Evaluating Similarity Measures for Dataset Search

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

published in
Web Information Systems Engineering – WISE 2020
2020

DOI (link to publisher)
10.1007/978-3-030-62008-0_3

document version
Publisher's PDF, also known as Version of record

Evaluating Similarity Measures for Dataset Search

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. Dataset search engines help scientists to find research datasets for scientific experiments. Current dataset search engines are query-driven, making them limited by the appropriate specification of search queries. An alternative would be to adopt a recommendation paradigm ("if you like this dataset, you'll also like..."). Such a recommendation service requires an appropriate similarity metric between datasets. Various similarity measures have been proposed in computational linguistics and informational retrieval. The goal of this paper is to determine which similarity measure is suitable for a dataset search engine. We will report our experiments on different similarity measures over datasets. We will evaluate these similarity measures against the gold standards which are developed for Elsevier DataSearch, a commercial dataset search engine. With the help of F-measure evaluation measure and nDCG evaluation measure, we find that Wu-Palmer Similarity, a similarity measure which is based on hierarchical terminologies, can score quite good in our benchmarks.

Keywords: Semantic similarity · Ontology-based similarity · Dataset search · Data science · Google Distance

1 Introduction

Sharing of datasets is becoming increasingly important in all branches of modern science [1,6,9]. Search engines dedicated to finding datasets that fill the needs of a scientist are now emerging rapidly, and similarity metrics for datasets are an important building block of such dataset search engines.

A scientific dataset is a set of data used by scientists or researchers for scientific experiments and scientific analysis. Usually, scientific datasets are categorized into three type: experimental datasets, computational datasets, and observational dataset [3].

Dataset search engines can help scientists to find such research datasets more efficiently. Dataset search engines are now emerging rapidly: DataSearch engine[1] (Elsevier), Dataset Search[2] (Google), Mendeley Data[3] (Mendeley) just to name

[1] https://datasearch.elsevier.com/.
[2] https://toolbox.google.com/datasetsearch.
[3] https://data.mendeley.com/.

Z. Huang et al. (Eds.): WISE 2020, LNCS 12343, pp. 38–51, 2020.
https://doi.org/10.1007/978-3-030-62008-0_3

a few. Elsevier's DataSearch engine is one of the most popular dataset search engines to date.

Although dataset search engine can be very helpful for scientists, the datasets returned by such search engines are strictly dependent on the appropriate specification of search queries. An alternative approach is the recommendation paradigm [2], where a search engine recommends datasets to a scientist based on similarity to datasets that are already known to be of interest to the researcher. Whereas the accuracy of queries is a limiting factor on dataset search, the quality of the similarity measure is crucial to dataset recommendation.

The goal of this paper is to answer which similarity measure is more suitable for a dataset search engine. In order to meet this goal, we propose a novel evaluation measure to evaluate the performance of a similarity measure for dataset search engines. The gold standard we used for evaluation is the gold standard ranking from a commercial dataset search engine (See footnote 1). However, this gives a gold standard for ranking, and not for similarity. To evaluate dataset *similarity* measures, we use of the similarity measures to reconstruct a ranking of datasets for a given query and then compare the reconstructed ranking to the gold standard ranking to get the accuracy of this reconstructed ranking. Usually, this accuracy is measured through the F-measure [4] and normalized Discounted cumulative gain (nDCG) measure [7]. We also propose a new F-measure to help us evaluate similarity measures because of the particularity of our gold standard ranking. In our experiments, we test our evaluation measures in Elsevier DataSearch engine with evaluating three measures (Wu-Palmer measures [12], Resnik measures [11] and Normalized Google Distance [5]). Then we using the evaluation measure to evaluate which similarity measure perform better in Elsevier DataSearch engine for these three similarity measures.

The main contributions of this paper are (1) to provide a new approach to evaluate similarity measures for dataset search engines, (2) to introduce two new kinds of F-measures (Brave and Cautious), and (3) to find out which similarity measure performs well on bio-medical datasets.

2 Preliminaries

In this section, we will introduce the similarity measures we used in this paper and the evaluation measures we used for evaluating the quality of our experiments results.

2.1 Similarity Measures

Various similarity measures have been proposed in computational linguistics and informational retrieval, such as topological similarity measures (for instance, Wu-Palmer Similarity measure [12] and Resnik Similarity measure [11]) and Statistical similarity measures (for instance Normalized Google Distance [5]). In NLP domain, word2vec is a popular measure to calculate the similarity between two terms.

Wu-Pamler Similarity. Wu-Palmer similarity measure [12] is a semantic similarity measure between two concepts based on the ontology structure. We use the Wu-Palmer Similarity measure in this paper because Wu-Palmer measure is an popular edge-based topological similarity measure. Wu-Palmer similarity between two concepts C_1 and C_2 is

$$Sim(C_1, C_2) = \frac{2 * N3}{N1 + N2 + 2 * N3} \tag{1}$$

where C_3 is the least common superconcept of C_1 and C_2, $N1$ is the number of nodes on the path from C_1 to C_3, $N2$ is the number of nodes on the path from C_2 to C_3 and $N3$ is the number of nodes on the path from C_3 to root.

Resnik Similarity. Resnik similarity measure [11] is a node-based topological similarity measure between two concepts based on the notion of information content, which combines the path based measure and the relative depth measure. We use the Resnik measure because most other node-based topological similarity measures are more or less based on the Resnik measure. In this measure, a function $p(c)$, which is the probability of encountering an instance of concept c, is introduced. $p(c)$ is computed as follows:

$$p(c) = \frac{\sum_{n \in words(c)} count(n)}{N} \tag{2}$$

where $words(c)$ is the set of concepts which are subsumed by concept c and N is the total number of nouns observed on the ontology structure.

Then the Resnik semantic similarity of two concepts C_1 and C_2 is defined as follows:

$$sim(C_1, C_2) = -\log p(C_3) \tag{3}$$

where C_3 is the least common super-concept of C_1 and C_2.

Normalized Google Distance. Normalized Google Distance (or Google Distance) [5] is a semantic similarity measure based on the number of hits from Google search engine. Different from Wu-Palmer and Resnik measures, Google Distance is a statistical similarity, and we use Google Distance measure as baseline to compare with two ontology-based similarity measures. For every two concepts x and y, the Google Distance between x and y is

$$NGD(x, y) = \frac{max\{logf(x), logf(y)\} - logf(x, y)}{logM - min\{logf(x), logf(y)\}} \tag{4}$$

where x and y are terms; $f(x)$ is the number of Google hits number of x; $f(x, y)$ is the number of Google hits for x and y; and M is the total number of web pages searched by Google multiplied by the average number of singleton search terms occurring on pages (estimated to be $25 * 10^9$).

Word2vec. Word2vec approach can produce word embedding and be used to calculate the similarity between two words. There are several popular NLP tools can implement word2vec algorithm. In this paper we use Gensim tools [10] for

word2vec approach. The pretraining model we used for word embedding is the wiki-data[4]. Wikipedia can cover most concepts from every domain. So wiki-data is a suitable choose as the pretraining model for word2vec.

2.2 Evaluation Measures for Information Retrieval

Here, we will shortly introduce F-measure and nDCG measures, which we use in this paper.

F-measure. F-measure (also F-score or F1-score) is a measure of a test's accuracy [4]. F-measure considers two aspects: relevant and retrieved. Relevant always means document or dataset selected by given standard. Retrieved means the one selected by approach under evaluation. F-measure is defined as follow:

$$Precision = \frac{True\ pos}{True\ pos + False\ pos}, \qquad Recall = \frac{True\ pos}{True\ pos + False\ neg} \tag{5}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{6}$$

where $True\ pos$ is relevant and retrieved document/dataset; $False\ neg$ is relevant and not-retrieved one; $False\ pos$ is not-relevant and retrieved one; $True\ neg$ is not-relevant and not-retrieved one.

nDCG. Discounted cumulative gain (DCG) is a measure of ranking quality [7]. Normalized Discounted cumulative gain (nDCG) is a normalized measure based on DCG measure [8]. nDCG through top rank position p is defined as follows:

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(i+1)}, \quad IDCG_p = \sum_{i=1}^{|Rel_p|} \frac{2^{rel_i} - 1}{\log_2(i+1)}, \quad nDCG_p = \frac{DCG_p}{IDCG_p} \tag{7}$$

where DCG_p considers the list of documents (ordered by approach under evaluation); $IDCG_p$ considers the list of documents (ordered by given standard); rel_i is the relevant score in position i, which sometimes means the gold standard score of position i.

3 Similarity Measure and Evaluation Measure

3.1 Similarity Between Sets of Concepts

These similarity measures above calculate the similarity between two concepts (or between two terms representing those concepts). Then we also introduce the similarity measure to calculate the similarity between two *sets* of terms. The similarity between two sets A and B of terms is:

$$Sim(A, B) = \frac{sum\{Sim(a, b) | a \in A, b \in B\}}{(|C(A)| * |C(B)|)} \tag{8}$$

where $|C(A)|$ means the number of concepts in set A, and $|C(B)|$ means the number of concepts in set B.

[4] https://dumps.wikimedia.org/enwiki/.

3.2 Evaluation Measure

In this part, we will introduce the gold standard used for our experiments, the ranking reconstruction and the Caution/Brave F-measure for evaluating experiment's accuracy.

Gold Standard Ranking. For our evaluation, we have obtained a gold standard from the Elsevier DataSearch engine (See footnote 1). The gold standard consists of a set of queries together with the ranked results returned by the search engine for these queries. Expert scientist users had been invited to judge these results by giving a score to every search result. The range of judgement score is from −100 to 100, with the score 0 meaning that the experts cannot judge if this result is similar to the query. The gold standard aggregates these expert judgments into four levels: Likely satisfaction (which means the dataset is an excellent match for the query), possible satisfaction, possible dissatisfaction and likely dissatisfaction, according to the following score range:

- Likely dissatisfaction (level 3): from −100 to −51;
- Possible dissatisfaction (level 2): from −50 to −1;
- Possible satisfaction (level 1): from +1 to +50;
- Likely satisfaction (level 0): from +51 to +100.

Ranking Reconstruction. As described above, we have been given a gold standard for the *ranking* of datasets as query results, whereas we want to measure the *similarity* between datasets. In order to evaluate the similarity measures, we use each of the three similarity measure introduced above to "re-construct" a derived ranking. We can then compare these "derived rankings" with the given gold standard ranking, and find out which of our measures produces a better ranking.

Caution and Brave F-measure. As usual, we use the F-measure to evaluate our experiments of similarity metric. But because our gold standard gives us four categories of answer qualities (as described above), we redefine the original definition of the F-measure, into two more specific measure: the Cautious F-measure and the Brave F-measure.

For the Brave F-measure, we consider both dissatisfaction categories (possible dissatisfaction and likely dissatisfaction) as negative, and similarly we consider both satisfaction as positive.

For the Cautious F-measure, we consider only the stronger categories (*likely* (dis)satisfaction) as positive (resp. negative), while leaving the less pronounced *possible* (dis)satisfaction out of consideration.

In order to calculate these Cautious and Brave F-measures, precision and recall are defined in this paper as follows: For the cautious F-measure, the relevant number is the number of "likely satisfaction" results in the Gold Standard Ranking, denoted as $rel_{caution}$. For the brave F-measure, the relevant number is

the number of "likely satisfaction" or "possible satisfaction" results in the Gold Standard Ranking, denoted as rel_{brave}. The retrieved number is the number of all results in the Reconstructed Ranking, donated as ret_num.

Then the brave/caution F-measure can be defined as follow:

$$precision_{caution} = \frac{\{rel_{caution}\} \cap \{ret\_num\}}{\{ret\_num\}}, \quad recall_{caution} = \frac{\{rel_{caution}\} \cap \{ret\_num\}}{\{rel_{caution}\}}. \tag{9}$$

$$precision_{brave} = \frac{\{rel_{brave}\} \cap \{ret\_num\}}{\{ret\_num\}}, \quad recall_{brave} = \frac{\{rel_{brave}\} \cap \{ret\_num\}}{\{rel_{brave}\}}. \tag{10}$$

$$F - measure_{caution} = \frac{2 * precision_{caution} * recall_{caution}}{precision_{caution} + recall_{caution}} \tag{11}$$

$$F - measure_{brave} = \frac{2 * precision_{brave} * recall_{brave}}{precision_{brave} + recall_{brave}} \tag{12}$$

nDCG with Gold Standard Ranking. In our gold standard introduced above, for every result in gold standard ranking list, the relevant score of it could be 0, 1, 2, 3 for level 0, level 1, level 2, level 3, respectively. So in this paper, we use nDCG with our gold standard ranking through all position in list.

4 Experiments and Results

In this section we introduce our experimental evaluation of the three similarity metrics.

4.1 Data Selection

For practical reasons, we restrict our experiment to the 3039 bio-medical datasets in Elsevier's DataSearch engine (See footnote 1). All the queries and ranked answers we used are given by the Gold Standard ranking obtained in Elsevier Data Search product testing.

Queries. In the Elsevier Gold Standard, 18 queries are listed as in the biomedical domain. These are listed in Table 1.

Ontology. In the biomedical domain, the MeSH terminology (Medical Subject Headings)[5] is an appropriate choice, since it is designed to capture biomedical terminology in the scientific domain.

Individual Datasets. Our datasets are characterized as a set of meta-data fields, and stored in JSON format (see Fig. 1). The restriction to meta-data seems appropriate, because in many real-life cases, the actual contents of the dataset might be entirely numerical, or encoded in some binary format, and hence not accessible for similarity measurements. To stay as close to this real-life situation as possible, we restricted ourselves to the meta-data fields only. Each dataset in the collection is data from an actual publication or scientific experiment.

[5] http://www.nlm.nih.gov/mesh.

Table 1. The corpus of 18 queries

Query	Content	Query	Content
E2	Protein Degradation mechanisms	E54	Glutamate alcohol interaction
E7	Oxidative stress ischemic stroke	E66	Calcium signalling in stem cells
E8	Middle cerebral artery occlusion mice	E67	Phylogeny cryptosporidium
E17	Risk factors for combat PTSD	E68	HPV vaccine efficacy and safety
E26	Mab melting temperature	E78	c elegans neuron degeneration
E28	Mutational analysis cervical cancer	E79	mri liver fibrosis
E31	Metformin pharmacokinetics	E80	Yersinia ruckeri enteric red mouth disease
E35	Prostate cancer DNA methylation	E89	Electrocardiogram variability OR ECG variability
E50	EZH2 in breast cancer	E94	Pinealectomy circadian rhythm


```
{
"id":"57525251:NEURO_ELECTRO",
"externalId":"3449",
"containerTitle":"Localization and function of the Kv3.1b ......",
"source":"NEURO_ELECTRO",
"containerDescription":"The voltage-gated potassium channel......",
"publicationDate":"2005",
"dateAvailable":"2005",
"containerURI":"http://neuroelectro.org/article/3449",
"firstImported":"2017-03-14T13:07:32.096Z",
"lastImported":"2017-03-14T13:07:32.096Z",
"containerKeywords":["Potassium Channels", "Voltage-Gated"......],
"authors":["Mark L Dallas","David I Lewis","Susan A Deuchars"......],
"assets":.....
}
```


Fig. 1. Meta-data fields in JSON

4.2 Extract Terms from Query or Dataset

As explained earlier, in order to calculate the similarity between a query and a dataset, we extract concepts and then consider the similarity between these two sets of concepts as the similarity between Query and Dataset.

MeSH has the following structure: a group of synonymous terms are grouped in a MeSH *concept*, and several MeSH concepts which are synonymous with each other are grouped in a **Descriptor**. A Descriptor is named by the preferred term of the preferred concept among all the concepts in this descriptor[6]. For example, for the MeSH descriptor Cardiomegaly:

[6] See also https://www.nlm.nih.gov/mesh/concept_structure.html.

```
Cardiomegaly                          [Descriptor]
    Cardiomegaly                        [Concept, Preferred]
        Cardiomegaly                      [Term, Preferred]
        Enlarged Heart                    [Term]
        Heart Enlargement                 [Term]
    Cardiac Hypertrophy                 [Concept, Narrower]
        Cardiac Hypertrophy               [Term, Preferred]
        Heart Hypertrophy                 [Term]
```

For each descriptor, we extract from the text-content of queries and datasets all the terms that occur with all concepts grouped under that descriptor. In the above example, extraction of any of the five terms Cardiomegaly, Enlarged Heart, Heart Enlargement, Cardiac Hypertrophy and Heart Hypertrophy would result in the annotation of the query or dataset with the descriptor Cardiomegaly. Since hierarchical relationships in MeSH are at the level of the descriptors, these are indeed suitable for calculating the ontology-based similarity measures.

4.3 Similarity Experiments

We calculate the similarity between a query and a dataset by using the Google distance measure, the Wu-Palmer measure, the Resnik measure and Word2vec. As described above, we will obtain these four similarity values, and use the ranking induced by these values with the ranking from the Elsevier gold standard, and thereby compare the performance of these metrics.

The sets of extracted MeSH descriptors are problematic in two distinct ways. Firstly, the number of MeSH descriptors varies widely between datasets, ranging from 1 to 110 per dataset, with an average of 15. Secondly, some of the extract MeSH descriptors are "noisy" and do not express the main meaning of the dataset. To correct both of these problems, we only consider the best 3, 4, 5 or 6 best scoring MeSH descriptor for each dataset, both balancing the number of descriptors across datasets, and avoiding the influence of poor descriptors (which are indicative of the dataset, and not of the quality of the similarity measure, which is what we are interested in).

4.4 Experimental Results

We will use the gold standard ranking from Elsevier Data Search engine to evaluate our reconstruction results.

First off, we reconstructed the similarity ranking as explained above. Two examples of such reconstructed rankings are shown in Fig. 2. Each column in these figures shows the datasets ranked according to their similarity under the indicated similarity metric, using the top 3, 4, 5 or 6 MeSH descriptors. We colored the result as green (resp. red) if this dataset is categorized as likely satisfaction (resp. dissatisfaction) in the gold standard ranking.

We can evaluate our similarity results by this visualization. From the right hand side table in Fig. 2, we can easily infer that the reconstructed ranking for query E80 is quite good, because most results from the top of the gold standard ranking (the green cells) are still located in top of the reconstructed ranking. This shows that the ranking based on the four similarity measures between query E80 and the datasets is similar to ranking given by the experts. The left hand side of

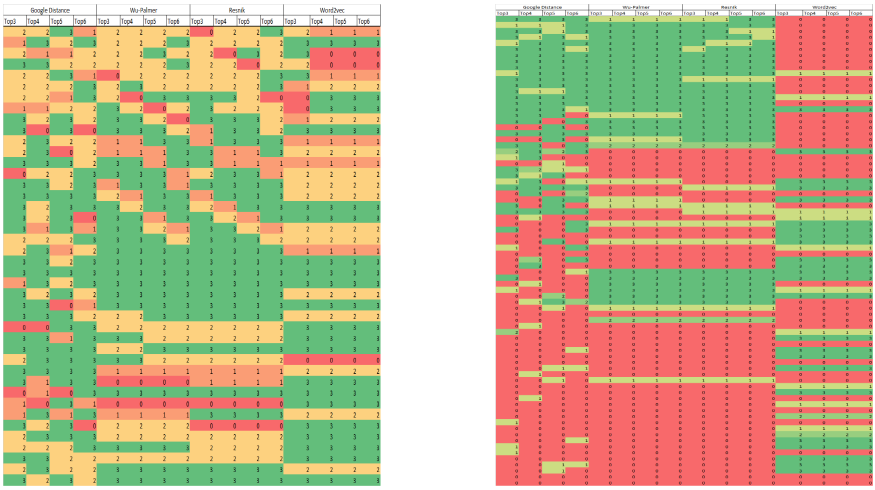


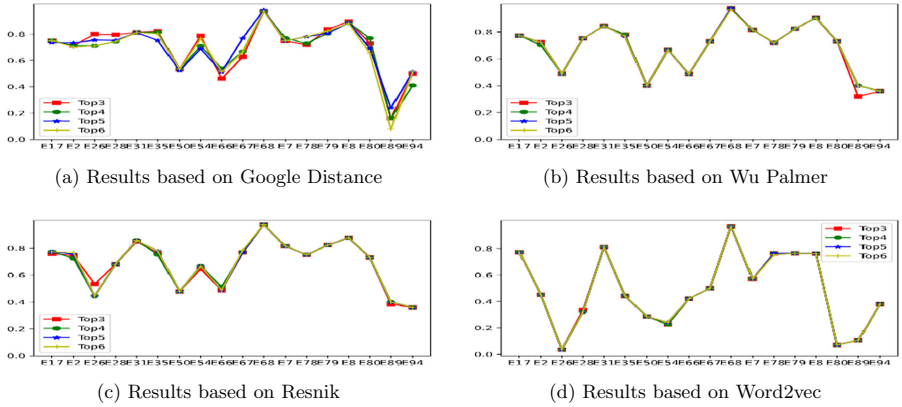
Fig. 2. Queries E79 (left) and E80 datasets ranking reconstructed with Google Distance, Wu-Palmer and Resnik similarity measures. (The tables are of different length because the gold standard contained more expert ratings for E80) (Color figure online)

Table 2. Cautious and Brave F-scores for queries E79 and E80 for the Google Distance, Wu-Palmer, Resnik and Word2vec

Google Distance				Wu-Palmer				Resnik				Word2vec			
Query	Top	Cautious	Brave	Query	Top	Cautious	Brave	Query	Top	Cautious	Brave	Query	Top	Cautious	Brave
E79	3	**0.55**	**0.84**	E79	3	**0.59**	**0.82**	E79	3	**0.59**	**0.82**	E79	3	**0.40**	**0.76**
E79	4	0.50	0.81	E79	4	**0.59**	**0.82**	E79	4	**0.59**	**0.82**	E79	4	**0.40**	**0.76**
E79	5	**0.55**	0.81	E79	5	**0.59**	**0.82**	E79	5	**0.59**	**0.82**	E79	5	**0.40**	**0.76**
E79	6	0.50	0.82	E79	6	**0.59**	**0.82**	E79	6	**0.59**	**0.82**	E79	6	**0.40**	**0.76**
E80	3	0.71	0.73	E80	3	**0.70**	**0.73**	E80	3	**0.70**	**0.73**	E80	3	**0.07**	**0.07**
E80	4	**0.78**	**0.77**	E80	4	**0.70**	**0.73**	E80	4	**0.70**	**0.73**	E80	4	**0.07**	**0.07**
E80	5	0.71	0.69	E80	5	**0.70**	**0.73**	E80	5	**0.70**	**0.73**	E80	5	**0.07**	**0.07**
E80	6	0.69	0.65	E80	6	**0.70**	**0.73**	E80	6	**0.70**	**0.73**	E80	6	**0.07**	**0.07**

Fig. 2 shows that for query E79, the similarity ranking by the metrics does not correspond to the ranking given by the experts, since for this query, most results in top of gold standard ranking are located in middle of reconstructed ranking.

The visual results from Fig. 2 are stated numerically in Tables 2 by computing the cautious and brave F-scores for all four similarity measures, again on queries E79 and E80 as examples. Again, we see that there is barely any difference between using 3, 4, 5 or 6 MeSH descriptors, with differences never bigger than ± 0.01. However, these tables do reveal a difference in behaviour when graded on cautious or brave F-score. Using the cautious F-score, all three similarity metrics performed better on query E80, while for the brave F-Score, all three similarity metrics (except Word2vec) performed better on E79. Word2vec measure perform worse because it's hard for wiki-data to cover the bio-medical concepts as many as MeSH ontology and Google search engine do.



(a) Results based on Google Distance

(b) Results based on Wu Palmer

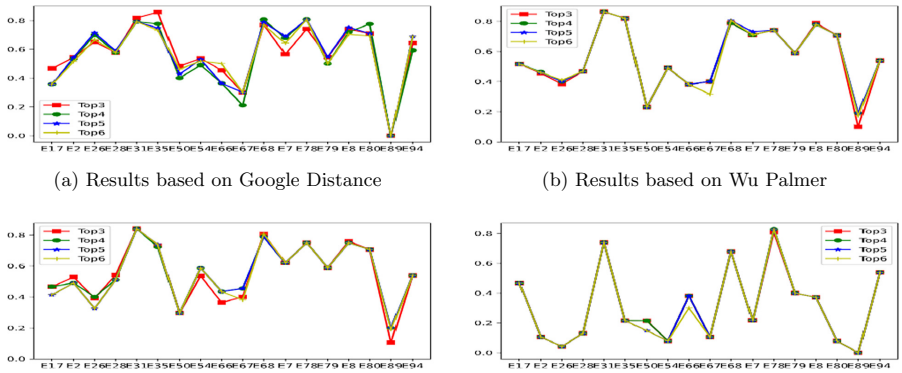
(c) Results based on Resnik

(d) Results based on Word2vec

Fig. 3. Brave F-scores of all queries based on Google Distance, Wu Palmer, Resnik and Word2vec

Table 3. Best F-score counts

Scenario	Best performance count-number				
	Google Distance	Wu-Palmer	Resnik	Wu-Palmer = Resnik	Word2vec
Caution and Brave	**7**	**5**	3	2	1
Caution	**8**	2	4	3	1
Brave	**7**	**4**	3	3	1



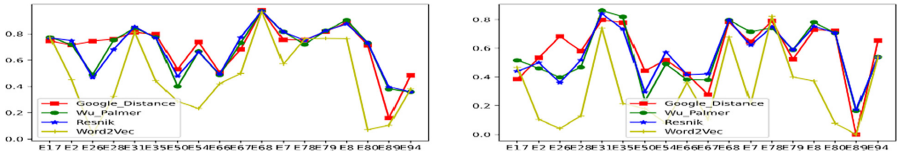
(a) Results based on Google Distance

(b) Results based on Wu Palmer

(c) Results based on Resnik

(d) Results based on Word2vec

Fig. 4. Caution F-scores of all queries based on Google Distance, Wu Palmer, Resnik and Word2vec



(a) Average brave F-scores of all queries (b) Average caution F-scores of all queries

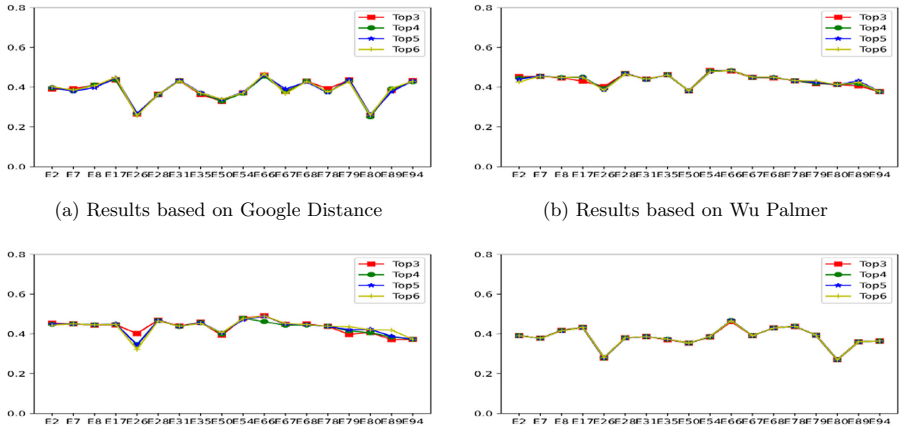
Fig. 5. Average brave and caution F-scores of all queries

F-measure Results. Of course queries E79 and E80 are just examples to illustrate our findings. We are actually interested in which similarity measures scores better across our entire corpus of 18 queries. For this purpose, we collected data as shown in Fig. 3, 5a, 4, 5b, which tabulates the both brave and cautious F-score for all metrics on all queries when computing similarity based on the top 3, 4, 5, 6 MeSH descriptors, as well as the average one based on top 3, 4, 5, 6 Mesh descriptors.

In Fig. 3 and Fig. 4, we can roughly see the difference among every results of different top MeSH descriptors. In this two figures, we can easily find that the differences among every results of different top MeSH descriptors are not big. So we can go straight to average F-scores to see the differences among every approaches based on different similarity measures in Fig. 5a and Fig. 5b. From Fig. 5a and 5b, we can easily know that approach based on Word2vec have a lower score than other approaches. To clearly know which approach performs best, we also show an overall tabulation. The overall tabulation of these findings is shown in Table 3. The table lists how often each similarity measure has the highest F-score across the 18 queries, separating the cases for scoring highest on cautious, scoring highest on brave F-score, and scoring highest on both. The final column state the number of cases where the Wu-Palmer and Resnik metrics resulted in an equal highest score (this rarely happened with Google Distance and Word2vec, so we do not include columns for that).

For F1 score results, this final table shows conclusively that the Google Distance similarity measure outperforms the other two similarity measures in the task of reconstructing the gold standard search ranking based on measuring the similarity between query and dataset. However, Google Distance measure only performs better than Wu-Palmer measure in the scenario of brave F-measure, and shares best performance with Wu-Palmer measure in other scenarios.

nDCG Results. We also use nDCG measure, a standard evaluation measure for information retrieval, to evaluate our experiment results. In Fig. 6, we can see the nDCG scores of all reconstructing approaches based on Google Distance, Wu Palmer, Resnik and Word2vec of different top MeSH descriptors. We can easily find that there is no clear difference among different top MeSH descriptors for each approach. So we know that the difference on top MeSH descriptors would not impact the final results.



(a) Results based on Google Distance

(b) Results based on Wu Palmer

(c) Results based on Resnik

(d) Results based on Word2vec

Fig. 6. nDCG scores of all queries based on Google Distance, Wu Palmer, Resnik and Word2vec

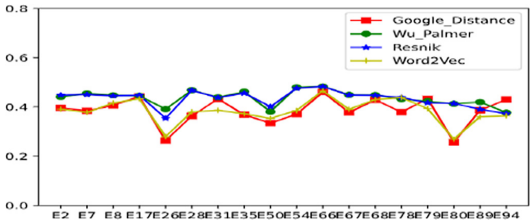


Fig. 7. Average nDCG scores of all queries

Table 4. Average nDCG score for each query.

Query	GoogleDistance	WuPalmer	Resnik	Word2Vec	Query	GoogleDistance	WuPalmer	Resnik	Word2Vec
E2	0.3969	0.4411	**0.4487**	0.3911	E54	0.3711	**0.48**	0.4762	0.3862
E7	0.3845	**0.4553**	0.4504	0.3781	E66	0.4609	**0.4835**	0.4833	0.466
E8	0.4066	**0.4476**	0.4451	0.4163	E67	0.3779	**0.4493**	0.4486	0.3908
E17	0.4429	0.4454	**0.448**	0.4336	E68	0.429	**0.4486**	0.447	0.4302
E26	0.2636	**0.3911**	0.354	0.2796	E78	0.379	0.4324	**0.4383**	0.4381
E28	0.3633	**0.4684**	0.4678	0.3807	E79	**0.4334**	0.4234	0.4172	0.3934
E31	0.433	**0.4394**	0.4383	0.3866	E80	0.2572	0.413	**0.4144**	0.2694
E35	0.3686	**0.4622**	0.456	0.3727	E89	0.3866	**0.4202**	0.3908	0.3605
E50	0.3345	0.3811	**0.4016**	0.3535	E94	**0.4305**	0.3772	0.373	0.3642

To find out which approach performs best in nDCG scores, we also collect the average nDCG scores for every approach, by taking average of all the scores of all different top MeSH descriptors for each approach. Average nDCG score results are shown in Fig. 7. According to this figure, we can intuitively know that approaches based on both Wu Palmer and Resnik outperform approaches based

on Google Distance and Word2vec. To find out the winner of nDCG scores among these two approaches, we also collect the full average scores shown in Table 4. In Table 4, we can know that approach based on Wu Palmer has best average nDCG scores on 11 queries of all 18 queries. So we can say that Wu Palmer can score best in our nDCG benchmark.

5 Discussion and Conclusion

In modern science, sharing of datasets is becoming increasingly important. Search engines dedicated to finding datasets that fill the needs of a scientist are now emerging rapidly, and similarity metrics for datasets are an important building block of such dataset search engines. In this paper, we have reported experiments on four important similarity measures for datasets. Using a gold standard from a commercial search engine in experiments on biomedical datasets, we have found that the Wu-Palmer Similarity metric outperformed the other three candidates in nDCG benchmark, although it performed a bit worse than Google Distance and scores secondary in F-measure benchmark.

Future work would of course involve the extension of our results to other candidate similarity measures, including those based on embedding the biomedical domain vocabulary in a high-dimensional vector space.

Our test datasets are limited to the datasets in biomedical domain, because of the availability of a gold standard for this biomedical domain. Future work will have to show whether our conclusion can be extended to cover datasets in other domains. To this end, the development of similar gold standards is an important and urgent task for the community.

Acknowledgements. 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. We are grateful to our colleagues in Elsevier for sharing their dataset, and to all of our colleagues in the Data Search project for their valuable input.

References

1. Bauchner, H., Golub, R., Fontanarosa, P.: Data sharing: an ethical and scientific imperative. J. Am. Med. Assoc. **12**(315), 1238–1240 (2016)
2. Bobadilla, J., Ortega, F., Hernando, A., Gutiérrez, A.: Recommender systems survey. Knowl.-Based Syst. **46**, 109–132 (2013)
3. Borgman, C.L., Wallis, J.C., Mayernik, M.S.: Who's got the data? Interdependencies in science and technology collaborations. Comput. Supported Coop. Work (CSCW) **21**(6), 485–523 (2012). https://doi.org/10.1007/s10606-012-9169-z
4. Chinchor, N.: MUC-4 evaluation metrics. In: Proceedings of the 4th Conference on Message Understanding, MUC4 1992, pp. 22–29. Association for Computational Linguistics, New York (1992)
5. Cilibrasi, R.L., Vitanyi, P.M.: The google similarity distance. IEEE Trans. Knowl. Data Eng. **19**(3), 370–383 (2007)

6. Editorial: Benefits of sharing. Nature **530**(7589), 129 (2016). https://doi.org/10.1038/530129a
7. 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)
8. 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
9. McNutt, M.: Data sharing. Science **351**, 1007 (2016). https://doi.org/10.1126/science.aaf4545
10. Řehůřek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45–50. ELRA (2010)
11. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. CoRR abs/cmp-lg/9511007 (1995). http://arxiv.org/abs/cmp-lg/9511007
12. 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)