CITREC: An Evaluation Framework for Citation-Based Similarity Measures based on TREC Genomics and PubMed Central

Bela Gipp, Norman Meuschke, Mario Lipinski
National Institute of Informatics, Tokyo

Abstract
Citation-based similarity measures such as Bibliographic Coupling and Co-Citation are an integral component of many information retrieval systems. However, comparisons of the strengths and weaknesses of measures are challenging due to the lack of suitable test collections. This paper presents CITREC, an open evaluation framework for citation-based and text-based similarity measures. CITREC prepares the data from the PubMed Central Open Access Subset and the TREC Genomics collection for a citation-based analysis and provides tools necessary for performing evaluations of similarity measures. To account for different evaluation purposes, CITREC implements 35 citation-based and text-based similarity measures, and features two gold standards. The first gold standard uses the Medical Subject Headings (MeSH) thesaurus and the second uses the expert relevance feedback that is part of the TREC Genomics collection to gauge similarity. CITREC additionally offers a system that allows creating user-defined gold standards to adapt the evaluation framework to individual information needs and evaluation purposes.

Keywords: Test Collection, Benchmark, Similarity Measure, Citation, Reference, TREC, PMC
Citation: Editor will add citation with page numbers in proceedings and DOI.

Acknowledgements: We thank Corinna Breitinger for her valuable feedback and gratefully acknowledge the support of the National Institute of Informatics Tokyo.
Contact: Bela@Gipp.com, N@Meuschke.org, Lipinski@Sciplore.org

1 Introduction

The large and rapidly increasing amount of scientific literature has triggered intensified research into information retrieval systems that are suitable to support researchers in managing information overload. Many studies evaluate the suitability of citation-based[1], text-based, and hybrid similarity measures for information retrieval tasks (see tables 1-3 on pages 2 and 3).

However, objective performance comparisons of retrieval approaches, especially of citation-based approaches, are difficult, because many studies use non-publically available test collections, different similarity measures, and varying gold standards. The research community on recommender systems has identified the replication and reproducibility of evaluation results as a major concern. Bellogin et al. suggested the standardization and public sharing of evaluation frameworks as an important strategy to overcome this weakness (Bellogin et al., 2013). The Text REtrieval Conference (TREC)[2] series is a major provider of high quality evaluation frameworks for text-based retrieval systems.

Only a few studies evaluating citation-based similarity measures for document retrieval tasks are as transparent as the studies evaluating text-based similarity measures using standardized evaluation frameworks. Citation-based studies often use only partially suitable test collections or a gold standard that is questionable. As a result, studies on citation-based measures often contradict each other.

To overcome this lack of transparency, we provide a large-scale, open evaluation framework called CITREC. The name is an acronym of the words citation and TREC. CITREC allows evaluating the suitability of citation-based and text-based similarity measures for document retrieval tasks. CITREC prepares the publicly available PubMed Central Open Access Subset (PMC OAS) and the TREC Genomics ’06 test collection for a citation-based analysis and provides tools necessary for performing evaluations. All components of the framework are available under open licenses[3] and free of charge at:
www.sciplore.org/projects/citrec

[1] We use the term citation to express that a document is cited. The term reference to denote works listed in the bibliography, and in-text citation to denote markers in the main text linking to references in the bibliography. We use the common generalizations citation analysis or citation-based for all approaches that use citations, in-text-citations, references or combinations thereof for similarity assessment.

[2] http://trec.nist.gov

[3] GNU Public License for code, Open Data Commons Attribution License for data

We divide the presentation of CITREC as follows. Section 2 shows that studies evaluating citation-based similarity measures for document retrieval tasks often arrive at contradictory results. These contradictions are largely attributable to the shortcomings of the test collections used. Section 2 additionally examines the suitability of existing datasets for evaluating citation-based and text-based similarity measures. Section 3 presents the evaluation framework CITREC, which consists of data parsers for the PMC OAS and TREC Genomics collection, implementations of similarity measures, and two gold standards that are suitable for evaluating citation-based measures. CITREC also includes a survey tool for creating user-defined gold standards, and tools for statistically analyzing results. Section 4 provides an outlook, which explains our intention to include additional contributions, such as similarity measures and results.

2 Related Work

2.1 Studies Evaluating Citation-based Similarity Measures

Tables 1-3 summarize studies that assess the applicability of citation-based or hybrid similarity measures, i.e. measures that combine citation-based and text-based approaches, for different information retrieval tasks related to academic documents. Footnote 4 explains abbreviations we use in the three tables.

Table 1 lists studies that evaluate citation-based or hybrid similarity measures for topical clustering, i.e. grouping of topically similar documents in the absence of pre-defined subject categories. Clustering is an unsupervised machine learning task, i.e. no labeled training data is available. A clustering algorithm learns the features that best possibly separate data objects (in our case documents) into distinct groups. The groups, called clusters, provide little to no information about the semantic relationship between the documents included in the cluster.

Study	Similarity Measures	Gold Standard	Test Collection
(Jarneving, 2005)	Bibliographic Coupling, Co-Citation	Similarity of title keyword profiles for all clusters	7,239 Science Citation Index records
(Ahlgren and Jarneving, 2008)	***cit.:*** Bib. Coup. ; ***text:*** common abstract terms	1 expert judgment	43 Web of Science records
(Ahlgren and Colliander, 2009)	***cit.:*** Bib. Coup ; ***text:*** cosine in tf-idf VSM, SVD of tf-idf VSM ; ***hybrid:*** linear comb. of dissimilarity matrices, "free combination" of transformed matrices		
(Janssens et al., 2009)	***cit.:*** "second order" journal-cross citation (JCC) ; ***text:*** LSI of tf-idf VSM ; ***hybrid:*** linear combination of similarity matrices	External: Thomson Reuters Essential Science Indicators Internal: Mean Silhouette Value, Modularity	Web of Science records covering 1,869 journals in (Liu et al., 2009) and 8,305 journals in (Janssens et al., 2009, Liu et al., 2010)
(Liu et al., 2009)	***cit.:*** JCC ; ***text:*** tf-idf VSM ; ***hybrid:*** ensemble clustering and kernel fusion alg.		
(Liu et al., 2010)	***cit.:*** Bib. Coup., Co-Cit., 3 variants of JCC (regular, binary, LSI) ; ***text:*** 4 variants of VSM (tf, idf, tf-idf, binary), LSI of tf-idf VSM ; ***hybrid:*** various weighted variants of hybrid clustering algorithms		
(Shibata et al., 2009)	***cit.:*** Bibliographic Coupling, Co-Citation, direct citation	Self-defined topological criteria for cluster quality	40,945 records from Science Citation Index
(Boyack and Klavans, 2010)	***cit.:*** Bib. Coup., Co-Cit., direct citation ; ***hybrid:*** comb. of Bib. Coup. with word overlap in title and abstract	Jensen-Shannon divergence, grant-to-article linkages	2,153,769 MEDLINE records
(Boyack et al., 2012)	***cit.:*** regular Co-Citation, 3 variants of proximity-weighted Co-Citation	Jensen-Shannon divergence	270,521 full text articles in the life sciences

Table 1: Studies evaluating citation-based and hybrid similarity measures for topic clustering.

[4] **alg.** – Algorithms | **Bib. Coup.** – Bibliographic Coupling | **cit.** - citation-based similarity measures | **Co-Cit.** - Co-Citation | **comb.** – combination | **idf** - inverse document frequency | **JCC** - journal cross citation | **LSI** - latent semantic indexing | **SVD** - single value decomposition | **text** - text-based similarity measures | **tf** - term frequency | **VSM** - vector space model

Table 2 lists studies that evaluate citation-based or hybrid similarity measures for topic classification, i.e. assigning documents to one of several pre-defined subject categories. Opposed to topic clustering, topic classification is a supervised machine learning task. Given pre-classified training data, a classifier learns the features that are most characteristic for each subject category and applies the learned rules to assign unclassified objects to the most suitable category.

Study	Similarity Measures	Gold Standard	Test Collection
(Cao and Gao, 2005)	***hybrid:*** iterative combination of class membership probabilities returned by text-based and citation-based classifiers	Classification of Cora dataset (created by text-based classifiers)	4,330 full text articles in machine learning
(Couto et al., 2006)	***cit.:*** Bib. Coup., Co-Cit., Amsler, cosine in tf-idf VSM ; ***hybrid:*** statistical evidence combination, Bayesian network approach	1st level terms of ACM classification	6,680 records from ACM Digital Library
(Zhu et al., 2007)	***cit. and text:*** SVM of citations or words ; ***hybrid*:** various factorizations of the similarity matrices	Classification of Cora collection (created by text-based classifiers)	4,343 records from Cora dataset
(Li et al., 2009)	***cit.:*** SimRank for citation and author links ; ***text:*** cosine in tf-idf VSM; ***hybrid*:** "link-based content analysis" measure	1st level terms of ACM classification	5,469 records from ACM Digital Library

Table 2: Studies evaluating citation-based and hybrid similarity measures for topic classification.

Table 3 lists studies that evaluate citation-based similarity measures for retrieving topically related documents, e.g., to give literature recommendations. Except for the study (Eto, 2012), all studies in Table 3 identify related papers within specific research fields. Thus, the scope of studies in Table 3 is narrower and more centered on particular topics than the scope of studies listed in Table 1 and Table 2.

Study: Objective	Similarity Measures	Gold Standard	Test Collection
(Lu et al., 2007): Literature recommendation	***cit.:*** new "authority" and "maximum flow" measure, CCIDF (CiteSeer measure) ; ***text:*** VSM using words and noun phrases	Relevance judgments of 2 domain experts	23,371 CiteSeer records on neural networks
(Yoon et al., 2011): Identify topically similar articles	***cit:*** SimRank, rvs-SimRank, P-Rank, C-Rank	Prediction of references in a textbook	23,795 DBLP records on database research (references from MS Academic Search)
(Eto, 2012): Identify topically similar articles	***cit.:*** 3 variants of "spread Co-Citation" measure	Overlap in MeSH terms	152,000 full text articles in biomedicine
(Eto, 2013) Identify topically similar articles	***cit.:*** regular Co-Citation, 5 variants of proximity-weighted Co-Citation	21 expert judgments	13,551 CiteSeer records incl. full texts on database research
To appear: Evaluation of similarity measures for topical similarity by the authors of this paper	***cit.:*** Bibliographic Coupling, Co-Citation, Amsler, Co-Citation Proximity Analysis, Contextual Co-Citation ; ***text:*** cosine in tf-idf VSM	Information Content analysis derived from MeSH thesaurus	approx. 172,000 articles from the PubMed Central Open Access Subset

Table 3: Studies evaluating citation-based sim. measures for identifying topically related documents.

The studies summarized in the three preceding tables demonstrate that researchers evaluate different sets of citation-based or hybrid similarity measures for a variety of retrieval tasks. An additional, currently evolving field of research is using citation-based similarity assessments to detect plagiarism (Gipp et al., 2014, Pertile et al., 2013).The datasets and gold standards used for evaluating citation-based measures vary widely and are often not publicly available, reducing the comparability and reproducibility of results. In Section 2.2, we discuss the shortcomings of the test collections used for prior studies in detail.

2.2 Shortcomings of Existing Test Collections

Most studies listed in the tables of Section 2.1 address different evaluation objectives. However, even studies that analyze the same research question often contradict each other. Examples are the publications “Comparative Study on Methods of Detecting Research Fronts Using Different Types of Citation” (Shibata et al., 2009) and “Co-citation Analysis, Bibliographic Coupling, and Direct Citation: Which Citation Approach Represents the Research Front Most Accurately?” (Boyack and Klavans, 2010). While the first study concludes:

> “Direct citation, which could detect large and young emerging clusters earlier, shows the best performance in detecting a research front, and co-citation shows the worst.”

The second study contradicts these findings:

> “Of the three pure citation-based approaches, bibliographic coupling slightly outperforms co-citation analysis using both accuracy measures; direct citation is the least accurate mapping approach by far.”

We hypothesize that the contradicting results of prior studies evaluating citation-based similarity measures are mainly due to the use of datasets or gold standards that are only partially suitable for the respective evaluation purpose.

2.2.1 Datasets

The selection of datasets is one of the main weaknesses of prior studies. Most studies we reviewed used bibliographic records obtained from indexes like the Thomson Reuters Science Citation Index / Web of Science, CiteSeer, or the ACM Digital Library. Bibliographic records comprise the title, authors, abstract, and bibliography of a paper, but lack full texts and thereby information about in-text citations. An increasing number of recently proposed Co-Citation-based measure like the Co-Citation Proximity Analysis (Gipp and Beel, 2009) consider the position of in-text citations. Consequently, these measures cannot be evaluated using collections of bibliographic records.

The use of small scale datasets is another obstacle to objective performance comparisons of citation-based similarity measures. Intuitively, smaller datasets provide less input data for analyzing citations, which decreases the observable performance of citation-based similarity measures. Especially the number of intra-collection citations, i.e. citations between two documents that are both part of the collection, decreases for small datasets. This decline significantly affects the performance of Co-Citation-based similarity measures, which can only compute similarities between documents if these documents are co-cited within other documents included in the dataset. Therefore, the ratio of intra-collection citations to total citations is an important characteristic, which we term self-containment.

The dependency of citation-based similarity measures on dataset size limits the informative value of prior studies. Conclusions drawn on results obtained from studies using the available small-scale test collections are likely not transferable to larger datasets with different characteristics.

2.2.2 Gold Standards

Defining the perceived ideal retrieval result, the so-called ground truth, is an inherent and ubiquitous problem in Information Retrieval. Relevance is the criterion for establishing this ground truth. Relevance is the relationship between information or information objects (in our case documents) and contexts (in our case topics or problems) (Saracevic, 2006). In other terms, relevance measures the pertinence of a retrieved result to a user’s information need.

In agreement with Saracevic, we define relevance as consisting of two main components - objective topical relevance and subjective user relevance. Topical relevance describes the “aboutness” (Saracevic, 2006) of an information object, i.e. whether the object belongs to a certain subject class. Subject area experts can judge topical relevance fairly well. User relevance, on the other hand, is by definition subjective and dependent on the information need of the individual user (Lachica et al., 2008, Saracevic, 2006).

The goal of Information Retrieval is to provide the user with documents that help satisfy a specific information need, i.e. the results must be relevant to the user. Yet, the subjective nature of relevance implies that in most cases a single accurate ground truth does not exist. For assessing the performance of information retrieval systems, researchers can only approximate ground truths for topical and user relevance. We use the term gold standard to refer to a ground truth approximation that is reasonably accurate, but not as objectively definitive as a ground truth.

Existing studies commonly use small-scale expert interviews or an expert classification system, such as the Medical Subject Headings (MeSH), to derive a gold standard. Using a classification system as a gold standard is suitable for finding similar documents, but unsuitable for identifying related documents, because classification systems do not reflect academic significance (impact), novelty, or diversity. Gold standards based on expert judgments do not share these shortcomings. Nonetheless, currently only small-scale test collections exist, because creating a comprehensive high quality test collection requires considerable resources.

The nonexistence of an openly available, large-scale test collection that features a comprehensive gold standard of the quality comparable to the existing standards for text-based retrieval systems makes most prior evaluations of citation-based similarity measures irreproducible. The test collections used in prior studies commonly remained unpublished and insufficiently documented. To overcome this non-transparency, we developed the CITREC evaluation framework. In Section 2.3, we analyze the suitability of datasets that we considered for inclusion in the CITREC framework.

2.3 Potential Datasets for CITREC

This Section analyzes existing datasets regarding their suitability for compiling a large-scale, openly available test collection that allows comparing the performance of citation-based and text-based similarity measures for document retrieval tasks.

2.3.1 Test Collection Requirements

An ideal test collection for evaluating citation-based and text-based similarity measures for document retrieval tasks should fulfill the following eight requirements.

First, the test collection should comprise scientific full texts. Full text availability is necessary to compare the retrieval performance of most text-based and some Co-Citation-based similarity measures. Recent advancements of the Co-Citation approach, such as Co-Citation Proximity Analysis (CPA) (Gipp and Beel, 2009) consider how close to each other the sources are cited in the text. Therefore, these approaches require the exact positions of citations within the full text to compute similarity scores.

Second, the test collection should be sufficiently large to reduce the risk of introducing bias by relying on a non-representative sample. Bias may arise, for example, by including a disproportionate number of very recent or very popular documents. Receiving citations from other documents requires time. This delay causes the citation counts for very recent documents to be lower regardless of their quality or relevance. Therefore, very recent documents are rarely analyzable by Co-Citation-based similarity measures. On the other hand, popular documents are likely to have more citations, which may cause citation-based results to score disproportionately.

Third, the documents of the test collection should cover identical or related research fields. Selecting documents from related subject areas increases the likelihood of intra-collection citations, thus increases the degree of self-containment, which improves the accuracy of a citation-based analysis.

Fourth, expert relevance judgments, or their approximation, should be obtainable for large parts of the dataset underlying the test collection. The effort of gathering comprehensive human relevance judgments for a large test collection and multiple similarity measures exceeds our resources. This necessitates choosing a dataset for which a form of relevance feedback is already available. We view expert judgments from prior studies or manually maintained subject classification systems as the best approach to approximate topical relevance using pre-existing information.

Fifth, the documents of the test collection should be available in a format that facilitates parsing of in-text citations and references. Parsing in-text citation and references from PDF documents is error prone (Lipinski et al., 2013). Parsing this information from plain text or from documents using structured markup formats such as HTML or XML is significantly more accurate.

Sixth, the documents of the test collection should use endnote-based citation styles to facilitate accurate parsing of citation and reference information. Endnote-based citation styles use in-text citation markers that refer to a single list of references at the end of the main text. The list of references exclusively states the metadata of the cited sources without author remarks. Endnote-based citation styles are most prevalent in the natural and life sciences. The social sciences and humanities tend to use footnotes for citing sources. Combining multiple references and including further remarks in one footnote are also common within these disciplines. Such discrepancies impede accurate automatic parsing of references in texts from the social sciences or humanities. Parsing citation and references formatted in endnote-based style is more accurate than parsing footnote style references.

Seventh, unique document identifiers, which increase the accuracy of the data parsing process, should be available for most documents of the test collection. Assigning unique identifiers and using them when referencing a document is more widespread in the natural and life sciences than in the social sciences and humanities. Examples of identifiers include Digital Document Identifiers (DOI), or identifiers assigned to documents included in major collections, e.g., arXiv.org for physics, or PubMed for biomedicine and the life sciences. Unique document identifiers facilitate the disambiguation of parsed reference data and the comparison of references between documents.

Eighth, the test collection should consist of openly accessible documents to facilitate the reuse of the collection for other researchers, which increases the reproducibility and transparency of results.

In the Sections 2.3.2 - 2.3.7, we discuss the suitability of seven datasets for meeting the requirements we derived in this Section:

a) Full text availability
b) Size of the collection
c) Self-containment of the collection
d) Availability of expert classifications or relevance feedback
e) Availability of structured document formats
f) Use of endnote-based citation styles
g) Availability of unique document identifiers
h) Open Access

2.3.2 Web of Science and Scopus

Thomson Reuters's Web of Science (WoS) and Elsevier's Scopus are the largest commercial citation indexes. WoS includes 12,000 journals and 160,000 conference proceedings[5], while Scopus includes 21,000 journals and 6.5 million conference papers[6]. Both indexes cover the sciences, social sciences, arts, and humanities, and both offer document metadata, citation information, topic categorizations, and links to external full-text sources. Studies suggest that data accuracy in WoS and other professionally managed indexes is approx. 90% with most discrepancies being attributable to author errors, while processing errors by the index providers are rare (Buchanan, 2006). We assume that the data in Scopus is comparably accurate as in WoS. Both indexes require subscription and do not allow bulk processing.

2.3.3 DBLP

DBLP is an openly accessible citation index that offers document metadata and citation information for approx. 2.8 million computer science documents[7]. DBLP data is of high quality and available in XML format. Full texts or a comprehensive subject classifications scheme are not available.

2.3.4 INEX 2009 Collection

The Initiative for the Evaluation of XML Retrieval (INEX)[8] offers test collections for various information retrieval tasks. For their conference in 2009, the INEX built a test collection by semantically annotating 2.66 million English Wikipedia articles. INEX derived the semantic annotations from linking words in the articles to the WordNet[9] thesaurus and exploiting features of the Wikipedia format, such as categorizations, lists, or tables (Geva et al., 2010). The INEX collection contains 68 information needs with corresponding relevance judgments based on examining over 50,000 articles. The INEX collection articles are formatted in XML and offer in-text citations and references. Because volunteers regularly check and edit Wikipedia articles for correctness and completeness, we expect citation data in Wikipedia to be reasonably accurate, yet we are not aware of any studies that have investigated this question. Citations between Wikipedia articles occur frequently. This characteristic of Wikipedia increases the self-containment of the INEX collection. Whether citations between Wikipedia articles are equally rich in their semantic content as academic citations is unclear. Due to Wikipedia's broad scope, we expect minimal overlap in citations of external sources.

2.3.5 Integrated Search Test Collection

The Integrated Search Test Collection (iSearch)[10] is an evaluation framework for information retrieval systems provided free of charge by the Royal School of Library and Information Science, Denmark (Lykke et al., 2010). The collection consists of 143,571 full text articles with corresponding metadata records from arXiv.org, additional 291,246 arXiv.org metadata records without full texts, 18,443 book metadata

[5] As of September 2014, source: http://wokinfo.com/products_tools/multidisciplinary/webofscience/
[6] As of September 2014, source: http://www.elsevier.com/online-tools/scopus/content-overview
[7] As of November 2014, source: http://dblp.uni-trier.de/
[8] https://inex.mmci.uni-saarland.de/
[9] http://wordnet.princeton.edu/
[10] http://itlab.dbit.dk/~isearch

records and 65 information needs with corresponding relevance judgments based on examining over 11,000 articles. All articles and records in the collection are in the field of physics.

2.3.6 PubMed Central Open Access Subset

PubMed Central (PMC) is a repository of approx. 3.3 million full text documents from biomedicine and the life sciences maintained by the U.S. National Library of Medicine (NLM)[11]. PMC documents are freely accessible via the PMC website. The NLM also offers a subset of 860,000 documents formatted in XML for bulk download and processing, the so-called PubMed Central Open Access Subset (PMC OAS)[12].

Data in the PMC OAS is of high quality and comparably easy to parse, because relevant document metadata, in-text citations, and references are labeled using XML. Many documents in the PMC OAS have unique document identifiers, especially PubMed ids (PMID). Authors widely use PMIDs when stating references, which facilitates reference disambiguation and matching. A major benefit of the PMC OAS is the availability of Medical Subject Headings, which we consider partially suitable for deriving a gold standard. We describe details of MeSH and their role in deriving a gold standard in Section 3.3.1.

2.3.7 TREC Genomics Collection

The test collection used in the Genomics track of the TREC conference 2006 comprises 162,259 Open Access biomedical full text articles and 28 information needs with corresponding relevance feedback (Hersh et al., 2006). The articles included in the collection are freely available in HTML format[13] and cover the same scientific domain as the PMC OAS. The TREC Genomics (TREC Gen.) collection offers comparable advantages regarding the use of unique document identifiers and availability of MeSH for most articles. In comparison to the XML format of documents in the PMC OAS, the HTML format of articles in the TREC Gen. collection offers less markup labeling document metadata and citation information. However, PMIDs are available that allow retrieving this data in high quality from a web service. In addition, parsing the HTML files of the TREC Gen. collection is still significantly less error prone than processing PDF documents.

2.4 Datasets Selected for CITREC

Table 4 summarizes the datasets we presented in Sections 2.3.2 - 2.3.7 by indicating their fulfillment of the eight test collection requirements we derived in Section 2.3.1.

	WoS	Scopus	DBLP	PMC OAS	TREC Gen.	INEX	iSearch
a) Full text availability	No	No	No	Yes	Yes	Yes	Yes
b) No. of records in millions[14]	>40	~50	~2.8	~0.86	~0.16	~2.66	~0.16
c) Self-containment	Good	Good	Good	Good	Good	Good	Good
d) Expert classification / relevance feedback	Yes	Yes	No	Yes (MeSH)	Yes	Yes	Yes
e) Structured document format	No	No	No	Yes	Yes	Yes	No
f) Endnote citation styles	partially		Yes	Yes	Yes	Yes	Yes
- Reference data available	Yes	Yes	No	Yes	Implicit	Implicit	Yes
- In-text citation positions	No	No	No	Implicit	Implicit	Implicit	Implicit
g) Unique document identifiers	Yes	Yes	Yes	Yes, for most doc.		No	Yes
h) Open Access	No	No	Yes	Yes	Yes	Yes	Yes

Table 4: Comparison of potential datasets.

We regard the PMC OAS, TREC Gen., INEX, and iSearch collections as most promising for our purpose. All four collections offer a high number of freely available full texts. Except for iSearch, all collections provide structured document formats. TREC Gen., INEX, and iSearch offer a gold standard based on

[11] as of January 2015, source: http://www.ncbi.nlm.nih.gov/pmc/

[12] http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

[13] http://ir.ohsu.edu/genomics/

[14] As of November 2014

specific information needs and experts' relevance feedback. The PMC OAS collection allows deriving a gold standard from the MeSH classification.

Due to limited resources, we excluded the INEX and iSearch collection from our new test collection. The reason for excluding the INEX collection is that Wikipedia articles are fundamentally different from the academic documents in the other collections. Evaluating citation-based similarity measures for information retrieval tasks related to Wikipedia articles is an interesting future task. However, for our first test collection, we chose to focus on academic documents, which represent the traditional area of application for citation analysis. We plan to extend CITREC to include the INEX or other collections based on Wikipedia in the future. We excluded the iSearch collection, because it does not offer full-texts in a structured document format.

Consequently, we established a new, large-scale test collection by adapting the PMC OAS and the TREC Gen. collection to the needs of a citation-based analysis. Both collections offer structured document formats, which are comparably easy to parse, and a wide availability of unique document identifiers. Both characteristics are important when aiming for high data quality. A major benefit of both collections is the availability of relevance information that is suitable for deriving a gold standard. For the PMC OAS, we use the MeSH classification to compute a gold standard. For the TREC Gen. collection, we derive a gold standard from the comprehensive relevance feedbacks that domain experts provided for the original evaluation. We describe both gold standards and the other components of the CITREC evaluation framework in Section 3.

3 CITREC Evaluation Framework

The CITREC evaluation framework consists of the following four components:

a) *Data Extraction and Storage* – contains two parsers that extract the data needed to evaluate citation-based similarity measures from the PMC OAS and the TREC Genomics collection, and a database that stores the extracted data for efficient use;

b) *Similarity Measures* – contains Java implementations of citation-based and text-based similarity measures;

c) *Information Needs and Gold Standards* – contains a gold standards derived from the MeSH thesaurus, a gold standard based on the information needs and expert judgments included in the TREC Genomics collection, and code for a system to establish user-defined gold standards;

d) *Tools for Results Analysis* – contains code to statistically analyze and compare the scores that individual similarity measures yield.

The subsections 3.1 - 3.3 introduce each component. Additional documentation providing details on the components is available at www.sciplore.org/projects/citrec.

3.1 Data Extraction and Storage

Given our analysis of potentially suitable datasets described in Section 2.3, we selected the PMC OAS and the TREC Genomics collection to serve as the dataset for the CITREC evaluation framework. Both collections require parsing to extract in-text citations, references, and other data necessary for performing evaluations of citation-based similarity measures. We developed two parsers in Java, each tailored to process the different document formats of the two collections. The parsers extract the relevant data from the texts and store this data in a MySQL database, which allows efficient access and use of the data for different evaluation purposes.

In the case of the PMC OAS, extracting document metadata and reference information such as authors, titles and document identifiers is a straightforward task, due to the comprehensive XML-markup. We excluded documents without a main text (commonly scans of older articles), and documents with multiple XML body tags (commonly summaries of conference proceedings). Additionally, we only considered the document types *brief-report, case-report, report, research-article, review-article* and *other* for import. The exclusions reduced the collection from 346,448 documents[15] to 255,339 documents.

The extraction of in-text citations from the PMC OAS documents posed some problems to parser development. Among these challenges was the use of heterogeneous XML-markups for labeling in-text citations in the source files. For this reason, we incorporated eight different markup variations into the parser. The bundling of in-text citations, e.g., in the form "[25–28]", was difficult to process because some

[15] The National Library of Medicine regularly adds documents to the PMC OAS. At the time of processing, the collection contained 346,448 documents. As of Nov. 2014, the collection has grown to approx. 860,000 documents (see Table 4)

source files mix XML markup and plain text. Different characters for the separating hyphen and varying sort orders for identifiers increased the difficulty of accurately parsing bundled citations. An example of a bundled citation with mixed markup is:

[<xref ref-type="bibr" rid="B1">1</xref> - <xref ref-type="bibr" rid="B5">7</xref>]

To record the exact character, word, and sentence-level at which in-text citations appear within the text, we stripped the original document of all XML and applied suitable detection algorithms. We used the SPToolkit by Piao, because it was specifically designed to detect sentence boundaries in biomedical texts (Piao and Tsuruoka, 2008). For the detection of word boundaries, we developed our own heuristics based on regular expressions. The same applies for the detection of in-text citation groups, e.g., in the form "[1][2][3]". A detailed description of the heuristics is available at www.sciplore.org/projects/citrec.

In the case of the TREC Genomics collection, processing the data required for analysis was more challenging, because the source documents offered less exploitable markup. We retrieved document metadata, such as author names and title, by querying the PMIDs in the collection to the SOAP-based Entrez Programming Utilities[16] (E-Utilities) web-service. Entrez is a unified search engine that covers data sources related to the U.S. National Institute of Health (NIH), e.g., PubMed, PMC, and a range of gene and protein databases. The E-Utilities are eight server-side programs that allow automated access to the data sources covered by Entrez.

We could obtain data for 160,446 of the 162,259 articles in the TREC Gen. collection. Errors in retrieving metadata resulted from invalid PMIDs. The problem that approx. 1% of the articles in the TREC Gen. collection have invalid PMIDs was known to the organizers of the TREC Gen. track (Hersh et al., 2006). We excluded documents that caused errors.

The developed TREC Gen. parser relies on heuristics and suitable third-party tools to obtain in-text citation and reference data. The TREC Gen. collection states references in plain text with no further markup except for an identifier that is unique within the respective document. We used the open source reference parser ParsCit[17] to itemize the reference strings.

For the PMC OAS and the TREC Gen. collection, we queried the E-utilities to obtain the MeSH information necessary to derive the thesaurus-based gold standard (see Section 3.3.1). MeSH are available for 172,734 documents (67%) in the PMC OAS and 160,047 document (99%) in the TREC Gen collection. The parsers for both collections include functionality for creating a text-based index using the open source search engine Lucene[18].

3.2 Similarity Measures

The CITREC framework provides open-source Java code for computing 35 citation-based and text-based similarity measures (including variants of measures) as well as pre-computed similarity scores for those measures to facilitate performance comparisons. Table 5 gives an overview of the similarity measures and gold standards included in CITREC.

Approach	Measures Implemented in CITREC
Citation-based	Amsler (standard and normalized)
	Bibliographic Coupling (standard, normalized)
	Co-Citation (standard and normalized)
	Co-Citation Proximity Analysis (various versions)
	Contextual Co-Citation (various versions)
	Linkthrough
Text-based	Lucene *More Like This* with varying boost factors for title, abstract, and text
Expert-based (gold standards)	Medical Subject Heading (MeSH) Relevance Feedback (TREC Genomics)

Table 5: Similarity measures and gold standards included in CITREC.

[16] http://www.ncbi.nlm.nih.gov/books/NBK25501/
[17] http://aye.comp.nus.edu.sg/parsCit/
[18] http://lucene.apache.org/core/

For each of the 35 similarity measures, we pre-computed similarity scores and included the results (one table with scores per measure) in a MySQL database. The database and the code are available for download at www.sciplore.org/projects/citrec.

Aside from classical citation-based measures, such as Bibliographic Coupling and Co-Citation, we also implemented more recent similarity measures, such as Co-Citation Proximity Analysis, Contextual Co-Citation and Local Bibliographic Coupling. These recently developed methods consider the position of in-text citations as part of their similarity score. Text-based measures in our framework use Lucene's *More Like This* function. We also included a similarity measure based on MeSH, which we describe in Section 3.3.1. We invite the scientific community to contribute further similarity measures to the CITREC evaluation framework.

3.3 Information Needs and Gold Standards

As we showed in Section 2.2, studies that evaluate citation-based similarity measures address different objectives and employ heterogeneous gold standards. In this Section, we present three options for defining information needs and gold standards that we implemented as part of the CITREC framework.

The first option, which we explain in Section 3.3.1, does not define specific information needs, but uses Medical Subject Headings to derive an implicit gold standard concerning the topical relevance of any document having MeSH assigned.

The second option, which we present in Section 3.3.2, uses the information needs of the TREC Genomics collection and employs the corresponding expert feedback to derive a new gold standard that is suitable for citation-based similarity measures.

For evaluation purposes that cannot be served by either of these two options, we developed a web-based system to define individual information needs and gather feedback that allows users of CITREC to derive customized gold standards. We explain this system in Section 3.3.3.

3.3.1 Medical Subject Headings

Medical Subject Headings are a poly-hierarchical thesaurus of subject descriptors. Experts at the U.S. National Library of Medicine (NLM) maintain the thesaurus and manually assign the most suitable descriptors to documents upon their inclusion in the NLM's digital collection MEDLINE (U. S. National Library of Medicine, 2014). We view MeSH as an accurate judgment of topical similarity given by specialists, which makes it partially suitable for deriving a gold standard for topical relevance. We include a gold standard derived from the MeSH-thesaurus to enable researchers to gauge the ability of citation-based and text-based similarity measures to reflect topical relevance. Multiple prior studies followed a similar approach by exploiting MeSH to derive measures of document similarity (Batet et al., 2010, Eto, 2012, Lin and Wilbur, 2007, Zhu et al., 2009).

A major advantage when deriving a gold standard using MeSH descriptors is that most documents in the CITREC test collection have been manually tagged with MeSH descriptors. Due to time and cost constraints, most other test collections can collect human relevance feedback only for a small fraction of the included documents.

However, MeSH descriptors also have inherent drawbacks. One drawback is that commonly a single reviewer assigns MeSH descriptors and hence categorizes documents into fixed subject classes even prior to the general availability of the documents to the research community. This categorization expresses topical relatedness only, but cannot reflect academic significance, which requires appreciation of the document by the research community. Another weakness of MeSH is that the reviewer assigns MeSH descriptors at a single point in time. After this initial classification, the MeSH descriptors assigned to a document remain unaltered in most cases. Hence, MeSH descriptors can be incomplete in the sense that they only reflect the most important topic keywords at the time of review. MeSH may not adequately reflect shifts in the importance of documents over time, which is especially crucial for newly evolving fields. An example of this effect can be seen in documents on sildenafil citrate, the active ingredient of Viagra. British researchers initially synthesized sildenafil citrate to study its effects on high blood pressure and angina pectoris. The positive effect of the substance in treating erectile dysfunction only became apparent during clinical trials later on. Therefore, earlier papers discussing sildenafil citrate may carry MeSH descriptors related to cardiovascular diseases, while the MeSH descriptors of later documents are likely in the field of erectile dysfunction. A similarity assessment using MeSH may therefore not reflect the relationship between earlier and later documents covering the same topic.

To derive the gold standard, we followed an approach used by multiple prior studies, which derived similarity measures from MeSH. The idea is to evaluate the distance of MeSH descriptors assigned to the documents within the tree-like thesaurus. We use the generic similarity calculation suggested by Lin (Lin, 1998), in combination with the assessment of information content (IC), for quantifying the similarity of concepts in a taxonomy proposed by Resnik (Resnik et al., 1995). The MeSH thesaurus is essentially an annotated taxonomy, thus Resnik's measure suits our purpose.

Intuitively, the similarity of two concepts c_1 and c_2 in a taxonomy reflects the information they have in common. Resnik proposed that the most specific superordinate concept $c_s(c_1, c_2)$ that subsumes c_1 and c_2, i.e. the closest common ancestor of c_1 and c_2, represents this common information. Resnik defined the information content (IC) measure to quantify the common information of concepts. Information content describes the amount of extra information that a more specific concept contributes to a more general concept that subsumes it. To quantify IC, Resnik proposed analyzing the probability $p(c)$ of encountering an instance of a concept c. By definition, concepts that are more general must have a lower IC than the more specific concepts they subsume. Thus, the probability of encountering a subsuming concept c has to be higher than that of encountering all its specializations $s(c)$ (Resnik et al., 1995). We assure that this requirement holds by calculating the probability of a concept c as:

$$p(c) = \frac{1 + |s(c)|}{N}$$

where N is the total number of concepts in the MeSH thesaurus. According to Resnik's proposal, we quantify information content using a negative log-likelihood function in the interval [0,1]:

$$IC(c) = -\log p\ (c)$$

Lin's generic similarity measure uses the relation between the information content of two concepts and their closest subsuming concept $c_s(c_1, c_2)$. It calculates as:

$$sim(c_1, c_2) = \frac{2 \times IC(c_s(c_1, c_2))}{IC(c_1) + IC(c_2)}$$

We used Lin's measure, since it performed consistently for various test collections, while other measures differed significantly in prior studies. Lin's measure solely analyzes the similarity of two occurrences of concepts. MeSH descriptors can occur multiple times within the thesaurus. To determine the similarity of two specific MeSH descriptors m_1 and m_2, we have to compare the sets of the descriptors' occurrences O_1 and O_2. Each set represents all occurrences of the descriptors m_1 and respectively m_2 in the thesaurus. We use the average maximum match, a measure that Zhu et al. proposed, for this use case (Zhu et al., 2009). For each occurrence o_p of the descriptor m_1 with $o_p \in O_1$, the measure considers the most similar occurrence $o_q$of the descriptor m_2 with $o_q \in O_2$ and vice versa as:

$$sim(m_1, m_2) = \frac{\sum_{o_p \in O_1} max(sim(o_p\ , o_q)) + \sum_{o_q \in O_2} max(sim(o_q\ , o_p))}{|O_1| + |O_2|}$$

To determine the similarity of two documents d_1 and d_2, we use the average maximum match between the sets of MeSH descriptors M_1 and M_2 assigned to the documents. To compute the similarity between individual descriptors in the sets M_1 and M_2, we consider the set of occurrences $O(m_p)$ and $O(m_q)$ of the descriptors $m_p \in M_1$ and $m_q \in M_2$.

$$sim(d_1, d_2) = sim(M_1, M_2) = \frac{\sum_{O(m_p) \in M_1} max(sim(O(m_p), O(m_q))) + \sum_{O(m_q) \in M_2} max(sim(O(m_q), O(m_p)}{|M_1| + |M_2|}$$

We only include the so-called major topics for calculating similarities. Major topics are MeSH descriptors that receive a special accentuation by the reviewers that assign MeSH for indicating that these terms best describe the main content of the document. Experiments by Zhu et al. showed that focusing on major topics yields more accurate similarity scores (Zhu et al., 2009). If a document has more than one major topic assigned to it, we take the average maximum match between the sets of major topics assigned to two documents as their overall similarity score.

The following example illustrates the calculation of MeSH-based similarities for two descriptors in a fictitious MeSH thesaurus. The left tree in Figure 1 shows the thesaurus that includes eight MeSH descriptors ($m_1 \ldots m_8$). One descriptor (m_4) occurs twice. To distinguish the variables used in the following formulas, we display the occurrences ($o_1 \ldots o_8$) of individual descriptors in the tree on the right.

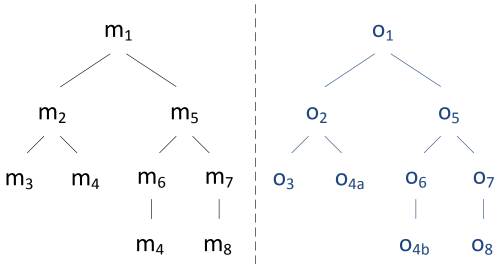

Figure 1: Exemplified MeSH taxonomy descriptors (left), occurrences (right).

The information contents of descriptors in the example calculate as follows. The total number of nodes N equals 9. Thus, the probabilities of occurrence are:

$$p(o_3) = p(o_{4a}) = p(o_{4b}) = p(o_8) = \tfrac{1}{9}\,;\, p(o_6) = p(o_7) = \tfrac{2}{9}\,;\, p(o_2) = \tfrac{3}{9}\,;\, p(o_5) = \tfrac{5}{9}\,;\, p(o_1) = 1.$$

The respective information contents are:

$IC(o_3) = IC(o_{4a}) = IC(o_{4b}) = IC(o_8) = 0.95$; $IC(o_6) = IC(o_7) = 0.65$; $IC(o_2) = 0.48$; $IC(o_5) = 0.26$; $IC(o_1) = 0$.

Let there be four documents $d_I, d_{II}, d_{III}, and\ d_{IV}$ with the following sets of MeSH descriptors assigned to them: $d_I \coloneqq \{m_3\}$; $d_{II} \coloneqq \{m_4\}$; $d_{III} \coloneqq \{m_6\}$; $d_{IV} \coloneqq \{m_3, m_7\}$ We exemplify the stepwise calculation of similarities for individual occurrences, descriptors, and lastly documents. Note that we use $o_s(o_n, o_m)$ to denote the closest common subsuming occurrence of o_n and o_m.

$$sim(o_{4b}, o_7) = \frac{2 \times IC\big(o_s(o_{4b}, o_7)\big)}{IC(o_{4b}) + IC(o_7)} = \frac{2 \times IC(o_5)}{IC(o_{4b}) + IC(o_7)} = \frac{2 \times 0.55}{0.95 + 0.65} = 0.69$$

$$sim(m_4, m_7) = sim(\{o_{4a}, o_{4b}\}, \{o_7\}) = \frac{\sum_{o_p \in \{o_{4a}, o_{4b}\}} \max\Big(sim\big(o_p, o_q\big)\Big) + \sum_{o_q \in \{o_7\}} \max\Big(sim\big(o_q, o_p\big)\Big)}{|\{o_{4a}, o_{4b}\}| + |\{o_7\}|}$$

$$= \frac{sim(o_{4a}, o_7) + sim(o_{4b}, o_7) + max\big(sim(o_{4a}, o_7), sim(o_{4b}, o_7)\big)}{2+1} = \frac{0 + 0.69 + max(0\,, 0.69)}{2+1} = \frac{1.38}{3} = 0.46$$

$$sim(d_{\mathrm{II}}, d_{\mathrm{IV}}) = \mathrm{sim}(M_{II}, M_{IV}) = sim(\{m_4\}, \{m_3, m_7\})$$

$$= \frac{\sum_{O(m_p) \in M_{\mathrm{II}}} \max\Big(sim\Big(O\big(m_p\big), O\big(m_q\big)\Big)\Big) + \sum_{O(m_q) \in M_{\mathrm{IV}}} \max\Big(sim\Big(O\big(m_q\big), O\big(m_p\big)\Big)\Big)}{|M_{\mathrm{II}}| + |M_{\mathrm{IV}}|}$$

$$= \frac{max\big(sim(m_4, m_3), sim(m_4, m_7)\big) + sim(m_4, m_3) + sim(m_4, m_7)}{1+2}$$

$$= \frac{max(0.33{,}0.46) + 0.33 + 0.46}{1+2} = \frac{1.25}{3} = 0.42$$

Table 6 lists the resulting MeSH-based similarities for all four documents in the example.

	$\mathbf{D_I}$	$\mathbf{D_{II}}$	$\mathbf{D_{III}}$	$\mathbf{D_{IV}}$
$\mathbf{D_I}$	■	0.33	0	0.67
$\mathbf{D_{II}}$	0.33	■	0.54	0.42
$\mathbf{D_{III}}$	0	0.54	■	0.27
$\mathbf{D_{IV}}$	0.67	0.42	0.27	■

Table 6: MeSH-based similarities for the example.

3.3.2 TREC Genomics

The organizers of the TREC Genomics track asked domain experts to define 28 information needs, i.e. questions comparable to: “What effect does a specific gene have on a certain biological process?”. Text passages contained within the document collection must provide an answer to the defined information needs. The organizers selected the text passages they presented to the expert judges by pooling the top-ranked text passages retrieved by each of the participating systems until 1,000 unique text passages were collected for each of the 28 information needs (28,000 text passages total). Since a document can contain more than one relevant passage, the experts judged passages from 11,638 documents.

The TREC Genomics track evaluated retrieval performance on three levels - the passage level, the aspect level, and the document level. Aspects were defined as “answers that covered a similar portion of a full answer to the topic question” (Hersh et al., 2006).

Depending on the research objective, one can derive different gold standards from these three performance measures. We use the information needs defined in the TREC Genomics collections and implemented a gold standard on the document level, because citation-based similarity measures, such as Bibliographic Coupling or Co-Citation operate on document level. All documents, which according to experts are relevant to a certain topic ID form a cluster. We sort the documents in each cluster according to the number of relevant passages they contain, as judged by the experts. Comparing the n documents that yield the highest scores using a certain similarity measure with the top-n scoring documents of a cluster can serve as a performance measure for relevance. A similarity measure performs well if its highest scoring documents comprise a high percentage of the highest scoring documents in the cluster.

Ensuring that similarity measures can only identify documents for which a comparison to the gold standard is feasible requires limiting the collection to documents that have expert judgments. The TREC Gen. collection includes expert judgments for 11,638 documents. Depending on the evaluation objective, this limitation of the test collection can be an unacceptable restriction.

Users of CITREC should also be aware that the pooling step of the TREC Genomic track may have introduced a bias to the gold standard. The TREC Genomics track exclusively evaluated text-based systems, which consequently delivered the results that became the input for the pooling step. Buckley et al. showed that the document pool of TREC 2005 and the gold standard derived from it exhibited a bias in favor of relevant documents that contained topic keywords in their title (Buckley et al., 2007). If the gold standard of the TREC Genomics collection exhibits a similar bias, it may punish citation-based similarity measures when retrieving and prominently ranking documents without topic title words.

3.3.3 User Survey

Because the gold standards described in the Sections 3.3.1 and 3.3.2 can be unsuitable for certain evaluations, we developed a web-based system for conducting surveys to create customized information needs and gold standards. The code of the system is open-source and included in the CITREC framework. The system allows users of CITREC to define information needs and create a gold standard by pooling, viewing, and rating document recommendations, which the system generates based on the different implemented similarity measures.

The survey tool offers three modes for selecting input documents for the pooling step. First, the user can instruct the system to choose documents from the test collection at random. Second, the user can specify a predefined list of documents as the input. Defining a list of input documents allows evaluating the performance of measures for documents that exhibit specific characteristics, e.g., frequently or rarely cited documents. Third, user study participants can self-select documents from the database. This last option allows participants can choose documents that fit their expertise.

During the rating step, user study participants can select their favorite and least favorite recommendations. To avoid bias, the individual similarity measure that generated the recommendation remains invisible to the study participants. As a second measure to avoid bias, the system randomizes the sequential display position of recommendations that originate from a particular similarity measure.

3.4 Tools for Results Analysis

A qualitative comparison of scores generated by different similarity measures is difficult, because the measures use ordinal scales that typically do not have an upper limit. Therefore, we cannot normalize measures to allow for meaningful direct comparisons. Co-Citation-based document similarities are a typical example of this problem. The Co-Citation measure considers two documents increasingly similar the more sources cite both documents together. Deriving a Co-Citation count that equals perfect similarity is impossible, because Co-Citations do not have an upper bound. Furthermore, the scale for Co-Citation-based similarity is ordinal. This means, one cannot determine the degree to which a document A is more similar to a document B than to a document C, given that document A was co-cited one, two or n times more frequently with document B than it was co-cited with document C.

There are different evaluation approaches that can partially deal with the outlined problems. The CITREC framework includes code for performing set-based comparisons of the top-n documents ranked according to a similarity measure. Comparing the ratio of relevant documents among the top-n similar documents identified by different similarity measures offers valuable and easily comprehensible information on the performance of the similarity measures. Additionally, CITREC includes code for calculating the Kendall’s tau rank correlation coefficient. The coefficient allows comparing the similarity of ordered datasets. For example, the coefficient allows comparing a gold standard result set to the result set as determined by a similarity measure if both result sets are ordered according to similarity score. We welcome researchers to add further analysis tools to CITREC.

4 Conclusion

Our review of prior work showed that citation-based similarity measures are important for many information retrieval tasks such as topic classification (see Table 1 on page 2), topic clustering (see Table 2 on page 3), literature recommendation (see Table 3 on page 3), and plagiarism detection (Gipp et al., 2014, Pertile et al., 2013). However, even for the basic citation-based similarity measures introduced more than 30 years ago, no consensus exists on their performance, e.g., for the suitability of using these measures in recommender systems.

We view the lack of a large-scale test collection and the absence of an accepted gold standard necessary for evaluating citation-based measures as the main reasons for the contradicting results of prior performance evaluations. For text-based similarity measures, comprehensive high quality evaluation frameworks that fulfill both criteria exist, e.g., provided by the Text REtrieval Conference (TREC) series.

This paper presents a framework, coined CITREC, and addresses the drawbacks of the currently used test collections. CITREC extends the PMC OAS and TREC Genomics ’06 collections by providing:

a) citation and reference information that includes the position of in-text citations;
b) code and pre-computed scores for 35 citation-based and text-based similarity measures;
c) two gold standards based on MeSH descriptors and the relevance feedback gathered for the TREC Genomics collection;
d) a web-based system that allows evaluating similarity measures on their ability to identify documents that are relevant to user-defined information needs;
e) tools to statistically analyze and compare the scores that individual similarity measures yield.

The purpose of the CITREC framework is to facilitate the evaluation of citation-based similarity measures. In a paper soon to be published, we use CITREC to perform a large-scale comparison of citation-based similarity measures regarding their ability to identify topically related documents.

Currently, CITREC exclusively contains biomedical and life science literature, which is notable for its rigorous and detailed presentation of prior work and for citing sources. Additional unique characteristics of biomedicine and the life sciences are the availability of the comprehensive literature index PubMed, which includes the MeSH classification scheme, and the common practice of authors to state PubMed identifiers for the references in their papers. Due to these characteristics of the subject areas covered by CITREC, evaluation results obtained using the framework in its current state may not be representative for other domains.

Our goal is to extend CITREC to include literature from other academic domains and sources, e.g., Wikipedia. The similarity measures and tools for deriving a gold standard implemented in CITREC are not yet complete and will be extended. We invite the academic community to contribute extensions to CITREC, e.g., by making available implementations of additional similarity measures or sharing results. To ease collaboration, we published all components of CITREC under open licenses (GNU Public License for code, Open Data Commons Attribution License for data) and offer them free of charge at:

www.sciplore.org/projects/citrec.

References

Ahlgren, P. and Colliander, C. (2009). Document–document Similarity Approaches and Science Mapping: Experimental Comparison of Five Approaches. *Journal of Informetrics*, 3(1):49–63.

Ahlgren, P. and Jarneving, B. (2008). Bibliographic Coupling, Common Abstract Stems and Clustering: A Comparison of Two Document-document Similarity Approaches in the Context of Science Mapping. *Scientometrics*, 76:273–290. 10.1007/s11192-007-1935-1.

Batet, M., Sánchez, D., and Valls, A. (2010). An Ontology-based Measure to Compute Semantic Similarity in Biomedicine. *Journal of Biomedical Informatics*, 44(1):118–125.

Bellogin, A., Castells, P., Said, A., and Tikk, D. (2013). Workshop on Reproducibility and Replication in Recommender Systems Evaluation - RepSys. In *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys), Hong Kong, China*, pages 485–486.

Boyack, K. W. and Klavans, R. (2010). Co-Citation Analysis, Bibliographic Coupling, and Direct Citation: Which Citation Approach Represents the Research Front Most Accurately? *Journal of the American Society for Information Science and Technology*, 61(12):2389–2404.

Boyack, K. W., Small, H., and Klavans, R. (2012). Improving the Accuracy of Co-citation Clustering Using Full Text. In *Proceedings of 17th International Conference on Science and Technology Indicators*.

Buchanan, R. A. (2006). Accuracy of Cited References: The Role of Citation Databases. *College & Research Libraries*, 67(4):292–303.

Buckley, C., Dimmick, D., Soboroff, I., and Voorhees, E. (2007). Bias and the Limits of Pooling for Large Collections. *Inf. Retr.*, 10(6):491–508.

Cao, M. and Gao, X. (2005). Combining contents and citations for scientific document classification. In *AI 2005: Advances in Artificial Intelligence*, volume 3809 of *Lecture Notes in Computer Science*, pages 143–152. Springer, Berlin Heidelberg.

Couto, T., Cristo, M., Gonçalves, M. A., Calado, P., Ziviani, N., Moura, E., and Ribeiro Neto, B. (2006). A Comparative Study of Citations and Links in Document Classification. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 75–84. ACM.

Eto, M. (2012). Spread co-citation relationship as a measure for document retrieval. In *Proceedings of the Fifth ACM Workshop on Research Advances in Large Digital Book Repositories and Complementary Media*, pages 7–8, Maui, Hawaii, USA. ACM.

Eto, M. (2013). Evaluations of context-based co-citation searching. *Scientometrics*, 94(2):651–673.

Geva, S., Kamps, J., Lethonen, M., Schenkel, R., Thom, J., and Trotman, A. (2010). Overview of the INEX 2009 Ad Hoc Track. In Geva, S., Kamps, J., and Trotman, A., editors, *Focused Retrieval and Evaluation*, volume 6203 of *Lecture Notes in Computer Science*, pages 4–25. Springer.

Gipp, B. and Beel, J. (2009). Citation Proximity Analysis (CPA) - A new approach for identifying related work based on Co-Citation Analysis. In Larsen, B. and Leta, J., editors, *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09)*, volume 2, pages 571–575, Rio de Janeiro (Brazil). International Society for Scientometrics and Informetrics. ISSN 2175-1935.

Gipp, B., Meuschke, N., and Breitinger, C. (2014). Citation-based Plagiarism Detection: Practicability on a Large-scale Scientific Corpus. *Journal of the American Society for Information Science and Technology*, 65(2):1527–1540.

Hersh, W., Cohen, A. M., Roberts, P. M., and Rekapalli, H. K. (2006). TREC 2006 Genomics Track Overview. In *Proceedings of the 15th Text Retrieval Conference*.

Janssens, F., Zhang, L., De Moor, B., and Glänzel, W. (2009). Hybrid Clustering for Validation and Improvement of Subject-classification Schemes. *Information Processing and Management*, 45:683–702.

Jarneving, B. (2005). A Comparison of Two Bibliometric Methods for Mapping of the Research Front. *Scientometrics*, 65(2):245–263.

Lachica, R., Karabeg, D., and Rudan, S. (2008). Quality, Relevance and Importance in Information Retrieval with Fuzzy Semantic Networks. In *Proceedings of the 4th International Conference on Topic Maps Research and Applications*, volume 7 of *Leipziger Beiträge zur Informatik*, Leipzig, Germany.

Li, P., Li, Z., Liu, H., He, J., and Du, X. (2009). Using link-based content analysis to measure document similarity effectively. In *Advances in Data and Web Management*, volume 5446 of *Lecture Notes in Computer Science*, pages 455–467. Springer Berlin Heidelberg.

Lin, D. (1998). An Information-Theoretic Definition of Similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML'98)*, pages 296–304, Madison, Wisconsin, USA.

Lin, J. and Wilbur, W. J. (2007). PubMed Related Articles: a Probabilistic Topic-based Model for Content Similarity. *BMC Bioinformatics*, 8(1):423.

Lipinski, M., Yao, K., Breitinger, C., Beel, J., and Gipp, B. (2013). Evaluation of Header Metadata Extraction Approaches and Tools for Scientific PDF Documents. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, pages 385–386, New York, NY, USA. ACM.

Liu, X., Yu, S., Janssens, F. A. L., Glänzel, W., Moreau, Y., and De Moor, B. (2010). Weighted Hybrid Clustering by Combining Text Mining and Bibliometrics on a Large-Scale Journal Database. *Journal of the American Society for Information Science and Technology*, 61(6):1105–1119.

Liu, X., Yu, S., Moreau, Y., De Moor, B., Glänzel, W., and Janssens, F. A. L. (2009). Hybrid Clustering of Text Mining and Bibliometrics Applied to Journal Sets. In *Proceedings of the SIAM International Conference on Data Mining*, pages 49–60, Sparks, NV, USA.

Lu, W., Janssen, J., Milios, E., Japkowicz, N., and Zhang, Y. (2007). Node similarity in the citation graph. *Knowledge and Information Systems*, 11(1):105–129.

Lykke, M., Larsen, B., Lund, H., and Ingwersen, P. (2010). Developing a Test Collection for the Evaluation of Integrated Search. In *Advances in Information Retrieval*, volume 5993 of *Lecture Notes in Computer Science*, pages 627–630. Springer.

Pertile, S. L., Rosso, P., and Moreira, V. P. (2013). Counting Co-occurrences in Citations to Identify Plagiarised Text Fragments. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, volume 8138 of *Lecture Notes in Computer Science*, pages 150–154. Springer.

Piao, S. and Tsuruoka, Y. (2008). A Highly Accurate Sentence and Paragraph Breaker. Online Source. Retrieved jan 28, 2014 from: http://text0.mib.man.ac.uk:8080/scottpiao/sent_detector.

Resnik, P. et al. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *International Joint Conference on Artificial Intelligence*, volume 14, pages 448–453.

Saracevic, T. (2006). *Relevance: a Review of the Literature and a Framework for Thinking on the Notion in Information Science. Part II*, volume 30 of *Advances in Librarianship*, chapter 1, pages 3–71. Emerald Group Publishing Limited.

Shibata, N., Kajikawa, Y., Takeda, Y., and Matsushima, K. (2009). Comparative Study on Methods of Detecting Research Fronts Using Different Types of Citation. *Journal of the American Society for Information Science and Technology*, 60:571–580.

U. S. National Library of Medicine (2014). Introduction to MeSH - 2014. Online Source. Retrieved aug 30, 2014 from: http://www.nlm.nih.gov/mesh/introduction.html.

Yoon, S.-H., Kim, S.-W., and Park, S. (2011). C-Rank: a Link-based Similarity Measure for Scientific Literature Databases. *arXiv.org Computing Research Repository*, abs/1109.1059:1–11.

Zhu, S., Yu, K., Chi, Y., and Gong, Y. (2007). Combining Content and Link for Classification Using Matrix Factorization. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'07)*, pages 487–494, Amsterdam, The Netherlands.

Zhu, S., Zeng, J., and Mamitsuka, H. (2009). Enhancing MEDLINE Document Clustering by Incorporating MeSH Semantic Similarity. *Bioinformatics*.

Table of Figures

Table of Tables

Citation for this Paper

Citation Example:

B. Gipp, N. Meuschke, and M. Lipinski. CITREC: An Evaluation Framework for Citation-Based Similarity Measures based on TREC Genomics and PubMed Central. In *Proceedings of the iConference 2015*, Newport Beach, California, Mar. 24-27 2015. URL http://ischools.org/the-iconference/.

Bibliographic Data:

RIS Format	**BibTeX Format**

TY - CONF
AU - Gipp, Bela
AU - Meuschke, Norman
AU - Lipinski, Mario
T1 - CITREC: An Evaluation Framework for Citation-Based Similarity Measures based on TREC Genomics and PubMed Central
T2 - Proceedings of the iConference 2015
AD - Newport Beach, California
Y1 - 2015/march 24-27
UR - http://ischools.org/the-iconference/
	@INPROCEEDINGS{Gipp15b,
	author = {Gipp, Bela and Meuschke, Norman and Lipinski, Mario},
	title = {CITREC: An Evaluation Framework for Citation-Based Similarity Measures based on TREC Genomics and PubMed Central},
	booktitle = {Proceedings of the iConference 2015},
	year = {2015},
	address = {Newport Beach, California},
	month = Mar # { 24-27},
	url = {http://ischools.org/the-iconference/}
	}