Document-to-document relevance assessment for TREC Genomics Track 2005

Olga Giraldo [1], María Fernanda Cadena [1,2,3], Andrea Robayo-Gama [1,2,3], Dhwani Solanki [1,4], Tim Fellerhoff [1,5], Lukas Geist [1,6], Rohitha Ravinder [1,4], Muhammad Talha [1,6], Dietrich Rebholz-Schuhmann [1,7] and Leyla Jael Castro [1]

[1] *ZB MED Information Centre for Life Sciences, Gleueler Str. 60, Cologne, 50931, Germany*
[2] *Institute of Molecular Medicine and Cell Research, University of Freiburg, Stefan-Meier-Str. 17, Freiburg im Breisgau, 79104, Germany*
[3] *Facultad de Farmacia y Bioquímica, Universidad de Buenos Aires, Junín 956,Buenos Aires, C1113AAD, Argentina*
[4] *Bonn-Aachen International Centre for Information Technology (B-IT), University of Bonn, Friedrich-Hirzebruch-Allee 6, Bonn, 53115, Germany*
[5] *Heinrich-Heine University Düsseldorf, Universitätsstraße 1, Düsseldorf, 40225, Germany*
[6] *Hochschule Bonn-Rhein-Sieg, Grantham-Allee 20, Sankt Augustin, 53757, Germany*
[7] *University of Cologne, Albertus-Magnus-Platz, Cologne, 50923, Germany*

Abstract

Here we present a doc-2-doc relevance assessment performed on a subset of the TREC Genomics Track 2005 collection. Our approach includes an experimental set up to manually assess doc-2-doc relevance and the corresponding analysis done on the results obtained from this experiment. The experiment takes one document as a reference and assesses a second document regarding its relevance to the reference one. The consistency of the assessments done by 4 domain experts was evaluated. The lack of agreement between annotators may be due to: i) The abstract lacks key information and/or ii) Lack of experience of the annotators in the evaluation of some topics.

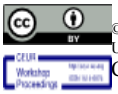
Keywords

Relevance assessment, literature manual curation, document similarity

1. Introduction

The TREC Genomics Track 2005 [1] provides a collection of document-to-topic relevance assessments for Medline abstracts. This collection has commonly been used also for document-to-document related tasks [2,3,4]; however, to the best of our knowledge, no analysis has been done regarding the suitability of such collection for such tasks, e.g., similarity or relevance assessment between a pair of documents. Our doc-2-doc relevance analysis aims at filling this gap. We take one document as a "reference article" while a second one is evaluated wrt its relevance to the referenced document. In this experiment the user is engaged in the evaluation process as a way to achieve better results.

Proceedings Semantic Web Applications and Tools for Healthcare and Life Sciences, February 13–16, 2023, Basel, Switzerland
EMAIL: ljgarcia@zbmed.de (A. 10)
ORCID: 0000-0003-2978-8922 (A.1); 0000-0002-5915-8895 (A.2); 0000-0002-8725-1317 (A.5); 0000-0002-2910-7982 (A.6); 0000-0002-1018-0370 (A.9); 0000-0003-3986-0510 (A.10)



2. Methodology

Selection of topics and documents. TREC 2005 Genomics track included a total of 50 topics. For the document-to-document relevance assessment we wanted to have 15 documents to assess per each reference document, at least 10 documents judged as definitely relevant wrt the TREC topic, and no more than 80 relevant articles (either definitely or partially relevant). The goal was having a sample covering about 10% of the relevant articles for the document-to-document assessment task. This gave us a total of 16 topics which were further reduced to 8 topics due to time constraints and expertise of the annotators on the different TREC topics. These topics contained a total of 42 reference documents and 630 documents to be assessed.

Development of an in-house annotation tool and a corpus of documents. The tool [5] presents a corpus of documents organized by topics. The corpus is based on the TREC 2005 Genomics track with some pre-processing to obtain the reference documents and the ones to be assessed against them. The annotation workflow is shown on Figure 1.

Figure 1: Assessments steps in the relevance assessment tool

Training sessions. Virtual sessions were organized with the participants in the evaluation of documents to train them in the use of the tool, and to solve doubts. The meetings were carried out by using Zoom. The participants were 4 domain experts with expertise in life science and/or bioinformatics.

Relevance assessment by domain experts. In the tool, the documents were organized by topics. Where each topic includes "N" number of reference articles. Each reference article includes 15 documents (or evaluation articles), to be assessed. Only title and abstract are available. The relevance assessment possible values are as follows: i) Relevant to the reference article, meaning "Yes, the user wants to get a hold of the full-text as it is definitely relevant to their research". ii) Partially relevant to the reference article, meaning "Looks promising but not sure yet. The user will keep the PMID just in case, as a maybe". iii) Non-relevant to the reference article, meaning "Not worth giving it a second look at all".

Analysis of results. Here the consistency of the assessments done by the 4 domain experts was evaluated. We focused on inter-annotator agreement.

3. Results

Documents assessed. A total of 630 "evaluation articles" classified in 8 topics were assessed by 4 annotators. The evaluation articles are distributed into 42 reference articles (15 documents per reference article). The full data is available online [6].

Inter-annotator agreement. Annotators rated the documents into three categories (2 definitely relevant, 1 partially relevant, 0 non-relevant). We observed that the four annotators all rated as "definitely relevant" 35 of the 630 assessed documents evaluated (5.56%). Similarly, the four annotators agreed on the rating as "partially relevant" on 6 documents (0.95%); and others 123 documents (19.52%) were rated as "non-relevant", giving us a total agreement among the four annotators for 164 articles (26.03%). The Fleiss Kappa results are distributed into three levels of agreement: "Poor", with values from -0.1708 to 0.1885. "Fair", with values from 0.2214 to 0.375; and "Moderate" with values from 0.4564 to 0.5328. From the 42 reference articles, 24 got a Fleiss Kappa corresponding to "Poor", 14 corresponding to "Fair", and 5 corresponding to "Moderate". A table summarizing the results about the inter-annotator agreement is available online [7].

Table 1
Agreement across annotators

Relevance categories	Full agreement among four annotators	Agreement among three annotators
Definitely relevant	35	72
Partially relevant	6	40
Non-relevant	123	109

4. Discussion and conclusions

This work is about the analysis done to the results obtained in an experiment focused on evaluating the relevance between two articles. The experiment takes one document as a reference and assesses a second document regarding its relevance to the reference one. The methodological aspects involved the participation of four domain experts, who used an in-house annotation tool tailored to the initial TREC data and the task at hand. The lack of agreement between annotators may be due to: i) The abstract lacks key information. For example, the objective, main results or conclusions. In this case, the reader has to search the entire document for the missing information. ii) Lack of experience of the annotators in the evaluation of some topics. In this case, the reader must search for more information on the web on a topic to better understand the document to be evaluated. Both implications are time consuming. In order to overcome those limitations and improve the results, we propose as a future work to extend the time required in the evaluation tasks and/or extend the number of annotators to cover the lack of experience in some topics.

5. Acknowledgements

This work is part of the STELLA project funded by DFG (project no. 407518790). This work was supported by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A).

6. References

[1] Hersh W, Cohen A, Yang J, Bhupatiraju RT, Roberts P, Hearst M. TREC 2005 Genomics Track Overview. : 26.

[2] Lin J, Wilbur WJ. PubMed related articles: a probabilistic topic-based model for content similarity. BMC Bioinformatics. 2007;8: 423. doi:10.1186/1471-2105-8-423

[3] Garcia Castro LJ, Berlanga R, Garcia A. In the pursuit of a semantic similarity metric based on UMLS annotations for articles in PubMed Central Open Access. Journal of Biomedical Informatics. 2015;57: 204–218. doi:10.1016/j.jbi.2015.07.015

[4] Wei W, Marmor R, Singh S, Wang S, Demner-Fushman D, Kuo TT, Hsu CN, Ohno-Machado L. Finding Related Publications: Extending the Set of Terms Used to Assess Article Similarity. AMIA Jt Summits Transl Sci Proc. 2016 Jul 20;2016:225-34.

[5] Talha M, Geist L, Fellerhoff T, Ravinder R, Giraldo O, Rebholz-Schuhmann D, et al. TREC-doc-2-doc-relevance assessment interface. Zenodo; 2022. doi:10.5281/zenodo.7341391

[6] Giraldo O, Solanki D, Cadena F, Robayo-Gama A, Rebholz-Schuhmann D, Castro LJ. Document-to-document relevant assessment for TREC Genomics Track 2005. Zenodo; 2022. doi:10.5281/zenodo.7324822

[7] Giraldo O, Solanki D, Rebholz-Schuhmann D, Castro LJ. Fleiss kappa for doc-2-doc relevance assessment. Zenodo; 2022. doi:10.5281/zenodo.7338056