In the pursuit of a semantic similarity metric based on UMLS annotations for articles in PubMed Central Open Access

Leyla Jael Garcia Castro[1,*], Rafael Berlanga[1] and Alexander Garcia[2]

[1]Temporal Knowledge Bases Group. Department of Computer Languages and Systems, Universitat Jaume I, 12071, Castelló de la Plana, Spain.

[2] Linking Data LLC, Denver, Colorado, United States.

[*] Corresponding author al278693@uji.es

Abstract

Motivation: Although full-text articles are provided by the publishers in electronic formats, it remains a challenge to find related work beyond the title and abstract context. Identifying related articles based on their abstract is indeed a good starting point; this process is straightforward and does not consume as many resources as full-text based similarity would require. However, further analyses may require in-depth understanding of the full content. Two articles with highly related abstracts can be substantially different regarding the full content. How similarity differs when considering title-and-abstract versus full-text and which semantic similarity metric provides better results when dealing with full-text articles are the main issues addressed in this manuscript.

Methods: We have benchmarked three similarity metrics –BM25, PMRA, and Cosine, in order to determine which one performs best when using concept-based annotations on full-text documents. We also evaluated variations in similarity values based on title-and-abstract against those relying on full-text. Our test dataset comprises the Genomics track article collection from the 2005 Text Retrieval Conference. Initially, we used an entity recognition software to semantically annotate titles and abstracts as well as full-text with concepts defined in the Unified Medical Language System (UMLS®). For each article, we created a document profile, i.e., a set of identified concepts, term frequency, and inverse document frequency; we then applied various similarity metrics to those document profiles. We considered correlation, precision, recall, and F1 in order to determine which similarity metric performs best with concept-based annotations. For those full-text articles available in PubMed Central Open Access (PMC-OA), we also performed dispersion analyses in order to understand how similarity varies when considering full-text articles.

Results: We have found that the PubMed Related Articles similarity metric is the most suitable for full-text articles annotated with UMLS concepts. For similarity values above 0.8, all metrics exhibited an F1 around 0.2 and a recall around 0.1; BM25 showed the highest precision close to 1; in all cases the concept-based metrics performed better than the word-stem-based one. Our experiments show that similarity values vary when considering only title-and-abstract versus full-text similarity. Therefore, analyses based on full-text become useful when a given research requires going beyond title and abstract, particularly regarding connectivity across articles.

Availability: Visualization available at ljgarcia.github.io/semsim.benchmark/, data available at http://dx.doi.org/10.5281/zenodo.13323

Keywords: Semantic similarity, scientific publications, similarity metrics, semantic annotations, related articles.

1 INTRODUCTION

The scientific literature is nowadays distributed in electronic form; publishers make PDF and HTML versions available over the web. Although an improvement over previous channels of distribution, the knowledge remains embedded in unstructured natural language text surrounded by meta-data information. Searching within collections of documents largely remains a keyword-based experience [1]. In Life Sciences, advanced queries against the PubMed repository often rely on the use of Boolean operators; however, the lack of support for queries based on semantic annotations limits the retrieval results that can be obtained [1]. Unlike Web documents, for which there is an

explicit linking structure, scientific papers lack such highly hyperlinked arrangement [2]; this makes it difficult to use Web search technologies based on link analysis. Search and retrieval should move from finding documents to finding relationships, facts, and actionable intelligence [3]; all this remains difficult as the core information contained in scientific publications is encoded in natural language within monolithic documents. Scientific papers are naturally related to each other in ways beyond sharing authors or bibliographic references; concept-based relations are also important when establishing the associations across collections of documents. However, such relations are usually hidden for practical purposes.

Co-citation analysis is a measure of the relatedness across documents; if at least one other document cites two documents in common, these documents are said to be co-cited. The more co-citations two documents receive, the higher their co-citation strength, and the more likely they are related [4, 5]. However, co-citation analysis does not provide enough information regarding the concept-based connectivity tissue between articles. Despite the existence and wide usage of standardized public resources such as Medical Subject Headings (MeSH) [6], the Systematized Nomenclature of Medicine Clinical Terms (SNOMED) [7], and the Unified Medical Language System (UMLS) [8], concept-based connectivity across scientific publications in the biomedical domain is still underexploited. We argue that the concept-based connectivity tissue could be revealed by analyzing the semantics of the articles; for instance, by using semantic annotations and similarity metrics. Semantic similarity is a measure used to estimate the likeness between documents or terms based on their meaning, i.e., semantic, rather than their syntactic [9]. When applied to documents, the similarity can be estimated by measuring the distance between the relevant concepts contained in both articles; such an approach, using word stems or complete words rather than concepts, has been reported in the literature [10, 11]. Relevancy usually takes into account the occurrence in the collection as well as the frequency in the analyzed documents. Semantic similarity plays an important role in a variety of text processing tasks [12], including document classification [13-16], information extraction [17], and information retrieval [18, 19].

Our investigation focuses on semantic similarity across full-text articles. We are interested in finding out which similarity metric should be used with UMLS annotations on full-text documents; we are also interested in the differences between title-and-abstract versus full-text based similarity approaches. Here we present a systematic analysis on different similarity metrics based on word stems and concepts identified in title and abstract, as well as concepts identified in the full-text. Particularly, we consider the PubMed Related Articles metric (PMRA) [11], BM25 [20-22], and Cosine Similarity [23, 24]. The method we follow considers a well-known collection of articles manually grouped according to relevance judgments; such a collection is used with a base line to assess how similarity metrics are capturing relations across articles that are, in principle, relevant to each other. We have performed experiments using the test collection from the Text Retrieval Conference 2005 (TREC-05) Genomics Track [25]. Our experiments aim to determine which similarity metric works best with our annotations in terms of correlation, precision, and recall regarding a baseline. Such baseline consists of a similarity matrix generated by applying the PubMed Related Articles algorithm on word stems (PMRA-stems) extracted from titles and abstracts [11]. Furthermore, we explore variations on title-and-abstract similarity versus full-text. Our results indicate that similarities based on annotations covering solely concepts identified in title and abstract differ from those covering the full-text; thus, in-depth similarity analyses may benefit from full-text annotations.

This article is organized as follows: in the Materials and Methods section, we introduce our test dataset, an overview of our method, and some definitions. Then we provide a detailed explanation of the different tasks carried on as part of our method. In the Results section, we present our findings, including a comparison between title-and-abstract based similarities against full-text based ones. We then discuss related work as well as our findings, particularly the best performing algorithm for UMLS-based annotations. We finish with conclusions and future work. Three appendixes providing additional information related to materials and results are included at the end of the article.

2 MATERIALS AND METHODS

Similarity between two articles (c, d) is defined as the probability of being interested in article C given a known interest in article D. Such probability is calculated based on the terms identified in article D as well as those present in article C [11]. A term is either a single

word like “phosphorylation” or several words associated with a single idea like “adenosine triphosphate (ATP)”. As similarity metrics, we used PMRA [11], BM25 [20-22], and Cosine similarity [23, 24]. Other well-known similarity metrics such as Latent Semantic Analysis (LSA) [26] and Topic Modeling [27] have not been considered. LSA heavily relies on singular value decomposition (SVD), a computationally intensive algorithm. SVD is still difficult to update as new documents appear, although new and more efficient algorithms have appeared since the first implementation of SVD. Different from PMRA and BM25, LSA relies on a Gaussian distribution. The topic modeling approach describes a collection as a list of topics and assigns a small number of these topics to each article within the collection. The topic modeling could be considered a probabilistic version of LSA [10].

PMRA is a ranking measure used to calculate the “Related articles” in the PubMed interface; it is usually accepted as a *de facto* standard as it has been selected by the National Institutes of Health (NIH) for PubMed. Similar to PMRA, BM25 is also a Poisson-based model; it is used for ranking matching documents according to their relevance regarding a given query. In our case, such a query corresponds to the article for which an interest has been already expressed. The Cosine Similarity corresponds to the inner product space that measures the cosine angle between two vectors; for the case of document similarity, such vectors comprise the relevant terms in the document. We calculated the similarity for each article (from article 1 to article 4240) against all other articles as well as itself, disregarding the topic they belong to. In such a way, we obtained similarity square matrixes of 4240 X 4240 where each row represents an article *A* while cells contain the similarity between all other articles and article *A*. As similarity matrixes depend on the terms contained in the documents, in order to obtain such a matrix, it is first necessary to profile all the participating documents. A document profile is a vector with all the relevant terms identified in the article with their term frequency *tf* and inverse document frequency *idf*; depending on the approach, terms can be identified solely in title and abstract or in the full-text. As *tf*, we use the raw frequency of a term in a document, i.e., the number of times that such a term occurs in the profiled document.

PubMed Central (PMC) [28] is a free full-text digital repository of biomedical literature provided by the National Center for Biotechnology Information (NCBI); currently, it includes more than 1600 journals and 2.9 million articles. PMC-OA consists of a subset of PMC where articles are still protected by copyright but are also available under the Creative Commons license; i.e., a more liberal redistribution is allowed, which makes the collection ideal for text-mining purposes. UMLS [8] is a collection of multiple controlled vocabularies in the biomedical domain; its meta-thesaurus (version 2012AA) covers more than 2.5 million concepts from over 150 terminological resources, including Medical Subject Headings, NCI Thesaurus, and some others. Both PMC-OA and UMLS are within the most comprehensive knowledge resources in the biomedical domain; therefore, we have narrowed our research question to publications available in PMC-OA and concepts comprised in the 2012AA UMLS release.

2.1 Materials

We used the test collection provided by TREC-05 Genomics Track, which comprises a ten-year subset of MEDLINE. This test collection includes 34,633 unique PubMed identifiers (PMID) grouped in 50 topics corresponding to different information needs. For each PMID within a particular topic, human assessors, i.e., domain experts, assigned a relevance judgment depending on whether the PMID was not relevant, partially relevant, or relevant for the topic, i.e., information need. 4584 PMIDs corresponding to 4232 unique articles were categorized as relevant or partially relevant; a summary of all the topics can be found in Appendix A. We used the NCBI’s Entrez Programming Utilities (*e-Utils)* web services [29] to retrieve title and abstract for those 4232 PMIDs judged to be relevant or partially relevant. *e-Utils* are a set of web services providing programmatic access to databases hosted by NCBI, particularly PubMed and PMC. Our title-and-abstract test dataset comprises those articles for which the retrieval was successful, i.e., 4240 articles, and their corresponding topics.

Additionally, we selected articles from the initial test dataset for which there is an entry in PMC-OA; i.e., full-text is freely available. From the 4240 articles, 94 can be mapped to a PMC identifier; however, only for 62 of them it was actually possible to retrieve the full-text using NCBI’s *e-Utils* web services. Those 62 articles correspond to our full-text test dataset. In Fig. 1 we present a graphical summary of the process followed in order to build our test datasets, i.e., title-and-abstract test dataset and full-text test dataset.

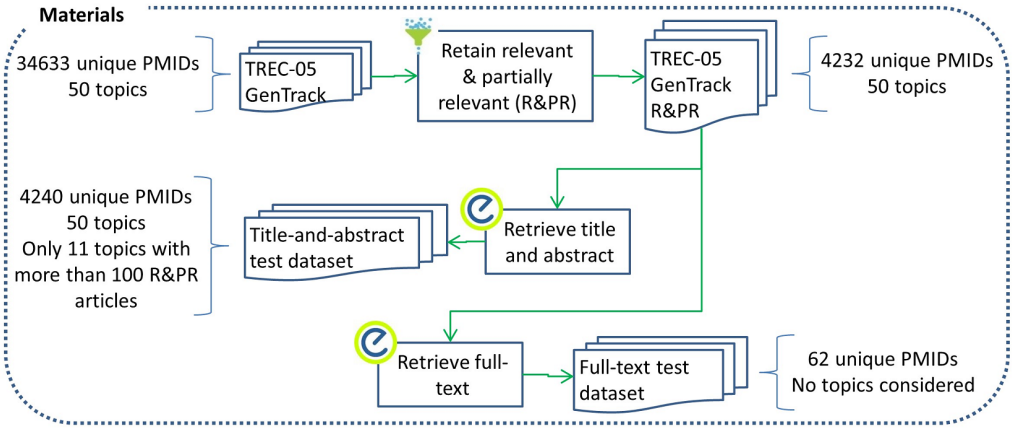

Fig. 1. Materials processing in order to build our test datasets for title-and-abstract as well as full-text.

2.2 Methods

Our workflow can be summarized in the following main tasks; a detailed explanation is provided in this subsection:

- Baseline generation using PMRA-stems algorithm, including word stem extraction, and stem-based document profiling;
- Annotation-based similarity matrixes generation using PMRA, BM25 and Cosine algorithms, also including annotation, and annotation-based document profiling;
- Full-text annotation-based similarity matrixes;
- Correlation, precision, recall, F1 score, and scattered plot analyses.

2.2.1 Baseline similarity matrix

Our baseline consists of a similarity matrix obtained by applying the PMRA-stems algorithm on all PMIDs in our test dataset, see Fig. 2. In order to build our test dataset, we initially retrieved the TREC-05 collection, particularly relevant and partially relevant documents per topic. Then, we retrieved the title and abstract for those documents, which were later used to generate document profiles based on word stems. A word stem is that part left after taking off the ending, for instance the stem “pigment” covers words such as “pigmented”, “pigment”, “pigmentations”, and “pigmentation”. For PMRA-stems, a relevant term included in a document profile consists of a stem present in title or abstract. In order to obtain stems from our test collection, we used the Porter’s algorithm [22] as provided by the author (http://tartarus.org/martin/PorterStemmer/java.txt). The *tf* for each stem within a document was also calculated. Once all stems were processed, we proceeded to calculate the *idf*. Similar to the pre-calculated related citations offered in PubMed [30], stems found in titles were accounted for twice, while stems found in the abstract were accounted for once. Different from the algorithm used in PubMed [30] but similar to the initial evaluation of such an algorithm [11], we did not consider MeSH terms for our baseline definition.

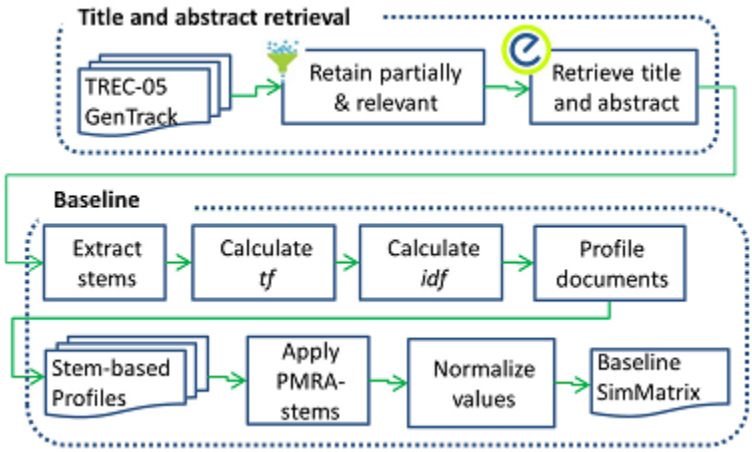

Fig. 2. Generation of our baseline similarity matrix.

Profile documents obtained from stems in title and abstract were used to generate our baseline similarity matrix. The PMRA-stems formula used is the one reported by [11] and presented in the Equation 1. Similarity values were normalized row by row, i.e., article by article, to values between [0, 1], with 1 corresponding to the similarity for article(row) X article(row); i.e., 1 is totally similar while 0 is totally dissimilar. We use the optimal values found in [11] for the PMRA constants, i.e., μ=0.013 and λ=0.022. PMRA also considers the length of the document in words, represented by *l(a)* in Equation 1; *tf(t,a)* corresponds to the term frequency of the term –stem, *t* in the document *a*, while *idf(t)* corresponds to the inverse document frequency for the term *t* in our corpus, i.e., test collection.

$$pmra\_sim(c,d) = \sum_{t=1}^{N} w(t,c) \times w(t,d)$$
$$w(t,a) = (1 + (\frac{\mu}{\lambda})^{tf(t,a)-1} \times e^{-(\mu-\lambda)\times l(a)})^{-1} \times \sqrt{idf(t)} \quad (1)$$

2.2.2 Similarity metrics based on annotations over title and abstract

Using the same input as that of the baseline, we semantically annotated title and abstract for all articles in our test dataset; see Fig. 3. A semantic annotation consists of a term associated with a concept coined in a controlled vocabulary. We performed the semantic annotation with the Concept Mapping Annotator (CMA) [31]. CMA aims to automatically identify biological entities by associating expressions in the text with entries in a given controlled vocabulary, i.e., lexicon. For this work, CMA was configured to deal with the UMLS® Meta-Thesaurus (2012AA release). More specifically, a lexicon was extracted from the MRCONSO file by cleaning the Meta-thesaurus entries and by rejecting those not appearing in PMC-OA. Additionally, we enriched this lexicon with words that appear in the concept definitions but have no entry in the Meta-thesaurus, i.e., they do not have any Concept Unique Identifier (CUI) associated. For each of these words, a CUI was automatically generated, and the most likely semantic type according to its occurrences within UMLS was assigned. We refer to these added concepts as UMLS-derived concepts. The total number of entries in this lexicon is around 2,037,998, from which 97,286 were new concepts.

Same as MetaMap [32], CMA assigns a score to each annotation related to the similarity between the lexicon associated with each concept and the chunk of text where the annotation occurs. In this way, it is possible to select a threshold to specify the minimum level of confidence of the generated annotations. In our case, we used a low setting in order to induce high recall. CMA also provides *tf* and *idf* statistics of the annotations. We used these annotations to profile the documents in our test dataset. As in the stems case, terms in the title were accounted for twice while terms in the abstract were accounted for once.

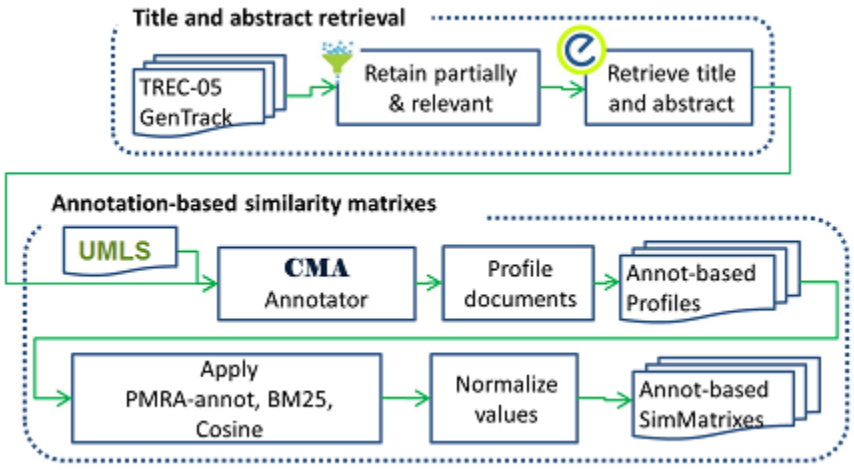

Fig. 3. Annotation-based similarity matrixes generation using title and abstract.

Using the document profiles as input, we generated the similarity matrix for three algorithms: PMRA-annotations, BM25 –see Equation 2, and Cosine –see Equation 3. We use the suffixes "stems" and "annotations" in order to distinguish whether stems or annotations were used to identify terms when using PMRA; for BM25 and Cosine only annotations were used. For BM25, we used values from 1.2 to 2.0 for the constant k, and values 0.75 and 1.0 for constant b. In Equation 2 $l(c)$ refers to the length in words of the document c, while $avgl$ is the average document length in our corpus. As we did for the baseline, all similarity values were normalized to values between [0, 1]. We obtained in this way one similarity matrix for PMRA-annotations, eighteen for BM25, and one more for Cosine.

$$bm25\_sim(c,d) = \sum_{t=1}^{N} idf(t) \times \frac{tf(t,c) \times (k+1)}{tf(t,c) + k\,(1 - b + b \times l(c)/avgl)} \tag{2}$$

$$cosine\_sim(c,d) = \cos(\Theta) = \frac{c \cdot d}{||c|| \times ||d||}$$
$$cosine\_sim(c,d) = \frac{\sum_{t=1}^{N} tf(t,c) idf(t) \cdot tf(t,d) idf(t)}{\sqrt{\sum_{t=1}^{N} tf(t,c) \cdot idf(t)^2} \cdot \sqrt{\sum_{t=1}^{N} tf(t,d) \cdot idf(t)^2}} \tag{3}$$

In order to select the best performing algorithm based on UMLS semantic annotations, we performed a correlation analysis for every annotation-based similarity matrix against our baseline as well as a precision and recall analysis for all similarity matrixes; see Fig. 4. We used the Pearson correlation algorithm, a measure of linear correlation that assigns values between [-1, 1] where -1 represents a total negative correlation, 0 represents no correlation, and 1 represents a total positive correlation. A negative correlation between variable x and y indicates that while x values increase, y values decrease; a positive correlation indicates that x values increase as y values increase. We calculated correlations for the whole corpus as well as correlations depending on the topic.

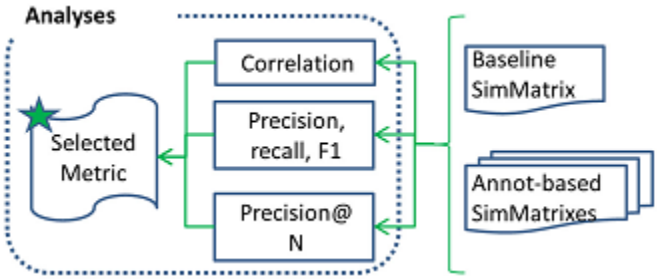

Fig. 4. Correlation and precision and recall analyses process. The star is used to indicate a final result from our method.

We also performed analyses based on precision and recall for all similarity matrixes, including our baseline. In order to assess the precision and recall, we used the topics from the TREC-2005 articles to define a gold standard. Articles belonging to a topic and categorized as relevant or partially relevant in TREC-2005 were considered as the relevant documents set for precision and recall metrics. We first separated the articles according to the topics for which they were categorized as relevant or partially relevant in TREC2005. Therefore, for a topic with *T* articles, we had similarity matrixes of *T* X 4240; only *T* of the 4240 articles were known to be relevant or partially relevant in TREC-2005. From our gold standard with respect to precision and recall, it follows that precisely those *T* articles should have been scored with the highest values of similarity. Any article in our test dataset not belonging to a particular topic should have been scored with a low similarity metric.

For each article in a topic, we calculated the precision for the highest five similarity values, then the highest fifteen, then the highest twenty-fifth, and so on until reaching the number of articles in the topic; such precision is referred to as P@N –see Equation 4, with N between [5, *T*]. P@N for a particular topic and a particular similarity metric is the mean for all its articles. For the corpus, we only analyzed P@5, which was calculated as the mean of the topic means.

$$P@N = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{N} \tag{4}$$

Additionally, we carried on analyses for precision –see Equation 5, recall –see Equation 6, and F1 score –see Equation 7, for different thresholds, i.e., similarity values between [0.1, 0.9] with incremental steps of 0.1. Precision measure corresponds to the portion of retrieved articles indeed relevant or partially relevant while recall corresponds to the portion of the relevant instances –or partially relevant in our case, that are retrieved. F1 score combines precision and recall providing a weighted average of them; the best scores are closer to 1 while the worst scores are closer to 0. In the equations, the variable *th* refers to the similarity threshold above which documents are retrieved. Similar to P@N, we also separated the articles according to the topics they belong to; precision, recall, and F1 for a topic is calculated as the mean of its articles, while for the whole corpus it is calculated as the mean of the topic means.

$$precision(th) = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|} \tag{5}$$

$$recall(th) = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{relevant\ documents\}|} \tag{6}$$

$$F1(th) = 2\ \frac{precision(th) \times recall(th)}{precision(th) + recall(th)} \tag{7}$$

2.2.3 Similarity metrics based on full-text annotations

We used CMA to annotate the 62 full-text articles. Later, we used those annotations to generate the document profiles; two sets of document profiles were generated. The first set of document profiles comprised annotations only on title and abstract, just as previous annotations on title were accounted for twice. The second set contained annotations for the full-text; in this case, all terms, whether in the title, abstract, or body, were accounted for once. We used a scattered plot to analyze the differences between the two approaches; such differences were also considered in the selection of the best performing algorithm. This whole process is summarized in Fig. 5.

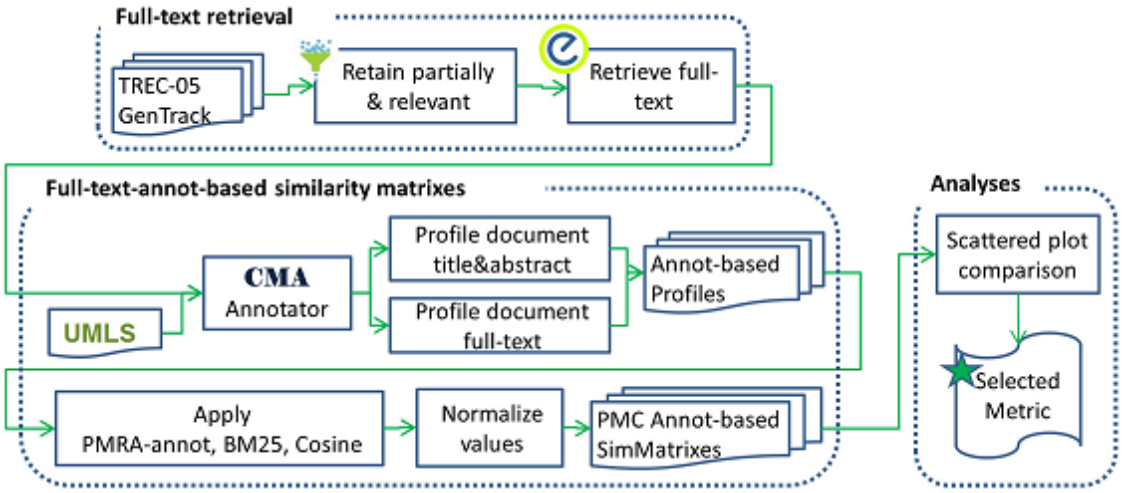

Fig. 5. Similarity matrixes for full-text articles and scattered plot analysis. The star is used to indicate a final result from our method.

3 RESULTS

Here we present our results. Document profiles were generated for stems and annotations on title-and-abstract for the test dataset, i.e., 4240 articles, as well as for annotations on title-and-abstract and full-text for the full-text test dataset, i.e., 62 articles. For the similarity matrixes obtained by applying the analyzed similarity algorithms on title-and-abstract, we present analyses based on correlation, precision, recall, and F1 score. For the similarity matrixes obtained from full-text articles, we present analyses based on dispersion and common statistical values such as mean, median, and standard deviation.

3.1 Document Profiles

3.1.1 Profiles from title and abstract

We found a total of 13,157 stems and 17,487 semantic annotations corresponding to UMLS concepts in titles and abstracts for articles in our test dataset. For both stems and semantic annotations, the term with a highest coverage was "gene" –UMLS concept C0017337; coverage here refers to the number of articles where the term occurred. However, when working with stems, "gene" was found in about 56% of the articles while it was only found in about 38% of the articles when working with semantic annotations. Some other coincidences in terminology but with differences in the article coverage are, for instance, "proteins" occurred in 49% of the articles in stems and 26% in annotations, "cell" occurred in 47% in stems and 33% in annotations, and "studi/study" occurred in 37% in stems and 30% in annotations. These coverage discrepancies are produced due to multi-word concepts including these words. Some coincidences in both terminology and percentages are also present, for instance "human/homo sapiens", occurring in 27% in stems and 28% in annotations, or "role/social role", occurring 28% in stems and 27% in annotations. We will analyze these results further in the Discussion section. The coverage for stems and semantic annotations can be found in Appendix B.

3.1.2 Profiles from full-text

For those articles with full content available and retrieved, a total of 62, the differences between the profiles when annotating only title-and-abstract versus full-text are wider. We found a total of 1,419 UMLS concepts only in title and abstracts while 6,023 were found when annotating the whole content. Only 12 concepts were found in more than 20% of the articles for the title-and-abstract case; many more were found for the full-text case –a total of 308 concepts. With a coverage above 50% of the articles, only one concept was found for the title and abstracts while 22 were found in the full content. In both cases, most common concepts are UMLS-derived concepts (i.e., concepts

that are mentioned but not defined in UMLS). The coverage for UMLS concepts in title-and-abstract as well as full-text can be found in Appendix C.

3.2 Analyses for document profiles obtained from title and abstract

3.2.1 Correlation analyses

PMRA-annotations was the similarity algorithm that correlated best to PMRA-stems with a correlation mean of 0.91537. In Fig. 6, we present the Pearson's correlation results for the twenty similarity matrixes analyzed; this correlation includes all topics in our corpus. Correlation for BM25 can be split in two groups: a first group with correlations between averages of 0.03 and 0.06 corresponding to b=1.0, and a second group with correlations between averages of 0.003 and 0.02 corresponding to b=0.75. Cosine similarity correlation average was 0.15, slightly higher than that of BM25 but still much lower than that of PMRA.

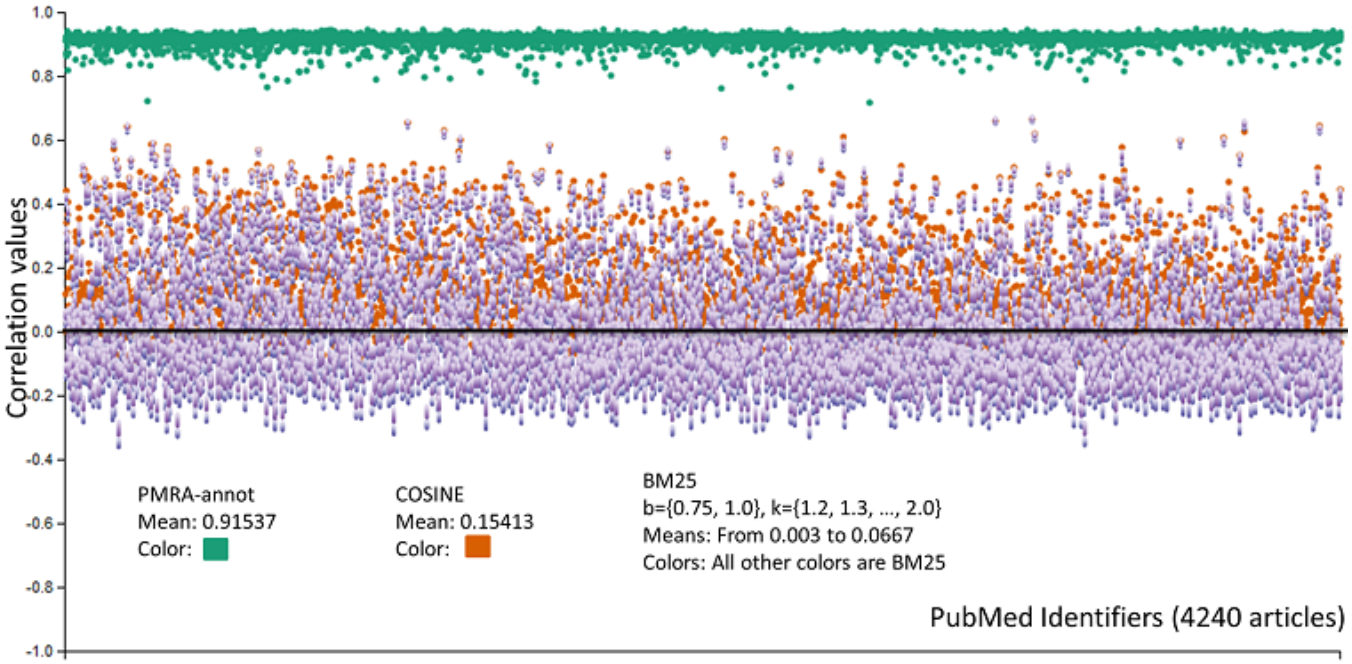

Fig. 6. Pearson's correlation for the annotation-based similarity matrixes against PMRA-stems. Axis X corresponds to the 4240 articles in the test dataset, while axis Y corresponds to the correlation values. Means per algorithm are also provided.

Correlation varies from topic to topic, although PMRA-annotations is consistently the best correlated metric regardless of the topic. In order to illustrate how correlation varies, we present three thumbnails in Fig. 7 corresponding to topic 117 with 653 articles providing information about the role of the gene Apolipoprotein E (ApoE) in the disease Alzheimer's Disease, topic 120 with 331 articles providing information on the role of the gene nucleoside diphosphate kinase (NM23) in the process of tumor progression, and topic 108 with 191 articles describing the procedure or methods for identifying in vivo protein-protein interactions in time and space in the living cell. While for some topics BM25 and Cosine exhibit a high correlation, e.g., topic 117, for others the correlation is poor, e.g., topic 108. Both global and by-topic correlations are available at [33], where further instructions on how to use the visualization script can be found.

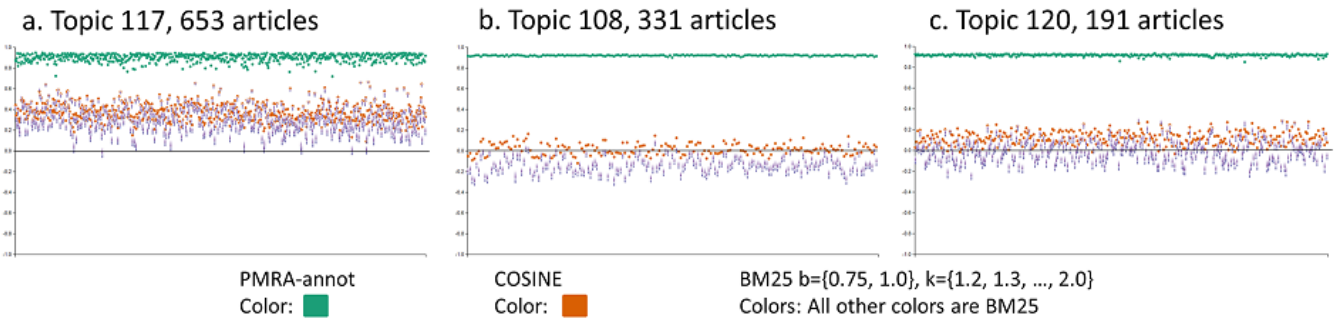

Fig. 7. Pearson's correlation for three of the eleven topics with more than 100 relevant or partially relevant articles.

3.2.2 Precision, recall, and F1 analyses

Our initial analysis regarding precision aims to mimic the related article search feature in PubMed. While reading a document, PubMed interface presents on the right side the top five related citations, i.e., articles, in PubMed. Therefore, we analyzed the top five results for each article in the twenty similarity matrixes plus the baseline similarity matrix, i.e. the precision at five or P@5. The global precision per algorithm was obtained as the mean of precisions across all articles for that algorithm; results are presented in Fig. 8 where BM25 variations have been grouped as they exhibited similar values.

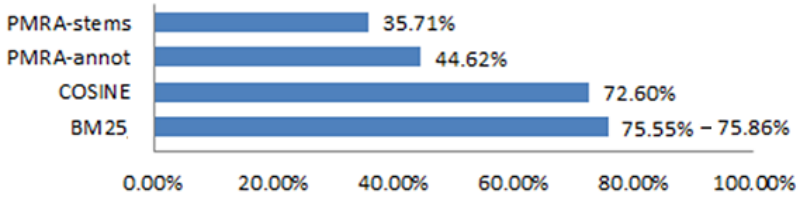

Fig. 8. P@5, top five precision for the analyzed algorithms.

We also analyzed P@N for those eleven topics with more than 100 relevant and partially relevant articles (from now on referred to as selected topics; see Fig. 9). In all cases, BM25 family got the highest precision for N between [5, T] where T is the total number of relevant and partially relevant articles in the topic. Similar to findings at a global level, PMRA-annotations surpassed PMRA-stems while Cosine tendency was closer to BM25.

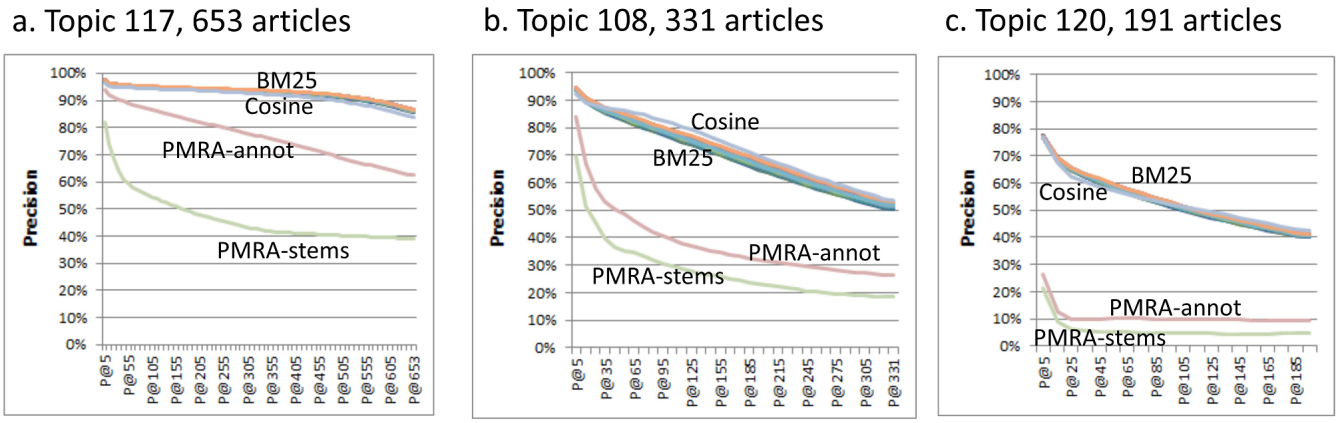

Fig. 9. P@N for three of the eleven topics with more than 100 relevant or partially relevant articles.

In order to assess which similarity metric performs best for high similarity values, we carried on further analysis using thresholds between [0.1, 0.9] with incremental steps of 0.1. Such analyses included precision, recall, and F1 score at a global level, i.e., comprising all topics, and for each selected topic; Fig. 10 presents results at a global level. Global values for precision, recall, and F1 per similarity algo-

rithm were obtained as the mean of all values for the corresponding metric across all articles for that algorithm. BM25 family displayed low precisions for thresholds below 0.2 but quickly increased for thresholds above 0.3, reaching values between 0.8 and 1, i.e., 80% and 100%. PMRA-annotations and PMRA-stems presented similar tendencies regarding each other, displaying low precisions for thresholds below 0.5 with an exponential-like increment for thresholds above 0.6. Cosine behavior depicted a soft diagonal from 0.4; the higher the threshold, the higher the precision.

The recall showed complementary tendencies for all cases. BM25 family displayed a recall of 0.6 for the threshold 0.1, it quickly decreased to 0.2 for thresholds above 0.2. It steadied at 0.1 for thresholds above 0.4. PMRA-stems and PMRA-annotations curves looked almost the same; a recall between 0.9 and 1.0 was displayed for thresholds between 0.1 and 0.3. The recall for PMRA fell 1.5 points per threshold from 0.3 to 0.7, reaching the lowest value of 0.1 for thresholds of 0.8 and 0.9. Cosine behavior depicted a soft descending diagonal from 0.7 to 0.1 for thresholds between 0.1 and 0.7; it then remained 0.1 for thresholds 0.8 and 0.9.

F1 combines precision and recall; thus, we focused on this score, which was usually under 50% except for some TREC topics such as topic 117 –proving information about the role of the gene Apolipoprotein E (ApoE) in the disease Alzheimer's Disease. BM25 family presented better F1 values for thresholds below 0.3, quickly decreasing for thresholds between 0.3 and 0.4, then steadily continuing to lower. The best F1 results for Cosine Similarity were located for thresholds between 0.2 and 0.5. PMRA-annotations and PMRA-stems exhibited similar tendencies, with the highest F1 values for thresholds above 0.5. Both presented lower F1 values for lower thresholds, softly increasing from 0.5 to 0.7 and then softly slowing down; although such slowdown was faster for topics such as 117.

Results indicate that BM25 and Cosine are good metrics to retrieve a few relevant documents –although not necessarily similar to the query document, whereas PMRA-stems and PMRA-annotations promote highly similar documents to the query document. Somehow the metrics are complementary: if the main interest is in diversity, BM25 and Cosine might be better than PMRA; conversely, if the interest is in finding out highly similar documents PMRA would be a better choice. In addition, PMRA-annotations metric performs better than PMRA-stems in all metrics. PMRA-stems and PMRA-annotations also exhibit better recall than BM25 and Cosine; thus, PMRA has better coverage for relevant articles than the other two. PMRA also shows a lower concentration of documents in the corresponding metric space, as it occurs with BM25. Results also show that BM25 similarity values are mainly concentrated around low values; see Fig. 10 and Fig. 11.

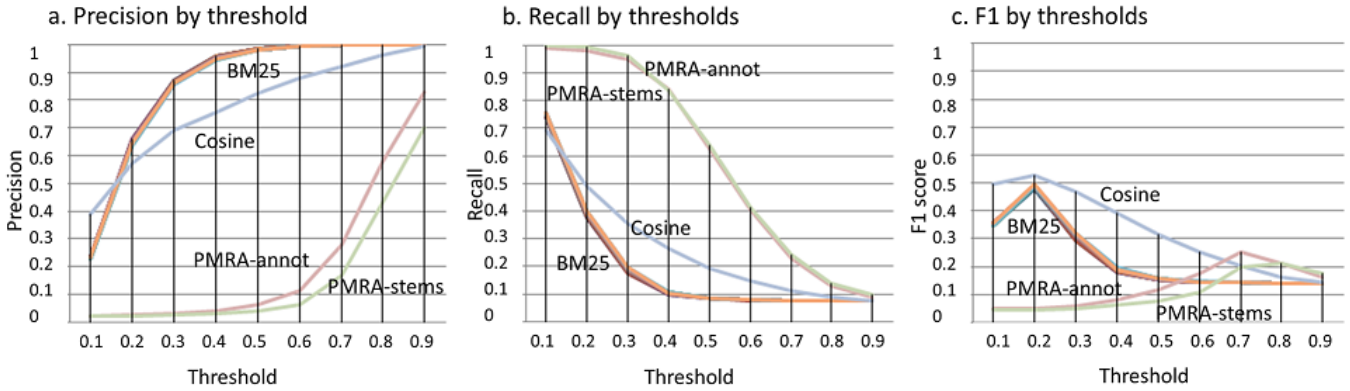

Fig. 10. Global precision, recall, and F1 for thresholds from [0.1, 0.9]

We also calculated the confidence intervals for precision and recall for thresholds from [0.1, 0.9]; we established a confidence level of 90%. With regard to the precision values, the confidence interval for PMRA-stems and PMRA-annotations increases from the threshold 0.5, exhibiting a margin of error of about 0.1. The Cosine has a lower margin of error on the edges, i.e., for low and high thresholds. For the intermediate values the margin of error is close to 0.08. BM25 exhibits the lowest margin of error, about 0.05. As for recall values, the margin of error decreases in all cases to approximately 0.04. For PMRA-stems, PMRA-annotations, and BM25, the margin of error is lower around the edges, while for Cosine this is true only for the upper edge, i.e., higher thresholds.

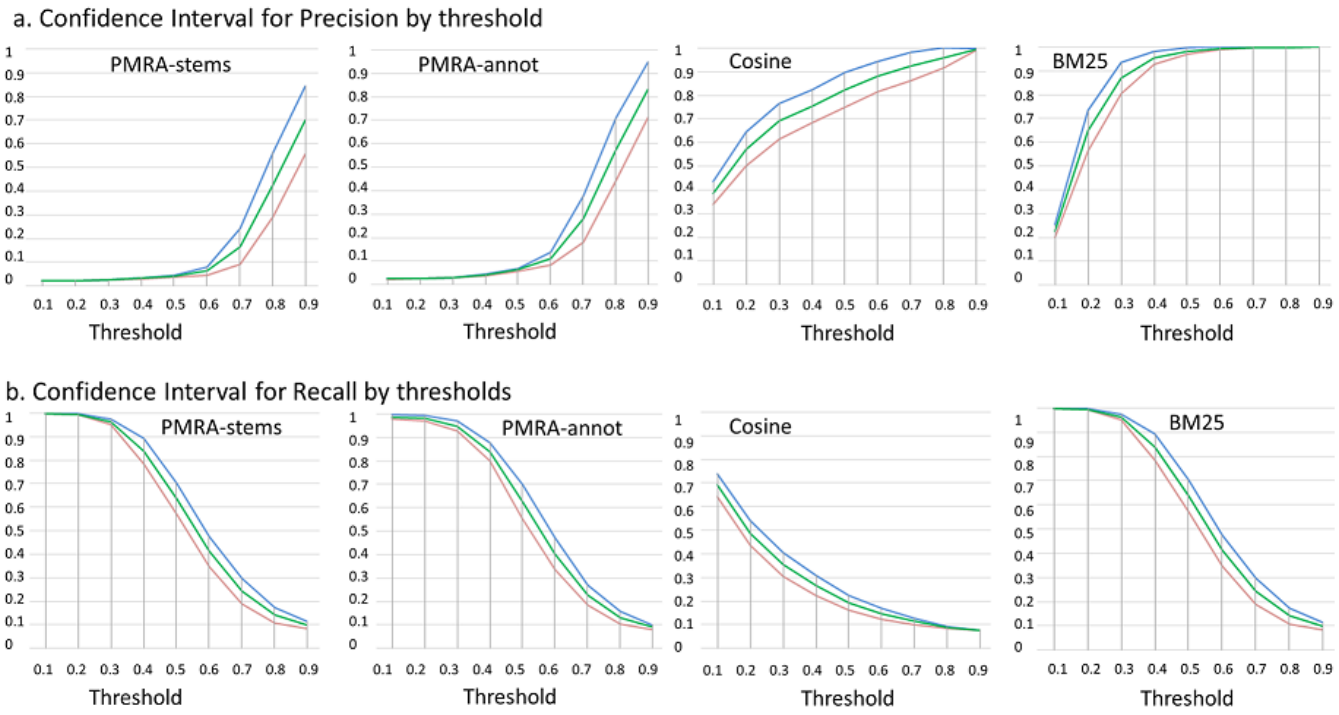

Fig. 11. Confidence intervals for precision and recall for thresholds from [0.1, 0.9]

3.3 Analyses for document profiles obtained from full-text

3.3.1 Scattered plot analyses for full-text versus title-and-abstract

Regardless of the metric, PMRA, BM25, or Cosine, similarities vary for title-and-abstract versus full-content based annotations. However, based on our findings, it is not clear how they exactly differ one from another, i.e., whether or not the differences follow a pattern or formula. PMRA presents more dispersion than BM25 and Cosine; nonetheless, for all of them, it was possible to find articles for which the similarity was close between title-and-abstract and full-content. In some cases, the similarity for full-content was higher than that for title-and-abstract, but the opposite case was also present. Fig. 12 shows the scattered plots for PMID:12537560 “Bayesian analysis of gene expression levels: statistical quantification of relative mRNA level across multiple strains or treatments” [34]; similarities are calculated given an interest in PMID:12049663 “Bayesian analysis of gene expression levels: statistical quantification of relative mRNA level across multiple strains or treatments” [35]. These two articles were randomly selected; both of them belong to the same TREC topic whose information need was defined as “Describe the procedure or methods for normalization procedures that are used for microarray data”. This TREC topic comprises 19 of the 62 full-text articles analyzed and corresponds to the topic with more articles from the full-text dataset. Both articles were published in the journal Genome Biology in the same year –2002, but in different volumes; they do not cite each other. Scattered plots for all the 62 PMC-OA analyzed articles are available at a GitHub repository [33], where instructions on how to use the visualization script may also be found.

Cosine shows the highest dispersion level, disregarding whether the articles belong to the same topic or not. As previously mentioned, BM25 metric shows a high concentration around low similarity values. Regarding PMRA-annotations on full-text, the articles belonging to a TREC topic different from the query article, i.e., PMID:12049663 in our example, are usually clustered on values around 0.4. The similarity values for articles belonging to the same TREC topic do not show a clear behavior, sometimes are in the diagonal but some other times are more dispersed. PMRA-annotations method on full-text shows the best correlation for articles within the same TREC topic. The tendencies described in this paragraph apply for all the 62 full-text articles.

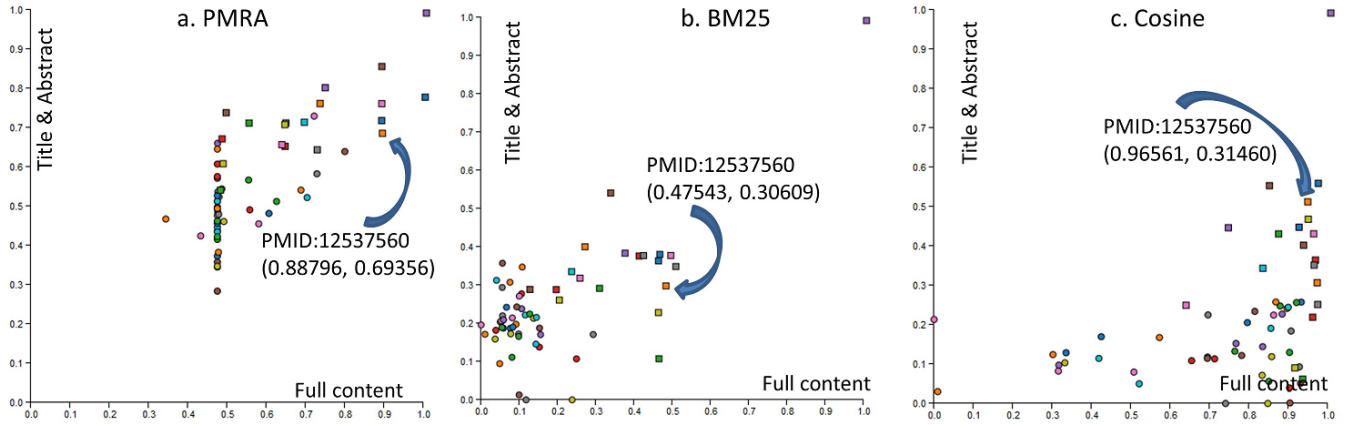

Fig. 12. Title-and-abstract versus full-text annotations for PMID:12537560 given an interest in PMID:12049663. Squares correspond to articles within the same TREC topic as the selected article, while circles represent articles from any other topic.

Furthermore, we calculated the mean, median, and standard deviation for the similarities calculated on title-and-abstract as well as on full-text for the considered algorithms; see Fig. 13. As similarity values were similar for all the considered parameters for BM25, we performed the analyses using the default parameter settings ($k = 1.2, b = 0.75$). For all three algorithms, both mean and median were close to each other suggesting a symmetric distribution. The standard deviation was between [0.10, 0.25], suggesting values not too far away from the mean. PMRA-annotations and Cosine similarity values exhibit a higher variation for title-and-abstract versus full-text than BM25. In BM25 the mean for title-and-abstract is quite similar to the mean for full-text. Values calculated with PMRA-annotations are greater for title-and-abstract than for full-text. The opposite happens for values calculated with Cosine, values for title-and-abstract are usually above those for full-text.

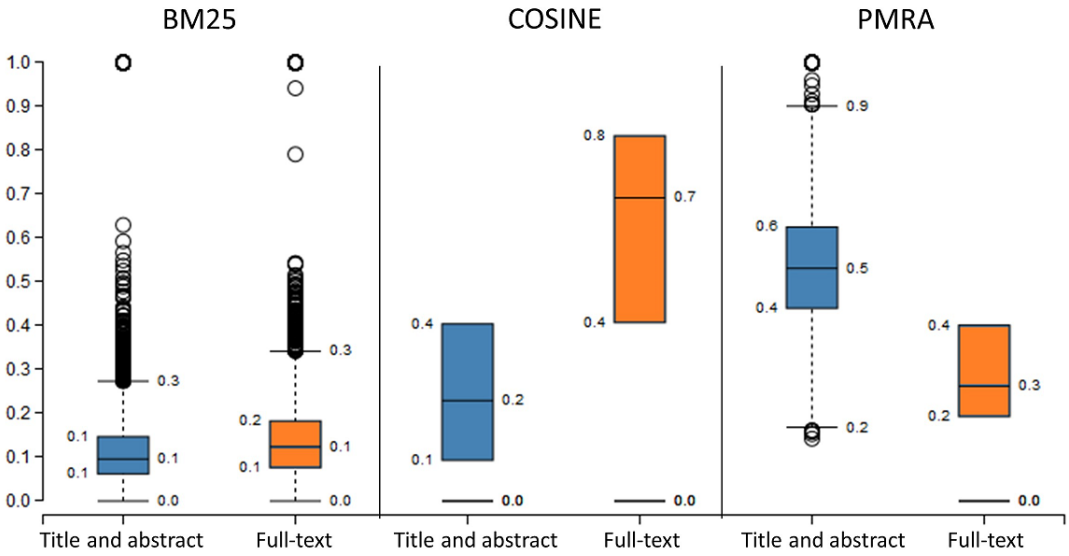

Fig. 13. Box-plots for mean, median, and standard deviation calculated for the similarity values according to title-and-abstract (blue) and full-text (green) for our full-text test dataset, i.e., 62 articles. Box-plots show outliers, maximum, third quartile, median, first quartile, and minimum values.

4 DISCUSSION

Citing articles is one of the most direct ways to interlink and relate articles to one another [5]; articles sharing citations are considered similar to some extent. In addition to citations, text-based approaches such as *tf-idf* and LSA have also been used to find out how similar one article is to another [10]. Clustering approaches have also been explored. For instance, Lewis [36] groups articles by using a keyword-

based method; initial results are ranked and ordered by a sentence-alignment algorithm. In the same vein, McSyBi [37] groups articles according to a set of topics gathered from information at title and abstract level. Different from Lewis, McSyBi enables the use of MeSH terms or UMLS semantic types in order to modify the clusters; thus, users can analyze the data from different perspectives. Different from the previously mentioned approaches, we are working with a semantically annotated dataset in contrast to plain text articles. Similar to McSyBi, we use UMLS concepts in order to calculate similarity between articles.

Using semantic annotations in scientific publications opens up wider possibilities. While working with our test dataset, we found that for both stems and annotations, the term "gene" was the most common in titles and abstracts; the stem "express" was found as the fourth most common stem. The word "gene" is part of the UMLs concept C0017262, i.e., "gene expression"; therefore, while all occurrences of "gene expression" in a text would have been associated with two different stems "gene" and "express", when working with annotations, "gene expression" would be considered a single term. In the biomedical domain, "gene expression" results are more significant than those for "gene" and "expression" separately. In fact, the word "expression" not related to genes may not be interesting. Another interesting finding involves the stem "human" and the concept "homo sapiens". When dealing with annotations, both expressions are associated with "homo sapiens", which is not captured with stems.

The association between terms and articles using controlled vocabularies instead of stems has been already explored [38-40]. In fact, the version of PMRA implemented in the PubMed repository, used to identify the articles related to the one currently being read, includes MeSH terms. Other efforts identifying concepts associated with controlled vocabularies in scientific publications are the Resource Index [38] and Europe PubMed Central [39]. The Resource Index is an ontology-based index covering more than twenty biomedical resources. Text contained in title and abstract is annotated with concepts coined in ontologies; such annotation is used to improve the search and retrieval. Europe PubMed Central is based on PMC; however, it offers not only PubMed abstracts but also full text articles, patent abstracts, clinical guidelines, and biomedical research grants. It uses text-mining technologies, particularly Whatizit [40], in order to identify and highlight gene names, organisms, and diseases in abstracts; searching by those concepts is also possible. To our knowledge, the identified concepts are not currently being used by either of these efforts in order to find related articles, nor are they applied to full-text collections.

In this manuscript, we have focused on similarity across PubMed articles; such similarity has been studied before. PMRA experimentation [11] defines the related document search problem as the retrieval of documents that a user may also want to examine given a known interest to a certain document. Documents of interest are similar in terms of topics or concepts; in PMRA, concepts are word stems taken from the title and abstract. Similar to PMRA experiments, we also use the TREC-2005 Genomic Track test collection. While PMRA focuses on document ranking, our concern is related to similarity metrics with values between [0, 1.0]. Thus, we transform PMRA ranking metric into a similarity metric by normalizing the obtained rank values; i.e., the ranking value between a document and itself turns to a similarity value of 1.0 while the rest of the values are calculated by cross-multiplication.

Similar to PMRA experimentation, we analyze P@5; however, the precision is applied to different datasets. In the experiments reported at [11], BM25 with default parameter settings ($k = 1.2, b = 0.75$) is used to retrieve related articles; then PMRA is used to re-rank the top 100 documents. The precision is calculated on the top 5 documents of those 100. In our experiments, we use all the related and partially related articles from the TREC collection, regardless the TREC topic. We calculate the similarity between any document pair in that TREC subset; i.e., our precision calculation takes into account 4,230 documents. Different from the results reported at [11], our results show a better performance regarding P@5 for BM25, independently of the parameter settings. In addition to the difference between the datasets used to calculate P@5, other differences could come from the method used to extract the stems from titles and abstracts; however, this point is not clear as the extraction of stems is not detailed by PMRA [11]. PMRA with annotations, i.e., PMRA-annotations, shows an improvement for P@5 regarding PMRA-stems; however, it is still about 30% lower than BM25 and Cosine metrics.

PMRA, BM25, LSA, self-organizing mapping, and topic modeling have been analyzed and compared over a set of two million biomedical publications in [10]. In this study, the authors use the ordered list of related articles provided by PubMed and normalized the results to a similarity value between [0, 1.0] by assigning a value of 1.0 to the first on the list, 0.98 to the second one, 0.96 to the third one, and so on;

that is how PMRA similarity values are calculated. In contrast to this work, we use the PMRA formula as originally defined at [11]. Moreover, we do not aim at massively clustering documents with these metrics, but at analyzing their behavior over the TREC topics. Clusters in [10] are based on the grant-to-articles linkages indexed in MEDLINE reporter, which is based on a previous study of the authors [41]. Boyack and colleagues found that PMRA and BM25 applied to title-and-abstract had the highest precisions; precisions for PMRA were slightly higher than those for BM25. Using TREC topics as pre-defined clusters for our test collection. We found that precisions calculated for BM25 and Cosine were much higher than those calculated for PMRA-stems and PMRA annotations methods. The maximum F1 for Boyack's clusters was found around a recall of 0.6. In our case F1 showed two tendencies, (i) a maximum value for recalls around 0.6 for BM25 and Cosine, and (ii) a maximum F1 value for recalls around 0.3 for PMRA-stems and PMRA-annotations. A fair comparison between the two studies is hard to achieve as we do not calculate clusters, and, even using TREC topics as clusters, articles are grouped based on different criteria.

In addition to P@5, we also analyzed precision and recall for thresholds from [0.1, 0.9]. Taking into account solely the precision, BM25 and Cosine perform better than PMRA-stems and PMRA-annotations; however, PMRA-annotations performs better than PMRA-stems. The picture changed when involving the recall. The F1 score gives us a combined picture of these two values. When looking for similar articles, users are more interested in those with the higher similarity values. PubMed can return more than 100 related articles, but unless a crawler is used, human users will mainly focus on the top results; in fact, PubMed initially displays only the top 5. Therefore, we are more interested in those similarity metrics that perform better for higher similarity values. F1 score shows that both PMRA-stems and PMRA-annotations perform better than BM25 and Cosine for similarities above a threshold of 0.6.

Similar to other authors [42, 43], we are interested in information extraction from full-text documents. Thus, beyond global and topic-based analysis, we also explored differences between title-and-abstract and full-text based similarities. Shah and colleagues [42] analyze full-text articles in order to determine where the keywords are. Their findings show that abstracts contain the best ratio of keywords per total words; however, the consideration of other sections is valuable in order to go deeper into biologically relevant data. Additional sections can contribute keywords not present in the abstract as the distribution along the narrative is heterogeneous. In fact, other sections contribute with much more relevant information regarding, for instance, gene names, anatomical terms, and organism names. Although we have not considered differences across sections, like Shah and colleagues, we have found that the information provided by the narrative beyond title and abstract does have an impact on further analysis.

With a different purpose, uncovering potential duplicate citations, Sun and colleagues [44] have also explored similarity for full-text articles. Different from the present work, similarity was measured in terms of words rather than stems or concepts found in the text. The text similarity algorithm eTBLAST was used. Findings show that articles that are highly similar based on words found in the abstract are likely to be highly similar based on words found in the full-text as well. Their experiments exhibited a high sensitivity (99%), i.e., recall, and a low specificity (20%). However, in order to truly uncover duplications, a comprehensive analysis of the full text is still necessary. Similarly, we have found that when using PMRA-annotations, similarity values could be alike regardless of using title-and-abstract or full-text, but they could also be different. The scattered plot analyses show that for about 50% of the 62 full-text articles, the similarity values based on title-and-abstract are close to 0 while they are above 0.5 when based on full-text.

In a previous work [43], we had already explored relations across full-text articles. Eleven articles randomly selected from a Biotea repository were analyzed regarding MeSH and UMLS concepts found along their full content. Biotea [45] provides MeSH annotations, among others, for articles in PMC-OA; annotations in Biotea are delivered as statements following the model proposed by the Resource Description Framework (RDF) [46]. The CMA annotator [31] was used in order to extract UMLS concepts. Our findings then showed that it is possible to find connections between articles beyond those reported as related articles in PubMed; however, new connections were only found for those articles without MeSH terms reported in PubMed. This current work is a step forward; not only have we worked with a wider collection, but we have also reached a better understanding of full-text based annotations and their impact on similarity between articles.

Similarity values calculated with PMRA-annotations fluctuate between [0.4, 0.7] for concepts in title-and-abstract and between [0.1, 0.7] for concepts in full-text. Similarity values based on title-and-abstract are consistently greater than those based on full-text. The difference fluctuates around 0.3. Thus, an article with a similarity value of 0.9 based on title-and-abstract could report a similarity value of 0.6 based on full-text. Similarity values calculated with the Cosine metric show greater variation range, from ca. 0 to 0.7 either based on title-and-abstract or full-text. Unlike PMRA-annotations, the values for title-and-abstract are usually greater than those for full-text. Variations might be significant for knowledge retrieval and knowledge discovery. For instance, in the biomedical domain, it has been reported that more linked data, associated with controlled vocabularies, could lead to the identification of novel associations [47], such as pathways associated with a particular disease or drug or the evaluation of hypotheses against experimental data [48].

Based on our results, we find that PMRA-annotations is the algorithm that best adjusts to the UMLS based on annotations. The PMRA-annotations present the best correlation to PMRA-stems independently of the TREC topic. The F1 score for PMRA-annotations consistently increases as the similarity values do; therefore, PMRA-annotations metric works better than BM25 and cosine for high similarity values. In addition to analyses carried over the whole test dataset collection, i.e., 4240 articles, we also performed analyses at a TREC topic level. Once again, PMRA-annotations shows the highest correlations while BM25 and Cosine show more variability depending on the topic. The PMRA-annotations are consistent for F1 scores at TREC topic level. PMRA-annotations metric also presents a significant variation between title-and-abstract versus full-text based similarities; however it is not too disperse as Cosine nor it concentrates similarities on low values as BM25. PMRA-annotations based on full-text also exhibits a better correlation for articles within the same TREC topic than Cosine or BM25. As mentioned before, full-text analysis opens up further analysis.

Finally, we have identified some limitations of our approach. We have worked with a well-known collection of articles, TREC-2005. This collection is restricted to genomics; thus, it is unknown whether the study results are broadly generalizable. However, due to the detailed description that we have included for the method we followed, it should be possible to reproduce the experiment with a different corpus. The lack of manually curated corpora in the biomedical domain, including a significant number of full-text articles, makes it difficult to evaluate the validity of similarity metrics. Also, PMRA has been tested within the scope of PubMed; it is not well understood how this metric could behave for articles in other domains. The use of an annotator such as CMA also imposes some constrains; different results might be obtained with other annotators such as Whatizit and the NCBO Annotator. By the same token, results would likely vary if a vocabulary different from UMLS is used. In addition, getting full-text could be challenging, and processing it is computationally expensive. However, within the digital publication, the importance of processing full-text has already been acknowledged. For instance, Europe PubMed Central offers an advance search feature enabling users to specify a section type. The exceptions to copyright in the United Kingdom for research purposes [49] also illustrate the value of full content. These exceptions allow researchers to copy material, e.g., from scientific publications, without infringing the copyright; such material can then be processed by text and data mining tools.

5 CONCLUSIONS AND FUTURE WORK

We have presented a benchmark on different similarity metrics applied to scientific articles. Such metrics have been applied to stems and concepts found either in title and abstract or the full content. We have worked with the TREC-05 Genomics track test collection where articles have been manually grouped regarding relevance judgments. We narrowed the collection to only relevant and partially relevant articles for all 50 relevance judgments. Our baseline consisted of the similarity for such articles according to a normalized version of the PMRA algorithm as reported in the literature [11].

Similar to previous works [10, 11], our findings show that PMRA is the similarity metric that performs best for PubMed articles, not only when applied to stems or words but also when applied to semantic annotations –particularly UMLS-based annotations. Our results show that depending on the aim, e.g., similarity, coverage, or diversity, different metrics could be needed. Furthermore, a better integration

of BM25, Cosine and PMRA might be useful. In the past, an integration between BM25 and PMRA has already been studied [11]. There, PMRA was used to re-rank articles for which BM25 has been already applied.

Using semantic annotations is a step forward when it comes to similarity between articles. Stems are simpler and straightforward but an in-depth analysis requires the precision provided by controlled vocabularies. When working with annotations, PMRA shows a high precision for high similarity values, i.e., those corresponding to the top articles in the PubMed related articles list. Thus, rather than calculating similarities from scratch, it would be possible to take the PubMed related articles list and enrich it with concept-based connections by applying PMRA-annotations metric. Our future work will go in that direction.

Additionally, similarity values based on terms found solely in title-and-abstract versus those found in full-text vary regardless the similarity metric used; a more significant variation was found for Cosine and PMRA-annotations. The exact nature, i.e., dispersion formula, of such variation is beyond the scope of this article; however, our findings show that there is indeed a significant difference. Although working with full-text articles consumes more resources, having concepts identified along the whole article opens up new and interesting possibilities. For instance, it becomes possible to analyze the similarity between a pair of documents from different perspectives such us how similar they are regarding a particular UMLS group or a particular section. We intend to use PMRA based on annotations for full-text articles in order to better understand the liaison between articles. We will focus on semantically linking articles within PMC-OA. By doing so, we aim at the long-term goal to contribute to recommendation systems.

A concept-based approach also makes it possible to explore and analyze documents from a semantic perspective. Concepts are related, and relations range from the common ones, e.g., *is-a* and *part-of*, to the more complex and domain specific ones, e.g., *transcripts-to* or *inhibits*. These relations could then be harnessed so that ancestors are included in the similarity metric. For instance, a concept *A*, not explicitly present in the text, could be taken into account for the similarity score if a descendant concept *D* is found in the text. The *tf-idf* of *A* could be smoothed in order to reduce the impact of this ancestor expansion. Special attention should be paid to ancestors originally present in the text as well as to generic concepts. Such concepts could introduce noise to the similarity score. Our preliminary work in this regard, not discussed in this manuscript, shows that including ancestors does not have a significant impact in the final similarity score.

Another issue to be studied is the weighting schema for multi-word annotations. This subject has been reported in the literature by multiple authors, e.g., Damerau [50], Frantzi, Anadiadou, & Mima [51], and Deane [52]; a comparative evaluation on term recognition algorithms can be found at [53]. Part of-speech tagging and weighting schemas for multi-word terms could improve the identification of chunk boundaries such as noun phrases in CMA; thus it could also improve the coverage of concepts associated to multi-word terms.

Finally, approaches such as the one presented in this manuscript could benefit from corpora that include a significant number of full-text articles. To date, PMC-OA provides about 1 million articles, i.e., less than 5% of the 24 million provided by PubMed and about 30% of the 3.5 million provided by PubMed Central. The use of test collections such as TREC-2005 is a common approach used in Information Retrieval; manually curated collections are usually preferred as gold standards. From 4240 relevant and partially relevant articles in TREC-2005, only 62, i.e., less than 2%, correspond to full-text articles freely available in PMC-OA. The generation of manually curated corpora including full-text remains a challenge.

A. APPENDIX - TREC05 TOPICS

The following table, Table A.1, presents a summary of the TREC05 topics.

Table A.1. Information needs in TREC-05 Genomics track with more than 100 relevant and partially relevant articles with available information for title and abstract using the NCBI's *e-Utils* web services. Topic numbers are taken from the TREC collection while the number of articles corresponds to those retrieved from NCBI's *e-Utils* web services. Please be aware that the last column is not a sum of the 4th and 5th columns as it was not possible to retrieve title and abstract for all articles in TREC-2005.

Topic code	Description	Non-relevant articles	Partially relevant articles	Relevant articles	Relevant and partially relevant articles with data for title and abstract
117	Role of the gene Apolipoprotein E (ApoE) in the disease Alzheimer's Disease	385	182	527	653

146	Mutations of hypocretin receptor 2 and its/their role in narcolepsy	388	67	370	421
114	Role of the gene APC (adenomatous polyposis coli) in the disease Colon Cancer	375	169	210	346
120	Role of the gene nucleoside diphosphate kinase (NM23) in the process of tumor progression	182	122	223	331
126	Role of the gene P53 in the process of apoptosis	1013	117	190	307
142	Sonic hedgehog mutations and its/their role in developmental disorders	257	120	151	263
108	Procedure or methods for identifying in vivo protein-protein interactions in time and space in the living cell	889	127	76	191
107	Procedure or methods for normalization procedures that are used for microarray data	294	114	76	189
111	Role of the gene PRNP in the disease Mad Cow Disease	473	93	109	185
109	Procedure or methods for fluorogenic 5'-nuclease assay	210	14	165	175
106	Procedure or methods for chromatin IP (Immuno Precipitations) to isolate proteins that are bound to DNA in order to precipitate the proteins out of the DNA	1061	125	44	158

B. APPENDIX - COVERAGE FOR PROFILES FROM TITLE AND ABSTRACT

The coverage for stems is shown in Table B.1 while the coverage for semantic annotations is shown in Table B.2.

Table B.1. Article coverage for stems found in title-and-abstract for articles within the full-text test dataset. Only the first 10 stems for each interval are shown, whenever more terms are available they are indicated as […]. Intervals where no concept was found are omitted. No concept exhibited a coverage above 60%; thus, such intervals are omitted.

Stems	**Number of stems (total = 13157)**	**Article coverage**
Gene	1	[55%, 60%)
Protein / cell	2	[45%, 50%)
Express	1	[40%, 45%)
Not / result / studi / diseas / suggest	5	[35%, 40%)
Associ / mutant	2	[30%, 35%)
Activ / role / human / develop / increas / factor / function	7	[25%, 30%)
Alzheim / two / differ / type / specif / famili / effect / patient / level / analysi […]	12	[20%, 25%)
Involv / induc / data / cancer / indic / report / dna / demonstr / allel / beta […]	28	[15%, 20%)
Signific / apolipoprotein / apo / receptor / mediat / more / process / mutant / transcript / examin […]	72	[10%, 15%)
Chang / test / progress / reveal / presenilin / reduc / site / presenc / time / cellular […]	170	[5%, 10%)
Part / approxim / four / essenti / us / damag / sever / reaction / togeth / transgen […]	12857	[0%, 5%)

Table B.2. Article coverage for UMLS concepts found in title-and-abstract for articles within the title-and-abstract test dataset. Only the first 10 terms for each interval are shown, whenever more terms are available they are indicated as […].. No concepts exhibited a coverage above 40%; thus, such intervals are omitted.

UMLS concepts	**Number of concepts (total = 17487)**	**Article coverage**
Genes	1	[35%, 40%)

Gene Expression / Cells / Experimental Result / Study	4	[30%, 35%)
Homo sapiens / Social Role / Mutation / Proteins	4	[25%, 30%)
Alzheimer's Disease / Increase / Relationships / suggestion / Patients / Induce (action)	6	[20%, 25%)
Levels (qualifier value) / physiological aspects / human data / Effect / High / Family / Entity Determiner - specific / Specific qualifier value / Alleles / Biologic Development […]	15	[15%, 20%)
NOS activity (molecular function) / APOE gene / Mutant / Mediate / Apolipoprotein E / Significant / Most / Observed / Apolipoproteins E measurement (procedure) / Container status – Identified […]	48	[10%, 15%)
Age / Disease Response / Familial / Add - instruction imperative / presence / Numbers / Biological Assay / Anabolism / Cultured Cell Line / Genotype […]	152	[5%, 10%)
Biological / DICOM Derivation / Derivation / Derived value / Evaluation / Protein Overexpression / Encode (action) / Role Class - part / Hereditary / Cessation of life […]	17257	[0%, 5%)

C. APPENDIX - COVERAGE FOR PROFILES FROM FULL TEXT

In Table C.1, we show the coverage for concepts found in title-and-abstract while in Table C.2 we show concepts found in full-text for the 62 articles comprising our full-text test dataset.

Table C.1. Article coverage for UMLS concepts found in title-and-abstract for articles within the full-text test dataset. Only the first 10 terms for each interval are shown, whenever more terms are available they are indicated as […].Intervals where no concept was found are omitted. No concept exhibited a coverage above 45%; thus, such intervals are omitted.

UMLS concepts	**Number of concepts (total = 1385)**	**Article coverage**
Gene Expression	1	[40%, 45%)
Cells / Proteins / human data	3	[35%, 40%)
Genes / Social role / experimental results	3	[30%, 35%)
Study / Levels (qualifier value) / suggestion / increase / physiological aspects	5	[20%, 25%)
Induce (action) / Microarray / Homo sapiens / Mutation / Complex / complex (molecular entity) / Relationships / High / Biological Models / investigates […]	15	[15%, 20%)
Antigen-Presenting Cells / Mechanism (attribute) / Direct (qualifier) / Report (document) / Microtubules / Drug Interactions / CTNNB1 gene / beta catenin / NOS activity (molecular function) / Array […]	34	[10%, 15%)
Normalize / Patients / Epithelial cell count (procedure) / Type:Finding:Point in time:Form:Nominal / dynamic / research study / Packaging Case / expression level / Biochemical Pathway / Anabolism […]	115	[5%, 10%)
PSEN1 gene / Presenilin-1 / Signal Transduction / Fragment / Presenilins / Genotype / Staining method / Breast Carcinoma / regulation of nitric-oxide synthase activity / Gene Deletion Abnormality / Colon structure (body structure) […]	1209	[0%, 5%)

Table C.2. Article coverage for UMLS concepts found in full-text for articles within the full-text test dataset. Only the first 10 terms for each interval are shown, whenever more terms are available they are indicated as […] Intervals where no concept was found are omitted. No concept exhibited a coverage above 95%; thus, such intervals are omitted.

UMLS concepts	**Number of concepts (total = 5979)**	**Article coverage**
Gene Expression	1	[90%, 95%)
Research study / genes / specimen / human data	4	[70%, 75%)
Cells / Proteins / Patient observation / Detected (finding) / Mental concentration	5	[60%, 65%)
Generation (action) / Comparison / Analysis	3	[55%, 60%)
Tracer / Figs - dietary / Induce (action) / biology (field) / expression level / Mus / Mutant / Binding (Molecular Function)	8	[50%, 55%)

Mutation / Incubated / Wild Type / Wild Type Unspecified - zebrafish / Minute of time / Antibodies / Immunoglobulins / Drug Interactions / vision table / Intensity […]	15	[45%, 50%)
in vivo / Transcription, Genetic / Cultured Cell Line / Affect (mental function) / Body tissue / Assessed / Staining method / Correlation / Binding action / in vitro […]	22	[40%, 45%)
Experimental Result / Tissue membrane / Fragment / metaplastic cell transformation / Science of Statistics / Human body / Regulation of biological process / Folded structure / Chemical Probe / Washed […]	23	[35%, 40%)
Molecule / Localized / Manual reduction / Fluorescence / Total / Purifying / reflecting / Parameter Value / Population Parameter / Neoplasms […]	36	[30%, 35%)
Packaging Case / Most / Large / Synthesis / Overlap / Conjugated / tri-(deoxyguanylic acid-deoxycytidylic acid) / Similarity / Activation action / Biologic Development […]	65	[25%, 30%)
Cell type / Research Activities / research / Groups / Social group / Specialty Group / Role Code - Group / Primary malignant neoplasm / Family history of cancer / Patients / Family […]	127	[20%, 25%)
Investigational / What subject filter - Order / Order [PK] / Order (document) / Risk Codes - Biological / size / Phase / Embryo / sporadic / Biological Processes […]	218	[15%, 20%)
Potential / Disease Progression / Variable (uniformity) / Malignant neoplasm of large intestine / COLON CANCER (allelic variant) / Alleles / Drosophila <fruit fly, subgenus> / cisplatin/ifosfamide / inactivation […]	359	[10%, 15%)
Immunohistochemical / Neoplastic Cell / Deterioration of status / Ectopic (qualifier value) / Personnel Turnover / Dominant-Negative Mutation / Tissue Adhesions / Disc - Body Part / Disk Device Component / Manifest […]	732	[5%, 10%)
Conflict (Psychology) / Event / Tissue Dissection / Moderation / Mutant Proteins / Binding Activity [MoA] / Dorsal / Genetic Screening (procedure) / Growth Factor / Abnormal degeneration […]	4361	[0%, 5%)

ACKNOWLEDGEMENTS

The authors acknowledge the support from the members of Temporal Knowledge Bases Group at Universitat Jaume I.
Funding: LJGC and AGC are both self-funded, RB is funded by the "Ministerio de Economía y Competitividad" with contract number TIN2011-24147. We also want to thank Melissa Carrion for her invaluable contribution as a proof reader; her iterative involvement has allowed us to continuously improve this manuscript.

REFERENCES

1. Doms A, Schroeder M: **GoPubMed: exploring PubMed with the Gene Ontology**. *Nucleic Acids Research* 2005, **33**(suppl 2):W783-W786.
2. Aleman Meza B: **Ranking documents based on relevance of semantic relationships**. In.; 2007.
3. Sheth A, Arpinar IB, Kashyap V: **Relationships at the Heart of Semantic Web: Modeling, Discovering, and Exploiting Complex Semantic Relationships**. In: *Enhancing the Power of the Internet.* Springer Berlin Heidelberg; 2003: 63-94.
4. Small H: **Co-citation in the scientific literature: A new measure of the relationship between two documents**. *Journal of the American Society for Information Science* 1973, **24**(4):265-269.
5. Hummon NP, Dereian P: **Connectivity in a citation network: The development of DNA theory**. *Social Networks* 1989, **11**(1):39-63.
6. Rogers F: **Medical subject headings**. *Bulletin of the Medical Library Association* 1963, **51**:114-116.
7. Cornet R, de Keizer N: **Forty years of SNOMED: a literature review**. *BMC Medical Informatics and Decision Making* 2008, **8 Suppl 1**:S2.
8. Bodenreider O: **The Unified Medical Language System (UMLS): integrating biomedical terminology**. *Nucleic Acids Research* 2004, **32**(suppl 1):D267-D270.
9. Harispe S, Ranwez S, Janaqi S, Montmain J: **Semantic Measures for the Comparison of Units of Language, Concepts or Entities from Text and Knowledge Base Analysis**. *Clinical Orthopaedics and Related Research* 2013, **abs/1310.1285**.
10. Boyack KW, Newman D, Duhon RJ, Klavans R, Patek M, Biberstine JR, Schijvenaars B, Skupin A, Ma N, Börner K: **Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches**. *PLoS ONE* 2011, **6**(3):e18029.
11. Lin J, Wilbur WJ: **PubMed related articles: a probabilistic topic-based model for content similarity**. *BMC Bioinformatics* 2007, **8**(1):423.
12. Garla V, Brandt C: **Semantic similarity in the biomedical domain: an evaluation across knowledge sources**. *BMC Bioinformatics* 2012, **13**(1):261.
13. Bloehdorn S, Hotho A: **Ontologies for Machine Learning**. In: *Handbook on Ontologies*. Edited by Staab S, Studer R: Springer Berlin Heidelberg; 2009: 637-661.
14. Aseervatham S, Bennani Y: **Semi-structured document categorization with a semantic kernel**. *Pattern Recognition* 2009, **42**(9):2067-2076.
15. Garla VN, Brandt C: **Ontology-guided feature engineering for clinical text classification**. *Journal of Biomedical Informatics* 2012, **45**(5):992-998.

16. Bloehdorn S, Moschitti A: **Combined Syntactic and Semantic Kernels for Text Classification**. In: *Advances in Information Retrieval.* Edited by Amati G, Carpineto C, Romano G, vol. 4425: Springer Berlin Heidelberg; 2007: 307-318.
17. Stevenson M, Greenwood MA: **A semantic approach to IE pattern induction**. In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics; Ann Arbor, Michigan*. 1219887: Association for Computational Linguistics 2005: 379-386.
18. Angelos H, Giannis V, Epimenidis V, Euripides GMP, Evangelos M: **Information Retrieval by Semantic Similarity**. *International Journal on Semantic Web and Information Systems (IJSWIS)* 2006, **2**(3):55-73.
19. Sahami M, Heilman TD: **A web-based kernel function for measuring the similarity of short text snippets**. In: *Proceedings of the 15th international conference on World Wide Web; Edinburgh, Scotland*. 1135834: ACM 2006: 377-386.
20. Robertson SE, Rijsbergen CJv, Porter MF: **Probabilistic models of indexing and searching**. In: *Proceedings of the 3rd annual ACM conference on Research and development in information retrieval; Cambridge, England*. 636673: Butterworth & Co. 1981: 35-56.
21. Robertson SE, Walker S: **Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval**. In: *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval; Dublin, Ireland*. 188561: Springer-Verlag New York, Inc. 1994: 232-241.
22. van Rijsbergen CJ, Robertson SE, Porter MF: **New models in probabilistic information retrieval**. In: *British Library Research and Development Report, no 5587*. London: British Library; 1979.
23. Jannach D, Zanker M, Felfernig A, Friedrich G: **The cosine similarity measure**. In: *Recommender Systems: An Introduction.* Cambridge University Press; 2010: 360.
24. Armstrong J: **Cosine similarity: the similarity of two weighted vectors**. In: *Programming Erlang, 2nd edition.* 2nd. edn: The Pragmatic Programmers; 2013: 548.
25. Hersh W, Cohen A, Yang J, Bhupatiraju RT, Oregon Health & Science University, Roberts P, Biogen Idec Corporation: **TREC 2005 Genomics Track Overview**. In: *Text Retrieval Conference: 2005*.
26. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R: **Indexing by latent semantic analysis**. *Journal of the American Society for Information Science* 1990, **41**(6):391-407.
27. Valle D, Baiser B, Woodall CW, Chazdon R: **Decomposing biodiversity data using the Latent Dirichlet Allocation model, a probabilistic multivariate statistical method**. *Ecology Letters* 2014, **17**(12):1591-1601.
28. U.S. National Institutes of Health's National Library of Medicine: **PubMed Central**. In. Internet (http://www.ncbi.nlm.nih.gov/pmc/); 2000.
29. Sayers E: **E-utilities Quick Start (2008 Dec 12 [Updated 2013 Aug 9])**. In: *Entrez Programming Utilities Help [Internet]*. Bethesda (MD): National Center for Biotechnology Information (US); 2010-. Available from: http://www.ncbi.nlm.nih.gov/books/NBK25500/; 2008.
30. National Center for Biotechnology Information (US): **Computation of Related Citations**. In: *PubMed Help [Internet]*. Bethesda, MD; 2005.
31. Berlanga R, Nebot V, Jimenez-Ruiz E: **Semantic annotation of biomedical texts through concept retrieval**. *Procesamiento de Lenguaje Natural* 2010, **45**:247-250.
32. Aronson AR, Lang F-M: **An overview of MetaMap: historical perspective and recent advances**. *Journal of the American Medical Informatics Association* 2010, **17**(3):229-236.
33. Garcia Castro LJ: **Semantic similarity benchmark**. In. Internet (ljgarcia.github.io/semsim.benchmark/): Github; 2014.
34. Townsend J, Hartl D: **Bayesian analysis of gene expression levels: statistical quantification of relative mRNA level across multiple strains or treatments**. *Genome Biology* 2002, **3**(12):research0071.0071 - research0071.0016.
35. Pan W, Lin J, Le C: **How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach**. *Genome Biology* 2002, **3**(5):research0022.0021 - research0022.0010.
36. Lewis J, Ossowski S, Hicks J, Errami M, Garner HR: **Text similarity: an alternative way to search MEDLINE**. *Bioinformatics* 2006, **22**(18):2298-2304.
37. Yamamoto Y, Takagi T: **Biomedical knowledge navigation by literature clustering**. *Journal of Biomedical Informatics* 2007, **40**(2):114-130.
38. Jonquet C, Lependu P, Falconer S, Coulet A, Noy NF, Musen MA, Shah NH: **NCBO Resource Index: Ontology-Based Search and Mining of Biomedical Resources**. *Web Semant* 2011, **9**(3):316-324.
39. McEntyre JR, Ananiadou S, Andrews S, Black WJ, Boulderstone R, Buttery P, Chaplin D, Chevuru S, Cobley N, Coleman L-A *et al*: **UKPMC: a full text article resource for the life sciences**. *Nucleic acids research* 2011, **39**(Database issue):D58-65.
40. Rebholz-Schuhmann D, Arregui M, Gaudan M, Kirsch H, Jimeno A: **Text processing through Web Services: Calling Whatizit**. *Bioinformatics* 2007, **24**(2).
41. Boyack KW: **Linking grants to articles: Characterization of NIH grant information indexed in Medline**. In: *International Conference of the International Society for Scientometrics and Informetrics*. 2009: 730-741.
42. Shah P, Perez-Iratxeta C, Bork P, Andrade M: **Information extraction from full text scientific articles: Where are the keywords?** *BMC Bioinformatics* 2003, **4**(1):20.
43. Garcia Castro LJ, Berlanga R, Rebholz-Schuhmann D, Garcia CA: **Connections across scientific publications based on semantic annotations**. In: *SePublica: 2013; Motpellier, France*.
44. Sun Z, Errami M, Long T, Renard C, Choradia N, Garner H: **Systematic Characterizations of Text Similarity in Full Text Biomedical Publications**. *PLoS ONE* 2010, **5**(9):e12704.
45. Garcia Castro LJ, McLaughlin C, Garcia A: **Biotea: RDFizing PubMed Central in Support for the Paper as an Interface to the Web of Data**. *Biomedical semantics* 2013, **4 Suppl 1**:S5.
46. World Wide Web Consortium: **RDF 1.1 Primer**. In. Edited by Schreiber G, Raimond Y, Manola F, Miller E, McBride B. Internet (http://www.w3.org/TR/rdf11-primer/); 2014.

47. Dumontier M: **Semantic approaches for biomedical knowledge discovery**. In: *Discovery Science; Bled, Slovenia*. 2014: Keynote speech presented at Discovery Science Conference.
48. Callahan A, Dumontier M, Shah N: **HyQue: evaluating hypotheses using Semantic Web technologies**. *Journal of Biomedical Semantics* 2011, **2**(Suppl 2):S3.
49. Intellectual Property Office: **Exceptions to copyright: research**. In. Internet (https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/375954/Research.pdf): UK Goverment; 2014.
50. Damerau FJ: **Generating and evaluating domain-oriented multi-word terms from texts**. *Inf Process Manage* 1993, **29**(4):433-447.
51. Frantzi K, Ananiadou S, Mima H: **Automatic recognition of multi-word terms:. the C-value/NC-value method**. *Int J Digit Libr* 2000, **3**(2):115-130.
52. Deane P: **A nonparametric method for extraction of candidate phrasal terms**. In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics; Ann Arbor, Michigan*. 1219915: Association for Computational Linguistics 2005: 605-613.
53. Zhang Z, Iria J, Brewster C, Ciravegna F: **A Comparative Evaluation of Term Recognition Algorithms**. In: *International Conference on Language Resources and Evaluation: 2008; Marrakech, Morocco*.