just consider Top10 retrieved datasets;
- Our approach performs not well in AP evaluation, and we discuss the possible reason might be the disadvantage of AP evaluation.

6 Conclusion and Discussion

In this paper, we provided a new dataset retrieval approach based on ontology-based semantic similarity. We test this approach on 2016 bioCADDIE Dataset Retrieval Challenge, and evaluate it with average precision (AP) measure and normalized discounted cumulative gain (nDCG) measure. The evaluation result shows that our approach can perform well in dataset retrieval task.

We also have some future work to do. Other similarity measures of calculating similarity between test-content can be used, such as Word2vec [9] and normalized Google Distance (NGD) [2]. In this paper we also find some disadvantage of our approach, which would be improved in future work.

Acknowledgments. This work has been funded by the Netherlands Science Foundation NWO grant nr. 652.001.002, it is co-funded by Elsevier B.V., with funding for the first author by the China Scholarship Council (CSC) grant number 201807730060.

References

1. Cheng, J., et al.: A knowledge-based clustering algorithm driven by gene ontology. *J. Biopharm. Stat.* **14**(3), 687–700 (2004). <https://doi.org/10.1081/BIP-200025659>
2. Cilibrasi, R.L., Vitanyi, P.M.: The google similarity distance. *IEEE Trans. Knowl. Data Eng.* **19**(3), 370–383 (2007)
3. Järvelin, K., Kekäläinen, J.: IR evaluation methods for retrieving highly relevant documents. In: *Proceedings of the 23rd SIGIR Conference, SIGIR 2000*, pp. 41–48. ACM, New York (2000)
4. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.* **20**(4), 422–446 (2002). <https://doi.org/10.1145/582415.582418>
5. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. In: *Proceedings of the 10th Research on Computational Linguistics International Conference. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP)*, Taipei, Taiwan, pp. 19–33, August 1997. <https://www.aclweb.org/anthology/O97-1002>
6. Kunze, S.R., Auer, S.: Dataset retrieval. In: *2013 IEEE Seventh International Conference on Semantic Computing*, pp. 1–8 (2013)
7. Lin, D.: An information-theoretic definition of similarity. In: *Proceedings of the Fifteenth International Conference on Machine Learning, ICML 1998*, pp. 296–304. Morgan Kaufmann Publishers Inc., San Francisco (1998)
8. Manning, C.D., Raghavan, P., Schütze, H.: *Introduction to Information Retrieval*. Cambridge University Press, Cambridge (2008)

9. Mikolov, T., Corrado, G., Chen, K., Dean, J.: Efficient estimation of word representations in vector space, pp. 1–12, January 2013
10. Pekar, V., Staab, S.: Taxonomy learning: factoring the structure of a taxonomy into a semantic classification decision. In: Proceedings of the 19th International Conference on Computational Linguistics, COLING 2002, vol. 1, pp. 1–7. Association for Computational Linguistics, USA (2002). <https://doi.org/10.3115/1072228.1072318>
11. Pilehvar, M.T., Jurgens, D., Navigli, R.: Align, disambiguate and walk: a unified approach for measuring semantic similarity. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Sofia, Bulgaria, pp. 1341–1351. Association for Computational Linguistics, August 2013. <https://www.aclweb.org/anthology/P13-1132>
12. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, IJCAI 1995, vol. 1, pp. 448–453. Morgan Kaufmann Publishers Inc., San Francisco (1995)
13. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. In: Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics, pp. 133–138. Association for Computational Linguistics (1994)
14. Zhang, E., Zhang, Y.: Average Precision, pp. 192–193. Springer, Boston (2009). https://doi.org/10.1007/978-0-387-39940-9_482