International Journal of Informatics and Information System ISSN 2579-7069
Vol. 3, No. 1, March 2020, pp. 36-42

System of Information Feedback on Archive Using Term Frequency-Inverse Document Frequency and Vector Space Model Methods

Didit Suhartono [a,*], Khodirun [a]

[a] Informatics Engineering Program, Universitas Amikom Purwokerto
* corresponding author

Abstract

The archive is one of the examples of documents that important. Archives are stored systematically with a view to helping and simplifying the storage and retrieval of the archive. In the information retrieval (Information retrieval) the process of retrieving relevant documents and not retrieving documents that are not relevant. To retrieve the relevant documents, a method is needed. Using the Term Frequency-Inverse Document and Vector Space Model methods can find relevant documents according to the level of closeness or similarity, in addition to applying the Nazief-Adriani stemming algorithm can improve information retrieval performance by transforming words in a document or text to the basic word form. then the system indexes the document to simplify and speed up the search process. Relevance is determined by calculating the similarity values between existing documents by querying and represented in certain forms. The documents obtained, then the system sort by the level of relevance to the query.

Keywords: Archive; Information retrieval; TF-IDF; Vector Space Model.

1. Introduction

The archive is one of the examples of documents that have a value (important). The archive can be a letter, warranty, deed, charter, book, etc., which can be used as proof for action and decision. According to Kumoro [1] archives are essentially systematically stored with the intent to assist and facilitate the storage and rediscovery of the archives.

In ODLIS, the information retrieval is the actions, methods and procedures for rediscovering stored data, and then providing the required information. The purpose of the interview information is toretrieve the relevant documents and nottoretrieve the irrelevant documents. A document is considered relevant if a document matches a user's question.

However, to retrieve the relevant documents needed a method. The Matching String method and the Knuth-Morris-Pratt algorithm can perform data retrieval process based on word input that generates information quickly but Knuth-Morris-Pratt algorithm can only handle string problems that are exact string Matching, i.e. find a string in the document exactly the same as the string input and required other algorithms for more advance [2]. Another method that is Term Frequency-Inverse Document Frequencyis to do the weight calculations in each word (term) and with the Vector Space Model to see the level of proximity or similarity can find the relevant documents to the keyword entered by the user [3] by generating value precission between 0.346154 and 1.

In the current research used the Term Frequency-Inverse Document Frequency method and Vector Space Model as well as the addition of stemming algorithm. Stemming is used to improve the performance of information retrieval by transforming words in a document or text to its basic word form. According to [3] the use of steeming can reduce the size of the index on the information retrieval system, so that it can increase the precission value of the search results of the information retrieval.

By implementing this method, the system performs indexing of documents to facilitate and expedite the search process. Relevance is determined by calculating the value of similarity between the existing documents and the query, and represented in a specific form. The documents are obtained, then the system sorts them by their relevance level to the query.

Public Middle School in Nusawungu is one of the first public middle school in the area of Nusawungu Sub-district of Cilacap Regency, which became the easternmost district in Cilacap Regency. The archiving system in Public Middle

School in Nusawungu is good enough. In 2017 there were 301 incoming letters and 689 letters out and in 2018 there were 194 incoming letter and 579 letters out.

The first step is to do the entry letter or outgoing letter is to do the recording on the letter of the agenda or the exit letter book. Then the archive and incoming letter are stored and grouped by their respective code numbers. At the time of searching for incoming and outgoing letter archives The first step is to perform the relevant filing checks on the agenda of the letter of entry and exit letter. Then look for the archive in the archive storage.

This process is considered less precise because it has to be checked one by one every agenda number that is recorded in the book Agenda and the letter of the agenda. So if at any time leader or other interested parties want to see the archives that have been stored, it takes a long time and is not necessarily accurate.

Based on the above issues, authors are interested in designing and building a system that can help rediscover the files that are already stored.

2. Research Methods

2.1. Research Time and Location

This research was conducted in Public Middle School in Nusawungu which is located in Rawabangus Street, Danasri, Nusawungu, Cilacap regency, Central Java, 53283. The study began in November 2018 – March 2019.

2.2. Data Collection Methods

a. Interview

According to [4] interviews are used as data collection techniques when researchers want to conduct preliminary studies to find issues to be researched, and also when researchers want to know the things of the respondents are more profound and the response amount is a little/small.

In this case the researcher conducted an interview with an employee of governance in Public Middle School in Nusawungu, especially the field of public administration on the flow of acceptance to the filing of letters and problems encountered in the filing process and re-discovery of the archive data.

b. Observation

Observation as a data collection technique has a specific characteristic compared to other techniques, namely interviews and questionnaires. According to [4] observation is a complex process, a process composed of various biological and psychological processes. In this research, the author performs observations directly by observing the process of receiving letters, disposition of letters up to the filing of letters and re-discovery of letter archives.

c. Literature Studies

The study of the library is a study related to theoretical studies and other references related to the study, in addition to that the studies of the library are very important in conducting a study, this is because the research will not be separated from scientific literature [5][9][10].

Researchers do data collection and information through books, internet, scientific journals, and other literature, it is done to obtain information about the background of research, system development techniques, research literature Related support and basic research.

d. Documentation

Documentation is an event record that has passed. Documents are usually written in the form of one's writings, drawings, or works of a person [5].

Researchers perform documentation to obtain supporting data such as the Administration room, filing cabinet, entry letter and exit letter, disposition sheet and others found in Public Middle School in Nusawungu.

2.3. Tool and Research Materials

a. Tool

1) Hardware

Hardware used in the creation of web applications archival in the Public Middle School in Nusawungu is as follows:

Table 1. Hardware Used

Hardware	Unit
Processor Intel ® Core™ i3-4030U	1 Unit
RAM 6 GB DDR3 L Memory	
500 GB HDD	
Printer HP Deskjet 2130 Series	

2) Software

Software used in the creation of web applications archival in Public Middle School in Nusawungu is as follows:

Table 2. Software Used

No	Nama Software	Keterangan
1.	Windows 8.1 Pro	Operating System
2.	Sublime Text 3.0	Program Editor Application
3.	XAMPP	Supporting Software
4.	Web Browser	Display Application Results

b. Materials

1) Materials for application creation:

- Codeigniter Framework
- Bootstrap

2) Application content Material:

- Incoming letter
- Outgoing letter
- Sheet disposition

2.4. Research Concept

In this study the first step was the method of data analysis. Data analysis is the process of finding and structuring systematically the data obtained from interviews, field records, and other materials, so it can be easily understood, and the findings can be informed to others [5].

a. Research Framework

The framework of thinking is a series of charts describing the flow of the research process in the creation of a letter archival application in Public Middle School in Nusawungu.

1) Identification

The identification process is the author stage to identify problems that exist in the research object[13]. This stage is an important step to formulate a problem that will be the background in the research object. The problem identified is how to create an archive application by implementing an information gathering system to simplify the process of archiving and re-discovery of archival data[14].

2) Data Collection Techniques

This stage is a stage that researchers do to collect the data needed to complement all research materials. This stage is done with several techniques, namely by means of observation techniques, interviews, documentation and library studies.

3) System needs Analysis

After the data is collected, then analyze the necessary needs in the creation of systems both from hardware, software, user needs, and the process of analyzing the data.

4) Application design

The analysis stage has been completed, the next stage by designing application creation with the stages of the extreme programing development method.

5) Planning

Planning is done by collecting the need for the design of a lettering archive application by iterating repeatedly until all the needs for the application are fulfilled.

6) Design

Designing is done by creating a Unified Modeling Language (UML) that assists with the descriptor and design of software systems. The modelling created include create use case, activity diagram, class diagram, sequence diagram.

7) Coding

Coding is done using PHP programming examples language and is done gradually by researchers according to the needs in the built-in application.

8) Testing

Testing this app with acceptance testing and user testing. Acceptance testing is done by testing the feasibility of the overall software features, while unit testing is done by checking the suitability between the input and output of the application using blackbox testing.

9) App Launch

Release is the end point of the system development process with the Extreme Programmingmethod. In this part of the application that has been created and passed the test will be implemented to the state Public Middle School in Nusawungu which will later be used for archiving letter data using the website.

10) Report

This stage is the final stage of the research conducted, all the research results are reported in the form of a scientific writing. In this report there are also conclusions and suggestions for this study.

b. System development Methods

The method used for system development in this research is the method of Extreme Programming (XP) which is one of the methods that belong to Agile Methodology which uses Unified Modeling Language (UML), or also known as visual modelling.

1) Planning

The planning or planning phase begins with listening to an activity aimed at collecting the needs of the software to be developed and to get a viewof output, key features and functionality. These listening activities are commonly referred to as a series of stories or user Stories describing the output required of features, and the functionalities that will be built using the software to be developed. At this stage customers and developers work together to decide how to group the story into its continuation for further development by the development team[15].

2) Design

Design on XP is done by following the principle keep It Simple (KIS). Simpler designs are always more liked compared to complex designs. XP encourages using a CRC-card as an effective mechanism for thinking about software in an object-oriented context. A CRC (class-reponsibility-collaboration) card is used to identify and regulate classes that are relevant to the current quality improvement of the software, the CRC card is the only design work product that is produced as part of the rapid Programming software development process.

3) Coding

This XP stage begins with building a series of tests (Unit Test), after which the developer has to focus on implementation to pass the test. In XP also introduced the term Pair Programming where the process of writing a program is done in pairs. Two progammers cooperated together on the computer to write the program. By doing this will be gained real-time problem solving and Real-time quality assurance.

4) Testing

Testing is done by testing the code in unit testing. In XP there is also acceptance test or commonly called customer test. This test is done by The customer that focuses on the features and functionality of the system as a whole. Acceptance Test comes from user Stories that have been implemented.

Unit testing in the archives application in the administration of Public Middle School in Nusawungu using blackbox testing. The blackbox Trial method focuses on the functional needs of the software. Therefore, blackbox test allows software developers to create a set of input conditions that will train all the functional requirements of a program, the blackbox trial is an approach to finding errors.

Then user acceptance testing done by the user and essentially focuses on the feasibility test features – software features and functionality. The feasibility test comes from a user response that has been implemented as part of a software release by filling out a questionnaire.

3. Discussion

The application of the TF-IDF and Vector Space Model Methods in the information gathering system in the section of Public Middle School in Nusawungu uses Extreme Programming method. With this method developers and clients can cooperate with each other for the success of a system. As previously explained, this method of Extreme Programming has four phases: planning, designing, coding, and testing.

3.1. Planning

At the planning stage, the author interviews to obtain information as a description of the feature or user story. The author meets with one of the employees in the Public Middle School in Nusawungu. In the interview is explained about what is the part of Public Middle School in Nusawungu, then talk about archived data and archiving process. In this case the archive document is a letter file. Then also told about the problems that exist and the needs of the system to be built and then done identification or analysis needs to build the application.

From the results of identification and data obtained can be concluded that it has not been a system of information gathering on the Archives in the Administration section of Public Middle School in Nusawungu. According to the results of the interview that the archive application is needed for the archiving process and for the re-discovery process can be easily done by implementing an information-gathering system. Then the author understands the business context, describingthe output, the required features, and the functionality to be built using the software to be developed.

3.2. Design

At the design stage make system design using a CRCcard (Class Responsibility Collaborator) because in the methodology of XP more emphasis on simple system design. Then for the modeling of systems used in this application is Unified Modelling Language (UML) which is the standard for visualization, designing, and documentation of a software or an object oriented application.

a. The data collected is preprocessing. Preprocessing includes the removal of a word that is considered insignificant (Stopword) and performed stemming, which is to change the word to its basic form by eliminating the initial suffix or ending. From this process will be generated more compaq word or term lists but still represent the documents being processed

b. After preprocessing, the next step is to take each word/term and calculate the number of appearances on a specific document.

c. Done word-weighted using TF/IDF formula

$$IDF_{(t)} = log\ (D/df_{(t)})$$

Where

DF (t) = number of documents containing the word to-t of the keyword

D = number of all documents contained in the data base

IDF = document frequency ratio on the T-word of the keyword

Calculation of TF-IDF using equation 2

$$TF\text{-}IDF(d,t) = TF(d,t) * IDF(t)$$

Where:

D = document to-D

t = Word to-t of a keyword

tf = frequency of number of words to-t of the keywords in the D-document

TF-IDF = weight of the D document against the T-keywords

IDF = document A ratio on the T-word of the keyword

d. Step indexing is done to save each word/term into the database with the attributes of the number of occurrences and the weight of each term.

e. Once data is processed, it will be done with the calculation of similarity (similarity) between the query of user requests and documents stored in the database. Calculations are done using the Vector Space Model. Then the result will be displayed some relevant documents with the query in sequential order based on similarity. Similarities between vector documents and vector queries will be calculated with the cosine approach. Measurement of Cosine Similarity using equations

3.3. Coding

Coding at this stage is encoding using the codeigniter framework. Frameworks can simply be interpreted like a library containing a set of functions or procedures and classes for a specific purpose and ready to use so as to simplify and accelerate the job of a Programmer. The Framework itself is MVC-based (Model View Controller) is a module – a small module that is separate from one with the other so that the program codes can be arranged properly and conditionally.

3.4. Testing

At this stage testing is conducted using Unit Testing Test writing the program code in the smallest unit individually. The Unit Testing phase is done every time you finish writing the program code then it is immediately done testing the code unit in the application features that have been created to know if the feature is already running as desired or not. Acceptance testing is done to determine whether the built-in system has fulfilled the acceptance criteria as well as determining whether the system is acceptable or not.

4. Conclusion and Suggestions

4.1. Conclusion

Based on the discussion and description in the previous chapters, it can be concluded as follows:

a. The information return system can be applied in the archive data.

b. With TF/IDF method The resulting data is more relevant.

c. The implementation of the archive information return system can simplify the discovery process of archival data.

d. This website-based application facilitates access for teachers, employees and school principals in the search for archival data.

4.2. Suggestion

a. This information retrieval system can be applied to other document searches.

b. Implement other methods to test effectiveness and accuracy in data retrieval.

References

[1] G. Eason, B. Noble, and I.N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955.

[2] S.Tan, Neighbor-weighted K-nearest neighbor for unbalanced text corpus, Expert Systems with Applications 28 (2005) 667–671.

[3] G.Salton, C.S.Yang, On the specification of term values in automatic indexing, Journal of Documentation, 29 (1973) 351-372.

[4] W.Zhang, T.Yoshida, A comparative study of TF-IDF, LSI and multi-words for text classification, Expert Systems with Applications 38 (2011) 2758–2765.

[5] H.Han, G.Karypis, V.Kumar, Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification, PAKDD (2001) 53-65.

[6] F.Sebastiani, Machine Learning in Automated Text Categorization, Consiglio Nazionale delle Ricerche, 2002.

[7] H.Jiang, P.Li, X.Hu, S.Wang, An improved method of term weighting for text classification, Intelligent Computing and Intelligent Systems,2009.

[8] J. T.-Y. Kwok, Automatic Text Categorization Using Support Vector Machine, Proceedings of International Conference on Neural Information Processing, (1998) 347-351.

[9] M.Miah, Improved k-NN Algorithm for Text Classification, DMIN (2009) 434-440.

[10] Y.Liao, V. Rao Vemuri, Using K-Nearest Neighbor Classifier for Intrusion Detection, Department of Computer Science, University of California, Davis One Shields Avenue, CA 95616.

[11] L.Wang , X. Zhao, Improved KNN classification algorithms research intext categorization, IEEE, 2012.

[12] M.Lan, C.L.Tan, J.Su, Y.Lu, Supervised and Traditional Term Weighting Methods for Automatic Text Categorization, IEEE Transactions on Pattern Analysis and Machine Intelligence, VOL. 31, NO. 4, 2009.

[13] K. Mikawa, T. Ishidat, M.Goto, A Proposal of Extended Cosine Measure for Distance Metric Learning in Text Classification, 2011.

[14] l.Wang , X. Li, An improved KNN algorithm for text classification, 2010.

[15] G. Guo, H.Wang, D.Bell, Y. Bi, K. Greer, KNN Model-Based Approach in Classification, (2003) 986 – 996.