Semantic Information Retrieval for Scientific Experimental Papers with Knowledge based Feature Extraction

Nur Rosyid Mubtadai[1], Ali Ridho Barakbah[2], Afrida Helen[3]
Graduate School of Engineering Technology
Department of Information and Computer Engneering
Politeknik Elektronika Negeri Surabaya
Email: rosyid@pens.ac.id[1], ridho@pens.ac.id[2], helen@pens.ac.id[3]

Abstrack – Along with the times, demands for information retrievals in scientific papers have also increased. Regarding experimental scientific papers, researchers have difficulty in searching for information on experimental scientific papers because information retrieval engines have limitations in the search process due to text mining-based feature extraction of the entire text, while experimental types of scientific paper have specific contents, which should have a different treatment in feature extraction. In this paper, we propose a new system for information retrieval on experimental scientific papers. This system consists of 4 main functions: (1) Specific content-based feature extraction, (2) Classification model, (3) Context-based subspace selection, and (4) Context-dependent similarity measurement. In feature extraction, our system extracts feature category in experimental scientific papers with specific content-based features, which are data, problem, method and result. To perform the applicability of our proposed system, we tested 77 papers in the dataset with the Leave-One-Out validation model with several classification algorithm (Nearest Neighbour, Naive Bayes, Support Vector Machine and Decision Tree) and on average performed 66.65% precision rate and accuracy of 76,18% precision rate. We also made the experiment on the similarity, our proposed system performed 79.17% accuracy rate
Keywords - Scientific experimental paper, Context-base subspace selection, Context-dependent similarity measurement.

Intisari **-** Seiring dengan perkembangan zaman permintaan pencarian informasi dalam makalah ilmiah juga meningkat. Mesin pencari informasi yang ada saat ini memiliki keterbatasan dalam proses pencarian berdasarkan ekstraksi fitur berbasis *text-mining* dari seluruh teks, sedangkan jenis makalah ilmiah eksperimental memiliki konten spesifik. Dalam makalah yang kami usulkan sistem untuk pengambilan informasi pada makalah ilmiah eksperimental. Sistem terdiri dari 4 fungsi: (1) Ekstraksi fitur berbasis konten, (2) Model klasifikasi, (3) Pemilihan subruang berbasis konteks, dan (4) Pengukuran kesamaan berdasar pada konteks. Dalam Pemilihan Subruang Berbasis Konteks, sistem melakukan pengurangan dimensi dengan pemilihan subruang berbasis konteks yang dipilih oleh pengguna. Untuk mendapatkan hasil pencarian akhir, kami mengukur kesamaan konteks dengan membangun metrik dataset berdasar konteks ke paper. Untuk melakukan penerapan sistem yang kami usulkan, kami menguji 77 makalah dalam dataset dengan model validasi *Leave-One-Out* dengan beberapa algoritma klasifikasi (Nearest Neighbor, Naive Bayes, Support Vector Machine, dan Decision Tree) dan rata-rata melakukan presisi 66,65% tingkat dan akurasi tingkat presisi 76,18%. Kami juga melakukan percobaan pada pengukuran kesamaan dengan memberikan queri paper dan konten yang diinginkan (data, hasil, metode, dan masalah) sebagai konteks yang diberikan oleh pengguna. Dalam percobaan pengukuran kesamaan, sistem yang kami usulkan memiliki tingkat akurasi 79,17%.
Kata Kunci - *Scientific experimental paper, Context-baseb subspace selection, Context-dependent similarity measurement.*

I. INTRODUCTION

The increasing use of the internet has caused text-based document growing up exponentially every time [3]. Similarly, the electronic data of scientific document papers increases and makes it easier to retrieve a scientific journal information. However, large data causes search results to become more numerous and can also make it difficult for users to determine the information retrieval. In the context of education, scientific document paper is one of main important references for higher education. One indication of the progress of higher education is to look at the quality and quantity of scientific papers produced. With the development of higher education institutions, researchers need scientific papers as a reference in their research, so that the need to search and retrieve appropriate scientific papers becomes important. However, the search engines that exist today are not designed to specifically look for scientific papers, but used to search for overall text-based document. This causes that when the search engine used to search for specific scientific papers, it may retrieve inappropriate search results.

In the search for scientific paper to produce accurate content, there are many obstacles, because the content is stored in the form of text-based document, consisting of unstructured data with high dimensional features of word. The scientific papers can be divided into two kinds, which are experimental and non-experimental scientific paper. The experimental scientific paper is a research paper composed directly on the object being examined. While non-experimental scientific paper is research paper written indirectly and led more to data collection and analysis. The experimental scientific paper consists of 4 main parts: (1) data content, (2) method, (3) result and (4) problem. For the experimental scientific paper, it can be classified based on those 4 main parts [1]. In this condition, a search engine needs to address the specific part of the experimental scientific paper to improve accuracy in the retrieval result. The users can determine more specifically their preferences for the retrieval of the experimental scientific paper, based on data content, method, result and problem related to papers they are interested in.

II. RELATED WORK

Because the availability of text data is increasing, to obtain information back from a text-based document is a widely need in the current era. Likewise, the scientific paper retrieval also increases along with the increasing number of researchers. This is due to a scientific study requiring documentation that aims to record research results. Another goal is to be published on the relevant forum in order to account for the results of the research [4]. The research documentation can be in the form of scientific journals or scientific papers. With the increasing number of electronic data in this study, this research in the field of classification of scientific data also attracts much attention. The scientific papers from research papers can be classified [5]. The documentation of this scientific paper is a collection of letters forming string words. This collection of strings is a classification feature. This results in these features consists of high dimensional features, so a retrieval process is needed to eliminate unnecessary data.

To retrieve a group of experimental scientific papers, it can be realized by classifying the experimental scientific papers. The process of this classification has been carried out for the topic of experimental scientific paper. The classification of experimental scientific papers is made by first obtaining information contained in these experimental scientific papers. The information that exists in each experimental scientific paper consists of problem, data, method and result. Based on these four main features, the experimental scientific papers that can be classified [1]. Several researchers tried to address the information retrievals of scientific paper. Liu et.al. [9] presented a research intelligent information retrieval system ontology-based in digital library by concluding that semantics retrieval technology would improve retrieval

quality extremely, and would be the preferred method to solving the lack of semantic relation in traditional retrieval technology. Suyan et.al. [10] constructed deep resolution and retrieval platform for large scale scientific and technical literature by applying methods of real-time construction of standardized Data Set for citation content analysis and proposing a scientific and technical literature retrieval platform based on Citation content. Yang et.all [11] applied Weighted Term Frequency to develop a scientific literature retrieval model and used simulated annealing algorithm to learn the weighted factor. Chikhi et.al. [12] proposed to integrate link and content information by exploiting the links semantics to enrich the textual content of documents. WeiDong et.al. [13] presented a scientific literature statistical analysis system based on DOM tree with the design idea, system structure, design and implementation of main model and the running result. Ma and Fang [14] proposed information cartography in scientific research domain which generates an information map for users to mitigate information overload in current systems. Saggion and Ronzano [15] presented an overview of approaches to the extraction of knowledge from scientific literature, including the in-depth analysis of the structure of the scientific articles, their semantic interpretation, content extraction, summarization, and visualization. However, the information retrieval of scientific experimental paper still remains problem due to difficulty to address the paper content consisting of data, result, method, and problem.

III. ORIGINALITY

In this paper, a new system for information retrieval on experimental scientific papers is presented. This system consists of 4 main functions: (1) Specific content-based feature extraction, (2) Classification model, (3) Selection of context-based subspaces, and (4) Measurement of similarities depending on the context. In the feature extraction, our system extracts feature categories in experimental scientific papers with certain content-based features, which are data, result, method, and problem. For the classification model, we use several classification algorithms, which are Nearest Neighbor, Naive Bayes, Support Vector Machine and Decision Tree, to classify certain content features from query papers into the aggregation of supervised documents. In the selection of context-based spaces, the system carries out dimension reduction by selecting context-based subspaces chosen by the user. To obtain the final search result, we make similarity measurements in context by constructing context-dependent dataset metrics to paper.

IV. SYSTEM DESIGN

In this section, we discuss the design of our proposed system for information retrieval on experimental scientific papers. The discussion is divided into 4 main parts of the system, which are (1) Specific content-based feature extraction, (2) Classification model, (3) Context-based subspace selection, and (4) Context-dependent similarity measurement. The system architecture of our proposed system is shown in Figure 1.

A. Specific Content-based Feature Extraction

In this part, we discuss how to make feature extraction of paper document. The discussion involves the paper document, text-mining process, supervised data of each paper document after text-mining process, and supervised document aggregation to create huge dimensional metric of supervised dataset.

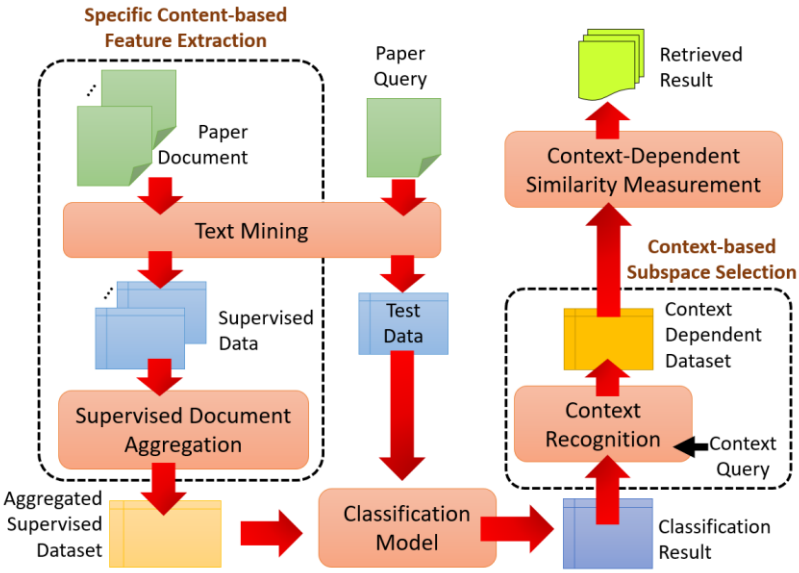


Figure 1. Architecture of our proposed system

1. Paper Document

The results of previous studies are stated that each experimental scientific paper can be ensured to have content: data, result, method and problem [1]. The content is written in sentences that are in the scientific paper. The position of the sentence in the paper is not grouped in certain chapters or sections. The sentences containing content can be in the title, abstract, introduction or other sub-chapter. So the sentence that describes the content can be in which part of the paper. In addition, in the paper there is a sentence describing the content (data, result, method and problem) used, there are also sentences that explain the method referred to. The sentence that describes the content that is used alone is called rhetorical sentence. The data obtained from previous research is rhetorical sentence data that has been classified based on existing content. The following is as screenshot of paper document that used in this paper.

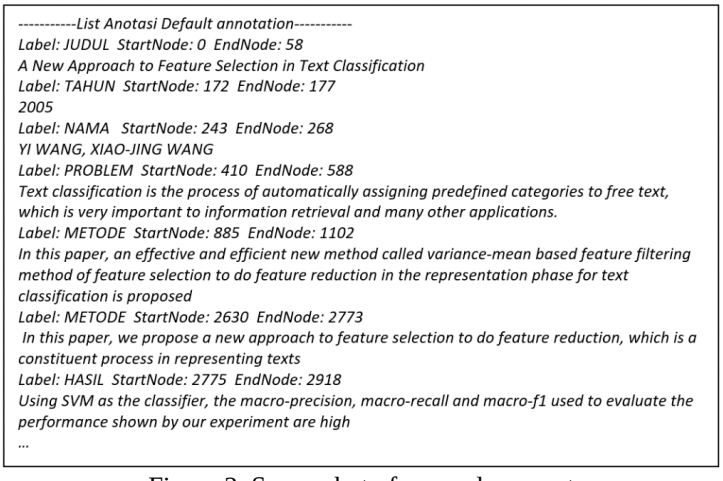


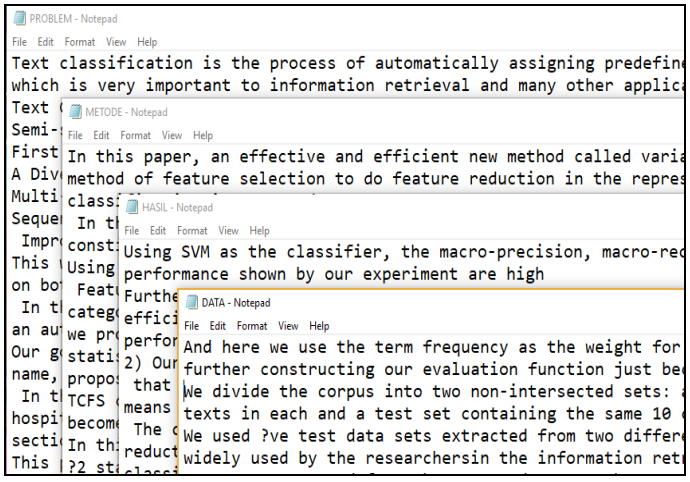
Figure 2. Screenshot of paper document

The paper documents then are prepared by separating into the content sentence which are data, result, method and problem sentences. The documents are grouped based on the original paper file. Each file will have 4 groups of content sentence: data, result, method and problem. We set these four groups of sentences to be used as supervised data and their group names to

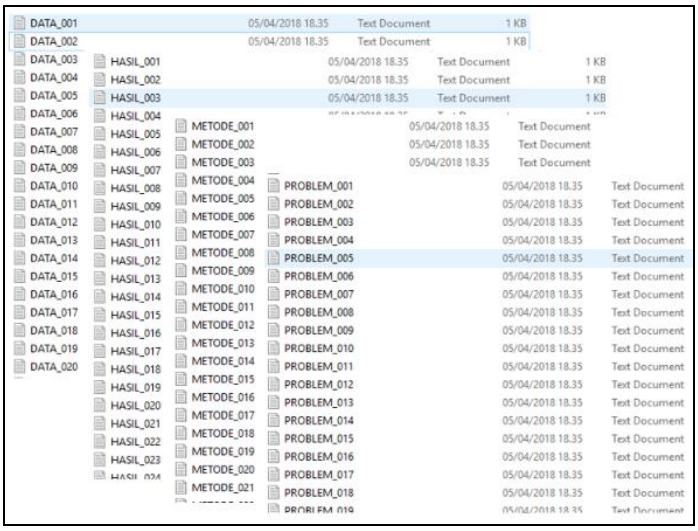
be used as class labels. Figure 3(a) shows the document separation into the content sentence, and Figure 3(b) shows the files representing the class labels.

2. Text Mining

To form a classification model, the next step is to do the text mining process. Text mining is mining that is carried out by a computer to get something new, something that is not known before or rediscover implicit information, which comes from information extracted automatically from different text data sources [6]. The text mining is a technique used to deal with problems of classification, clustering, information extraction and information retrieval [7]. Basically, many of the work processes of text mining adopt from Data Mining research, but the difference is that the patterns used by text mining are taken from a set of unstructured natural languages, while Data Mining patterns are taken from structured databases [8]. Based on the irregularity of the text data structure, the process of text mining requires several initial stages which in essence is to prepare so that the text can be changed to be more structured. These stages are as follows.



(a)



(b)

Figure 3. Screenshot of paper separation and class labelling

Tokenizing stage is the process of decomposing the description in the form of a sentence into a word and removing delimiter such as a period (.), Comma (,), space and number characters in the word [9].

The feature selection stage aims to reduce the dimensions of a collection of texts, or in other words delete words that are considered insignificant or do not describe the contents of the document so that the classification process is more effective and accurate [6] [7]. At this stage, the action taken is to eliminate stopword and stemming from the words that are affixed [6] [7]. Stopword is a vocabulary that is not a feature (unique word) of a document. Before the stopword removal process is done, a stopword list must be created. If included in the stoplist, the words will be removed from the description so that the words left in the description are considered as words that characterize the contents of a document or keywords. At this stage in addition to filtering, the separation of sentences is also carried out, each sentence becomes one file. Figure 4 shows screenshot of the paper document after applying the filtering.

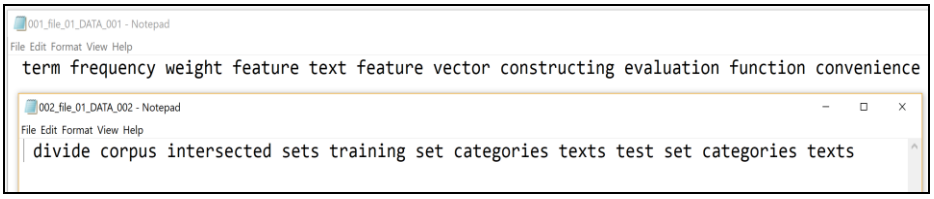


Figure 4. Screenshot of the paper document after applying the filtering

Stemming is the process of mapping and decomposing various forms (variants) from a word into its basic word form (stem). The purpose of the stemming process is to eliminate affixes, both in the form of prefixes, suffixes, and confixes in each word. If the affix is not removed, then every single word will be stored in a variety of different forms according to the affixes attached to it, so that it will add to the database load. This is very different if it removes the inherent affixes of each basic word, so one basic word will be stored once, even though the base word in the data source has changed from its original form and got various kinds of additions.

Tagging is the process of justifying words that are not properly written. This error is usually obtained from accidental writing or writing errors. In addition, the tagging process is also used as a substitute for non-standard words. Figure 5 shows screenshot the paper documents after stemming-tagging process.

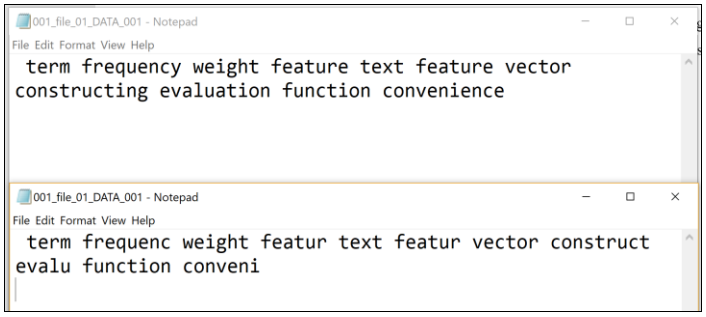


Figure 5. Screenshot the paper documents after stemming-tagging process

Term Frequency (TF) is to state the number of occurrences of a word in a sentence. The purpose of this process is to calculate the number of occurrences of the word. The results of the TF process are shown in Figure 6.

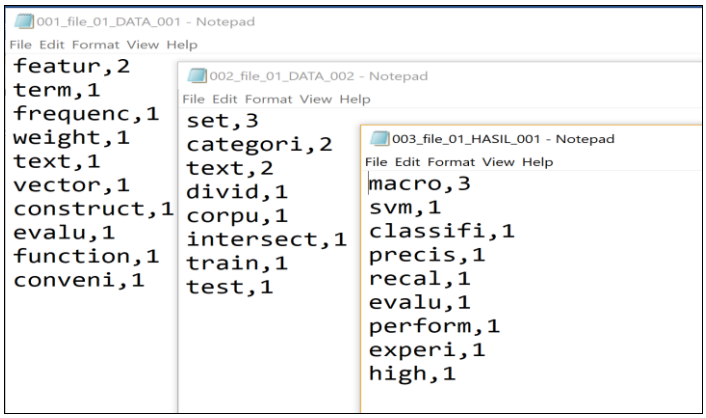


Figure 6. Screenshot of term frequency of the paper document after TF process

B. Supervised Data

Each sentence will be saved into one file and then, the files are grouped according to the type of sentence. The type of sentence according to the content to be searched for consists of: sentence data, sentence result, sentence method and sentence problem. Every existing data will be tokenizing process, filtering, stemming, tagging and TF (frequency term) to prepare data in the form of words along with the number of occurrences of words in the sentence. The paper data has been grouped into content categories so that it already has a class, as shown in Figure 7. To simplify labeling, the rules for making files are made. Each sentence will be sorted and stored in one file.

C. Supervised Document Aggregation

The next process is to create supervised document aggregation, or training data that have content labels. All files are collected and created tables that consist of a collection of words and the number of occurrences of words. Namely with a column consisting of: description of the origin of the file, number of sentence sequences, all words that appear and labels of content. And the line is the value of the appearance of the word.

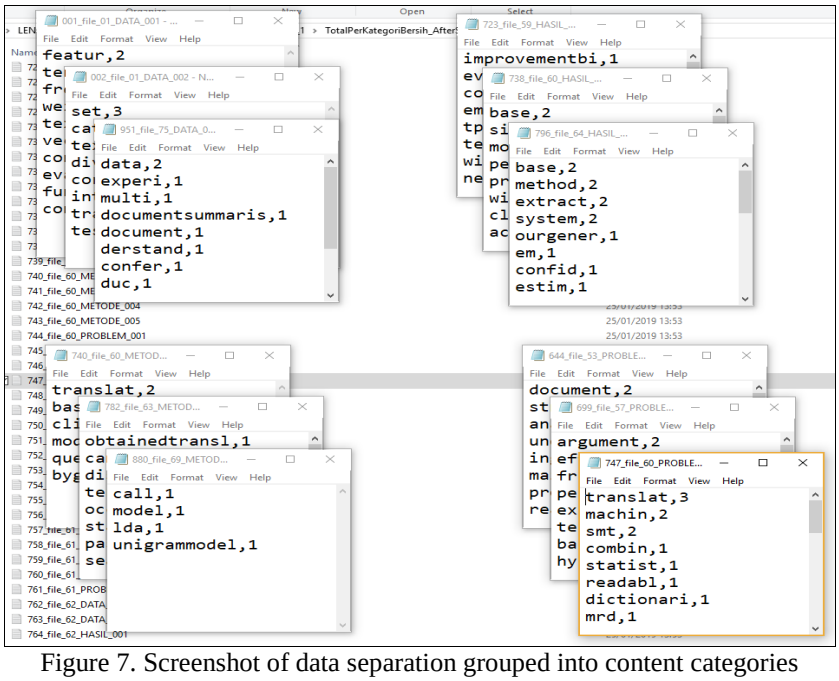


Figure 7. Screenshot of data separation grouped into content categories

The number of words that appeared were 2,926 words. The number of words that appear causes the computing process to be heavy. Then the remote data needs to be deleted. Remote data is data that has less contribution value in determining label class. This remote data is characterized by a small number of occurrences. In this paper, the amount considered small is that which has a smaller number equal to 2. This means that words with emergence values less than or equal to 2 will be omitted. By deleting data that appears one time and twice, the total number of words becomes 829 words. This means that the number of deleted columns is 2,097. Figure 8 shows screenshot of supervised document aggregation among the supervised data.

	A	B	C	D	E	AEY	AEZ
1	nama File	Label	featur	method	document	gre	
2	file_01	DATA	2	0	0	0	DATA
3	file_01	DATA	0	0	0	0	DATA
4	file_01	HASIL	0	0	0	0	HASIL
5	file_01	HASIL	0	1	0	0	HASIL
6	file_01	HASIL	0	1	0	0	HASIL
7	file_01	HASIL	0	2	0	0	HASIL
8	file_01	HASIL	1	2	0	0	HASIL
985	file_77	METODE	0	0	0	0	METODE
986	file_77	PROBLEM	0	0	0	0	PROBLEM
987	file_77	PROBLEM	0	0	0	0	PROBLEM
988			216	180	166	3	

Figure 8. Screenshot of supervised document aggregation

D. Classification Model

The supervised document data that has been deleted from the data outlier is used as the meta data for classification model. In this process, we describe the classification model in order to classify the paper query from the user. The classification model consists of 2 processes: classification algorithm and validation model of classification analysis.

Regarding the classification model, we use Nearest Neighbors Algorithm. Nearest neighbor is an approach to look for cases by calculating the closeness between new cases and old cases, which is based on matching weights of a number of features. k-NN algorithm is a method for classifying new objects based on k of their closest neighbors. This algorithm includes a supervised learning algorithm, where the results of new query instances are classified based on the majority of the categories on k-NN. The most appearing classes will be the class of classification.

k-NN method algorithm works based on the shortest distance from the query instance to the training data to determine k-NN. The sample training is projected into a large dimension space, where each dimension represents the features of the data. This space is divided into sections based on the training sample classification. A point in this space is marked as a class c if class c is the most common classification in the nearest neighbor k from that point. Near or near neighbors are usually calculated based on the Euclidean Distance which is represented as follows.

$$d(a,b) = \sqrt{\sum_{i=1}^{n}(a^2 + b^2)} \qquad (1)$$

where the matrix d(a, b) is the scalar distance of both vectors a and b of the matrix with dimensions d dimensions.

In the training phase, this algorithm only stores feature vectors and classification of sample training data. In the classification phase, the same features are calculated for testing data

(whose classification is unknown). The distance from this new vector to all training vector samples is calculated and the number of k fruits that are closest is taken. The new classification point is predicted to be included in the highest classification of these points.

In this paper, we apply nearest neighbor classification algorithm with different k=3,5,7, and 9. For comparing algorithms, we also use other classification algorithms, which are Naïve Bayes, Support Vector machine and Decision Tree.

The next step is classification analysis with validation model. Cross-validation is a way of validating a model to assess the results of data analysis. Cross-validation has several testing techniques. In this study the technique used was Leave-One-Out (LOO) technique. LOO Cross-validation technique is to experiment as much as the amount of data, with regard to each time all data will be considered as learning data as well as testing data.

E. Context-based Subspace Selection

In this part, we discuss to how involve the user's intention as a context for the retrieval by applying the context recognition. To find the similarity of paper based on content is done by calculating the similarity of paper query with a set of paper based on content. Each paper in a set of paper is carried out by the Mining process and classified. Every single sentence in a paper has a label, which are data, method, result, and problem. Therefore, we apply the context recognition process to involve the context given by user's intention for the query and reduce the dimensional space of features with the given context. After that the similarity of Paper queries is calculated with each Paper in the database. The measurement of the similarity of each paper based on the desired content. Figure 9 shows the given context by the user to the classification result in order to create context dependent dataset.

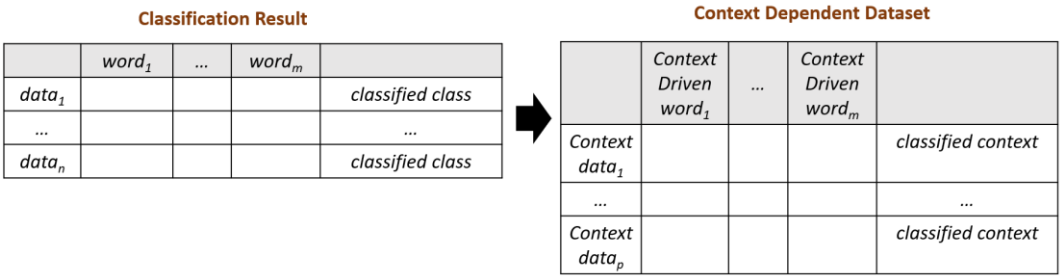


Figure 9. Context dependent dataset creation

F.Context-dependent Similarity Measurement

In this paper, the algorithm for measuring similarity uses the cosine similarity algorithm. Cosine similarity is the calculation of similarities between two n-dimensional vectors by looking for cosines from the angle between them and often used to compare documents in text mining with the formula.

$$similarity = cos\,(\emptyset) = \frac{A.B}{\|A\|\|B\|} = \frac{\sum_i^n A_i B_i}{\sqrt{\sum_i^n A^2}\sqrt{\sum_i^n B^2}} \quad (2)$$

where A dan B = vector of context features in the dataset

The input given to the system is the paper query and the type of content desired. Such as input in the form of paper1 and the desired content is 'method', then labeled the sentences in paper1 and selected sentences labeled 'method'. Likewise, all the papers in the dataset are selected sentences labeled 'method'. Then the value of the similarity between the paper queries and all the papers in the dataset is calculated. The similarity values are ranked from the most

similar to the non-similar ones. The ranking value will be used to display the 10 most similar papers. Figure 10 shows screenshot retrieved result by the 'Data' given context.

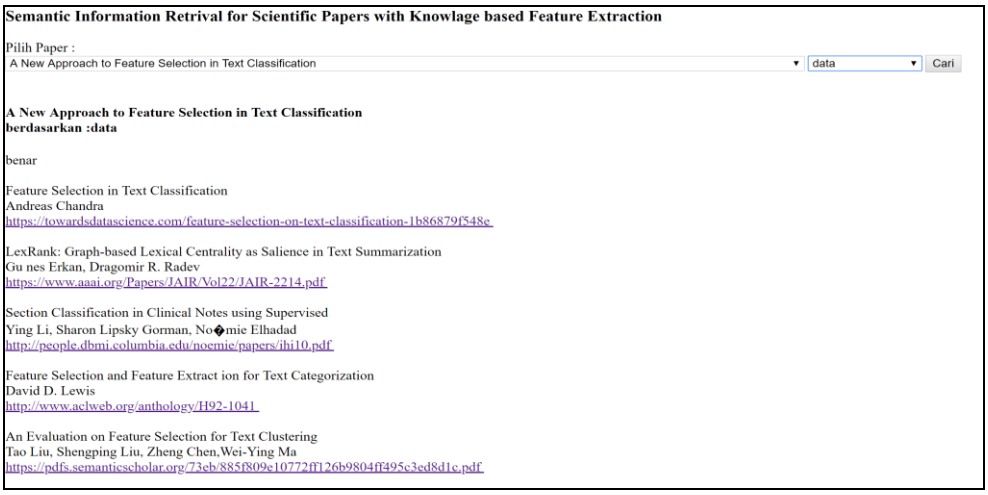


Figure 10. Screenshot retrieved result by the 'Data' given context

V. EXPERIMENT AND ANALYSIS

To see the applicability of our proposed system, we conducted a series of experimental study with two types of experiment, which are experiment on the classification model, and experiment on the similarity measurement.

A. Experiment on Classification Model

Here we use training data with number of 77 scientific experimental papers that have content consisting of data, result, method and problem. After the sentences in the 77 papers are separated based on the label, the data is ready for the text mining process. The result is that data can be completely separated based on the initial paper file. The sentence in each paper can also be separated based on the label of the sentence. After text-mining process, supervised dataset is created with 986 sentences used as learning data and 2,926 extracted features of word. For the classification analysis, we used Leave-One-Out algorithm for validation model to 77 scientific experimental papers. In the experimental comparison, we use several classification algorithms, which are 1-NN, 3-NN, 5-NN, 7-NN, 9-NN, Naïve Bayes, Support Vector Machine and Decision Tree. Table 1 shows the precision and accuracy of the result of experimental comparison among classification algorithms for each content of scientific experimental papers. From Figure 1, overall experimental results performed a good classification in the precision and accuracy. It manifested a good Specific Content-based Feature Extraction process to create Aggregated Supervised Dataset.

TABLE I
PRECISION AND ACCURACY OF THE RESULT OF EXPERIMENTAL COMPARISON AMONG CLASSIFICATION ALGORITHMS FOR EACH CONTENT OF SCIENTIFIC EXPERIMENTAL PAPERS

Classification Algorithm		Data	Result	Method	Problem	Average Result
1-NN	Precision	96.77%	84.75%	41.13%	53.66%	69.08%
	Accuracy	91.34%	74.44%	47.06%	79.21%	73.01%
3-NN	Precision	73.74%	64.76%	54.32%	69.33%	65.54%
	Accuracy	92.95%	76.57%	67.34%	81.85%	79.68%
5-NN	Precision	88.89%	70.49%	53.42%	66.67%	69.87%
	Accuracy	93.95%	80.43%	66.33%	81.34%	80.51%

7-NN	Precision	96.77%	84.75%	41.13%	53.66%	69.08%
	Accuracy	91.34%	74.44%	47.06%	79.21%	73.01%
9-NN	Precision	88.57%	71.25%	54%	68.13%	70.49%
	Accuracy	93.78%	80.63%	66.94%	82.25%	80.90%
Naïve Bayes	Precision	90.91%	62.78%	59.95%	42.27%	63.98%
	Accuracy	94.61%	78.30%	69.78%	75.86%	79.64%
Support Vector Machine	Precision	95%	88.24%	40.99%	60.87%	71.27%
	Accuracy	90.22%	75.56%	46.45%	79.41%	72.91%
Decision Tree	Precision	76.92%	0%	38.51%	100%	53.86%
	Accuracy	89.81%	70.24%	40.28%	78.85%	69.79%

B. Experiment on Similarity Measurement

In this experiment, paper retrieval is proceed by giving the paper query and the desired content (data, result, method, and problem) as a context given by the user. For experimental study, we choose 3 papers from the scientific experimental paper collection as paper queries, as shown in Table 2.

TABLE II
CHOSEN PAPERS FROM SCIENTIFIC EXPERIMENTAL PAPER COLLECTION AS PAPER QUERIES

Paper ID	Paper Title
1	A New Approach to Feature Selection in Text Classification
7	Statistical Section Segmentation in Free-Text Clinical Records
16	Feature selection and semi-supervised clustering using multiobjective optimization

For every experiment, we retrieve 4-top retrieved result. Table 3-5 shows correctness of the experimental results of each chosen paper in the context of data, result, method, and problem. From the experimental result, our proposed system performed 38 correct results and 10 incorrect results of retrieved papers, that gave 79.17% accuracy and 20.83% error rate

TABLE III
CORRECTNESS OF EXPERIMENTAL RESULT OF PAPER ID 1 IN THE CONTEXT OF DATA, RESULT, METHOD, AND PROBLEM

Content	Retrieved Result	Correctness
Data	Feature Selection in Text Classification	correct
	LexRank: Graph-based Lexical Centrality as Salience in Text Summarization	correct
	Section Classification in Clinical Notes using Supervised	correct
	Feature Selection and Feature Extract ion for Text Categorization	correct
Result	A Comparative Study on Feature Selection in Text Categorization	correct
	Feature Selection in Text Classification	correct
	A Logic-Based Approach to Relation Extraction	incorrect
	Text Clustering with Feature Selection by Using Statistical Data	correct
Method	Feature Selection in Text Classification	correct
	Text Clustering with Feature Selection by Using Statistical Data	correct
	Feature selection and semi-supervised clustering using multiobjective optimization	correct
	Clustering-Based Feature Selection in Semi-supervised Problems	correct
Problem	Section Classification in Clinical Notes using Supervised	correct
	High-Performing Feature Selection for Text Classification	correct
	Text Tiling: A Quantitative Approach to Discourse Segmentation	correct
	A Framework of Feature Selection Methods for Text Categorization	correct

In this experiment the selected content is data content. So in the paper query and all the papers in the database, sentences that explain about 'data' will be searched. The sentences are calculated by the number of words that appear, then stored in vector form. So each paper has a 'data' vector value. This value will be calculated using cosine similarity. Because paper is in more than one database, more than one data will be obtained. Every paper in the database has a cosine value for the paper query. This cossine value shows the closeness between the paper query and paper in the database. The cossine value will be sorted from large to small. The biggest value shows the paper that has similarities with the paper query. Then the order of the paper is displayed as a search result.

TABLE IV
CORRECTNESS OF EXPERIMENTAL RESULT OF PAPER ID 7 IN THE CONTEXT OF DATA, RESULT, METHOD, AND PROBLEM

Content	Retrieved Result	Correctness
Data	Causal Relation Extraction	correct
	An Efficient Linear Text Segmentation Algorithm Using Hierarchical Agglomerative	correct
	Implementation Of An Automated Text Segmentation System Using Hearsts Texttiling Algorithm	incorrect
	Using Query-Relevant Documents Pairs for Cross-Lingual Information Retrieval	incorrect
Result	Causal Relation Extraction	correct
	An Efficient Linear Text Segmentation Algorithm Using Hierarchical Agglomerative Clustering Agglomerative Clustering	correct
	Implementation Of An Automated Text Segmentation System Using Hearsts Texttiling Algorithm	correct
	Using Query-Relevant Documents Pairs for Cross-Lingual Information Retrieval	incorrect
Method	Survey on Feature Selection in Document Clustering	correct
	WikiTranslate: Query Translation for Cross-Lingual	correct
	Feature Selection and Feature Extract ion for Text Categorization	correct
	Accurate Query Translation for Japanese-English Cross-Language Information Retrieval	incorrect
Problem	A New Approach to Feature Selection in Text Classification	correct
	A Text Tiling Based Approach to Topic Boundary Detection in Meetings	correct
	High-Performing Feature Selection for Text Classification	incorrect
	Classifier Chains for Multi-label Classification	incorrect

Test in table IV, this data content uses query paper ID 7 with the title "Statistical Section Segmentation in Free-Text Clinical Records". The paper is analyzed by taking the top four data from the search results. The top four data means that the four papers are the most similar based on system calculations. The results are like in table IV. The results of the trial after analyzing the two top data have the same data, but the third and fourth order papers do not have the same data, but the system has the same data.

TABLE V
CORRECTNESS OF EXPERIMENTAL RESULT OF PAPER ID 16 IN THE CONTEXT OF DATA, RESULT, METHOD, AND PROBLEM

Content	Retrieved Result	Correctness
Data	Efficient semi-supervised feature selection by an ensemble approach	correct
	First Order Statistics Based Feature Selection: A Diverse and Powerful Family of Feature Selection Techniques	correct

	Generating and evaluating evaluative arguments	incorrect
	Automatic Extraction of Hierarchical Relations from Text	correct
Result	Efficient semi-supervised feature selection by an ensemble approach	correct
	Clustering-Based Feature Selection in Semi-supervised Problems	correct
	Text Clustering with Feature Selection by Using Statistical Data	correct
	A General Framework of Feature Selection for Text Categorization	correct
Method	Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts	correct
	Feature Selection in Text Classification	correct
	A New Approach to Feature Selection in Text Classification	correct
	First Order Statistics Based Feature Selection: A Diverse and Powerful Family of Feature	incorrect
Problem	Clustering-Based Feature Selection in Semi-supervised Problems	correct
	Classifier Chains for Multi-label Classification	correct
	Efficient semi-supervised feature selection by an ensemble approach	correct
	Abstract feature extraction for text classification	incorrect

The results of the second trial can be seen in table V above, that is by using the query paper with ID 16 with the title "Feature selection and semi-supervised clustering using multi-objective optimization" and search based on 'data' content. The trial produced three papers with the same data and one paper data that was not appropriate. An inappropriate paper is a paper entitled "Generating and evaluating evaluative arguments". The two papers if observed manually do have different data. On the search test based on the content of 'Results'. The trial resulted in all of the top four data having the same results as the Query paper.

VI. CONCLUSION

In this paper we present a new system for information retrieval on experimental scientific papers. This system consists of 4 main functions: (1) Specific content-based feature extraction, (2) Classification model, (3) Selection of context-based subspaces, and (4) Measurement of similarities depending on the context. In the feature extraction, our system extracts feature categories in experimental scientific papers with certain content-based features, which are data, problem, method, and result. To perform the applicability of our proposed system, we tested 77 scientific experimental papers in the dataset with the Leave-One-Out validation model with several classification algorithms (Nearest Neighbor, Naive Bayes, Support Vector Machine and Decision Tree) and on average performed 66.65% precision rate and accuracy of 76,18% precision rate. We also made the experiment on the similarity measurement by giving the paper query and the desired content (data, result, method, and problem) as a context given by the user. In the similarity measurement experiment, our proposed system performed 79.17% accuracy rate.

REFERENCES

[1] Afrida Helen, Pendekatan Penggunaan Section dan Judul untuk Klasifikasi Kalimat Retorik pada Makalah Ilmiah Eksperimental, *Doctoral Dissertation*, Institut Teknologi Bandung, March 2016.

[2] Bruno Trstenjak, Sasa Mikac, Dzenana Donko, K-NN with TF-IDF Based Framework for Text Categorization, *Procedia Engineering* 69 (2014) 1356 – 1364 , ScienceDirect,

24th DAAAM International Symposium on Intelligent Manufacturing and Automation, 2013.
[3] Igg Adiwijaya, Text Mining dan Knowledge Discovery, *Kolokium bersama komunitas datamining Indonesia & soft-computing Indonesia*, September 2006.
[4] A.D. Robert, How to Write and Publish a Scientific Paper, *Book 7th edition*, 2012.
[5] Stephany D Hubsy, Topic Classification of Blog Post Using Distant Supervision, *Procedings of the 13th Conference of the Europe Chapter of the Association for Computation Linguistics*, Page 28-36, Avigno France, Association fot Computation Linguistics, 2012.
[6] Ronen Feldman, James Sanger, The Text Mining Handbook: Advanced Analyzing Unstructured Data, *Cambridge University Press*, 2007.
[7] Michael W.Berry, Jacob Kogan, Text Mining: Application and Theory, *John Wiley and Son, Ltd.*, 2010.
[8] Jiawei Han, Micheline Kamber, Data Mining: Concepts and Techniques, *Morgan Kaufmann Publishers*, 2nd edition, ISBN 1-55860-901-6, March 2006.
[9] Lizhen Liu, Chengli Wang, Minhua Wu, Guoqiang He, Research of Intelligent Information Retrieval System Ontology-Based in Digital Library, *2008 IEEE International Symposium on IT in Medicine and Education*, Xiamen-China, December 12-14, 2008.
[10] Wu Suyan, Li Wenbo, Wu Jiangrui, Construction of Deep Resolution and Retrieval Platform for Large Scale Scientific and Technical Literature, *2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, Chengdu-China, April 20-22, 2018.
[11] Xi Quan Yang, Dian Yang, Ming Yuan, Xing Hua Lv, Scientific Literature Retrieval Model Based on Weighted Term Frequency, *2014 Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, Kitakyushu-Japan, August 27-29, 2014.
[12] Nacim Fateh Chikhi, Bernard Rothenburger, Nathalie Aussenac-Gilles, Combining Link and Content Information for Scientific Topics Discovery, *2008 20th IEEE International Conference on Tools with Artificial Intelligence*, Dayton-USA, November 3-5, 2008.
[13] Li WeiDong, Dong Yibing, Wang RuiJiang, Tian HongXia, Design and Implementation of Scientific Literature Statistical Analysis System on Three Retrieval Systems Based on DOM Tree, *2009 Asia-Pacific Conference on Information Processing*, Shenzhen-China, July 18-19, 2009.
[14] Tianmu Ma, Wei Fang, Information Cartography Based on Syncretic Representation of Scientific Papers, *2018 4th International Conference on Information Management (ICIM)*, Oxford-UK, May 25-27, 2018.
[15] Horacio Saggion, Francesco Ronzano, Scholarly Data Mining: Making Sense of Scientific Literature, *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Toronto-Canada, June 19-23, 2017.

ACKNOWLEDGEMENT

Peneliti mengucapakan terima kasih kepada Tim *Jurnal Inovtek Seri Informatika Polbeng* yang telah meluangkan waktu untuk merevisi jurnal guna menunjang penelitian ini dengan baik.