IMPLEMENTATION OF INFORMATION EXTRACTION TEXT DOCUMENTS SKRIPSI USING RULED BASED METHOD

Dimas Mustaqwa,Nelly Indriani

Program Studi Teknik Informatika
Fakultas Teknik dan Ilmu Komputer Universitas Komputer Indonesia
Jl. Dipati Ukur 114 Bandung
E-mail : dimasmustaqwa@gmail.com

ABSTRACT

Search for information one of them by reading a document. Documents available in softcopy form generally use pdf or doc format. The information available on pdf format documents can be read but the structure is not known in the document. One example is the structure of the document such as title, topic, focus, and comment. The existence of the structure on the document, search information is more easily obtained so there is no need to search the entire contents of the document. In this research will perform the extraction of information to find the structure on the cover report document document, abstract, and abstract. The extraction of information using rule based is a method that uses factual rules from the data being analyzed. For accuracy testing using 50 on the cover, abstrak, and abstract report document shows good enough, which means the extraction of information in this study can be used to extract The desired thesis data identity.

Keywords: Information extraction, rule based, text, thesis.

1. INRODUCTION

Search for information one of them by reading document. Documents available in softcopy form generally use pdf or doc format. Search on a document can be done if the document is already saved.

In the storage library of a document must be given an information or identity. Librarians include an identity in a document by filling in the necessary data into the system. The downside of this way is one of them is long if the number of documents that many should be stored, another problem that may arise is a mistake to type the identity of the document. The weakness can be handled by filling the document identity automatically, one way by extracting the document. Information extraction is the retrieval of facts and structured information from large text collection contents. The notions of fact here are the various entities that are taken into account. Briefly the extraction of information is a process of obtaining the structured facts from the available data. [1]

Extraction of documents can be done using a method. One method of extracting is rule based. Rule Based System is a computer program that processes information contained in a working memory with a set of rules contained in the knowledge base using an inference engine to generate new information. . [2] A rule-based method can be used if a document is a structured document. Document information can be obtained by finding the structure of a document one of them is thesis. Thesis is defined as the writing of scientific papers containing the results of comprehensive research systematically arranged based on the provisions of scientific research methods. Writing this thesis is intended as a training for students to pour his ideas in the form of scientific work. [3] The content of the structure taken from the thesis is the title of thesis, type of thesis, author name, nim, study program, faculty, university, abstract contents, and keywords in abstract. Understanding structured information is a sentence or text that can be divided into categories such as topics, focus, comments, background, and comparing old or new information. [4]

One research on information extraction on the rule base is titled "information extraction with rule-based method for evaluation of kinematics physics" extraction of information using rule-based derives the essence of the questions posed and displays the answers according to the questions asked. The use of rule-based data is the search for keywords in practice questions. Keywords search for questions, numbers, quantities, units, and formulas in physics lessons. From the keywords found will be completed or memeberikan appropriate answers from the data obtained. Accuracy results obtained for 90.6% - 95.4% in research evaluation of the physics of kinematics. [5] Another study entitled "rules-based information extraction algorithm" uses the extraction

of information on the Reporting Result (LHP) document on the Local Government Financial Statement (LKPD) which the extraction results are grouped into several classifications. Accuracy results in LHP LKDP documents are 89.77% and 98.27%. [6]

Based on the background, then in this study will be extracted information to obtain the structure of information on the thesis document and do the identity assignment on the document automatically. The process of extracting information will use the rule based method. The result of extracting the information done is to classify the contents of the document into several classifications. This study aims to find the title of thesis, type of thesis, author name, nim, study program, faculty, university, abstract contents, and keywords on the abstract. Research data will analyze and build information extraction system using rule based to get identity from thesis data which will extraction.

2. METHOD INFORMATION EXTRACTION

System analysis can be defined as the decomposition of a whole system into its component parts with a view to identifying the necessary requirements in order to build an application to know the presentation of the analytical accuracy of the method used. System analysis about information extraction with rule based method to identify unstructured thesis data become structured divided into several parts that can be seen in the following Figure 1.

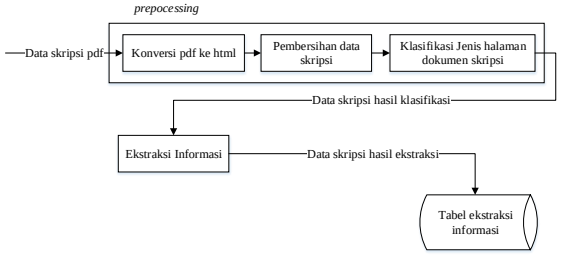


Figure 1. Flow Process of Thesis Data Extraction

Before performing system analysis, the analysis is aimed at making rule based on thesis cover, abstract, and abstract data.

2.1. Keywords and Rules Analysis

In this study using thesis data format .pdf. The analysis focuses on searching keywords and rules that will be used to extract thesis data information. Thesis data to be analyzed ie cover data, abstract Indonesian and abstract english. Thesis cover, abstract, and abstract data can be seen in Figure 2, Figure 3, and Figure 4.

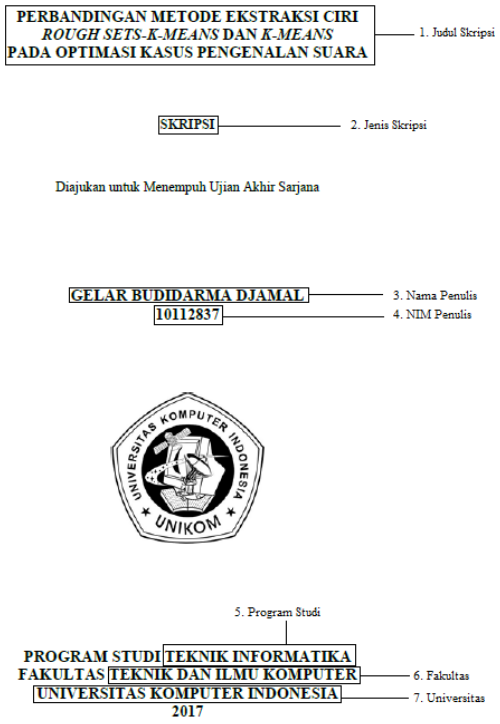


Figure 2. Cover Thesis Data

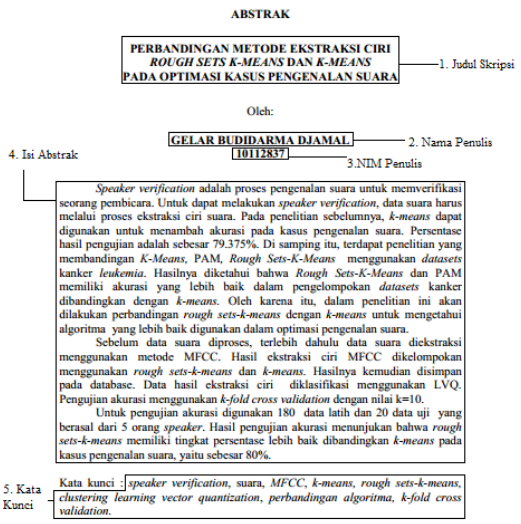


Figure 3. Abstract Thesis Data

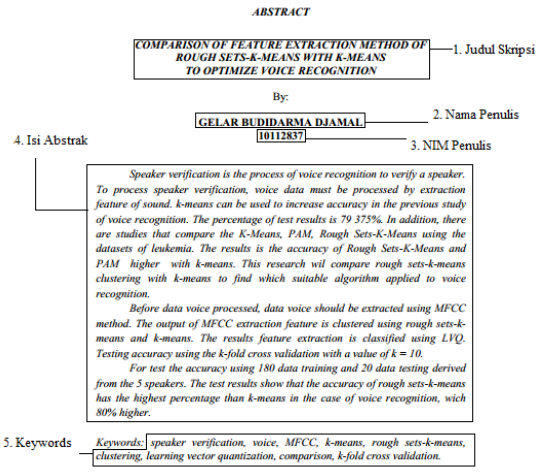


Figure 4. Abstract Thesis Data

Thesis data sample will be identified thesis data. The result of thesis data and data extraction rules obtained from data identification can be seen in table 1 as follows.

Table 1. Keyword Result and Thesis Data Extraction Rules

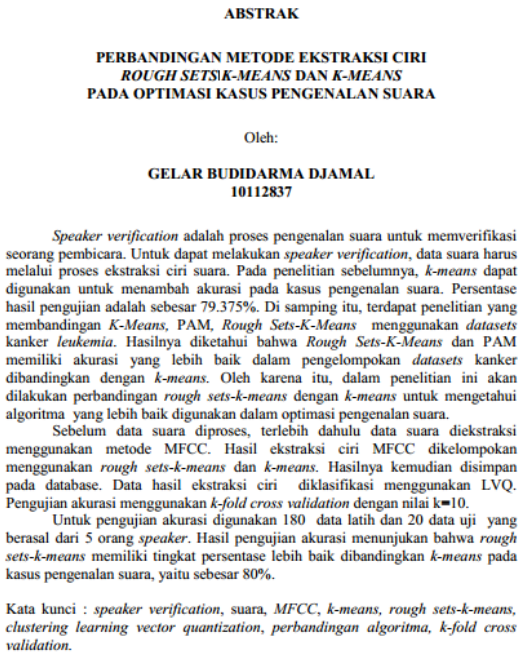
Identification	Keywords	Information
Data Skripsi Cover		
Judul skripsi	-	Take the thesis title from the beginning of the word to find the last type of thesis keyword from the thesis data
Jenis Skipsi	Skripsi, Tesis, Disertasi	Search for the last word "Thesis, Thesis, Dissertation" from thesis data
Nama Penulis	Sarjana	Looking for the last word "Bachelor" from the thesis data takes the word after the word "Bachelor" to find the number or if there is a different position "Techniques and Computer Science" take the word after the word "Engineering and Computer Science" in the first position
NIM		Figures after the author name of 8 digits
Program Studi	Program	Looking for the last word "Program" from the thesis data takes the word after the word "Program" to find the keyword faculty
Fakultas	Fakultas	Search for the last word "Faculty" from the thesis data to take the word after the word "faculty" to find the keyword of the University
Kampus	Universitas	Searching for the last word "University" from the thesis data takes the word after the word "University" to find the numbers
Data Skripsi Abstrak dan Abstract		
Judul Indonesia atau Judul Inggris	-	After the word "Abstract or Words Abstract" to find the beginning of the keyword Author name
Nama Penulis	Oleh, By	Finding the last "By, By" word from the thesis data takes the word after the word "By, By" to find the number
NIM	-	The number after the author's name
Isi Abstrak Indonesia atau Bahasa Inggris	-	After Nim and up to find the last word of Indonesian or English words from thesis data
Kata Kunci Indonesia atau Inggris	Kata Kunci, Keywords, Keyword	Search for the last word "Keywords, Keywords, Keywords" from the thesis data takes after the keyword to the end of the word srkipsi data

2.2 Preprocessing

Based on the data analysis done need to do preprocessing phase. The preprocessing stage that will be done is the conversion of pdf to html, thesis data cleaning, classification of thesis document type.

2.2.1 Converting pdf to html

Converting pdf to html is the process of converting pdf-formatted data to html. This is done so that data can be analyzed. As an illustration the conversion of pdf to html is shown in figure 5 and figure 6 below.



ABSTRAK

PERBANDINGAN METODE EKSTRAKSI CIRI *ROUGH SETS*\K-*MEANS* DAN *K-MEANS* PADA OPTIMASI KASUS PENGENALAN SUARA

Oleh:

GELAR BUDIDARMA DJAMAL
10112837

Speaker verification adalah proses pengenalan suara untuk memverifikasi seorang pembicara. Untuk dapat melakukan *speaker verification*, data suara harus melalui proses ekstraksi ciri suara. Pada penelitian sebelumnya, *k-means* dapat digunakan untuk menambah akurasi pada kasus pengenalan suara. Persentase hasil pengujian adalah sebesar 79.375%. Di samping itu, terdapat penelitian yang membandingan *K-Means*, PAM, *Rough Sets-K-Means* menggunakan *datasets* kanker *leukemia*. Hasilnya diketahui bahwa *Rough Sets-K-Means* dan PAM memiliki akurasi yang lebih baik dalam pengelompokan *datasets* kanker dibandingkan dengan *k-means*. Oleh karena itu, dalam penelitian ini akan dilakukan perbandingan *rough sets-k-means* dengan *k-means* untuk mengetahui algoritma yang lebih baik digunakan dalam optimasi pengenalan suara.

Sebelum data suara diproses, terlebih dahulu data suara diekstraksi menggunakan metode MFCC. Hasil ekstraksi ciri MFCC dikelompokan menggunakan *rough sets-k-means* dan *k-means*. Hasilnya kemudian disimpan pada database. Data hasil ekstraksi ciri diklasifikasi menggunakan LVQ. Pengujian akurasi menggunakan *k-fold cross validation* dengan nilai k=10.

Untuk pengujian akurasi digunakan 180 data latih dan 20 data uji yang berasal dari 5 orang *speaker*. Hasil pengujian akurasi menunjukan bahwa *rough sets-k-means* memiliki tingkat persentase lebih baik dibandingkan *k-means* pada kasus pengenalan suara, yaitu sebesar 80%.

Kata kunci : *speaker verification*, suara, *MFCC*, *k-means*, *rough sets-k-means*, *clustering learning vector quantization*, *perbandingan algoritma*, *k-fold cross validation*.

Figure 5. Data Skripsi Abstrak berupa pdf

Sesudah Konversi
i   ABSTRAK   PERBANDINGAN METODE   EK STRAKSI CIRI   ROUGH SETS K - MEANS   DAN K - MEANS   PADA OPTIMASI KASUS PENGENALAN SUARA   Oleh: GELAR BUDIDARMA DJAMAL   10112837   Speaker verification   adalah proses pengenalan suara untuk memverifikasi seorang pembicara. Untuk dapat   melakukan speaker verification , data suara harus mela lui proses ekstraksi ciri suara . Pada   penelitian sebelumnya ,   k - means   dapat digunakan untuk menambah akurasi pada kasus pengenalan suara . Persentase   h asil pengujian adalah   sebesar 79.375% .   Di   samping itu ,   terdapat penelitian yang mem bandingan   K - Means, PAM , Rough Sets - K - Means   menggunakan datasets   kanker   leukemia . H asil nya   diketahui   bahwa   R ough S ets - K - M eans   dan PAM memiliki akurasi   yang

Figure 6. Results Conversion of abstract thesis data

2.2.2 Pembersihan Data Skripsi

The results of pdf to txt conversion provide data that is still not appropriate. New paragraphs appear even though the words are still in the same sentence. A few new paragraphs appear after the end of the sentence. Therefore it is necessary to clean up the thesis data. Here are the steps of cleaning thesis data.

1. If there are 3 blank lines then make 1 blank line.
2. If the new line is not empty then the new line is combined with the previous line.

The results of cleaning thesis data can be seen in figure 7 below.

Sesudah Pembersihan
i ABSTRAK PERBANDINGAN METODE EKSTRAKSI CIRI ROUGH SETS K-MEANS DAN K-MEANS PADA OPTIMASI KASUS PENGENALAN SUARA Oleh: GELAR BUDIDARMA DJAMAL 10112837 Speaker verification adalah proses pengenalan suara untuk memverifikasi seorang pembicara. Untuk dapat melakukan speaker verification, data suara harus melalui proses ekstraksi ciri suara. Pada penelitian sebelumnya, k-means dapat digunakan untuk menambah akurasi pada kasus pengenalan suara. Persentase hasil pengujian adalah sebesar 79.375%. Di samping itu, terdapat penelitian yang membandingan K-Means, PAM, Rough Sets-K-Means menggunakan datasets kanker leukemia. Hasilnya diketahui bahwa Rough Sets-K-Means dan PAM memiliki akurasi yang

Figure 7. Results of cleaning thesis data

2.2.3 Classification of Thesis Document Page Type

Classification of page type of thesis document on preprocessing is a classification of data type of thesis to be extracted whether the cover or abstract after the process of cleaning thesis data. Table classification page type of thesis document can be seen in table 2 below..

Table 2. Classification of page type of thesis document

Classification	Status data	Description
1	Data cover skripsi	Search for study program, faculty and university on thesis data.
2	Data abstrak skripsi	Search for abstract word, or abstract on thesis data

Classification	Status data	Description
3	Bukan data skripsi cover atau abstrak.	No words found in the search were performed on the classification

2.3 Data Information Extraction

Extraction Information will be done using a rule base based on the keywords and rules that have been analyzed. Stages to be performed on the extraction of information is the extraction of thesis cover data and abstract thesis data extraction.

1.3.1 Data Extraction Cover Thesis Information

The extraction of cover data information using rule based can be seen in figure 8 to figure 14.

a. Extraction of thesis title

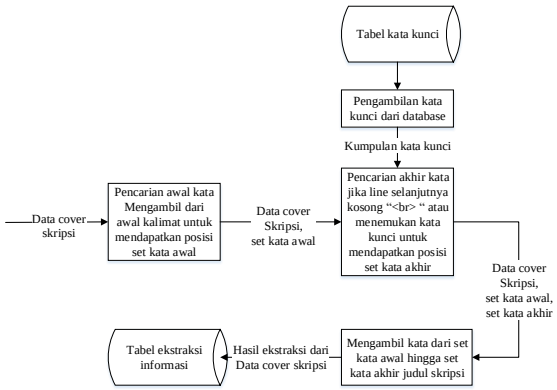


Figure 8. Extraction of thesis title

b. Extraction of thesis type

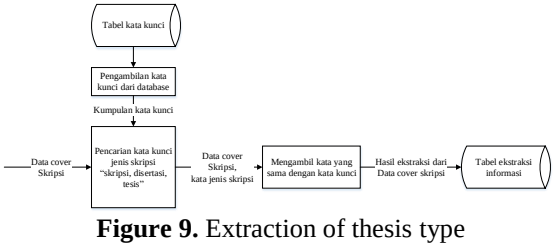


Figure 9. Extraction of thesis type

c. Writer name extraction

Figure 10. Writer name extraction

d. Nim extraction

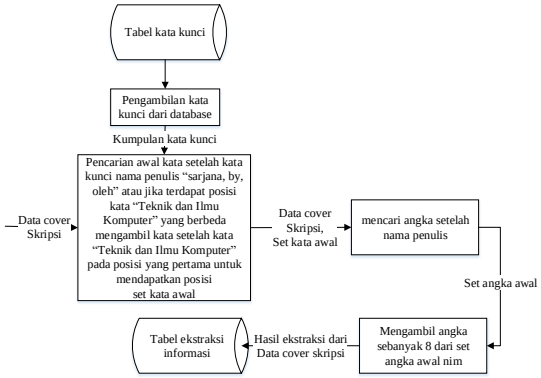


Figure 11. nim extraction

e. Extraction of study program

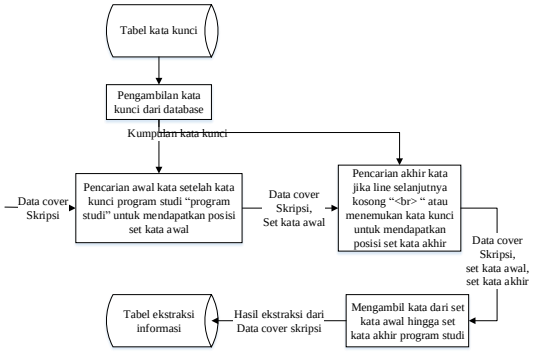


Figure 12. Extraction of study program

f. Faculty extraction

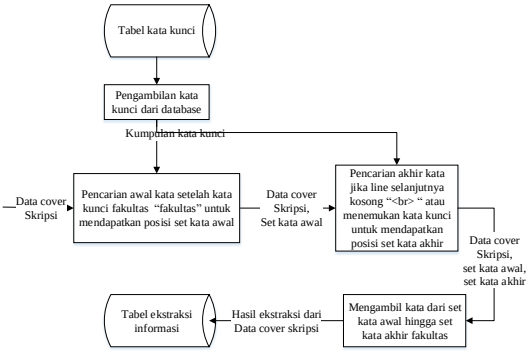


Figure 13. faculty extraction

g. University extraction

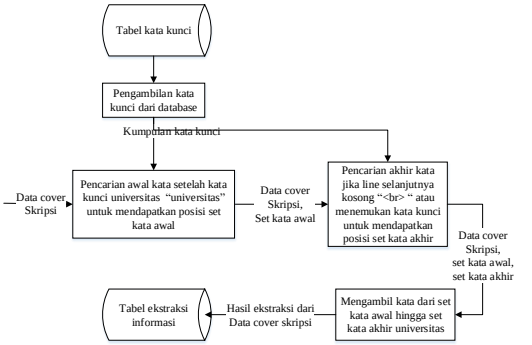


Figure 14. University extraction

1.3.2 Ekstraksi Informasi Data Abstrak Skripsi

The extraction of abstract data information using rule based can be seen in figure 15 to figure 19.

a. Extraction of thesis title

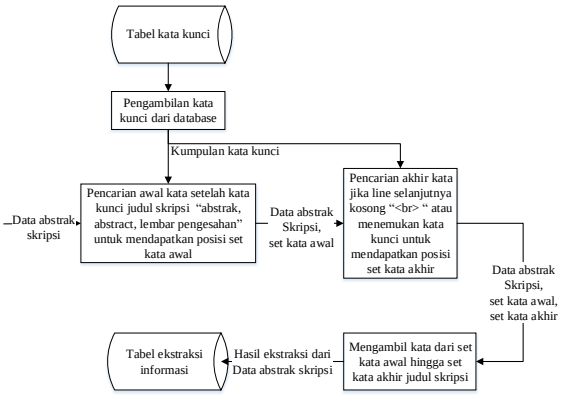


Figure 15. Ekstraksi judul skripsi

b. Writer name extraction

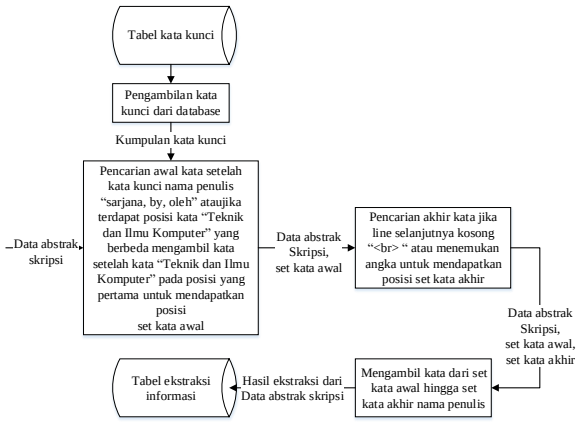


Figure 16. Writer name extraction

c. Nim extraction

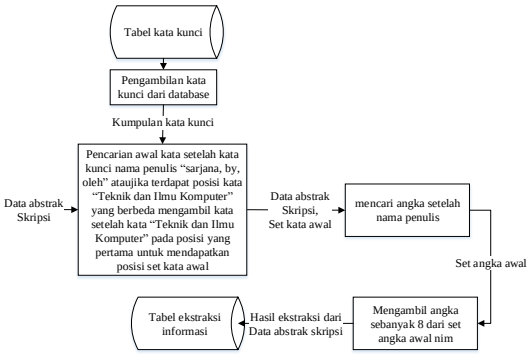


Figure 17. nim extraction

d. Extraction of abstract contents

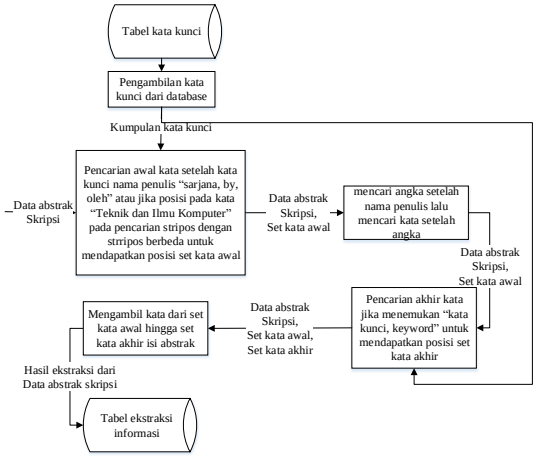


Figure 18. Extraction of abstract contents

e. Ekstraksi kata kunci

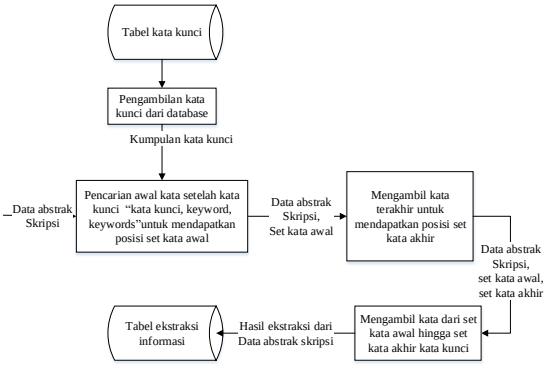


Figure 19. Keyword extraction

3. TESTING AND DISCUSSION

Extraction testing is a stage that has a goal to determine the performance of the method of feature selection used on the built system that is matching data extraction information with manual search. Divided into 3 tests of comparative document testing with system data on 50 thesis cover data, abstract skripsi data, and abstract thesis data.

3.1 Comparison of Thesis Cover Data

Comparison of cover thesis data is a comparison test document with system data on thesis cover data. The results of the comparison made are seen in table 3 below.

Tabel 3. The results of comparative document testing with cover thesis data system data

No	Testing Thesis Cover Data	Graduation year	**Description**	Conclusion
1	gelar budidarma djamal	2017	Extraction is appropriate	Successful Extraction
2	idwar halid	2017	Extraction is appropriate	Successful Extraction
3	narji jaariah	2017	Extraction is appropriate	Successful Extraction
4	faisal muslim	2017	Extraction is appropriate	Successful Extraction
5	dede juniawan suri	2017	Extraction is appropriate	Successful Extraction

3.2 Comparison of Thesis Data Abstract

Comparison of abstract skripsi data is a comparison test of document with system data on abstract skripsi data. The results of the comparison made are seen in table 4 below.

Tabel 4. The results of comparative document testing with abstract thesis data system data

No	**Testing Thesis Abstract Data**	**Graduation year**	**Description**	**Conclusion**
1	gelar budidarma djamal	2017	Extraction is appropriate	Successful Extraction
2	idwar halid	2017	Extraction is appropriate	Successful Extraction
3	narji jaariah	2017	Extraction is appropriate	Successful Extraction
4	faisal muslim	2017	Extraction is appropriate	Successful Extraction
5	dede juniawan suri	2017	Extraction is appropriate	Successful Extraction

3.3 Comparison of Thesis Data Abstract

Comparison of abstract thesis data is a comparison test document with system data on abstract thesis data. The results of the comparison made are seen in table 5 below.

Tabel 5. The results of comparative document testing with abstract thesis data system data

No	**Testing Thesis Abstract Data**	**Graduation year**	**Description**	**Conclusion**
1	gelar budidarma djamal	2017	Extraction is appropriate	Successful Extraction
2	idwar halid	2017	Extraction is appropriate	Successful Extraction
3	narji jaariah	2017	Extraction is appropriate	Successful Extraction
4	faisal muslim	2017	Extraction is appropriate	Successful Extraction

No	Testing Thesis Abstract Data	Graduation year	Description	Conclusion
5	dede juniawan suri	2017	Extraction is appropriate	Successful Extraction

4. CONCLUSION AND SUGGESTIONS

4.1 Conclusions

The conclusions obtained from the research that has been done is known that from 50 documents thesis cover, abstract, and abstract there is no extraction that failed or not appropriate so that the extraction accuracy done on 3 cover thesis document, abstract, and abstract thesis that is 100%. It is concluded that the extraction of this research information can be used to extract the thesis data on the cover, abstract, and abstract documents.

4.2 Suggestions

Based on the results of research that has been done, the problems that emerged in this study the results of the conversion from pdf to html find irregular symbols so that the extraction of information made unsuccessful. As for suggestions for further study as follows..

1. Use a more appropriate library so as to reduce errors on extraction.

DAFTAR PUSTAKA

[1] J. P. a. R. Yangarber, “Informastion Extraction: Past, Present and Future,” 2013.

[2] B. G. Buchanan, Principles of Rule-Based Expert Systems, 1982.

[3] A. D. Nindyati, “Buku Panduan Penulisan Skripsi atau Tugas Akhir Universitas Parmadina,” 2010.

[4] M. Krifka, Basic notions of information structure, 2008.

[5] M. H. Taqwim, “Ekstraksi Informasi Dengan Metode Rule - Based untuk Evaluasi Pemahaman Fisika Kinematika,” *Jurnal Informatika,* 2016.

[6] A. Ismaya, “Algoritma Ekstraksi Informasi Berbasis Aturan,” *Jurnal Informatika,* 2014.