View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Proceeding of the Electrical Engineering Computer Science and Informatics

Obtaining Reference's Topic Congruity in Indonesian Publications using Machine Learning Approach

Sam Farisa Chaerul Haviana
Department of Informatics Engineering
Universitas Islam Sultan Agung
Semarang, Indonesia
sam@unissula.ac.id

Imam Much Ibnu Subroto
Department of Informatics Engineering
Universitas Islam Sultan Agung
Semarang, Indonesia
imam@unissula.ac.id

***Abstract*—There are some criteria on how an article is categorized as a good article for publications. It could depend on some aspect like formatting and clarity, but mainly it depends on how the content of the article is constructed. The consistency of the topic that the article was written could show us how the authors construct the main idea in the article content. One indication that shows this consistency is congruity in the article's topic and the topic of literature or reference cited in the document listed in the bibliography. This works attempting to automate the topic detection on the article's references then obtain the congruity to the article title's topic through metadata extraction and text classification. This is done by extracting metadata of an article file to obtain all possible reference title using GROBID than classify the topic using a supervised classification model. We found that some refinements in the whole approach should be considered in the next step of this work.**

Keywords—Text classification, Bibliography extraction, Reference congruity

I. INTRODUCTION

Indonesia is experiencing an increase in publications number in the last several years. Figure 1 shows the increase of numbers of documents produced by Indonesian researchers published in journals and conferences, at least in the last three years, it is significantly emerges.

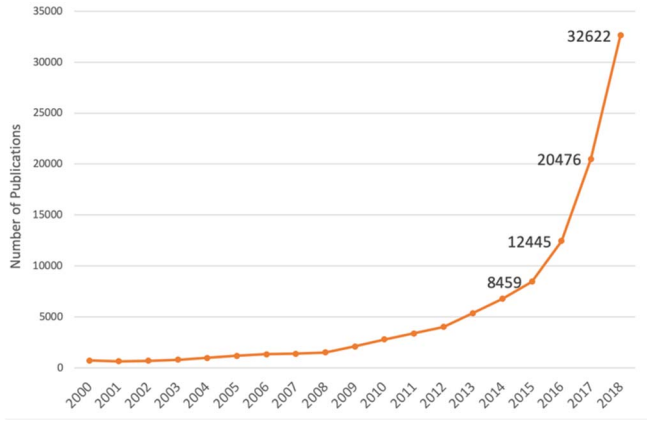

Fig. 1. Number of Indonesian publications by year [1]

Even though there are positive effects by the increase of quantity but there is also a major drawback in case of quality. This increase of publication quantity does not mean also the increase of quality of publications. More numbers of publications the more possibilities that lesser good articles were published. This has to be a concern for the authors and moreover for the policymakers. The efforts to distinguish the good articles from the lesser ones need to be done, even though in general there is review process that has been done by the reviewer when the articles were sent to the publisher for publication.

One of the decisive things of good research papers or publications is the selection of reference. As reference playing the vital role for convincing the readers about the research main idea. In [2] the summary of various roles of reference has been classified as shown in Table I.

TABLE I. SUMMARIZES VARIOUS ROLES OF REFERENCES AND THEIR CLASSIFICATIONS [2]

Roles
Support of an argument by referencing an authoritative piece of writing and/or research
Development of parallel or branching opinions without disturbing the flow of the main text
Provision of details in order to check the genuineness and quality of the references cited
Shortening of the paper by referring readers elsewhere for details

References assist in validating a paper, improve its readability, and direct the interested reader to other appropriate material [3]. The consistency of reference ideas, direction or topics of reference become important in producing good research papers of publications. This could be not an easy task for reviewers to clearly determine whether the article's references belong to the main topic or direction of the article. This study proposing an approach in automating the finding of congruity of article topic and its reference's topic through metadata extraction then later processed using a machine learning approach. This includes text processing in the metadata and text classification using several known methods.

II. METHODOLOGY

A. Related Works and Tools

The first step for getting information of a PDF document is by extracting the document's metadata. Extracting includes the process of finding the structure of the document's content and parsing the text. In the past few years, there have been some good studies and approaches to extracts PDF document metadata. Those results some techniques or combination of techniques including CRF (Conditional Random Fields), PDF content parsing, and machine learning approach. Some of them were introduced in [4]-[9] and in many other studies. There are also some good tools built and implemented for PDF

document's metadata extraction as stated in [10] and other tools that this study evaluates shown in Table II.

GROBID (Generation Of Bibliographic Data) [11], and CERMINE (Content Extractor And Miner) [12] was evaluated in this works. GROBID was evaluated works better than the others as stated in [13]. The best performing out-of-the-box tool is GROBID, followed by CERMINE and ParsCit [14]. However, CERMINE was evaluated and considered adequate good for extracting metadata of PDFs documents. This study also evaluating those tools from the simplicity of usage and integration with other tools or used as a library for development. GROBID and CERMINE have better integration and more convenient for development rather than other tools. This study utilizes GROBID for implementation, as this tool stated the best in [13], simple, and more convenient to be integrated with other tools.

TABLE II. TOOLS FOR PDF DOCUMENTS METADATA EXTRACTION

Tools	Approaches Used
Docears PDF Inspector	Style information analysis
GROBID	CRF
Mendeley Desktop	SVM
ParsCit	CRF
PDF Meat	Queries Google Scholar, pdftotext
PDFSSA4MET	Structure/Syntax analysis of XML
SciPlore Xtract	Style information analysis of XML
SVMHeaderParse	SVM
Zotero	Queries Google Scholar
CERMINE	Layout, style and word pattern analysis

The next step proposed in this study was to classify the topic of articles as well as classify the topic of references in the bibliography. This can be achieved by using a classification method suitable for text classification. This study evaluates well-known classification methods SVM and Naïve Bayes for text classification. The article topic and the references topic predicted in classification then compared to get how many topics are the same topic or under the same topic cluster and how many topic of reference were too distant from the article topic.

B. Flowchart and Data

The proposed implementation flowchart for obtaining topic congruity in this study shown in Figure 2. For training purpose in classification, we randomly collected 1000 article's titles. This data was collected from Garuda (http://garuda.ristekdikti.go.id/) that indexed more than 750.000 articles of Indonesian publication and mostly written in Bahasa. The data consist of 10 selected topics that came from Garuda article's subjects. Some topics were considered in the same cluster, such as Computer Science & IT topics and Electrical & Electronics Engineering topics. In the result of classification, those topics were also considered as matched topics with article topic. The PDF articles as input were also downloaded randomly from Garuda.

We selected 100 PDF articles in each selected topics used in this study. Those PDF documents also consisted of various content structure or format. This content structure depends on the writing style of where the article published.

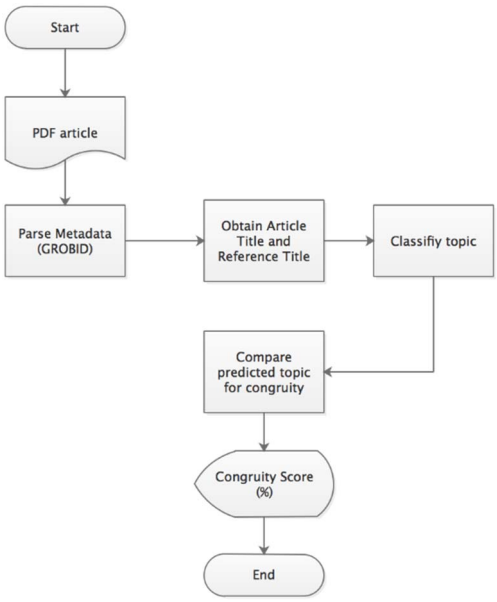

Fig. 2. Flowchart of implementation

Table III showing the topics and number of article's titles in each topic collected for training and testing the classification model.

TABLE III. SELECTED TOPICS

Topic	No. of Titles	Same Cluster with
Electrical & Electronics Engineering	100	Computer Science & IT
Computer Science & IT	100	Electrical & Electronics Engineering
Educations	100	-
Arts and Humanities	100	-
Law, Crime, Criminology and Criminal Justice	100	-
Economic, Econometrics & Finance	100	-
Public Health	100	-
Mechanical Engineering	100	-
Environmental Science	100	-
Chemical Engineering, Chemistry and Bioengineering	100	-

C. Metadata Extraction

Our implementation utilizes GROBID Application Programming Interface (API) to extract PDF articles metadata. We focus on getting the title of the article and the list of references. As GROBID generate TEI-encoded (http://www.tei-c.org/) XML document, we also need to parse this XML document into and obtain title and list of references as a text for later used as input in the classification process. Instead of parsing PDF document references only, we parse full document using GROBID *processFullText* API. Figure 3 shows an example of the XML node that consists of article title node. The article's title obtained from the *title* node under *titleStmt* node in the generated XML. Figure 4 shows a more complex node as it is shown the structure of bibliography of the document. The bibliographies can be obtained from *listBibl* node in the generated XML. Each reference can be obtained under *biblStruct* node. And for the title of each reference can be obtained title node inside the *biblStruct* node.

Note that each reference node may be resulting from a different structure of XML tag.

```
<titleStmt>
        <title
            level="a"
            type="main">

            TEKNIK SERANGAN PADA APLIKASI BERBASIS RUBY ON RAILS

        </title>
</titleStmt>
```

Fig. 3. XML node for article title

The child of *biblStruct* node can be resulting from a different tag which is *monogr* and or *analytic*. The generated XML does not always in the same structure of nodes. This makes the parsing process challenging.

```
<listBibl>

    <biblStruct xml:id="b12">
        <analytic>
            <title level="a" type="main">A framework for
                recognizing the simultaneous aspects of
                American sign language
            </title>
            <author>
                <persName xmlns="http://www.tei-c.org/ns/1.0">
                    <forename type="first">Vogler</forename>
                </persName>
            </author>
            <author>
                <persName xmlns="http://www.tei-c.org/ns/1.0">
                    <forename type="first">C</forename>
                    <surname>Metaxas</surname>
                </persName>
            </author>
        </analytic>
        <monogr>
            <title level="j">Computer Vision and
                Image Understanding
            </title>
            <imprint>
                <biblScope unit="volume">81</biblScope>
                <biblScope unit="issue">3</biblScope>
                <biblScope unit="page" from="358" to="384" />
                <date type="published" when="2001" />
            </imprint>
        </monogr>
    </biblStruct>

</listBibl>
```

Fig. 4. XML node list of references or bibliography

In order to handle this unpredicted XML structure, we utilize DOM parser to help to find the right node. This process then resulting text data of article title and an array of the text of reference title. This text data later used as input in the classification process.

D. Topic Classification

The classification model builds using a well-known algorithm in text classification. In this study, SVM and Naïve Bayes were utilized for classification method. The classification helps to find the topic for every reference title. We are using the title only as input for the classification, and this is quite challenging as the title of the document might be consist of only a few words. The classes found later compared with the topic class of the article itself. In SVM classifier we have evaluated four kernels, Linear, Polynomial, RBF and Sigmoid. Sigmoid kernel works slightly better than the other three in our evaluation with the gamma parameter is set to 0.

Naïve Bayes was also implemented and results in slightly different accuracy with SVM in our implementation. In average both methods were only differ very few points in accuracy and processing time. Figure 5 shows the flowchart of data pre-processing and building the classification model. For data pre-processing, all titles collected were transformed into a vector of word token count. Then using Term Frequency-Inverse Document Frequency (TF-IDF) before random split the data for training and testing. Random split divide data into 90% of training data and 10% of testing data. The validation is done to get the accuracy of the classification model by comparing actual topics to the predicted topics. Training and validation process has been done several times to get the confident accuracy of the model. The last step is to save the classification model for later implementation.

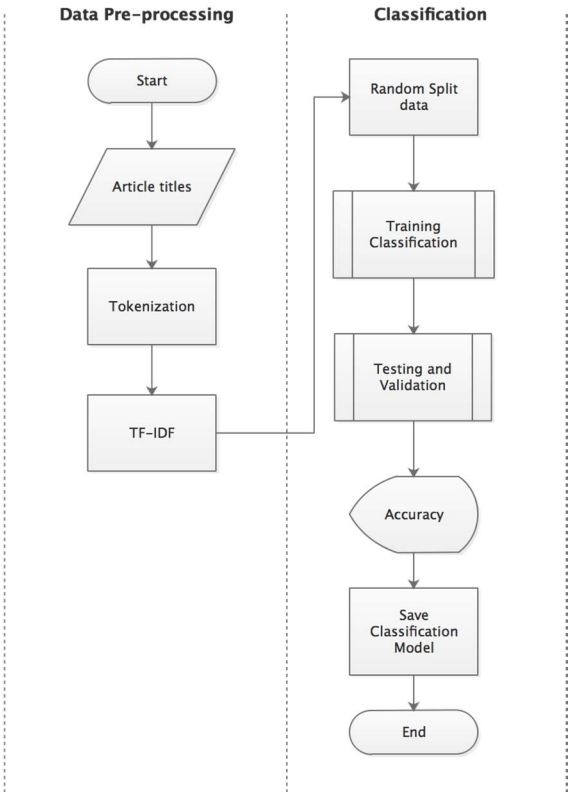

Fig. 5. Flowchart of data pre-processing and classification for topic classification

E. Results and Discussion

We build a web application for this study as shown in Figure 6.

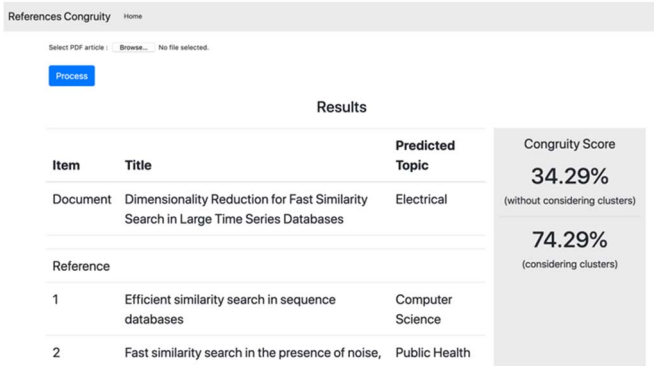

Fig. 6. Reference's topic congruity web application

Due to time efficiency and computation resource limitations, we limit the input of PDF file into one file per

process. This also helps us easily read the result of the classification and congruity score. The congruity score counts the matched topics or topic's cluster of article's references with the article topic then get the percentage by dividing the matched references with the number of parsed references. The data we are using is relatively small and this affects the classification. The pre-processing steps for feature extraction in title's text was also not adequate effective, as some topics provide more features in common. There are also two other factors that can affect accuracy in this work. First, the title of an article can be very short, consists of only two or three words. Second, there are many articles which bias in categorization, for example, some articles in Arts and Humanities could also categorize as Education, which is considering as not in the same cluster. Those factors lead to higher confusion in classification.

III. Conclusions

The proposed approach is still needed more refinement, as the result was below our expectation. The accuracy is very low, around 30% on the article topic without considering topic cluster, and around 70% on the article topic considering topic cluster. We conclude that this result implies more refinement before it is implemented.

Future Works

Considering this work as the first step, there is more to do in refining this proposed study. For feature extraction, we need some other features. Other than the article title, we looking forward to also classify the article based on its abstract. In term of metadata parsing, GROBID is doing very good in parsing well-known paper format but less good on many Indonesian paper format. More training on GROBID should give opportunity for better results. And last, we are looking forward to evaluating more classification methods to get more convincing accuracy.

Acknowledgment

We would like to thanks all faculty members of Fakultas Teknologi Industri Universitas Islam Sultan Agung for supporting. And we immensely grateful to Sinta and Garuda development team for providing research data to us.

References

[1] SINTA, "ASEAN Benchmarking," Science and Technology Index, 2019. [Online]. Available: http://sinta2.ristekdikti.go.id/home/benchmark. [Accessed: 22-Apr-2019].

[2] B. Eunson, Writing and presenting reports / Baden Eunson. John Wiley & Sons Brisbane, 1994.

[3] D. M. Taylor, "The appropriate use of references in a scientific research paper," Emergency Medicine, vol. 14, no. 2. pp. 166–170, 2002.

[4] M. Ohta, D. Arauchi, A. Takasu, and J. Adachi, "CRF-based bibliography extraction from reference strings focusing on various token granularities," Proc. - 10th IAPR Int. Work. Doc. Anal. Syst. DAS 2012, pp. 276–281, 2012.

[5] P. Yin, M. Zhang, Z. Deng, and D. Q. Yang, "Metadata extraction from bibliographies using bigram HMM," Proc. Int'l Conf. Asian Digit. Libr. (ICADL 2004). LNCS, vol. 3334, pp. 310–319, 2004.

[6] F. Peng and A. McCallum, "Accurate Information Extraction from Research Papers using Conditional Random Fields," Proc. Hum. Lang. Technol. Conf. North Am. Chapter Assoc. Comput. Linguist. HLTNAACL, vol. 2004, pp. 329–336, 2004.

[7] Z. Guo and H. Jin, "Reference metadata extraction from scientific papers," in Parallel and Distributed Computing, Applications and Technologies, PDCAT Proceedings, 2011, pp. 45–49.

[8] O. Saleem and S. Latif, "Information extraction from research papers by data integration and data validation from multiple header extraction sources," Proc. World Congr. Eng. Comput. Sci. WCECS 2012, vol. I, pp. 215–219, 2012.

[9] Y. Hu, H. Li, Y. Cao, L. Teng, D. Meyerzon, and Q. Zheng, "Automatic extraction of titles from general documents using machine learning," Inf. Process. Manag., vol. 42, no. 5, pp. 1276–1293, 2006.

[10] M. Lipinski, K. Yao, C. Breitinger, J. Beel, and B. Gipp, "Evaluation of header metadata extraction approaches and tools for scientific PDF documents," Proc. 13th ACM/IEEE-CS Jt. Conf. Digit. Libr. - JCDL '13, p. 385, 2013.

[11] P. Lopez, "GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2009, vol. 5714 LNCS, pp. 473–474.

[12] D. Tkaczyk, P. Szostek, M. Fedoryszak, P. J. Dendek, and Ł. Bolikowski, "CERMINE: Automatic extraction of structured metadata from scientific literature," Int. J. Doc. Anal. Recognit., vol. 18, no. 4, pp. 317–335, 2015.

[13] D. Tkaczyk, A. Collins, P. Sheridan, and J. Beel, "Machine Learning vs. Rules and Out-of-the-Box vs. Retrained: An Evaluation of Open-Source Bibliographic Reference and Citation Parsers," in Proceedings of ACM JCDL, 2018, p. 10.

[14] I. G. Councill, C. L. Giles, and M.-Y. Kan, "ParsCit: An open-source CRF Reference String Parsing Package," Int. Lang. Resour. Eval., no. 3, pp. 661–667, 2008.