International Journal of Computer Applications Technology and Research
Volume 7–Issue 08, 292-296, 2018, ISSN:-2319–8656

Measure the Similarity of Complaint Document Using Cosine Similarity Based on Class-Based Indexing

Syahroni Wahyu Iriananda
Electrical Engineering Department,
Faculty of Engineering,
Brawijaya University.
Malang, East Java, Indonesia

Muhammad Aziz Muslim
Electrical Engineering Department,
Faculty of Engineering,
Brawijaya University
Malang, East Java, Indonesia

Harry Soekotjo Dachlan
Electrical Engineering Department,
Faculty of Engineering,
Brawijaya University
Malang, East Java, Indonesia

Abstract: Report handling on "LAPOR!" (Laporan, Aspirasi dan Pengaduan Online Rakyat) system depending on the system administrator who manually reads every incoming report [3]. Read manually can lead to errors in handling complaints [4] if the data flow is huge and grows rapidly, it needs at least three days to prepare a confirmation and it sensitive to inconsistencies [3]. In this study, the authors propose a model that can measure the identities of the Query (Incoming) with Document (Archive). The authors employed Class-Based Indexing term weighting scheme, and Cosine Similarities to analyse document similarities. CoSimTFIDF, CoSimTFICF and CoSimTFIDFICF values used in classification as feature for K-Nearest Neighbour (K-NN) classifier. The optimum result evaluation is pre-processing employ 75% of training data ratio and 25% of test data with CoSimTFIDF feature. It deliver a high accuracy 84%. The k = 5 value obtain high accuracy 84.12%

Keywords: Complaints Document, Text Similarity, Class-Based Indexing, Cosine Similarity, K-Nearest Neighbour, LAPOR!

1. INTRODUCTION

The amount of incoming complaints and public opinion data on "LAPOR!" system (Online Peoples Complaint Service and Aspirations) can serve as a source of information to measure the performance of government service [1]. It processes an average of 900 reports every day, only 13 % - 14% of the reports, while about 86% remain subject to unknown and archived. The most used channel is via SMS s around 80 % - 90% report [2]. The report handling depends on the system administrator who reads every incoming report [3] . This can lead to errors in handling complaints [4], and if the data flow is very large it can take at least three days, this is sensitive to inconsistencies [3]. Limited administrators and high complaint report rates are a major cause of the lack of quality of service responsiveness characteristics [2]. A solution to that problem of complaints analysis is needed. It could help the "LAPOR!" Administrator in determining the category, so big data analysis becomes very important [2].

In this study the authors propose a model or approach that can measure and identify the similarity of document reports conducted in computerized that can identify the similarity between the Query (Q) with Document (D) collections,. This research employs Class-Based Indexing term weighting scheme, then compare with other term weighting schemes like TFIDF and TFICF. The weight values of TFIDF, TFICF, and TFIDFICF then converted into Cartesian coordinates and calculated similarities using the Cosine Similarity function to analyze the resemblance of text documents by obtaining similarity by measuring it in vector space model. Cosine Similarity value from those weighting scheme (CoSimTFIDF, CoSimTFICF, CoSimTFIDFICF) to be setup as a set of features for the classification process. Next is the process of classifying the text using the K-Nearest Neighbor (K-NN) method for document classification and predicting new document categories based on those features. This study aims to identify and evaluate text similarity using TFIDFICF (Class Indexing Based) method and Cosine Similarity.

Relevan research conducted [6] by utilizing TF.IDF.ICF for e-complaint classification of students using Centroid Based Classifier, combined with TF.IDF.ICF, Cosine Similarity, and Class Feature Centroid. [7] Categorize creative ideas on a company using K-NN and TF.IDF.ICF algorithms. [8] Classifying SambatOnline complaint of Malang City using K-NN algorithm, Cosine Similarity and Chi-Square than TFIDF. [9] Using the K-NN algorithm and TFIDF feature selection, and Categorical Proportional Difference (CPD). The same dataset is used [10] by employing the NW-K-NN algorithm, the term weighting TFIDF filter N-Gram, and Unigram on preprocessing. The experimental results [11] show that the classification of text can be/is used to evaluate the quality of service with the data text of handling a customer complaint (complaint). This method can solve the automatic evaluation problem in customer complaint handling management. [11]

2. METHOD

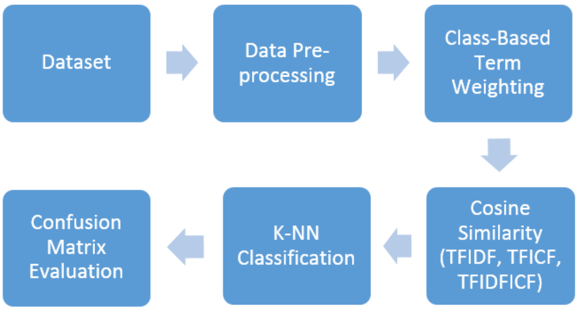

Dataset
Data Pre-processing
Class-Based Term Weighting
Cosine Similarity (TFIDF, TFICF, TFIDFICF)
K-NN Classification
Confusion Matrix Evaluation

Figure 1 Document Classification Framework

Generally the problem-solving framework can be seen in Figure 1, which consists of Data Collecting (Dataset), Text Preprocessing, Text Representation, Feature Selection (Feature Selection) includes common term weighting scheme (TFIDF) and Class-Based Indexing (ICF), K-NN classification, and Confusion Matrix evaluation.

This study uses three weighting schemes for comparison and evaluation to obtain weightings that have the most optimal results. Tests conducted are experiments on the process of Pre-processing that is through the sub-process Stemming and not using Stemming. Experiments with different term weightings using TF-IDF, TF-ICF, and TF-IDF-ICF along with Cosine Similarity variation experiments based on each term weighting. Experiments with variations in the amount of data, and variations in ratio of training data and data testing. Then on the final result will be evaluated the effect of both performances.

The method used to analyze the similarity between the newly reported incoming reports (Query) and the report that the administrator has processed (Document) is Cosine Similarity. The term-weighting results with TF-IDF, TF-ICF and TF-IDF-ICF are then converted into Cartesian coordinates and calculated using the Cosine Similarity function to obtain the angle of similarity and measure the vector distance. The textual classification process is based on the Cosine Similarity feature of TF-IDF (CoSimTFIDF), TF-ICF (CoSimTFICF), TF-IDF-ICF (CoSimTFIDFICF) using different weighting schemes. The greater the value of the three cosine similarity features that are close to the value of 1 (one), then the more like a Query (q) with Document collection (d). Method K-Nearest Neighbor (KNN) chosen to classify and predict the category of the Query.

2.1 Class-Based Indexing

A category-based term weighting scheme is proposed [12]. This research introduces Frequency Category Reverse *(Inverse Category Frequency)* in the term weighting scheme for text classification tasks. Two concepts are defined as *Class Frequency (CF)* is the number of categories in which the term *(t)* appeared and *Inverse Class Frequency (ICF)* whose formula is similar to IDF [12]. The next *Class-Based Indexing (ICF)* concept was developed by [13]. *Inverse Class Frequency (ICF)* pay attention to the occurrence of terms in the category/class set. Term rarely appears in many classes is a term that is valuable for classification. The less the occurrence of the term, the value will be greater or closer to the value of 1 (one), and conversely the more the occurrence of the term is the value of smaller or close to the value of 0 (zero). The importance of each term is assumed to have a proportion that is in contrast to the number of classes containing term. Accurate indexing also depends on the term importance of the class or the scarcity of terms in the whole class *(rare term)*. So we need class-based term weighting called inverse class frequency (ICF). However, ICF only takes into account the terminology of the class regardless of the number of terms in the document into the class. In this research we use traditional TF-IDF-ICF [13] The following formula of ICF is calculated by the formula:

$$ICFLog_i = Log_2\left(\frac{|C|}{cf_i}\right) \qquad (1)$$

Where C is the number of uh classes/categories in the collection (cf$_i$) is the number of classes/categories containing terms.

In classical VSM, which relies on document indexing, the digital representation criteria of the text in the document vector are the product of local parameters (frequency terms) and global parameters (IDF), ie TF.IDF. In the category of tasks correspond to the class frequency, term weighting scheme, the ICF (categorical global parameter) is multiplied by TF.IDF, generating TF.IDF.ICF. And the formula of traditional TF.IDF.ICF shown on equation (2).

$$W_{TF.IDF.ICF}(t_i, d_j, c_k) = tf_{(t_i,d_j)} \times \left(1 + \log\frac{D}{d(t_i)}\right) \times \left(1 + \log\frac{C}{c(t_i)}\right) \qquad (2)$$

Where C denotes the number of categories defined in the collection, c(t$_i$) is the number of categories in the collection where it occurs at least once, c(t$_i$)/C is known as CF, and C/c(t$_i$) is the ICF of term t$_i$.

2.2 Cosine Similarity Measure

$$\cos = \frac{\boldsymbol{Q} \bullet \boldsymbol{D}}{|\boldsymbol{Q}||\boldsymbol{D}|} = \frac{\sum_{i=1}^{n}\boldsymbol{Q}_i \; X \; \boldsymbol{D}_i}{\sqrt{\sum_{i=1}^{n}(\boldsymbol{Q}_i)^2} \; X \; \sqrt{\sum_{i=1}^{n}(\boldsymbol{D}_i)^2}} \qquad (3)$$

Where Q denote the vector of documents, D is the query vector. Q • D is the multiplication of dot vectors Q and vector D it's obtain inner product. |Q| is the length of vector Q (Magnitude of Q) while |D| is the length of vector D (Magnitude of D) then |Q||D| is the cross product between |Q| and |D|. The weight of each term in a document is non-negative. As a result the cosine similarity is non-negative and bounded between 0 and 1. Cos (Q$_i$D$_i$) = 1 means the documents are exactly similar and the similarity decreases as the value decreases to 0. An important property of the cosine similarity is its independence of document length. Thus cosine similarity has become popular as a similarity measure in the vector space model [14]

2.3 Preparation and Data Processing

In this study, main dataset using published "LAPOR!" complaint stream data that published on public data sharing portal http://data.go.id. This data can be freely downloaded at the open government data sharing *(Indonesia Open Government)*. This study uses several experimental scenarios, one of which is the dataset variation shown in Table 3. This scenario aims to investigate the effect of the number of rows of data on related processes.

Table 3 Partition Table Data Document (D) and Query (Q)

Dataset Series	90% (D)	10% (Q)	Amount of Data
Dataset100	90	10	100
Dataset200	180	20	200
Dataset300	270	30	300
Dataset400	360	40	400

In experiment, this research is done *Data Partition* or data partition. This is done by dividing the total number of data rows in the dataset on table 3 into two parts: **1) Dataset *Documents (D)*** employ 90%, **2) Dataset *Query (Q)*** use 10% as in table 3. After the process of Preprocessing with stemming and without stemming the dataset is further divided into two parts, ie 90% for the document dataset (D) and 10% used for the query dataset (Q). Data sharing is also done randomly *(random sampling)* thus obtained members dataset in table 3

3. RESULT AND DISCUSSION

3.1 Comparison of Cosine Similarity Based on Term Weighting

After a series of manual cosine similarity between TFIDF, TFICF, and TFIDFICF we can see the result of cosine similarity compared to figure 2. And in cosine simiality calculation it is found that document recommendation result based on Cosine Similarity value with TFIDF weighting scheme, TFICF, and TFIDFICF is document D5 with the largest cosine value is 0.705 or 70.5% based on TFICF weighting .

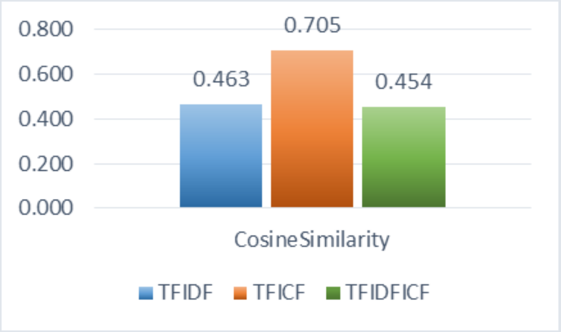


Figure 2 Chart of Cosine Similarity Based on the Term Weighting Scheme

2) Experimental Results Variation Preprocessing

Table 5 Table of Preprocessing Variations Testing Scenarios

Scenario	***Number of D ata***		***Evaluation Results (%)***			
Term Weighting	***Train***	***Test***	***A***	***P***	***R***	***F1***
TFIDF	***75***	***25***	**84.00**	**30.30**	**33.30**	**31.73**
TFICF	***75***	***25***	80.00	24.20	32.40	27.71
TFIDFICF	***75***	***25***	80.00	26.00	32.40	28.85
TFIDF	***90***	***31***	46.15	17.71	19.82	18.71
TFICF	***90***	***31***	**58.06**	**31.88**	**21.13**	**25.42**
TFICFICF	***90***	***31***	45.16	18.14	17.21	17.66

In this experiment aims to evaluate the performance of term weighting Class Indexing Based (TFIDFCF) compared to TFIDF term weighting performance, and TFICF. The dataset used is Dataset200, with 75% training data comparison ratio and 25% data testing. Using six categories and then labeled *(class)* as another name *(alias) is* shown in the following table:

Here are experimental results based on testing with variations of preprocessing stemming and without stemming. The evaluation used is Accuracy (A), Precision (P), Recall (R) and F1-Measure (F1) using macro average model, it is used considering multi-class classification .

Figure 5 Graph of evaluation testing by stemming process (in percent)

3) Variation Testing Weight Feature

Table 6 Test Results Weight TFIDF Feature

Evaluation	**Number of Datasets**			
	100	**200**	**300**	**400**
A	69.23	**84.00**	60.53	63.33
P	18.89	**30.30**	17.36	21.22
R	18.89	**33.33**	18.89	21.15
F	18.89	**31.70**	6 PM	18.47

Based on the results listed in table 6 , the best accuracy for CoSimTFIDF is at Dataset200 which is 84% with F-Measure 31.70%, then Dataset100 with 69.23% accuracy but F-Measure value is quite low at 18.89%, while the highest F-Measure values obtained from Dataset200 obtained the best K-NN classification results for the classification of complaint reports with CoSimTFIDF features. The precision and recall values in Dataset200 also show results with the highest values among other datasets.

Table 7 Test Results Weight TFICF Feature

Evaluation	**Number of Datasets**			
	100	**200**	**300**	**400**
A	76.92	**80.00**	71.05	61.67
P	12.82	24.17	**28.24**	13.57
R	16.67	**32.41**	20.37	19.96
F	14.49	**27.68**	19.72	16:00

Based on the results listed in table 7 , the best accuracy value for CoSimTFICF is on Dataset200 which is 80% with F-Measure 27.68%, then Dataset100 with 76.92% accuracy but F-Measure value is quite low ie 14.49%, Dataset200 obtained the best K-NN classification results for the classification of complaint reports with CoSimTFICF features. The precision also shows good results with 24.17% and the recall value of Dataset200 also shows the highest value of the other datasets of 32.41%.

Table 8 Test Results Weight TFIDFICF Feature

Evaluation	**Number of Datasets**			
	100	**200**	**300**	**400**
A	69.23	**80.00**	60.53	63.33
P	12.50	**25.99**	22.80	19.52
R	3pm	**32.41**	19.63	21.15
F	13.64	**28.7**	18.86	18.86

Based on the results listed in table 8 , the best accuracy for CoSimTFIDFICF is on Dataset200 which is 80% with F-Measure 28.7%, then Dataset100 with 69.23% accuracy but F-Measure value is low ie 13.64%, Dataset200 obtained the best K-NN classification results for the classification of complaint reports with CoSimTFICF features. The precision also shows good results with 25.99% and the recall value in Dataset200 also shows the highest value among other datasets of 32.41%.

4) Accuracy of Classification Process Based on Value k

The experiment uses Dataset200 with a preprocessing process using Stemming and CoSimTFIDF features.

Table 9 Accuracy (%) KNN Based on Value k

Value k Ratio	***1***	***2***	***3***	***4***	***5***
25/75	75.71	80.00	78.57	80.00	**80.00**
75/25	**83.33**	**83.33**	**83.33**	**83.33**	**83.33**
40/60	67.86	80.36	82.14	80.36	**82.14**
60/40	76.32	**84.21**	**84.21**	**84.21**	**84.21**

From the test, the result of K-NN algorithm accuracy with the test of 60% training ratio and 40% test data has a high accuracy level seen from the result of k = 1 is quite low, but increased 8% when testing k = 2 value until k-5 with a stable and equal value of 84.21%, while in this test found that the ratio of 75% training data and 25% test data resulted in an accuracy of 83.33% lower 6.7% of experiments with variations of preprocessing at Table 5.39 is 84% . This is very possible because sampling of trainer data and test data used is *random sampling* method. In this pen can be seen that with the value k = 5 all variations of the ratio of training data and test data has maximum value than other k values. Thus it can be concluded based on the test that has been done that the value k = 5 is the optimal value

5) Results Comparison of KNN Accuracy Based on Features and Dataset

Figure 6 KNN Accuracy Based on Features and Dataset

The experiment was done by determining the dataset used ie Dataset100 and Dataset200. With a comparison ratio of 75% training data and 25% (25/25) of data testing. The k value used is k = 5. In the first test feature used only Cosine Similarity feature based on TFIDF term weighting, then in the next test used Cosine Similarity feature based on TFICF, next CoSimTFIDFICF. The best accuracy result has been achieved using TFIDF-based Cosine Similarity (CoSimTFIDF) feature that 84% in Dataset200 increased 4% from both Cosine Similarity TFICF and TFIDFICF features. While on Dataset100 obtained the best accuracy value using CoSimTFICF feature that is 76,92% increase about 6% from both other features

6) Accuracy Results With Variations of Data Ratios

In this experiment using Dataset200 with variations of preprocessing process using stemming and without stemming. As has been found in Table 5.43 where the value of K that has optimal results is k = 5, then set in this test the classification of K-NN using the value k = 5. Comparative ratio of trainee data and test data for various results. The following is the result of accuracy testing based on the ratio of data and features in table 5.45

Table 10 Accuracy On Term Weighting variations

Pre processing	**Ratio (%)**	**Accuracy (%)**		
		CoSim TFIDF	**CoSim TFICF**	**CoSim TFIDFICF**
With Stemming	25:27	66.67	66.67	66.67
	75:25	**84.00**	**80.00**	**80.00**
	40:60	66.67	78.33	71.67
	60:40	60.00	75.00	70.00
No Stemming	25:27	54.50	64.86	60.36
	75:25	54.05	55.41	55.41
	40:60	52.81	60.67	60.67
	60:40	55.46	57.98	57.98

Based on these results it was found that the optimum accuracy result with preprocessing Stemming and best result of all features is 75% training data ratio and 25% test data on TFIDF feature-based Cosine Similarity that is 84%. Then CoSimTFICF feature with 40% training data ratio and 60% test data

4. CONCLUSIONS

A. Conclusion

In the test results that have been carried out, it was found that **1)** S kem *term* terming TFIDF has a significant influence on the accuracy of classification. **2)** Tests with variations of stemming process using TFIDF-based *Cosine Similarity* feature (CoSimTFIDF) by employing 75 training data and 25 test data resulted in the best K-NN algorithm accuracy of 84%, with 30.3% precision, 33.3% recall, and f-measure 31.73%. This result is better 35% than a preprocessing without stemming is about 58%. **3)** The test to investigate the values of k = 1,2,3,4 and 5 with 100 data of train and test with variation of training data ratio and different test data is value k = 5. The best accuracy value obtained is the ratio of 60:40 that is 84.21%.

B. Suggestion

Some things that can be developed for further research in the same scope include: **1)** We recommend that the addition of variations of preprocessing ie stopword list different languages eg Sundanese, Basaha Java, Slang / Slang and so forth. **2)** *Cross Validation* techniques should also be used to obtain the ratio of training data and proportional K-NN test data .

5. REFERENCES

[1] A. Sofyan And S. Santosa, "Text Mining Untuk Klasifikasi Pengaduan Pada Sistem Lapor Menggunakan Metode C4.5 Berbasis Forward Selection," *Cyberku J.*, Vol. 12, No. 1, Pp. 8–8, 2016.

[2] I. Surjandari, "Application of Text Mining for Classification of Textual Reports: A Study of Indonesia's National Complaint Handling System," in *6th International Conference on Industrial Engineering and Operations Management (IEOM 2016)*, Kuala Lumpur, Malaysia.

[3] A. Fauzan and M. L. Khodra, "Automatic multilabel categorization using learning to rank framework for complaint text on Bandung government," in *2014 International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA)*, 2014, pp. 28–33.

[4] S. Tjandra, A. A. P. Warsito, and J. P. Sugiono, "Determining citizen complaints to the appropriate government departments using KNN algorithm," in *2015 13th International Conference on ICT and Knowledge Engineering (ICT Knowledge Engineering 2015)*, 2015, pp. 1–4.

[5] W. H. Gomaa and A. A. Fahmy, "A Survey of Text Similarity Approaches," *Int. J. Comput. Appl.*, vol. 68, no. 13, pp. 13–18, 2013.

[6] M. A. Rosid, G. Gunawan, and E. Pramana, "Centroid Based Classifier With TF – IDF – ICF for Classfication of Student's Complaint at Appliation E-Complaint in Muhammadiyah University of Sidoarjo," *J. Electr. Electron. Eng.-UMSIDA*, vol. 1, no. 1, pp. 17–24, Feb. 2016.

[7] R. R. M. Putri, R. Y. Herlambang, and R. C. Wihandika, "Implementasi Metode K-Nearest Neighbour Dengan Pembobotan TF.IDF.ICF Untuk Kategorisasi Ide Kreatif Pada Perusahaan," *J. Teknol. Inf. Dan Ilmu Komput.*, vol. 4, no. 2, pp. 97–103, May 2017.

[8] C. F. Suharno, M. A. Fauzi, and R. S. Perdana, “Klasifikasi Teks Bahasa Indonesia Pada Dokumen Pengaduan Sambat Online Menggunakan Metode K-Nearest Neighbors (K-NN) dan Chi-Square,” *J. Pengemb. Teknol. Inf. Dan Ilmu Komput. Vol 1 No 10 2017*, Jul. 2017.

[9] N. H. A. Sari, M. A. Fauzi, and P. P. Adikara, “Klasifikasi Dokumen Sambat Online Menggunakan Metode K-Nearest Neighbor dan Features Selection Berbasis Categorical Proportional Difference,” *J. Pengemb. Teknol. Inf. Dan Ilmu Komput. Vol 2 No 8 2018*, Oct. 2017.

[10] A. A. Prasanti, M. A. Fauzi, and M. T. Furqon, “Klasifikasi Teks Pengaduan Pada Sambat Online Menggunakan Metode N-Gram dan Neighbor Weighted K-Nearest Neighbor (NW-KNN),” *J. Pengemb. Teknol. Inf. Dan Ilmu Komput. Vol 2 No 2 2018*, Aug. 2017.

[11] S. Dong and Z. Wang, “Evaluating service quality in insurance customer complaint handling throught text categorization,” in *2015 International Conference on Logistics, Informatics and Service Sciences (LISS)*, 2015, pp. 1–5.

[12] D. Wang and H. Zhang, “Inverse-Category-Frequency based supervised term weighting scheme for text categorization,” *J. Inf. Sci. Eng.*, vol. 29, no. 2, pp. 209–225, Dec. 2010.

[13] F. Ren and M. G. Sohrab, “Class-indexing-based term weighting for automatic text classification,” *Inf. Sci.*, vol. 236, pp. 109–125, Jul. 2013.

[14] A. Huang, “Similarity Measures for Text Document Clustering,” in *Proceedings of the New Zealand Computer Science Research Student Conference*, Christchurch, New Zealand, 2008, pp. 49–56.

International Journal of Computer Applications Technology and Research
Volume 7–Issue 08, 297-300, 2018, ISSN:-2319–8656

Security in Software Defined Networks (SDN): Challenges and Research Opportunities for Nigeria.

Abdulsalam S. Mustafa
Khazar University
Baku, Azerbaijan.

Donald Mkpanam
National Institute for Legislative and Democratic Studies, National Assembly
Abuja, Nigeria.

Ali Abdullahi
National Institute for Legislative and Democratic Studies, National Assembly
Abuja, Nigeria.

Abstract: In networks, the rapidly changing traffic patterns of search engines, Internet of Things (IoT) devices, Big Data and data centers has thrown up new challenges for legacy; existing networks; and prompted the need for a more intelligent and innovative way to dynamically manage traffic and allocate limited network resources. Software Defined Network (SDN) which decouples the control plane from the data plane through network vitalizations aims to address these challenges. This paper has explored the SDN architecture and its implementation with the OpenFlow protocol. It has also assessed some of its benefits over traditional network architectures, security concerns and how it can be addressed in future research and related works in emerging economies such as Nigeria.

Keywords: SDN; OpenFlow; Mobile Networks; Network Security; IoT; Big Data

1. INTRODUCTION

1.1 Background

Software Defined Networking (SDN) decouples the control plane from the data plane in order to enhance programmability and flexibility of the control and management of a network. Legacy networks are regarded as complex and rigid, difficult to scale and manage, and too costly but SDN provides a more innovative and dynamic network architecture that transforms traditional network architecture into rich service-delivery platforms [1]. SDN, places a layer of software over the network like a network operating system which interacts with all routers in the network. A major outcome of its design and development is its inherent security and simplified networking. Its emergence offers a robust environment for designing future networks that will be dynamic, cost effective, adaptable, and flexible, and suitable for high bandwidth use and dynamic nature of present applications [2].

The Architecture and Framework working group proposed a Software Defined Network model composed of the application plane, the controller plane and the data plane [3]. SDN developers aim to achieve scalability and agility in network management through separation of the control plane (the controller) which decides where packets are sent from the data plane (the physical network) which forwards traffic to its destination [3]. SDN increasingly uses elastic cloud architectures and dynamic resource allocation in its infrastructure goals [4]. SDN forwarding decisions are flow based in comparison to destination based traditional networks. Openflow was the most commonly used SDN protocol and presently, most companies have decided to adopt different protocols. Some of the protocols currently used are open network environment by Cisco and network virtualization platform by Nicira [3]. This paper's objective is to identify some of the challenges and research opportunities for SDN implementation in Nigeria. Some of the applications of SDN include data center, wide area backbone networks, enterprise networks, internet exchange points and home networks. Figure 1 below consists of the 3 distinct layers: application layer; control layer; and infrastructure or physical layer.

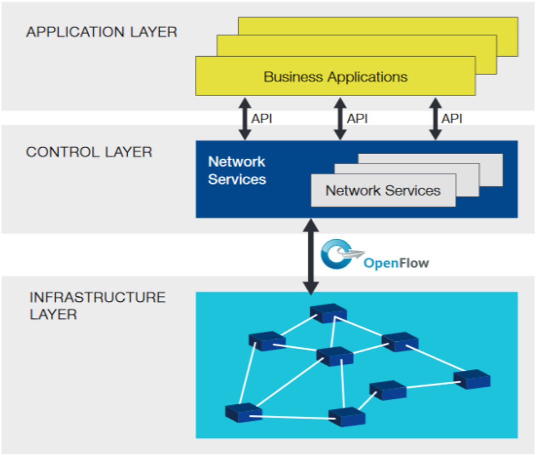


Figure 1. SDN Architecture [3]

1.2 Benefits of SDN

The separation of the control and data planes increases the flexibility of the network to adapt to evolving networks. One of the major benefits for operators and service providers is reduction in operation cost due to centralized management, efficiency in operations and existing hardware being fully utilized. The ability of the networking infrastructure to be programmable and manageable makes it scalable and more dynamic. Other expected benefits include increased network reliability and security discussed in this paper in addition to better user-experience due to SDN ability to adapt to dynamic user-needs. SDN is also expected to manage inflow of traffic from internet of things (IoT) devices by segmenting the traffic and helping to orgainse the data. Furthermore, SDN is expected to enable networks keep pace with the speed of change on a network without the need to continuously invest in new infrastructure or devices.

1.3 SDN and Mobile Networks

OpenFlow-based SDN offers several benefits for mobile networks, including wireless access segments, mobile backhaul networks, and core networks [5], [6]. SDN will enable carrier networks benefit from its architecture by incorporating innovative ways of managing and controlling the network [5], [6]. In addition, it will increase flexibility by enabling the smooth introduction of new service bundles and value-added services at a faster pace in Nigeria. The future 5G network aims to incorporate SDN principle in its framework and in network slicing concept. A recent position paper by Malik et.al. [5] on SDN based mobile networks suggests that it can simplify mobile networks and lower management costs. Furthermore, SDN in mobile networks is expected to provide maximum flexibility, openness, and programmability to future carriers without the need to make changes in user-equipment. In addition, SDN could provide mobile operators with greater control over their equipment and infrastructure and simplify network management.

2. LITERATURE REVIEW: RELATED WORKS

In a computer network, communication is effected between the network and its host through the use of switches and routers configured for data packet and routing functionality. Configurations on the devices occur through a process of translating high-level network policies into device-specific low-level commands, which is manually done through command-line or graphical user interfaces (GUI) [7]. This is vulnerable to security issues; exploitation; threats and attacks on the network including Denial of Service (DoS) attacks; compromised controller attacks (faulty or hijacked controllers); spoofing attacks; malicious interjection; traffic anomaly; and forwarding control link attacks.

Colville & Spafford [8] reveal, that lack of integrated network control creates network management challenges and the error-prone configuration process triggers network faults, bugs, and security lapses. Feldmann et al. [9] suggest that because of inflexibility, network innovation has essentially stagnated. However, SDN model frontally addresses this challenge by separating the packet forwarding functionality of the forwarding devices or data plane from the control element or control plane [6]. The separation technique which is technically called decoupling remains a key feature of SDN. Decoupling spawns innovative network architecture where the network switches functions such as basic forwarding devices and the control logic is implemented in a logically centralized controller [10].

Akhunzada et.al. [11], argue that the integrity and security of SDNs remain unproven regarding the placement of management functionality in a single centralized virtual server making it easier to compromise the whole network through a single point of failure. However, Medhi et.al. [12], claim that SDN provides a unique opportunity for effectively detecting and containing network security problems in home and office networks. The research findings of Medhi et.al. [12], reveal four prominent track anomaly detection algorithms which can be implemented in an SDN framework using Open flow compliant switches and NOX (open source development platform for C++ based SDN control applications) as a controller. They further indicated that these algorithms are significantly more accurate in detecting malicious activities in the home networks in comparison to the Internet Service Provider (ISP) [12].

SDN's major security issue is being self-secure. Kreutz et.al. [6] advocated incorporating security and dependability into the SDN architecture from the ground level up. According to them, SDN is susceptible to several threats such as forged traffic flow to attacking network entities; Denial of Service (DoS) attacks on switches, controllers and control plane communications [13]. Potential attacks on the interface between the controller and high-level applications, exploiting the weaknesses in Secure Socket Layer (SSL) and Transport Layer Security (TLS) protocol implementations in addition to switches in the network may be hijacked or exploited [14]. These are missing gaps that this study will attempt to address on the security issues in the evolution of SDN and its adoption by service providers in Nigeria.

Kreutz et.al. [15] proposed stringent authentication mechanisms and trust models which could counter common identity-based attacks as few of the potential solutions to the identified threats inherent in the current SDN is a monotony-regime [16]. Therefore, there is need to diversify the protocols, controllers, and tools employed and consequently reduce common implementation vulnerabilities, a major focus of this study. Shin et.al. [17] propose FRESCO, a security-specific application development framework for OpenFlow networks for securing the design of SDN. FRESCO simplifies transferring of the application programming interface (API) scripts to enabling the development of threat-detection logic and security monitoring as programming libraries [17].But Akhunzada et.al. [11] state that, FRESCO does not improve the security of the application and infrastructure layers of SDN.

As alternatives, Shirali-Shahrez and Ganjali [18], propose FleXam, a sampling extension for OpenFlow to enhance the security of SDN while Shing and Gu [16], propose CloudWatcher, a framework for monitoring clouds. Kreutz et.al. [13] propose L-IDS, a learning intrusion detection system to protect mobile devices in a specific location which they regard as a prominent solution for security enhancement. Also, Wang et.al. [19] offer a systematic approach to detecting and resolving conflicts in an SDN firewall by checking firewall authorization space and flow space using 'header space analyses' to investigating the effectiveness and efficiency of this approach in addressing security analyses threats.

Shin et.al. [17] suggest the use of connection migration, an extension to the data panel to reducing interactions between data and control panel to addressing DoS attacks on the southbound interface. This is like the approach proposed by Ying-Dare et. al. [20], for reducing the traffic overhead to the controller and providing NFV through an extended SDN architecture. Their evaluation show that in the extended SDN architecture, only 0.12 percent of the input traffic is handled by the controller extended, while 77.23 percent is handled on the controller in conventional architecture [20]. Akhunzada et. al. [11] also claim that, OpenWatch, an adaptive method of flow counting to detect anomalies in SDN is a credible solution for security analyses and is expected to improve the overall security of Network protocols such as OpenFlow. Ali et al. [7] points out, that as cyber-threats continue to evolve and become more sophisticated, the potentials of a highly configurable network attack is catastrophic, hence, the need to move away from the reactive strategy approach common to legacy networks. It is evident that there a lot of vulnerabilities which could target SDN. This study will address security threats to the configuration of SDN so as make it suitable for large scale adoption and deployment in Nigeria.

3. DISCUSSIONS AND FUTURE RESEARCH

As SDN is being adopted, there is demand for secured SDN solutions and a more adaptable secured framework. Several issues relating to SDN are currently actively being researched, however, there is also need for security vulnerability assessment because this is an important process that must be conducted to fully secure a system before its deployment. The Control plane in SDN handles configuration management of devices, responsible for routing decisions and monitoring the network. The controller is considered as a single point of failure [15], and it is a major security target. In this regard, there is need to investigate new security architectures for the controller to support more innovative security services and intelligent network defense systems.

FRESCO by Shin et.al. [17], is an extension of the research work done by Kreutz et.al. [15], which that makes it easy to make and deploy security services in SDN, however, they believe none of those works adopts or enforces the security of SDN itself [15]. Furthermore, there is need for research on creating more secured and resilient SDN controllers and approaches to addressing the security issues. This therefore generates the normative question of how innovative SDN-based security applications can potentially replace existing security applications? There is also the question of how new vulnerabilities in SDN controllers can be exploited through threat vectors and possible solutions and improvements to address these problems? In addition, there is the need to question the handling of malicious applications being developed and deployed on SDN controllers?

4. CONCLUSIONS

In this paper, we have discussed the concept of Software Defined Network framework which uses network virtualization to separate the control plane from the data plane. With a centralized control, SDN can easily manage and enable networks to adapt and cope with unpredictable traffic patterns which can place high demands on the limited network resources. However, in any network, security is a major concern, therefore, it is imperative to do an analysis of security challenges in SDN from the perspective of attacks on SDN controllers and development of a new mitigation technique and security model. To achieve this requires further research, data gathering, testing, consultations with specialists in this area and do more feasibility studies. The result of future research will provide more innovative methods for threat management and mitigation of attacks on SDN controllers while enhancing overall network security and management. This will support more secured networks and drive the adoption of SDN which is considered more cost effective and will be beneficial to emerging economies such as Nigeria.

5. ACKNOWLEDGMENTS

The authors would like to thank Khazar University and NILDS NASS for their support.

6. REFERENCES

[1] Liu, S., and Li, B. 2015. On Scaling Software-Defined Network in Wide-Area Networks. Tsinghua Since and Technology. 20(3). 221-232.

[2] Open Network Foundation 2015. Principles and Practices for Securing Software-Defined Networks. ONF TR-511.

[3] Anthony, L. 2015. Security Risks in SDN and Other New Software Issues. RSA Conference. Frost and Sullivan.

[4] Malik, M.S., Montanari, M., Huh, J.H., Bobba, R.B., and Campbell, R.H. 2013. Towards SDN enabled network control delegation in clouds. 43rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN).

[5] Open Network Foundation (ONF). 2013. SDN Architecture Overview.

[6] Ali, S.T., Sivaraman, V., Radford, A., and Jha, S. 2013. A Survey of Securing Networks Using Software Defined Networking. IEEE Transactions on Reliability. 3(64).

[7] Colville, J., and Spafford, G. 2010. Configuration Management for Virtual and Cloud Infrastructures. Gartner Inc.

[8] Feldmann, A., Kind, M., Maennel, O., Schaffrath, G., and Werle, C. 2013. Network Virtualization - An Enabler for Overcoming Ossification. Future Internet Technology. European Community in Information Technology (ERCIM) News.

[9] Open Network Foundation. 2013. OpenFlow-Enabled Mobile and Wireless Networks. ONF Solution Brief.

[10] Akhunzada, A., Ahmed, E., Gani, A., Khan, M.K., Imran, I. and Guizani, S. 2015. Security and Privacy in Emerging Networks: Securing Software Defined Networks: Taxonomy, Requirements, and Open Issues. IEEE Communications Magazine. 34-44.

[11] Mehdi, S.A., Khalid, J., and Khayam, S.A. 2011. Revisiting traffic anomaly detection using software defined networking. In Proceedings of 14th Int. Symposium on Recent Advances in Intrusion Detection (RAID). 6961. 161–180.

[12] Kreutz, D., Ramos, F.M.V., and Verissimo, P. 2013. Software-Defined Networking: A Comprehensive Survey. In Proceedings of the IEEE,103(1). 55-60.

[13] Dabbagh, M., Hamdaoui, B., Guizani, M., and Rayes, A. 2015. Software-Defined Networking Security: Pros and Cons. IEEE Communications Magazine, Communications Standards Supplement.

[14] Kreutz, D., Ramos, F.M.V., and Verissimo, P. 2013. Towards secure and dependable software defined networks, In Proceedings of the second ACM SIGCOMM Workshop on Hot topics in software defined networking. ACM, 55–60.

[15] Shin, S., and Gu, G. 2013. CloudWatcher: network security monitoring using OpenFlow in dynamic cloud networks. Springer. 92–103.

[16] Shin, S., Yegneswaran, V., Porras, P.A., and Gu, G. 2013. AVANT-GUARD: Scalable and Vigilant Switch Flow Management in Software-Defined Networks, In Proceedings of 2013 ACM SIGSAC Conference on Computer and Communications Security. 413–424.

[17] Shirali-Shahrez, S., and Ganjali, Y. 2013. FleXam: Flexible Sampling Extension for Monitoring and Security Applications in OpenFlow. In Proceedings of the second ACM SIGCOMM workshop on Hot topics in software defined networking. HotSDN'13. 167-168.

[18] Wang, Y., Wen, X., Chen, Y., Hu, C., and Shi, C. 2013. Towards a Secure Controller Platform for Openflow Applications, Proc. 2nd ACM SIGCOMM Workshop on Hot topics in Software Defined Networking, 171–72.

[19] Ying-Dar, L., Po-Ching, L., Chin-Hung, Y., Yao-Chun, W., and Yuan-Cheng, L. 2015. An Extended SDN Architecture for Network Function Virtualization with a Case Study on Intrusion Prevention, IEEE Network.

[20] Li, Y. 2014. Computer Networks 72. 74–98.

[21] Metzler, J. 2012, Understanding Software-Defined Networks, Information Week Reports. 1–25.

[22] Scott-Hayward, S., Natarajan, S., and Seker, S. 2016. A Survey of Security in Software Defined Networks. IEEE Communication Surveys and Tutorials. 18(1).

[23] Shin, S., Porras, P., Yegneswaran, V., Fong, M., Gu, G., and Tyson, M. 2013. FRESCO: Modular Composable Security Services for Software-Defined Networks. IOSC Network and Distributed System Security Symposium (NDSS).

[24] Anon. 2016. Software-Defined Networking (SDN) Definition. [online] Available at: http://www.opennetworking.org. [Accessed 5 Mar. 2007].

[25] Son, S., Shin, S., Yegneswaran, V., Porras, P.A., and Gu, G. 2013. Model Checking Invariant Security Properties in OpenFlow. In Proceedings of IEEE ICC. 74–79.

Energy-Aware Routing in Wireless Sensor Network Using Modified Bi-Directional A*

Nurlaily Vendyansyah
Departement of Electrical Engineering
University of Brawijaya
Malang, East Java, Indonesia

Sholeh Hadi Pramono
Departement of Electrical Engineering
University of Brawijaya
Malang, East Java, Indonesia

Muladi
Departement of Electrical Engineering
State University of Malang
Malang, East Java, Indonesia

Abstract: Energy is a key component in the Wireless Sensor Network (WSN)[1]. The system will not be able to run according to its function without the availability of adequate power units. One of the characteristics of wireless sensor network is Limitation energy[2]. A lot of research has been done to develop strategies to overcome this problem. One of them is clustering technique. The popular clustering technique is Low Energy Adaptive Clustering Hierarchy (LEACH)[3]. In LEACH, clustering techniques are used to determine Cluster Head (CH), which will then be assigned to forward packets to Base Station (BS). In this research, we propose other clustering techniques, which utilize the Social Network Analysis approach theory of Betweeness Centrality (BC) which will then be implemented in the Setup phase. While in the Steady-State phase, one of the heuristic searching algorithms, Modified Bi-Directional A* (MBDA *) is implemented. The experiment was performed deploy 100 nodes statically in the 100x100 area, with one Base Station at coordinates (50,50). To find out the reliability of the system, the experiment to do in 5000 rounds. The performance of the designed routing protocol strategy will be tested based on network lifetime, throughput, and residual energy. The results show that BC-MBDA * is better than LEACH. This is influenced by the ways of working LEACH in determining the CH that is dynamic, which is always changing in every data transmission process. This will result in the use of energy, because they always doing any computation to determine CH in every transmission process. In contrast to BC-MBDA *, CH is statically determined, so it can decrease energy usage.

Kata Kunci: Energy; Routing; Wirelss; Sensor, Network; Betweenness Centrality; Searching; Modified Bi-directional A*.

1. INTRODUCTION

Internet of Thing (IoT) is a concept whereby an object has the ability to transfer data over a network without requiring human-to-human or human-to-computer interaction. IoT has evolved from the convergence of wireless technologies, micro-electromechanical systems (MEMS), and the Internet[4]. Based on a survey of IoT analysis, 10 popular IoT applications are Smart Home (100%), Wearable (63%), Smart City (34%), Smart Grid (28%), Industrial Internet (25%), Connected Car 19%), Connected Health (6%), Smart Retail (2%), Smart Supply Chain (2%), and Smart Farming (1%). This application works automatically by utilizing Wireless Sensor Network (WSN) technology.

Although there are many WSN applications, this network has some limitations that should be considered when deciding what protocol to use. Some of these limitations are first, WSN is limited energy supply, WSN has limited energy supply, thus required energy-saving communication protocol. Second, Limited Computation, node sensors have limited computing capabilities so that WSN can not run sophisticated network protocols. Third Communication, limited bandwidth, so that often inhibit intersensor communication[4].

In contrast to traditional wireless networks such as cellular networks, prioritizing quality of service and bandwidth efficiency, energy consumption and network lifetime are important in wireless sensor networks (WSN). In this research, we apply clustering based routing protocol for WSN. Various protocol clustering has been widely developed[8],[9],[10],[11],[12],[13],[14],[15],[16], such as LEACH[5] and its various modifications, PEGASIS, TEEN and so on. The clustering process will generate nodes designated as cluster head (CH). CH is tasked to forward packet data to Base Station. This method will make CH overloaded, affecting energy usage. If one node or CH die, it will disrupt the work function of the network.

Low Energy Adaptive Clustering Hierarchy (LEACH) is one of the most popular WSN routing protocols. CH is selected periodically each time it sends data (per round), while the energy supply consumed by each large node is fixed[5].

2. THEORY

This chapter describes the supporting theories of this research, which will be described in detail in subsequent chapters.

2.1 Wireless Sensor Network

Wireless Sensor Network (WSN) is a collection of hundreds or thousands of wirelessly connected sensors. The sensor device contains a complex set of electronics capable of performing sensing functions, performing simple computing processes and having the ability to communicate with other peers (other sensor nodes) or directly communicate with the base station (BS). Deployment of sensor node can be either randomly or manually planted (static). Components of sensor node are generally shown Figure 1, and Figure 2 show the wireless sensor network architecture.

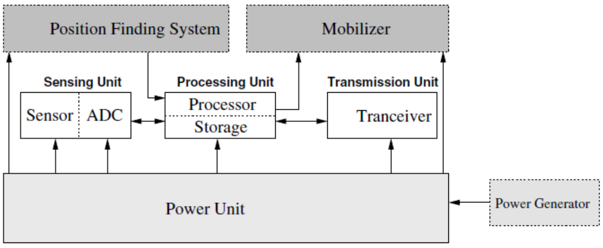


Figure 1. Components of *sensor node*

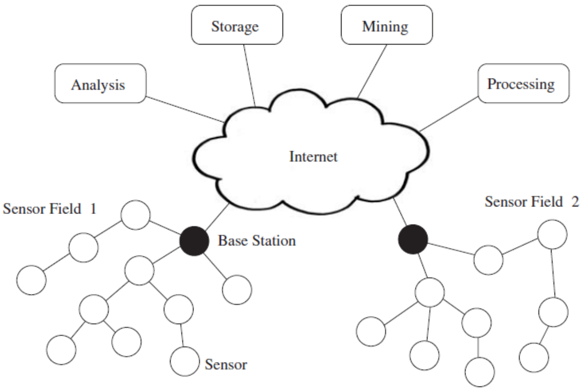


Figure 2. Architecture of *Wireless Sensor Network*[5]

In Figure 1 it can be explained that components of node sensor generally consist of four main parts, namely sensing unit, processing unit, communication unit, and power unit. Sensing unit consists of sensor and ADC (Analog Digital Converter). The function of the ADC is to change the data output from the sensor that is analog data into digital data which will be entered into a digital component that is microcontroller. Sensor classification and sensor samples are shown in Table 1.

Table 1. Classification and sample of sensors[1]

Type	**Example**
Temperature	*Thermistors, thermocouples*
Pressure	*Pressure gauges, barometers, ionization gauges*
Optical	*Photodiodes, phototransistors, infrared sensors, CCD sensors*
Acoustic	*Piezoelectric resonators, microphones*
Mechanical	*Strain gauges, tactile sensors, capacitive diaphragms, piezoresistive cells*
Motion, vibration	*Accelerometers, gyroscopes, photo sensors*
Flow	*Anemometers, mass air flow sensors*
Position	*GPS, ultrasound-based sensors, infrared-based sensors, inclinometers*
Electromagnetic	*Hall-effect sensors, magnetometers*
Chemical	*pH sensors, electrochemical sensors, infrared gas sensors*
Humidity	*Capacitive and resistive sensors, hygrometers, MEMS-based humidity sensors*
Radiation	*Ionization detectors, Geiger–Mueller counters.*

The Communication protocol for low power devices can be shown in Table 2.

Table 2. Communication protocol for low power devices

	GPRS/GSM 1xRTT/CDMA	***IEEE 802.11b/g***	***IEEE 802.15.1***	***IEEE 802.15.4***
Market name for standard	*2.5G/3G*	*Wi-Fi*	*Bluetooth*	*ZigBee*
Network target	*WAN/MAN*	*WLAN and hotspot*	*PAN and DAN (Desk Area Network)*	*WSN*
Application focus	*Wide area voice and data*	*Enterprise applications (data and VoIP)*	*Cable replacement*	*Monitoring and control*
Bandwidth (Mbps)	*0.0064 – 0.128+*	*11 – 54*	*0.7*	*0.020 – 0.25*
Transmission range (ft)	*3000+*	*1 – 300+*	*1 – 30+*	*1 – 300+*
Design factors	*Reach and Transmission Quality*	*Enterprise support, scalability, and cost*	*Cost, ease of use*	*Reliability, power, and cost*

2.2 Routing Protocol in WSN

In sensor networks, energy conservation, directly related to network lifetime, is relatively more important than network performance in terms of quality of data that can be transmitted (QoS). As the energy will be exhausted, the network may be needed to reduce the quality of the results in reducing dissipation energy at the node and thus can extend the network lifetime. Therefore, energy conservation is considered more important than network performance. In general, the division routing protocols in WSN can be shown Figure 3.

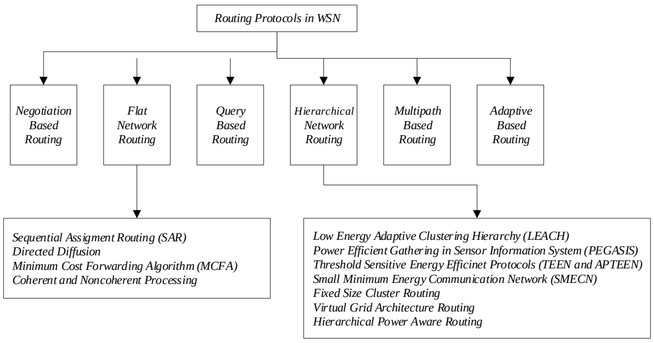

Routing Protocols in WSN
Negotiation Based Routing
Flat Network Routing
Query Based Routing
Hierarchical Network Routing
Multipath Based Routing
Adaptive Based Routing
Sequential Assigment Routing (SAR)
Directed Diffusion
Minimum Cost Forwarding Algorithm (MCFA)
Coherent and Noncoherent Processing
Low Energy Adaptive Clustering Hierarchy (LEACH)
Power Efficient Gathering in Sensor Information System (PEGASIS)
Threshold Sensitive Energy Efficinet Protocols (TEEN and APTEEN)
Small Minimum Energy Communication Network (SMECN)
Fixed Size Cluster Routing
Virtual Grid Architecture Routing
Hierarchical Power Aware Routing

Figure 3. *Routing Protocols in WSN*[4]

2.3 Modified Bi-Directional A*

The Modified Bi-Directional A * is algorithm uses heuristic functions with slight modifications. The heuristic function for n vertices in the forward search of Source (S) to Destination (G) is shown in equation (1)[6].

$$f = g\,(S, n) + \frac{1}{2}[h_s\,(n) - h_g\,(n)] \qquad (1)$$

While the heuristic function for n vertices in the search return (from Destination (G) to Source (S)) is shown equation (2)

$$f = g\,(G, n) + \frac{1}{2}[h_g\,(n) - h_s\,(n)] \qquad (2)$$

S	: origin node or initial state
G	: destination node or goal state
g (S, n)	: the actual cost of S to n
g (G, n)	: the actual cost from G to n
h_s (n)	: approximate cost from n to G
h_g (n)	: approximate cost from n to S

2.4 Heuristic Search

In the methods included in the heuristic search, heuristic functions play a decisive role. A function may be accepted as a heuristic function if the estimated cost generated does not exceed the actual cost. When a heuristic function gives an estimated cost that exceeds the actual cost (overestimate), the search process can get lost and make the heuristic search to be not optimal. The heuristic function is said to be good if it can provide approximate costs that are close to the actual cost. The closer the actual cost, the heuristic function more better. The heuristic function that can be used for the problem of finding the shortest route is a straight line distance on Cartesian coordinates which can be calculated using equation (3)[6].

$$h_s(n) = \sqrt{((x_n - x_s)^2 + (y_n - y_s)^2)} \quad (3)$$

With d_{ab} is the distance between node a and node b. x_a and y_a are the coordinate values of node a on the x and y axes respectively. x_b and y_b are the coordinate values of node b on the x and y axes respectively.

2.5 Betweeness Centrality

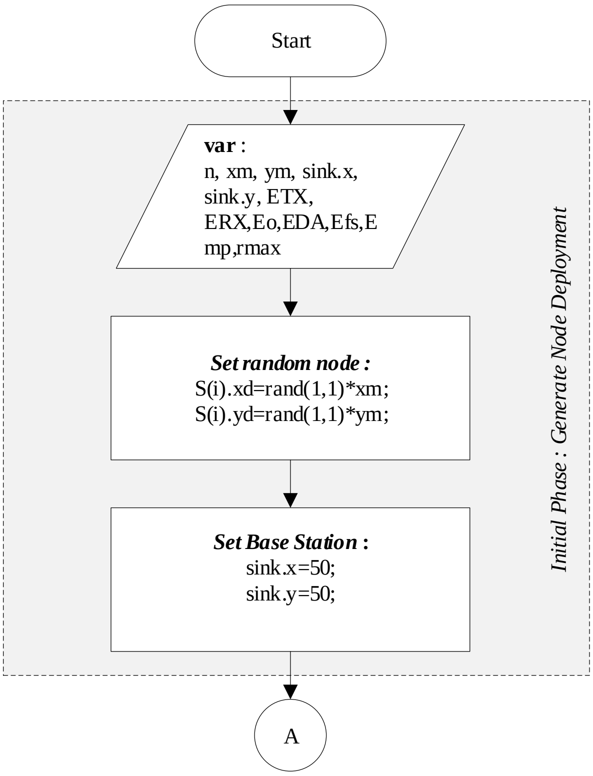
In graph theory and network analysis, the centrality indicator is used to identify the most important nodes in the graph. Usually used to identify the most influential people in social networks, the key key infrastructure on the Internet or urban networks, and the major disease spreaders.

Betweeness Centrality is an indicator of the centrality of the nodes on a computer network. BC or Betweeness Centrality works by counting the number of paths that pass through that node. Betweeness Centrality of the node (v) is formulated by equation (4)[7].

$$BC_i = \sum_{s \neq d} \frac{\sigma_{sd}(i)}{\sigma_{sd}} \quad (4)$$

3. Routing Protocol

In this chapter, we will explain in detail the designed routing protocol, which can be described in the step of process with a flowchart in Figure 4.



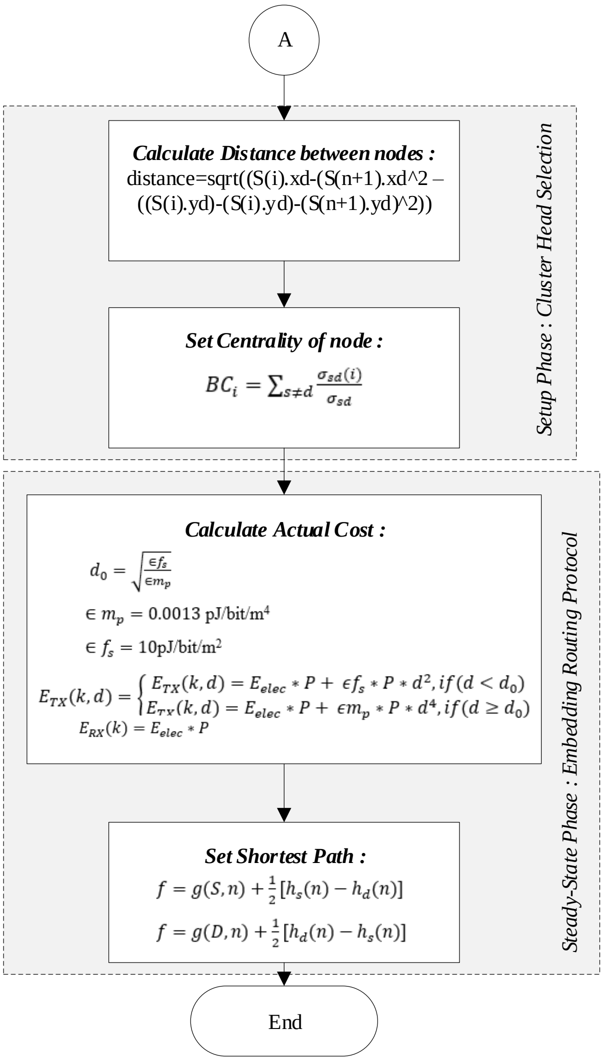

A
Calculate Distance between nodes :
distance=sqrt((S(i).xd-(S(n+1).xd^2 – ((S(i).yd)-(S(i).yd)-(S(n+1).yd)^2))
Set Centrality of node :
$BC_i = \sum_{s \neq d} \frac{\sigma_{sd}(i)}{\sigma_{sd}}$
Setup Phase : Cluster Head Selection
Calculate Actual Cost :
$d_0 = \sqrt{\frac{\epsilon f_s}{\epsilon m_p}}$
$\in m_p = 0.0013$ pJ/bit/m^4
$\in f_s = 10$pJ/bit/m^2
$E_{TX}(k,d) = \begin{cases} E_{TX}(k,d) = E_{elec} * P + \epsilon f_s * P * d^2, if(d < d_0) \\ E_{TX}(k,d) = E_{elec} * P + \epsilon m_p * P * d^4, if(d \geq d_0) \end{cases}$
$E_{RX}(k) = E_{elec} * P$
Set Shortest Path :
$f = g(S,n) + \frac{1}{2}[h_s(n) - h_d(n)]$
$f = g(D,n) + \frac{1}{2}[h_d(n) - h_s(n)]$
Steady-State Phase : Embedding Routing Protocol
End

Figure 4. *Flowchart BC-MBDA**

The variables used in this research are shown in Table 2.

Table 2. Identification of Operational Variables

No	Name of variable	Kind of Variable	Function
1.	n	Input	Number of node.
2.	xm	Input	Coordinat max value x axis of field dimension.
3.	ym	Input	Coordinat max value y axis of field dimension.
4.	sink.x	Input	Coordinat value x axis of Base Station.
5.	sink.y	Input	Coordinat value y axis of Base Station.
6.	E_{TX}	Input	Energi for Transmit packet data.
7.	E_{RX}	Input	Energy for Receive packet data.
8.	E_{elec}	Input	Energy of sensor node for computation process.
9.	E_o	Input	Initial energi for each node.
10.	E_{DA}	Input	Data aggregation energy.
11.	E_{fs}	Input	Energy free space loss (direct).
12.	E_{mp}	Input	Energi multipath (deflective).
13.	r_{max}	Input	Rounds max.
14.	x_d	Output	Coordinat value x axis of node.
15.	y_d	Output	Coordinat value y axis of node.
16.	d_o	Output	Distance of Base Station to Cluster Head.
17.	C	Output	Node to become Cluster head.
18.	E	Output	Energy of Cluster Head.
19.	first_dead	Output	Number dead node.

20.	distance	Output	Distance intersensor node.
21.	X	Output	Coordinat value x axis of cluster head.
22.	Y	Output	Coordinat value y axis of cluster head.
23.	min_dis	Output	Distance intercluster head.
24.	id	Output	Node id.
25.	Message	Input	Message size.

3.1 Initial Phase

This phase is the stage process of the node deployment, WSN is modeled in two-dimensional graphics by placing 100 nodes scattered randomly in a 100x100 (m) area. The base station is located at the coordinates (50,50). The placement of node and base station is static. The initial energy Eo = 0.5 Joule per node, we assuming all nodes are homogeneous and 4000bit message size.

3.2 Setup Phase

This phase is the stage process of the node deployment, WSN is modeled in two-dimensional graphics by placing 100 nodes scattered randomly in a 100x100 (m) area. The base station is located at the coordinates (50,50). The placement of node and base station is static. The initial energy Eo = 0.5 Joule per node, assuming all nodes are homogeneous and 4000bit message size.

3.3 Steady-State Phase

In this process the system determines the centrality of the node function as a cluster head in charge of forwarding the ndata packet to the Base Station. In this process the author uses Social Network Analysis (SNA) theory approach that is Betweeness Centrality (BC), which is calculated by using equation (4). The distance between nodes will be determined using the heuristic function, according to equation (3). The output of this process is the node designated as the cluster head.

4. EXPERIMENTAL RESULT

In this chapter, we will discuss the results of the research from the scheme designed in the previous chapter. Discussion of test results includes the results of each process from the process of generating node deployment, embedding routing protocol, and routing protocol performance analysis. Parameters used to determine the performance of routing protocol in this research are network lifetime, throughput, and residual energy. The experimental results will be compared with LEACH.

4.1 Network Model

In Figure 5 it can be seen the network model of the experiment. In this experiment obtained the results of 21 nodes designated as cluster head.

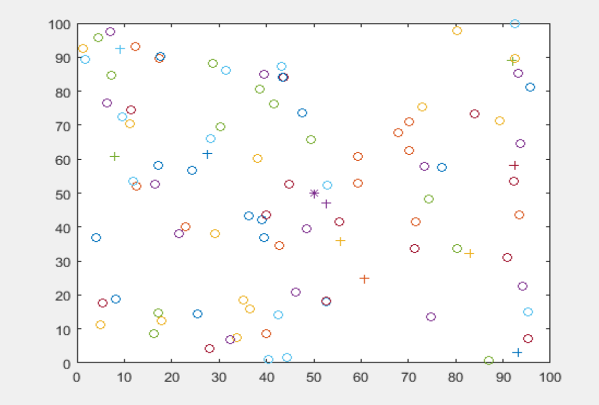


Figure 5. ***Node Deployment***

4.2 Performance Analysis

4.2.1 Network Lifetime

Figure 6 shows a graph of the simulation results of routing protocol performance for network lifetime. In this experiment the node will die after consuming the energy of 0.5 Joule. BC-MBDA * ability to show Network lifetime better 3.22473% than LEACH. This is influenced by the way LEACH always uses dynamic clustering in every round, which will certainly result in the use of energy in CH. Different with LEACH, BC-MBDA * uses a semi dynamic clustering strategy. CH is determined in the setup phase, and is static after being determined as CH. This will reduce the computation process, so energy usage can be saved.

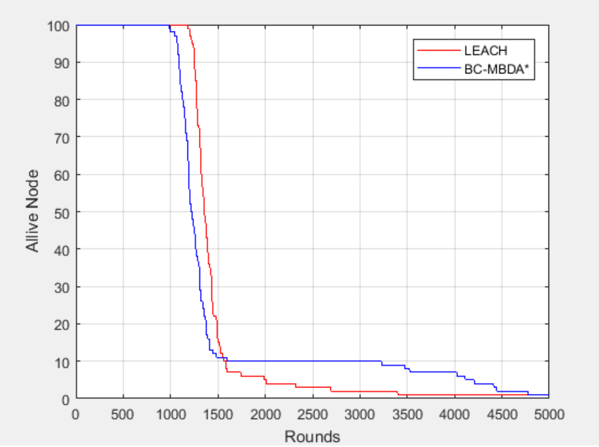


Figure 6. Network Lifetime analysis result

4.2.2 Throughput

To evaluate throughput performance, the amount of packet data received in BS is compared to the number of packets sent by the nodes on each round. The BC-MBDA* throughput performance can be shown with the graph in Figure 7. The graph shows that BC-MBDA * throughput performance is 1.692883% better than LEACH. This is influenced by the ability of CH in delivering packets to BS. When CH is overloaded, it causes CH to die because the energy availability has been exhausted. When CH is off then the node will not be able to forward packets to BS. This resulted in disruption of the performance of the WSN system.

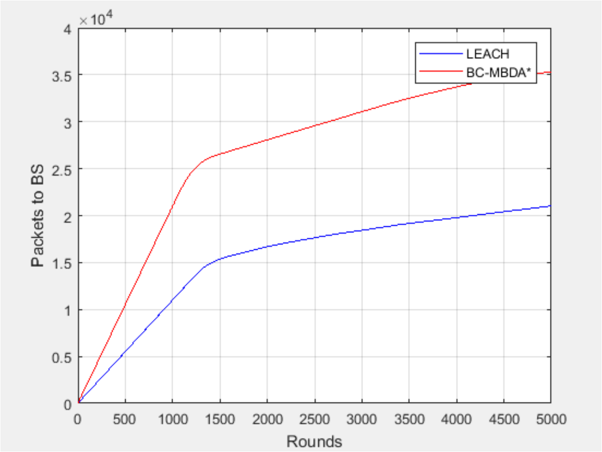


Figure 7. Throughput analysis result

4.2.3 Residual Energy

Figure 8 shows the results of the analysis of the number of dead nodes in the experiment. The experiment was carried out as much as 5000 rounds, with an initial energy of 0.5 Joule / node. So the total energy of 100 nodes is 50 Joules. The graph shows that BC-MBDA * is better 0.879459% than LEACH.

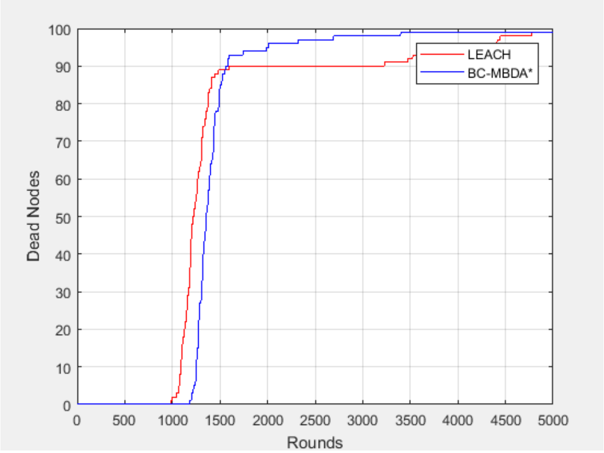


Figure 8. Residual Energy analysis results

5. CONCLUSION

Performance metrics analyzed in this research are network lifetime, throughput, and residual energy. The test results show that the performance of the designed routing protocol is better than LEACH. 1.692883% for Network Lifetime, 1.692883% for Throughput and 0.879459% for Residual Energy.

In the next study, we will develop a combination of K-NN method with MBDA*. In setup phase we will use K-Nearest Neighbors algorithm for cluster head selection, and for Steady-State Phase used MBDA*.

6. REFERENCES

[1] C, Shanti, and Sharmila D. "A self-organized location aware energy efficient protocol for wireless sensor networks." Computers and Electrical Engineering 41, 2015: 265-274.

[2] Dargie, Waltenegus, and Christian Poellabauer. Fundamentals of Wireless Sensor Networks : Theory and Practice. United Kingdom: A John Willey and Sons,Ltd, 2010.

[3] Abassi, Ameer Ahmed, and Mohamed Younis. "A survey on Clustering algorithms of wireless sensor networks." Computer Communication 30, 2007: 2826-2841.

[4] Mohammad Ilyas, Imad Mahgoub, Handbook of Sensor Networks : Compact Wireless and Wired Sensing Systems.

[5] A. Nayebi and H. Sarbazi-Azad, "Performance modeling of the LEACH protocol for mobile wireless sensor networks," Journal of Pararel and Distributed Computing, vol. 71, pp. 812-821, 2011.

[6] Suyanto, Artificial Intelligence Searching - Reasoning - Planning - Learning Edisi Revisi, Bandung: Informatika, 2011.

[7] Social Network Analysis : Theory and Application.

[8] Yu, Hu, and Wang Xiaohui. "PSO-based Energy-balanced Double Cluster-heads Clutering Routing for wireless sensor networks." *Procedia Engineering 15*, 2011: 3073-3077.

[9] Z. Yong and Q. Pei, "A Energy-Efficient Clustering Routing Algorithm Based on Distance and Residual Energy for Wireless Sensor Networks," *International Workshop on Information and Electronics Engineering (IWIEE)*, pp. 1882-1888, 2012.

[10] T. Amgoth and P. , "Energy-aware routing algorithm for wireless sensor networks," *Computer and Electrical Engineering 41*, pp. 357-367, 2015.

[11] Hui Li, Xia, and Zhi Hong Guan. "Energy-Aware Routing in Wireless Sensor Networks Using Local Betweenness Centrality." *International Journal of Distributed Sensor Networks, Hindawi Publishing Corporation*, 2013.

[12] Khalil, Enan A, and Bara'a A Attea. "Energy-aware evolutionary routing protocol for dynamic clustering of wireless sensor networks." *Swarm and Evolutionary Computation 1*, 2011: 195-203.

[13] Mahajan, Shilpa, Jyoteesh Malhotra, and Sandep Sharma. "An energy balanced QoS based cluster head selection strategy for WSN." *Egyptian Informatics Journal 15*, 2014: 189-199.

[14] Nadeem, Q, N Javaid, S. N Mohammad, M. Y Khan, S Sarfraz, and M Gull. "SIMPLE: Stable Increased-throughput Multihop Protocol for Link Efficiency in Wireless Body Area Network." *Broadband and Wireless Computing, Communication and Applications (BWCCA)*, 2013: 221 - 226.

[15] Nam, Su Man, and Tae Ho Cho. "A fuzzy rule-based path configuration method for LEAP in sensor networks." *Ad Hoc Networks 31*, 2015: 63-79.

[16] Nazir, Babar, and Halabi Hasbullah. "Energy efficient and QoS aware routing protocol for Clustered Wireless Sensor Network." *Computers and Electrical Engineering 39*, 2013: 2425-2441.

International Journal of Computer Applications Technology and Research
Volume 7–Issue 08, 306-311, 2018, ISSN:-2319–8656

ASSESSMENT OF THE EFFICIENCY OF CUSTOMER ORDER MANAGEMENT SYSTEM: A CASE STUDY OF SAMBAJO GENERAL ENTERPRISES JIGAWA STATE, NIGERIA.

Muhammad Yakubu,
Kampala University, Department of Business and
Management Studies,
Ggaba P.O.BOX 25454, Kampala, Uganda

Abubakar Yahaya Bakabe
St. Lawrance University Uganda Department of
Telecommunication Engineering,
Mengo P.O.BOX 24930, Kampala, Uganda

Abstract: The Supermarket Management System deals with the automation of buying and selling of good and services. It includes both sales and purchase of items. The project Supermarket Management System is to be developed with the objective of making the system reliable, easier, fast, and more informative.

Keywords: Employee, customers, managers, staffs, online order, supermarket.

1. Introduction

Order management system is the administration of business processes related to orders of goods and services from a customers to a seller. With the rapid growth of online retail world, when the number of orders placed, the number of shipments made is growing rapidly, a good order management system/ sales management system can help a retailer save a lot of time and make their life much easier. It includes the process of tracking a sales order from inception (when the order is placed from the buyer the administrator check the store for the available stocks, if stock ordered are available or not a notification sent to the buy when stock a not available the order will be cancel, if stock a available the buy will paid, then the order will be deliver to the buyer). Order processing is an important area for improvement to achieve efficiency in supply chain performance. For high performing organizations efficiency in the processing of customer orders is a distinguishing characteristic (Esker 2013). Although the manual system of order management has its own benefits, the electronic system seems to be more advantageous for being faster and efficient in processing of information; automatic generation of accounting documents like invoices, checks and statement of account. With the larger reductions in the cost of hardware and software and availability of user-friendly accounting software package, it is relatively cheaper like maintaining a manual accounting system. Many types of useful reports can be generated for management to make decisions, on timely information can be produced. However, the digital system has its own disadvantages as well for example, power failure, computer viruses and hackers are the inherent problems of using computerized systems. Once data been input into the system, automatically the output are obtained hence the data being input needs to be validated for accuracy and completeness, we should not forget concept of GIGO (Garbage In (Input) Garbage out (Output).

Lack of integration among different stakeholders in the supply chain causes fall in business performance and reduced profit and market share of the business (Whitepaper 2007). An order management system (OMS) automates and streamlines order processing for businesses. An OMS provides constantly updated inventory information, a database of vendors, a database of customers, a record of customer returns and refunds, information on billing and payments, order processing records, and general ledger information. Benefits of a well-implemented OMS include improved sales visibility, improved customer relations, and efficient order processing with a minimum of delays and backorders. Order management is important primarily in the retail industry, but also in the telecommunications, health care, pharmaceutical, financial, and securities sectors

2. Statement of the Problem

Business transactions yield a valuable data which provides information on that facilitates the monitoring and proper management of businesses. The information needs to be accurate, easy to retrieve and secure. Currently some supermarkets still use paper based kind of system for monitoring and managing their business. This paper based system is prone to errors, difficult to retrieve when it increases in volume, and it is insecure whereby any unauthorized one could access it, misplace it, tear it, and so forth. What's more, it enables employees as well as all the stakeholders of the business to waste their time as they wait processing their transactions.

It's also less accurate in comparison to computerized processing system. Besides, there is difficulties in which it would be hard to track stock products knowing how much you bought and how much products and goods you sold plus stock balance. It is difficult to track finance whereby it involves how much money you spent on purchases and expenses, how much you got from sales, how much you have in bank, how much you have in cash and lastly how much you have for profit/loss. Tracking customer orders, customer details and managing stuff details could be another challenge in the manual system. Employees always encounter a problem when they generate periodic reports from sales or purchases and profit/loss.

The situation in the study area (Sambajo General Enterprise) with regards to customer order management system was not different from those found in other places. Complains from both staff and customers with regards to difficulties experienced in managing customer orders are so frequent leading to frustrations, loss of customers, less productivity etc. Thus, this study was designed to look into issues relating to customer order management system with the aim of ascertaining its SWOT analysis so as to come up with a better system that bring to end the numerous difficulties faced by both staff and customers.

The current system in Sambajo General Enterprise customer's information collected where recorded manually with pen in the record book which included: name, email, residential address, phone number and business address if a business man is. There were a lot of problems with this manual system. The data collected varied from employee to another a day 300 to 400 attended the store. Most of the time the hand-writing was unclear when it was typed in a hurry, to Search for a customer details when needed is become a problems. They is a need to develop a system to record such information.

3. Requirements for the solution

There is needed to create a better solution for customer information record system. The solution will be used by employees and management as well as customers.

The new solution should provide a unified way of storing the customer data. In other words the information stored should not fluctuate depending on the employee who stored it. The data should be easier to find than with the manual bookkeeping.

The system should speed up the daily work with the employees working with the customers. The customer information usability should be enhanced for both the employees working with customers and for the management managing the daily work. The new system will also allow the customers to order online without visiting the supermarket

4. Materials and Methods

The study involved 155 respondents comprising of 55 staff of the Sambajo General Enterprise and 100 customers selected using the Purposive and Systematic Random sampling techniques. Descriptive Survey design was used deploying both qualitative and quantitative approaches. A closed ended questionnaire was used to collect quantitative data while one on one Interview was also used to collect qualitative data. Collected data was analyzed in SPSS Version 20 using Descriptive Statistics.

5.1 Analysis of results

5.2 Efficiency of the system in use

One of the fundamental data necessary for the design of a customer order management system is the evaluation of the efficiency of the current system in use. Quantitative data was collected in order to analyze efficiency of the current system at the store.

Table 1: Response of staff on the efficiency of the current system (paper based)

Categories of responders	**Number of responder**	**Frequency (%)**		
		Less efficient	Efficient	Very efficient
Managers of the supermarket	3	2 (66.6%)	1	0
Operational staff	37	20 (%)	17	0
Customer care services	15	14(%)	1	0
Customers	100	98 (%)	2	0
Total	**155**	**134 (86.5%)**	**21 (13.5%)**	**0**

The above table indicates responses of the respondents in respect to the efficiency of the current system in use i.e. paper based system. The table shows that out of the total respondents of 155, 134 of them (86.5%) agreed that the current system was less efficient while only 21 of them (13.5%) stated that the system was efficient. Among the 3 managers involved in the study, 2 said the current system is less efficient (66.6%), only 1 claimed that the system is efficient (33.3%) while on the part of the operational staff out of 37 of them, 20 stated that the current system was less efficient (54.0%) and 17 said it was efficient (45.9%).

With regards to the customer care service, 93.3% of the respondents stated that the current system was inefficient while only 6.6% claimed that the system was efficient. Finally, with regards to customers' responses on the same issue, 98% of them stated that the system was inefficient while only 2% felt the current system was efficient.

5.3 Ease of use of the system

Table 2: Responses in relation to ease of use of the current system

Categories of responders	Number of responders	Easy	Difficult
Managers of the supermarket	3	1 (33.3%)	2 (66.6%)
Operational staff	37	10 (27.0%)	27 (72.9%)
Customer care services	15	2 (13.3%)	13 (86.6%)
Customers	100	10 (10%)	90 (90%)
Total	**155**	**23 (14.8%)**	**134 (85.1%)**

The above table shows responses obtained from the different respondents on the ease of use of the current order system in use at the store. Based on the responses of the managers of the supermarket, 33.3% stated that the system was easy to use while the majority of them (66.6%) claimed that system was not easy for manipulation. On the part of the operational staff, 27.0% of them stated that the system was easy to operate while the majority of them (72.9%) were of the view that the system was not easy to operate. Besides, responses obtained from the customer service staff showed that 13.3% of them believed that the current system was fine while a good number of them was of the view that the system was not difficult to operate whereas the remaining 86.6% claimed that it was difficult to manipulate. Finally, on the part of the customers, only10% was of the view that the current system being used at the store could be easily used whereas the majority of them (90%) stated that the system was difficult to operate.

6. Feasibility Study

This was how problems inherent with the current manual system being used at the store were examined based on the requirements for the development of the proposed system were weighed in order to find out the possibility of overcoming them. This was looked at from three different perspectives; operational, technical and economic perspectives.

6.1 Technical feasibility

This was conducted with the aim of examining whether the operational and managerial staff of the Sambajo General Enterprise as well as the customers had the technical knowledge how to operate the new system proposed to be developed. From the results obtained from the feasibility studies involving Key Interview Informants (KIIs), all the operational staff (100%) as well as about 78% of the customers interviewed had the technical skills required to operate the system.

6.2 Operational feasibility

The operational feasibility study was conducted with the aim of studying whether the environment at the Sambajo General Enterprise was suitable to operate on the proposed system to be developed. It was discovered that as a result of using the manual order system, whenever customers wanted to shop at the store, they had to go to the Sambajo General Enterprise in person to select what they want to buy and then proceed to the available cashier on seat to pay for what they purchase. As result of this, the supermarket most of the time becomes congested with customers and it took longer time to serve the customers in queues. However, from the findings made from the study operational feasibility study conducted, the supermarket had all the requirements needed to support the new order system proposed to be developed such as the availability of enough space, stable power supply, security as well as sufficient hardware to maintain the new system to be designed. In addition, both the operational staff and customers were found to possess the knowledge and skills required to operate the new system when implemented.

6.3 Economic feasibility

From the interviews conducted with managerial staff of the supermarket, it was discovered that the supermarket can financially afford the cost of the system proposed to be designed and developed with regards to hardware, software, sufficient staff and security. Besides, that survey revealed that 80% of the customers use smart phones which they can utilize in operating the new system when installed.

7. System Requirements

These refer to the requirements needed for the system as a whole rather than individual component of it. These are services that can be grouped into functional and non-functional requirements. Functional requirements vary from system requirement however; non-functional requirements tend to be almost uniform for all types of systems. Examples of such requirements include; efficient utilization of system resources, time to load, ram required, disk space occupied, usability, maintainability to mention but a few.

7.1 Functional Requirements of the System

In the current system in use at the stores at present, all transactions were recorded on papers whereby a customer presents the commodities he/she wants to buy to the operational staff who then checks for the price of the commodity requested and writes the invoice one by one. From there, the customer moves to the cashiers table for payment. After payment, the purchased goods are brought out from the store and given to the buyer. More often such transactions tend to be so tiresome and stressful leading to errors in the day's records which arise from the manual recording. At the end of each day's business transactions, records of the daily transactions must be balanced before closing. Hence, it is sometimes very difficult to get the record of back date transaction, however when they is a new product in the market the price has to be write and place a side for checking by the operational staffs when customers brought, in some time.

Meanwhile, with the new system proposed to be developed, data from the keyboard can be captured which will be processed by the system such that transaction details can be displayed. With this system, all product details, sales, expenses, suppliers and other transactions conducted at store by the customers and staff can be tracked. The system can as well accept orders from the suppliers and hence calculate the expenses involved. It should provide security mechanism which could restrict any un-authorized access and must allow the administrators to manage and access all information entered by all users of the system. In addition, the system must have a search engine mechanism which allows the stakeholders of the system to search for relevant records and information such as searching customer's details. It also enables users to select and print out the reports generated by the system and makes it easy for users to backup data. The system must also have functions such as updating records, deleting, adding and so on.

7.2 Non-functional Requirements

When fully implemented, the new system must be able to possess the under listed criteria in order to be able to install with case.
The system must not occupy more than 500mb of disk space.
It must be easy to learn and use.
The system must be able to easily load.
The system failure rate not be more than five minutes a months.
The user must able to use the system without errors after five hours of training.
The system must be platform independent.

7.3 Hardware and Software Requirements

This system can run on any computer with minimum speed of 500mhz, ram of 512mb, and 2gb free of hard disk space, 1 eight port switches, 1 uninterruptible power supplies (ups) of 220-240 volts 100 meters of cables. This system can be run on any machine with windows xp or above like windows vista, windows 7.

7.4 Security Requirements

This specifies system behavior that disallows unauthorized users from accessing the system. Thus, unauthorized users should not be allowed to access the database; this can be enhanced by the use of passwords and user names. Relevant information should only be made available to authorized users (staff). Besides, security measures such as antivirus to prevent damage to the software should also be incorporated.

8. Development tools

8.1 PHP and MySQL can be use

PHP and MySQL can be used to develop. There are no licensing costs for these tools. The tools have wide community support and there are lots of examples and tutorials available even for complex application designs, it was noticed that PHP and MySQL are widely supported globally with the hosting service providers.

9. Advantages and disadvantages of the proposed system

It is believed that once the proposed system is developed and fully implemented, a lot of benefits can be driven from it. Some of these are; order can be made within few minutes depending on the number of items a customer wants and the customer will be able to keep track of orders placed and also organize the bill. The managers can as well manage the products and stock in the inventory using the new system. Other benefits of the new system are it saves time, enhances security, reduces congestion, enables large sales, and reduces stress on the part of the customers as well as the staff and so on. One of the disadvantages of the new system may be that of internet and power failure that can bring transactions to abrupt stop.

10. Discussion

Based on findings made with regards to the current system, it was revealed to be so inefficient based on the respondents' views. This finding agrees with that of Maxat *et al.,* who stated that in a study of a restaurant at Switzerland, the customers 46% stated that the Restaurant was in efficient when it come to the order management process. Similarly, with regards to the ease of use of the current system (paper based), it was revealed that those using it (customers, staff and managers) stated that the system was not easy to use and was so tiresome and time consuming. Besides, it was very liable to loss and destruction. It's a digital world, and every business owner is inundated with finding another solution to streamline work, and to take things "to the cloud." Going paperless has many advantages for business owners. Even so, there are risks that some business owners are wondering exactly what they should keep in a digital space, and what should they relegate to old-fashioned paper methods. Keeping everything stored in a digital format, whether on computer drives, flash drives or in cloud-based systems, is cheaper than printing and storing it on paper. This eliminates the cost of shredding services for paperwork with sensitive information. Some businesses have entire rooms and storage units devoted to archiving paper, Paperless systems eliminate this cost.

In addition, the study was also able to find out that the current system in use at the Sambajo General Enterprise was not easy to use and was too stressful and tiresome. Many studies have confirmed this further explaining the benefits of using the digital order system. When everything is stored digitally, versus on paper in files, accessibility becomes quick and easy. Employees, consumers and business owners have access to all data, contracts and consumer files with just a few mouse clicks. This eliminates having to locate the file or form, which saves everything at a go one time. Meanwhile, when information is stored on paper and locked in file cabinets, someone would need to physically have access to the papers to steal information. Hackers don't need to worry about this when everything is stored digitally. Business owners often get too busy to update software and virus protections, making it easier for hackers to install spyware, steal information or hijack company data.

Furthermore, the recent years have gone through a rapid growth of ecommerce and as a result of that the distribution of goods to consumers has reshaped (Birner, 2015). Consumers of online shopping, increasingly use numerous of internet enabled devices. A global standard for all online payments is under development by the consortium.

With regards to ease of use of the current system in use, the majority of the managerial staff, operational staff as well as the customers stated that the system was difficult to use and more stressful. This finding agrees with that of Resham *et al*. in his paper Volume 2 , Issue 1 January 2014 International Journal of Advance Research in Computer Science and Management Studies Research Paper Available online at: www.ijarcsms.com

11. Conclusion

Sales and Invoice Management is an important aspect of any organization that must be handled skillfully. The implementation of these processes has made the working more efficient, keeps the employees up-to-date and administrators are well informed to take important decisions. Existing systems have various missing aspects that is covered by the system, making it a better alternative. Another highlight of this system is that it extracts knowledge from data which goes beyond traditional charting tools to determine critical information about customer loyalty, product popularity, employee efficiency and overall organizational success. From the findings made by this study, it can be concluded that, considering the inefficiency of the current system being used at the Sambajo General Enterprise presently coupled with the fact that the system was found to be very difficult to use besides being too stressful and tiresome, the proposed system to be designed if well installed and implemented by the store, all operations of the store with regards to order management and customers details will be highly enhanced. In addition, customer satisfaction will also be improved.

12. Recommendations

Sambajo General Enterprise are encouraged to adopt this system for its efficient running of their operations which enables them to meet their user requirements, especially in the current age of technology where every activity in any organization should integrate computer into their business transactions in order to improve and enhance their services such as vast customer order service.

In fact, data is really vulnerable and very crucial which has to be given a further consideration. Therefore, it can get lost, by deleting intentionally or it can be infected by tragedy like failure of computer hardware or may be, say, a virus can mess it up. Hence, I strongly recommend you to back up all the relevant data for the system in order to avoid any unnecessary risks which could happen to them. Moreover, the system developer has really restricted some forms to be dealt with a concrete data entry. For instance, there are some forms which restricts user to fulfill the required information which he/ she can’t skip it due to validation. So, what I really recommend you is that, you need to fill all the required information as indicated by asterisk symbol on the form indicating you it’s a must to fill that particular field. If failure so, any data entered by user will not be saved hence error display message will pop up.

Finally, it’s also important thing to have a system developer to work with the system if any problems are encountered or may be, say, that if there’s need to educate some naïve users of how to be familiar with the system, this also recommended.

Reference

[1] Sunita B. and Jitender A. (2012). Classification and Feature Selection Techniques in Data Mining. *International Journal of Modern Engineering Research* (IJMER) Vol. 3, Issue. 6, pp-3348-3358.

[2] Esker. (2013). Automated Sales Order Processing for Order-to-Cash Performance (www.esker.com), Access Date: March 19, 2013.

[3] Finney, S. and Corbett, M. (2007) ERP Implementation: A Compilation and Analysis of Critical Success Factors. *Business Process Management Journal*, 13(3), pp. 329-347.

[4] Braude, E., Bernstein, M. (2011). *Software engineering: modern approaches*. John Wiley & Sons, Inc. United States of America.

[5] Bachelor's Thesis Degree in Business Information Technology 2012 Haag-Helia University of applied sciences by Timo Aho.

[6] Bilili, S. and Raymond, L. (1993) Information technology: threats and opportunities for small and medium-sized enterpris. *International Journal of Information Management,* 13(6) p. 439-448.

[7] Soon Nyean Cheong, Wei Wing Chiew, Wen Jiun Yap,"Design and Development of Multi Touchable E Restaurant Management System" ,in 2010 International Conference on Science and Social Research (CSSR 2010), December 5 -7, 2010, Kuala Lumpur, Malaysia

[8] Volume 2, Issue 1 January 2014 International Journal of Advance Research in Computer Science and Management Studies Research Paper Available online at: www.ijarcsms.com

[9] Bachelor's Thesis Degree in Business Information Technology 2012 Haag-Helia University of applied sciences by Timo Aho.

[10] Master Project Communication Systems Group (CSG) Department of Informatics (IFI) University of Zurich

[11] Binzmühle strasse 14, CH-8050 Zürich, Switzerland URL: http://www.csg.uzh.ch/

[12] Ardhariksa Zukhruf K Fakultas Ilmu Komunikasi Universitas Mercu Buana , Jl. Meruya Selatan No.01, Kembangan, Jakarta Barat 11650, Efforts to Communicate Corporate Identity through Company Website *International Journal of Science and Research* Volume 7 Issue 7, July 2018 www.ijsr.net

APPENDIX 1
QUESTIONNAIRES

The aim of this study is to investigate the challenges face by the Sambajo General enterprise for customer record and customer order management and suggest the solution for the enterprise. Therefore, I kindly request you to answer the following questions objectively.

1. What is your position in Sambajo General Enterprise? **(Tick where appropriate)**

□ Manager □ storekeeper □ Customer

2. Are you a computer literate?

□ Yes □ No

3. Did it take long time to sever one customer?

□ Yes □ No

4. Do you have back-up copies for your records?

□ Yes □ No

5. Does it take long to process a customer shopping order and print out a result slip?

□ Yes □ No

6. Did the current system face challenges as well as weakness of the current paper manual system?

□ Yes □ No

7. How do you view the current system of managing customer order shopping's?

□ Very easy to use Easy to use Difficult to use □

8. Did you encounter a problems when you're trying to storing, updating and retrieving customer orders?

□ Yes □ No

9. Do you think that Management Information System can help the challenged mentioned in?

□ Yes □ No

12 Which system between the following would you prefer?

Paper-based system □ computerize system □

13. How are the important records of transactions stored in the Supermarket's stock?

□ Yes □ No

14. Did the current system is up to standard for recording the customer details?

□ Yes □ No

Thank you for taking time to fill in this questionnaire. Your response is appreciated.

Integrated System for Vehicle Clearance and Registration

Okeke Ogochukwu C.
Department of Computer Science,
Chukwuemeka Odumegwu Ojukwu University,
Uli, Anambra State, Nigeria

Ezenwegbu Nnamdi Chimaobi
Department of Computer Science,
Chukwuemeka Odumegwu Ojukwu University,
Uli, Anambra State, Nigeria

Abstract: Efficient management and control of government's cash resources rely on government banking arrangements. Nigeria, like many low income countries, employed fragmented systems in handling government receipts and payments. Later in 2016, Nigeria implemented a unified structure as recommended by the IMF, where all government funds are collected in one account would reduce borrowing costs, extend credit and improve government's fiscal policy among other benefits to government. This situation motivated us to embark on this research to design and implement an integrated system for vehicle clearance and registration. This system complies with the new Treasury Single Account policy to enable proper interaction and collaboration among five different level agencies (NCS, FRSC, SBIR, VIO and NPF) saddled with vehicular administration and activities in Nigeria. Since the system is web based, Object Oriented Hypermedia Design Methodology (OOHDM) is used. Tools such as Php, JavaScript, css, html, AJAX and other web development technologies were used. The result is a web based system that gives proper information about a vehicle starting from the exact date of importation to registration and renewal of licensing. Vehicle owner information, custom duty information, plate number registration details, etc. will also be efficiently retrieved from the system by any of the agencies without contacting the other agency at any point in time. Also number plate will no longer be the only means of vehicle identification as it is presently the case in Nigeria, because the unified system will automatically generate and assigned a Unique Vehicle Identification Pin Number (UVIPN) on payment of duty in the system to the vehicle and the UVIPN will be linked to the various agencies in the management information system.

Keywords: Vehicle Clearance, Vehicle Registration, Treasury Single Account

1. INTRODUCTION

1.1 Background to the Study

The International Monetary Fund (IMF), recommended a unified financial policy for funds flow in Nigeria. Treasury Single Account (TSA) is a financial policy in use in several countries all over the world. It was proposed and partially implemented by the federal government of Nigeria in 2012 under the Jonathan Administration - and fully implemented by the Buhari’s Administration to consolidate all inflows from all agencies of government into a single account at the Central Bank of Nigeria. Efficient management and control of government's cash resources rely on government banking arrangements. Nigeria, like many low income countries, employed fragmented systems in handling government receipts and payments. Establishing a unified structure as recommended by the IMF, where all government funds are collected in one account would reduce borrowing costs, extend credit and improve government's fiscal policy among other benefits to government. The IMF also recommends the establishment of a legal basis to ensure its robustness and stability. The introduction of the Treasury Single Account policy therefore was vital in reducing the proliferation of bank accounts operated by ministries, departments and agencies (MDAs) towards promoting financial accountability among governmental organs. The compliance of the policy in Nigeria created challenges for majority of the MDAs. Commercial banks in Nigeria remitted over 2 trillion Naira worth of idle and active governments deposits with full implementation of this policy in 2016. Meanwhile, the bankers' committee of the country has declared their support for the policy. Through Remita, the integrated electronic payments and collections has enabled the Federal Government of Nigeria to take full control of over 3 trillion Naira ($15 billion) of its cash assets as at the end of the first quarter of 2016. The ongoing motivated the design and implementation of an integrated system that will unify all payments and collections involved in vehicle clearance and registration. The growth in computer technology development is increasing as long as more research are performed on daily basis. Information technology improvement has gone versatile over the world on different application in every country. Advancement in technology comes with the need for individuals and organizations to harness the power of information technology to make their duties easier. In this research, attention is given to how information technology can be harnessed to automate vehicle clearing and registration in Nigeria. Vehicle and plate number registration in Nigeria has been in existence for the past decade ago and the document have been manually operated which in turn has not helped to raise the efficiency of general automotive services in recent years. additionally, there are multiple documents issued by many agencies to a particular vehicle. This situation opens more gap for fraud since an officer from one agency cannot verify the authenticity of a document issued by another agency at any point in time. So, we propose an integrated system where proper interaction and collaborations will be created among the agencies saddled with vehicular administration and activities in Nigeria. These agencies are: Nigeria Customs Service (NCS), Federal Road Safety Commission (FRSC), Vehicle Inspection Office (VIO), States Boards of Internal Revenue (SBIR) and Nigeria Police Force (NPF). The system will also be required to give proper information about a vehicle starting from the exact date of importation to registration and renewal of licensing. Vehicle owner information will also be efficiently retrieved from the system by the Nigerian police when a crime is committed with such vehicle. Officers from any agency can verify vehicle documents issued by other agencies by visiting the unified web application. Also number plate will no longer be the only means of vehicle identification as it is presently the case in Nigeria, because the unified system will automatically generate and assigned a unique UVIPN (unique vehicle identification pin number) on prompt payment of duty in the system to the vehicle and the UVIPN will be linked to the various agencies in the management information system. Lastly, the system will also incorporate a vehicle maintenance alert menu which will regulate and timely alert vehicle owners when such vehicle is due for servicing to avoid vehicle breakdown. In this research work, a lot of observation have been conducted towards the existing application system in use that prompts various problems in processing data in order to identify the various problems that are been encountered in the registration of vehicles and plate number.

The proposed application system will allow automatic change of ownership when a vehicle is disposed through the vehicle and plate number registration code generated. The online registration site will also incorporate Nigeria Customs Service (NCS) for importation and clearing of the goods and direct it to Federal Road Safety Commission (FRSC). One unified web application is developed and all the information about a vehicle can be found on the website irrespective of the agency that issued the command.

1.2 The Statement of the Problem

The existing vehicle registration and plate number issuance system were analyzed and the following problems were found. They are:

a. The agencies involved in vehicle certification operate independently with different software systems under separate domains and servers. This situation hinders fast verification of documents.
b. The existence of uncertainty regarding whether the treasury will have sufficient funds to finance programmed expenditures may lead to sub-optimal behavior by budget entities, such as exaggerating their estimates for cash needs or channeling expenditures through off-budget arrangements.
c. Participation of unauthorized officers in the vehicle clearance and registration process.
d. There is delay in verifying the authenticity of vehicle documents because there is no online system where the information is stored.
e. Wrong charging of fees and exploitation by registration of officers.
f. Difficulty in tracing a record/information concerning a vehicle owner due to improper information keeping as a result of carelessness or volume in the size of the record kept.
g. Car buyers have reportedly been victims of deceit because there is no way these ones can confirm whether a particular vehicle with a particular engine number has been cleared of custom duty.
h. Loss of files and human error have led to denial of payment by licensing agencies.
i. Illegal extortion of funds by agents who are also officials of vehicle registration agencies.
j. Improper accounting of registration transactions.

These problems among other have motivated the construction of the new system.

1.3 Aim and Objectives of the Study.

The aim of this research is to design and implement a multipurpose online vehicle clearing and registration system. The specific objectives of this research are:

a. To simplify vehicle clearance and registration by integrating the procedure on one platform.
b. To build an integrated web application that will host the functions of all agencies that are involved in vehicle registration and licensing. All information and procedures will be constructed on one website.
c. To build an integrated system that facilitates efficient payment mechanism for vehicle clearance and registration.
d. To create a system that improves operational control during budget execution. When the treasury has full information about cash resources, it can plan and implement budget execution in an efficient, transparent, and reliable manner.
e. To develop a notification system for vehicle owners using sms and email on every registration process.

1.4 Significance of the Study

The new system will be of great significance because it will expedite the efficiency of principal licensing officers in the processing of vehicle registration data and documents online processing. The proposed web application will also improve the confidence of Federal Road Safety Commission (FRSC) and Vehicle owners since it produces accurate information timely. The new system will also develop a method that will allow easy storage and retrieval of vehicle and owner's registration information and online assessment at any time in the future. A highly accurate method of generating and assigning plate numbers will be featured. This will determine the easiest and fastest way to access vehicle owner's plate number, registration information and missing vehicles through the code generated. Officers on duty will not require original documents anymore. Instead documents can be verified on the website with an internet enabled device. Integration and single account will improve appropriation control. The TSA attached to the web application ensures that the government has full control over budget allocations, and strengthens the authority of the budget appropriation. When separate bank accounts are maintained, the result is often a fragmented system, where funds provided for budgetary appropriations are augmented by additional cash resources that become available through various creative, often extra-budgetary, measures.

2. LITERATURE REVIEW

2.1 Vehicle Clearance

The Comptroller-General of Customs, retired Col. Hameed Ali (2017), on Tuesday announced a code number for efficient and effective vehicle duty clearance verification. Ali disclosed this at a media stakeholders meeting in Abuja.

Ali (2017) said that the essence of the meeting was to have a roundtable with stakeholders to come up with solution to avoid causing hardship to Nigerians in regards to duty payment on old vehicles and verification. He said that customs had taken further step to ensure that Nigerians, who wanted to verify the authenticity of their customs duty clearance, could do so at the comfort of their homes with the use of their mobile phones.

“For effective and easy customs duty clearance verification, you can dial or send SMS to these numbers 094621597 with your vehicle C-number, the year you paid the duty and the port or location where the vehicle came through into the country. “Immediately all that information is given, just in five minutes you will get a response whether your vehicle duty clearance is genuine or not,’’ Ali said. He said the essence of the numbers was to ensure stress free verification, to motor dealers and innocent Nigerian vehicle owners.

According to him, for easy traffic flow, the last number which is 7 in the digits 094621597 can be either changed to 8 or 9, to get response faster with different customs personnel on duty at every point in time. Ali said that Nigerians misunderstood customs intention regarding duty payment on old vehicles, adding that the excise was actually meant for motor dealers.

He added that customs later decided to give innocent private vehicle owners, who after verification, might find out that their vehicles had no genuine duty clearance to take advantage of the 60 per cent rebate.

2.2 Vehicle Registration - AUTO REG

Auto-Reg. is an automated vehicle Licensing and Renewal system in Nigeria. It is a proprietary web based business solution developed by Coulterville Business Solutions PLC, to address the inefficiencies of the motor-vehicle administration system in Nigeria. However, Auto-Reg. succeeded in helping government generate accrued revenue by using a designated banking system for payment of tax and licensing fees but never solved the problem of security and inspection of vehicles. In Auto-reg., vehicle license is to be renewed annually and the system was designed to show the details of vehicle and expiration of licenses but has not addressed the issue of duty evasion, identification of theft vehicle by the police and above all the unification of all the agencies saddled with the responsibilities of vehicular activities in Nigeria for proper collaboration. Since the commencement of Auto-Reg., over one hundred thousand (100,000) cases of number plate duplications in the system have been discovered in Nigeria. (The nation newspaper, Nov. 2014) Auto Reg was deployed first in Lagos State in (February 2007), Oyo (June 2008), Delta (Jun, 2008), Anambra (Mar 2008), Abia (Dec 2008), Rivers (Jan 2009), Enugu (Sept, 2008), Niger (Oct, 2009), Kebbi (Nov, 2009), Borno (Jan, 2010) and Sokoto (Jan, 2010).

2.3 Federal Road Safety Commission.

The FRSC responsibility is to design and produce vehicle number plates by virtue of Section 5(g) and Section 10 sub section 3(f) of the Federal Road Safety Commission (Establishment) Act, 2007. After production, the number plates are handed over to the State through State Board of Internal Revenue (SBIRs) who now sell to the public. Nigerians have berated the FRSC for its handling of motor vehicle registration across the country, describing it as “cumbersome” and “exploitative.” In July 2009, the FRSC planned to restore the integrity of Unified Licensing Scheme (ULS) and National Vehicle Identification Scheme (NVIS), they also planned to maintain a credible database of all drivers in Nigeria and to develop a robust Information and Communication Technology (ICT) network. Indeed, the FRSC was one of the earliest federal agencies to embrace ICT. The idea behind ULS was to unify vehicle and driver licensing in Nigeria in order to create a national database so that authority/personnel would have instant access to vehicle or motorist’s records. Drivers sex, height, blood group, disability, health status etc. similar scheme has long operated in developed and even developing countries with positive implication for road safety management and crime control.

In 2011, The Federal Road Safety Commission in conjunction with the Joint Tax Board (JTB), commenced the issuance of new Number Plates in an attempt to harmonize all existing modes of licensing vehicles nationwide. This according to the Corps Marshall and Chief Executive is part of the commission's strategy towards restoring order and sanity in the nation's Motor Vehicle Administration Scheme. To this end FRSC introduced an Enterprise System called National Vehicle Identification System (NVIS) which is a unified system designed to automate the processes involved in the Number Plate Production and Vehicle Registration. NVIS is open to members of the public who are Vehicle Owners as well as representatives of States, Federal Ministries, Departments and Agencies. (http://nvis.frsc.gov.ng/). The NVIS incorporated only SBIR, VIO and FRSC.

Driver’s license remains a huge racket for road safety officers, revenue officials and the touts that litter licensing offices. Adeniji, K. (2013), hitherto, urban transport problems are becoming more and more acute in the cities in Nigeria

(Badejo Dele, 2013) summarized Features of Urban Transport System in the Nigerian cities, 95% of urban trips are by road. Out of this, about 70% of the urban trips are made by public transport. Inter modality of trips is limited to public transport journey by road based public transport. Ownership and organization of road public transport systems are characterized by haphazard and uncoordinated operators. Complete absence of comprehensive and integration of urban mass transit public transportation system. According to Torres Martinez, A. J (2001) due to poor condition of city roads which in turn shortens life span of motor vehicles and high cost of maintenance. Filani (2002) noted that the country has the lowest level of motorization in West Africa with four vehicles per 1000 inhabitants. To compound the problem further, the rate of vehicle growth is much lower than the population growth rate. Resulting from this mismatch is a general fall in the level of motorization in all parts of the country. Since 1982 and up till 1989/1990 there was a substantial reduction in new vehicle registration in all parts of the country. Olanrenwaju (2013) classified transportation infrastructure as one of the hard infrastructure that are basic, physical and organizational structures needed for the operation of a society or enterprise, or the services and facilities necessary for an economy to function effectively.

2.4 The Need for Technologies for Collecting Vehicle Registration Data.

Hogan, J.O. (2015) in a study conducted in discussing the effectiveness of police computer use and the problems that exist with the use. It was found in that study that the respondents in forty-four cities across the United States view computers as a major force in the fight against crime. This too could be applied in Nigeria if properly established and managed. According to the Minister of Finance, Mrs. Kemi

Adeosun, who disclosed this in a workshop in Abuja titled “FG TO USE TECHNOLOGY TO TACKLE SMUGGLING” noted that the country was losing billions of naira annually to the activities of smugglers and described the technology system as a powerful tool against the illicit and dangerous practice. She also said that there is a need to introduce technology as a platform that provides a form of identity for each vehicle that will be linked to proof of ownership and connected to a centralized database. She added that the programme was also expected to significantly boost vehicle security and ease of transfer of vehicles from one owner to the other. (The Punch Newspaper, 2017).

A variety of technologies have been tested and used by many law enforcement agencies in Nigeria. The technologies used in data collection and processing include a variety of systems such as Mobile Phones, optical storage disks, portable computers, and digital cameras. The current computer technologies allow shareholders to pay their collection/renewal bills at the designated banks or existing offices, electronically transfer the payment to the state agency account and provide deposit slips for the collection of receipts at the state agencies. The use of online error checks, and subsequently the needs for reentering Vehicle detailed data are not inevitable. At the beginning, these devices seem to be the best solution to all the registration problems because it tackles the issues of payment of vehicles registration dues into the government’s account. However, it still has its limitation, as they have not met up with the demands to the masses that spend endless time anxiously waiting for their demands to be met at the Licensing/Commission offices. Hence, the full computerization has not been effected as expected while technology and software programming has advanced in other countries. Shall we continue to wait for the criminals to get away with our stolen vehicles? Shall we keep spending endless time waiting on queues in which have been divulged are corrupt practices of officials based on personalities? Shall we spend endless time searching for owners of whose vehicles have been recovered when software can be developed to tackle such problem like these? The merit of automation is far reacting more than just saving time and holding down persons cost, automating gives vehicle management the means to truly streamline the vehicle registration processes. Automating manual processing tasks allow registration officers eliminate duplicate data entry, move towards a completely paperless environment and process multi - day function, emphasizing the use of technology in vehicle registration. Zhang, Y., Zhang, J., and Chen, J. (2016), opinion was that “in developing computerized system which can help vehicle licensing officers and offices to automatically register with ease, so that the process becomes an automatic day – to – day operation. The solution can help motor licensing officers and offices to improve registration by automating the manual based process, error caused by manual interventions can be reduced and electronic process support enables faster processing time. Meet regulatory demands Archive, email and documentation so that it is easily accessible, usable and quickly retrievable for legal demands. By reducing the administrative burden of paper management and error prone and repetitive data entry in the existing system. For a computerized system to work efficiently and effectively, a strong and reliable database is needed. According to Microsoft encyclopedia, database is a structured format for organizing and maintaining information that can be easily retrieved. Data is stored in a computer in such a way that the computer can easily retrieve and manipulate the data. A collecting of records describing information resources usually computerized. According to Ahmed Suleiman T. (2006), “there are many reason for vehicle registration, take for instance, if you just bought a vehicle and completed all the registration requirement and you are given your vehicle license, then on your way back from the village, you were attacked at gun point and the vehicle snatched from you, you reported to the nearest police station and if you are lucky, your vehicle will be found". It would be difficult for you to get your vehicle within a short period because of the existing system. According to Balogun, Segun A. (2006), states that in his Road Safety Practice in Nigeria that "the method of vehicle and plate number registration and identification has caused a lot of people pains, a pregnant woman died on the queue in her quest for vehicle registration." According to Dr. Ikechukwu David N. (1995), states that "our vehicle registration offices today are faced with potential rise and inefficiencies associated with manual i.e. paper based processes which are costly, prone to error and require mental and manual labor. Heightened regulation in the country is also placing these vehicle owners under pressure to meet litigation needs".

According to Oyeyemi, B O. (2003), states in his Stand in Road Traffic Administration states "the level of tediousness the system of vehicle registration and administration in Nigeria is so alarming that requires a new modified method that will be easy and simple." According to Manager E T. (2000), "most vehicle owner finds it difficult to register their vehicle on time due to the manual process which consumes time. For you to register your vehicle within a short period, you need to know one or two persons in the licensing office. This factor is peculiar to most Nigerian offices". According to Bishop, M (2003), vehicle crime accounts for a quarter of all recorded crime; it costs over £3 billion a year and causes immense distress and inconvenience to its victims to track their records. That is why there is need to setup a national target of reducing vehicle crime by 30% over the next five years in Nigeria. According to Dr. Marcellina Hembadoon A. (2006), "the vehicle plate number is very important because it is an identification mark that distinguishes vehicle from each other. It shows the country a vehicle belongs".

2.6 The Importance of Computer Usage in the Registration.

Computer plays vital role in the development of any company it also saves some of its complex problem that is been faced by man and processes voluminous data within a short period of time or at an incredible speed. Recent emphasis on information and data processing in most of our business has grown adversely as in the case of motor vehicle license and plate registration. In as much as motor vehicle registration has been in existent for ages now, the old system of registration has been in adoption which did not play a significant role on highway safety until the development of the new system of vehicle registration where a reflective sheeting which is more visible to read even in the dark. This new system of motor vehicle and plate number registration, which is the main focus of this project, came into existent on the 19th March 1997 and handled by the motor licensing officer. It was introduced to enforce strict compliance to traffic rules and regulation as well as providing a proper data as to the behavior of road users. The roles, which the introduction of computer system will play in this function, will about more efficiency, effectiveness and improve competence. "The FRSC responsibility is to design and produce vehicle number plates by virtue of Section 5(g) and Section 10 sub section 3(f) of the Federal Road Safety Commission (Establishment) Act, 2007. After production, the number plates are handed over to the States through States Boards of Internal Revenue (SBIRs) who now sell to the public." Nigerians have berated the FRSC for its handling of motor vehicle registration across the country, describing it as "cumbersome" and "exploitative."

"To register a vehicle, an applicant is expected to go to the Motor Licensing Office of the State Board of Internal Revenue (SBIR) where he would be guided on the process and procedure of vehicle registration. Alternatively, the applicants can apply online by visiting www.nvisng.org and fill form, submit the form, after which an item number will be automatically generated which will be taken to SBIR for necessary payment. The applicant will then be issued with necessary vehicle documents. These are Vehicle License, Certificate of Road Worthiness, Valid Insurance Certificate and Proof of Ownership Certificate. Binding classification advice can only be given by the Office of Regulations and Rulings. The importer submits a letter describing the product in detail and provides a sample to the CBP Information Exchange, National Commodity Specialist for a ruling. The importer generally receives a response within 30 days. While tariff classifications are binding, duty rates are not. The object is to promote import compliance, uniformity and accuracy in classification of products. The importer should keep in mind that the Binding Ruling Program is just that- binding. Once CBP issues their decision, it is legally binding and enforceable by law. While the initial ruling may be protested, once a decision is finalized it must be incorporated into the importing process. When submitting a ruling request, include the names, address and other identifying information of all interested parties including the manufacturer. Identify the ports in which the merchandise will be entered and provide a detailed description of the transaction. It always helps to submit a sample of the product when practical. Transport system represents a major interface between the location of activities and the general movement of people in an urban system (Ayeni, 1998).

Hitherto, urban transport problems are becoming more and more acute in the cities in Nigeria (Ogunsanya, 2002; Oyesiku, 2002; etc.) World Health Organization (2000) recently articulated that health concerns related to traffic and transportation have become a worldwide phenomenon and will likely become more of an issue in the future. Findings from other recent studies suggest that stress from transportation may represent an important factor that influences the well - being of urban population (Asiyanbola, 2004; Gee and Takeuchi, 2004). The trend of urbanization and city growth in developing countries are characterized by rapidity of urban increase, urbanization outpacing industrialization, and a high rate of urban population growth by natural increase and migration (Oyesiku, 2002). In Nigeria, urbanization has a fairly long history in its growth and development. Historical account shows that extensive urban development in Nigeria predates the British colonial administration. Early explorers, missionaries and merchants estimates of population of towns show the existence of substantial human settlements in this part of the world in the 19th century (Mabogunje, 1968). During this period, the major factors crucial to the growth and development of cities were trading, marketing and administration.

3. THE PROPOSED SYSTEM

3.1 Analysis of the New System

The purpose of the new system is to create a multipurpose platform that will facilitatet all the procedures by all the agencies in one web application. The new system is a client–server computer program in which the client (including the user interface and client-side logic) runs in a web browser. On importation, the vehicle owner registers the vehicle engine number and chassis on the web application. Other details that will be provided on the registration page for custom duty requires, personal details, Certificate of Entry, Payment Schedule, Engine Number, Receipt of Purchase, Terminal Delivery Order, Vehicle Releasing Invoice, passport photograph and photo of the vehicle. The custom officers at the administrative side of the web application will review the application and also check the vehicle information registered during importation. If verifications are successful, payment gateway will be generated for the user to pay online. Once payment is confirmed, payment confirmation documents will be printed from the website. These documents will acknowledge that the payment was actually made. The user can also request that the document be delivered at home with little additional charges. With the custom duty paid, the applicant can proceed to the SBIR page on the same website. On the SBIR section, the user will input the chassis number of the vehicle and then the system will verify if custom duty has been duly paid. If paid, the applicant can proceed, else the system will redirect the applicant to the custom duty page. After successful payment of the custom duty, an applicant must visit SBIR section. Here he applicant will be required to provide details of his driver's license. The name on the driver's license must be name that will be used in the registration. The picture of the driver's license (front and back sides) will be uploaded too. If the verification is successful, then the user can proceed to SBIR else, the user will be redirected to FRSC section to obtains a driver's license. To obtain a driver's license, the applicant must register at the FRSC section and pay online through the web application. After payment, the application proceeds to FRSC office for driving testing. If the applicant's driving ability is satisfactory, biometric data is captured and temporary driver's license will be issued. After some weeks, the applicant will be contacted for the permanent copy. If the applicant already has a driver's license, the application can proceed with registration on the SBIR section on the web application. On the SBIR section, the applicant will provide custom duty serial number, driver's license number and also fill the allocation of plate number form online. After completion of the form, the applicant will make payment online through the web application. After payment, the applicant must take the vehicle and the payment details to any Vehicle Inspection Office (VIO) so that the vehicle will be physically inspected. After successful inspection, the Vehicle Inspection Office (VIO), will issue verification code that will be used to finish registration on the website. Once verified, all documents and information will be forwarded to Nigerian Police section. The Nigerian Police will stamp and conduct a final verification. After the final verification, Proof of Ownership Certificate (POC), Vehicle Identification Tag (VIT), Vehicle Number Plate and other documents are sent to the nearest SBIR office from the address the applicant provided.

The system also allows an applicant to pay for registrations all at once. To do bulk registration, the applicant will visit 'Bulk Registration' Section. The applicant will select registrations desired and web payment page will be displayed. After payment, a payment code with chassis number will be generated. The applicant will use this printout to visit FRSC, VIO, and finally SBIR for collection of Proof of Ownership Certificate (POC), Vehicle Identification Tag (VIT), Vehicle Number Plate and other documents. Additionally, these documents can be renewed online.

The new system will be used to retrieve information about a particular vehicle using only the chassis number of the vehicle. An officer on the road can check from the web application if a vehicle has been cleared by all other agencies: NCS, FRSC, VIO, NPC and SBIR.

3.2 Description of Input and Output Documents

Since the computer will require data to produce output, input and output format is described. To register a vehicle, the user will to input details such as: full names, date of birth, gender, vehicle engine number, vehicle chassis number, driver license number, state of plate allocation, name of car, model of car, color of car, brand of car, scanned copy of driver's license, engine capacity of vehicle, year of manufacture, etc. After registration, the user will be given an identification number which will be used to track the progress of the registration. On return, the applicant will input the ID number issues to him on registration. The system has input specification for checking registration status, a separate link has been provided for officials in the respective agencies to verify a registration. The system has login for administrative entry to each agency.

3.3 Overview Description of the New System

The new system, which is web based has many benefits and it is designed to solve the problems noted in the existing system. The new system is a web application designed to enable vehicle owners to register their vehicles with government authorities in Nigeria. The purpose of motor vehicle registration is to establish a link between a vehicle and an owner or user of the vehicle. This link might be used for taxation or crime detection purposes. In Nigeria, vehicle registration procedure is conducted by five agencies namely: Nigeria Customs Service (NCS), Federal Road Safety Commission (FRSC), Vehicle Inspection Office (VIO), States Boards of Internal Revenue (SBIR) and Nigeria Police Force (NPF). These agencies have parts to play in vehicle registration and there are procedures too in the registration.

The new system is subdivided into several sub-programs called modules which can be debugged independently. The modules in the new system are:

a. NCS Module
b. FRSC Module
c. NPC Module
d. VIO Module
e. Vehicle registration module
f. Payment module
g. About module

3.4 High Level Model of the New System

The high level model of the new system is shown in figure 3.8.

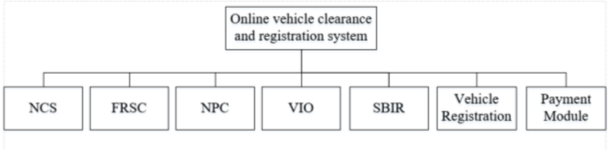


Figure 3.1: High Level Model of the New System

3.5 Overall Data Flow Algorithm

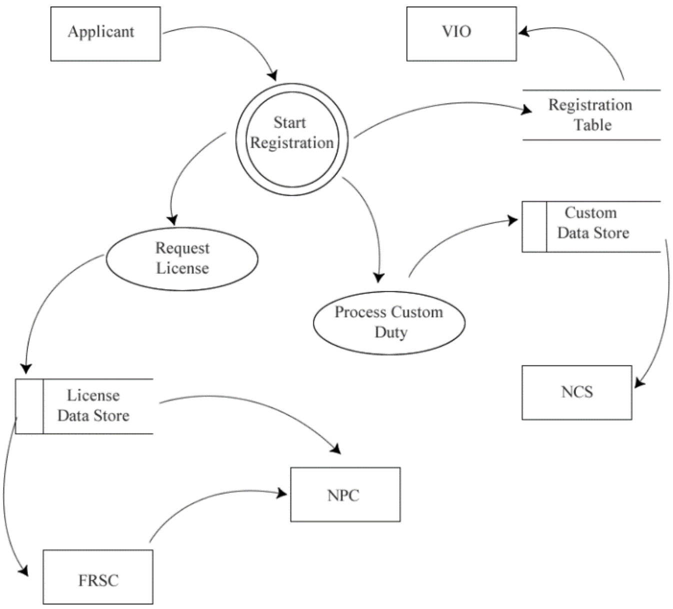


Figure 3.2: Data flow diagram of the new system

3.6 System Flowchart

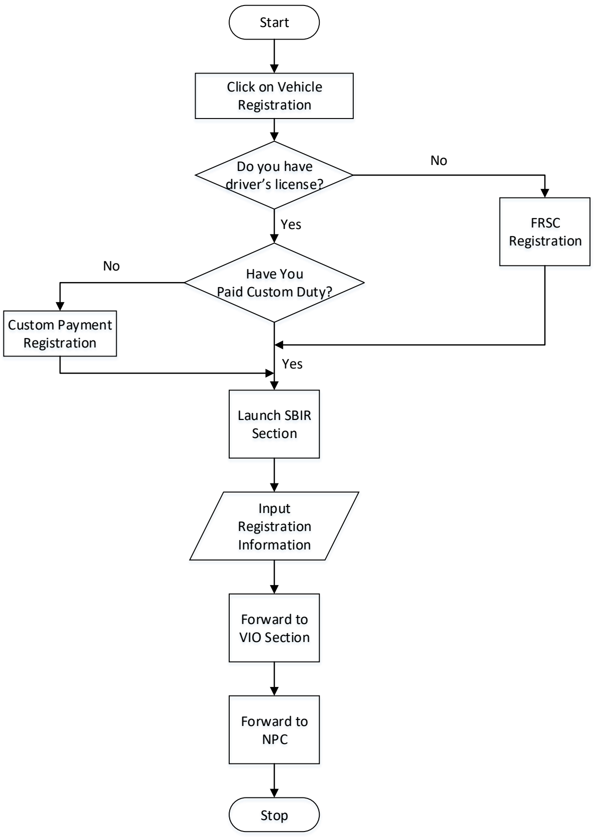

Start
Click on Vehicle Registration
Do you have driver's license?
No
FRSC Registration
Yes
Have You Paid Custom Duty?
No
Custom Payment Registration
Yes
Launch SBIR Section
Input Registration Information
Forward to VIO Section
Forward to NPC
Stop

Figure 3.3: system flowchart of the new system

3.7 Summary

Vehicle registration and plate number are performed casually via online and recording of vehicles information, which ranges from cars to buses and later to truck and heavy duty equipment. Vehicle registration in Nigeria began some years ago and the records have been essentially via net which in turn is not helped to raise the efficiency of general automotive services in recent years and voluminous load on Federal Road Safety Commission. A unified registration software was developed which collaborate with other agency involved in the Vehicle registration and management.

The federal government of Nigeria has identified economic development as a major for achieving the 2020 socio-economic development. The vehicle registration system is a must for any country that wants to be information and communication technology inclined and ready to reduce the vehicle crime rate and corruption in her system.

3.8 Recommendations

The following recommendations are made of the unified online vehicle clearance and registration system:

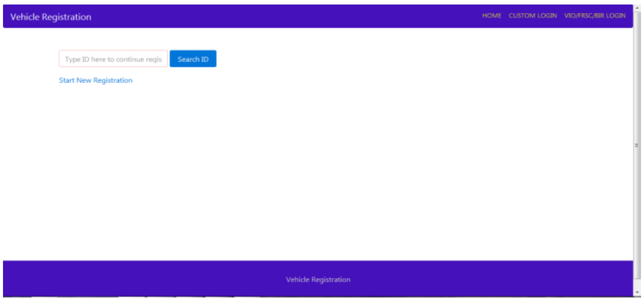
a. It is recommended that this system be used in vehicle registration in Nigeria.
b. The new system is recommended for officers on duty for checking of authenticity of vehicle documents and registration.
c. The new system can be used by the government and other stake holders in Nigeria to monitor the generation of revenue in the country.
d. The system is recommended for the general public for confirmation of services rendered by to them by public officers.
e. For further research, it is recommended that the researcher develop a mobile application for unified online vehicle clearance and registration system.

3.9 Contribution to Knowledge

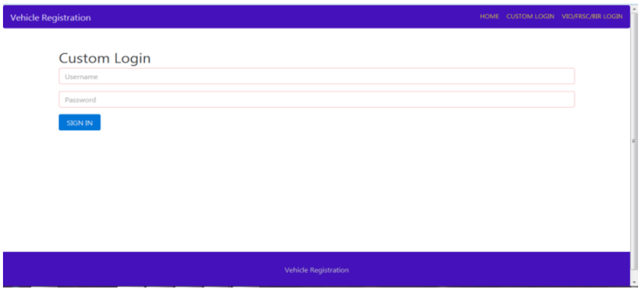
This research will change the way vehicle registration is done in Nigeria. It provides the facility for making vehicle registrations online. The system allows the applicant to select whether the payment is to be made once for all the registrations. The system can accept online payments which will forester faster registrations since the applicant can pay from home. The new system can be used on the road by officers on duty as a replacement for checking of papers. The general public will not be victims of fraud since they have the privilege of verifying vehicle registration made for them by another person.

3.10 Sample Outputs of the Proposed System

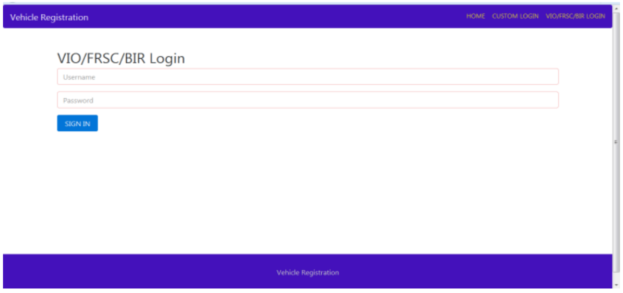
New Registration Page



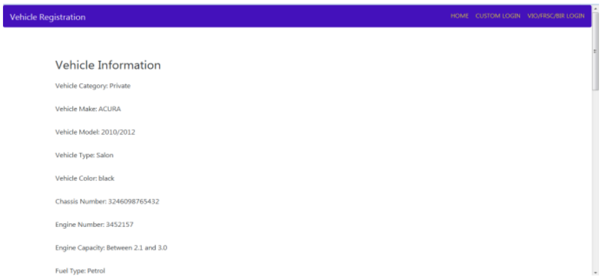
Custom Registration Page



Federal Road Safety Commission Registration Page



Vehicle Information Reregistered



Vehicle Information

Vehicle Category: Private

Vehicle Make: ACURA

Vehicle Model: 2010/2012

Vehicle Type: Salon

Vehicle Color: black

Chassis Number: 3246098765432

Engine Number: 3452157

Engine Capacity: Between 2.1 and 3.0

Fuel Type: Petrol

Glossary

Treasury Single Account (TSA) is a financial policy in use to consolidate all inflows from all agencies of government into a single account at the Central Bank of Nigeria.

Application System: It is a collection of procedures, method, instructions and equipment to produce information in a useful form.

Instructional Rules: Information can be defined as the process of gathering, transmitting, receiving, storing and retrieving data or several items put together to convey a desired message.

Vehicle Plate Number: This is a metallic or plastic plate attached to a motor vehicle for official identification purposes. The number is made up of alphanumeric characters or numbers.

Vehicle Registration: is the process where we add a vehicle’s details to the motor vehicle register and issue its registration plates. You have to license your vehicle regularly at least annually and you must display a current license label on your vehicle windscreen.

Vehicle Licensing: A regular fee paid to permit the use of one’s vehicle on the public roads. The fee helps to pay for road projects and road safety programs. Your vehicle must be both registered and licensed for you to legally drive it on the road.

Vehicle owner: is a person who has met up with the entire necessary requirement for owning a vehicle and has the right to drive it on public roads.

Vehicle Registration and Enquiry Software (VehRES) System: This is an application software that is a customizable data collection system, which can be used by law enforcement and motor vehicle agencies (i.e. Liaison Offices) nationwide.

Licensing office: A place where vehicle registration, licenses and other vehicle related documents are performed.

Licensing officers: Is a person who registers vehicles in the licensing office.

Vehicle: A mechanically propelled and wheeled object used for conveyance.

Computerization: Introduction of the use of computer in an application area by writing a program that will suit the work.
ICT: This is an acronym for Information and Communication Technology.

Federal Road Safety Commission (FRSC): They serve as law enforcement agency charged with responsibilities for, among others, policymaking, organization and administration of road safety in Nigeria.

E-Government: E – Government is a technology exercise, integrating individual database and websites of government.

Driver's license or **driving license** is an official document, which states that a person may operate a motorized vehicle, such as a motorcycle, car, truck or a bus, on a public roadway.

AutoReg Vehicle license: Is the automated vehicle license registration and renewal system, which is for all vehicle owners to register or renew their vehicle license with the state government. It is renewed annually; it shows the details of the vehicle owner and Vehicle details.

REFERENCES

A. O. and Adeniji, S. A. (1998), *“Sustaining Urban Public Transport in Nigeria: Critical issues and Remedies”* in Freeman and Jamet (eds.) Urban Transport Policy. Balkema, Rotterdam, pp. 775-781

Adeniji, K. (1993), *Transport subsidies in Nigeria*: A Synopsis of Workshop Proceedings, NISER, Ibadan and Friedrich Ebert Foundation, Germany.

Adesanya, S. (1996), “*Public transport operation in Nigeria*” In Bolade, T. and Adesanya Transport operations.

Agbola, T. (1989), “Perspective planning: the urban and regional planning dimensions” *The Nigerian Journal of Economic and Social Studies Vol. 31.*

Agbola. T. and Agbola. E.O. (1997), “*The Development of Urban and Regional Planning Legislations an0 d their impact on the Morphology of Nigerian Cities”*. *The Nigerian Journal of Economic and Social Studies*, Vol. 39, No. 1, p. 123-144.

Ahmed, S.T. (1991), “*Essentials of Vehicle Registration in Nigeria”.* Ibadan: University Press Plc.

Balogun, S.A (2006). *Road Safety Practice in Nigeria*. Nigeria: Resources Nig Ltd.

Barry Boehm (1996., "A Spiral Model of Software Development and Enhancement". In: *ACM SIGSOFT Software Engineering Notes* (ACM) 11(4):14-24, August 1986

Barry W. Boehm (2000). *Software cost estimation with Cocomo II: Volume 1*. Centers for Medicare & Medicaid Services (CMS) Office of Information Service

Geoffrey Elliott (2004) *Global Business Information Technology: an integrated systems approach*. Pearson Education. p.87.

Hina; Khalid, Hannan; Ahmed, Mukhtar; Sameer, Abu; Arif, Fahim (2015-09-01). *"Systematic Literature Review of Agile Scalability for Large Scale Projects"*. ResearchGate. **6** (9). *doi*:*10.14569/IJACSA.2015.060908*. *ISSN 2156-5570*.

Hogan, J.O. (1999). *Computer for Everyone.* India: Lone and Vikas Publishing House.

https://infoguidenigeria.com/register-car/

Ikechukwu, D.N. (1995). *Nigeria and Traffic Regulations*. Ibadan: Africana FEB publishers Ltd.

Jerry, N.A (2000). Benefits and Barriers: People with Disabilities and the *National information infrastructure*. Boston: Little, Brown and Company.

Lübke, Daniel and Tammo van Lessen. “Modeling Test Cases in BPMN for Behavior-Driven Development.” IEEE Software 33 (2016): 15-21.

Marcellina H.A. (2006). *Drivers and Passengers Conduct*. London: Macdonald and Evans Ltd.

Oyeyemi, B.O. (2003). *Stands in Road Traffic Administration* Ibadan: Clemeve Media Konsult.

Richard H. Thayer, Barry W. Boehm (1986). *Tutorial: software engineering project management*. Computer Society Press of the IEEE. p.130

Rowland, P.A and Raymond, B.N. (2005). *Usability study information collected* at 1995 CSUN. Boston: Pearson Ally and Bacon.

Suryanarayana, Girish (2015). *"Software Process versus Design Quality: Tug of War?"*. IEEE Software. **32** (4): 7–11. *doi*:*10.1109/MS.2015.87*.

Tunji, F.D. (2001). *Information Provision to Academic Research and Development Organizations in the 21st Century*. The information manager (Pg2(1), 1-9). London: Macdonald and Evans Ltd.

Whitten, Jeffrey L.; Lonnie D. Bentley, Kevin C. Dittman. (2003). *Systems Analysis and Design Methods*. 6th edition. ISBN 0-256-19906-X.

Williams, B.K and Sawyer, S.C (2003). *Using information Technology*. Complete Edition, (Pg 65-79). New York: Mc-Graw Hill http://shipsandports.com.ng/customs-service-announces-code-number-vehicle-duty-clearance-verification/

International Journal of Computer Applications Technology and Research
Volume 7–Issue 08, 327-330, 2018, ISSN:-2319–8656

A Strategy for Improving the Performance of Small Files in Openstack Swift

Xiaoli Zhang
School of Communication Engineering,
Chengdu University of Information Technology
Chengdu, China

Chengyu Wen
School of Communication Engineering,
Chengdu University of Information Technology
Chengdu, China

Zizhen Yuan
School of Communication Engineering,
Chengdu University of Information Technology
Chengdu, China

Abstract: This is an effective way to improve the storage access performance of small files in Openstack Swift by adding an aggregate storage module. Because Swift will lead to too much disk operation when querying metadata, the transfer performance of plenty of small files is low. In this paper, we propose an aggregated storage strategy (ASS), and implement it in Swift. ASS comprises two parts which include merge storage and index storage. At the first stage, ASS arranges the write request queue in chronological order, and then stores objects in volumes. These volumes are large files that are stored in Swift actually. During the short encounter time, the object-to-volume mapping information is stored in Key-Value store at the second stage. The experimental results show that the ASS can effectively improve Swift's small file transfer performance.

Keywords: Openstack; Swift object storage; high performance; small files; aggregated storage strategy.

1. INTRODUCTION

The popularity of the World Wide Web is largely responsible for the dramatic increase in Internet data during the past few years. Usually, social media, e-commerce, scientific experiments and other related fields will produce small files by the tens of millions every day. Global data volume is about double every two years, and will increase to 40ZB by 2020, according to IDC, a market-research firm [1][2]. It is worth noting that the largest proportion and fastest growing are small files. Typically, "a small file" refers to a file less than 1MB in size. The size of the small file ranges from a few KB to tens of KB [3-4]. Texts, pictures, and mails are often small files. The public climate system stores 450,000 climate model files. Their average size is 61 bytes [5]. Sharing photos is one of Facebook’s most popular feature. Users have uploaded over 65 billion photos by 2010 [6]. As the largest personal e-commerce website in the world, TAOBAO stores over 20 billion images, whose average size is only 15KB [7]. How to store and access large numbers of small files efficiently over time makes a new challenge to the storage architecture of the “big data era”.

Storing many small files requires a high performance, high availability, high scalability, security and manageable storage system. But although traditional RAID technology has high performance, it is not suitable for today's Internet environment due to its high cost [8]. NAS and SAN are also not suitable for storing large amounts of data because of their limited scalability [9]. The famous GFS (Global File System) consists of inexpensive PC servers and provides fault tolerance [10]. However, when the system stores small files, as the number of stored files grows rapidly, plenty of metadatas are generated on the metadata server. This results in poor file access performance. Facebook independently developed Haystack as its image storage dedicated storage system [11]. Nevertheless, it is limited in scalability because it refers to the central node design of GFS. To solve this problem, Amazon developed Dynamo storage system [12]. It adopts the method of no center node and relies on the hash algorithm to solve the file distribution problem. Similarly, Cassandra [13] and TAIR [14] are non-centralized storage system. Unfortunately, they are designed for the storage of large files and do not optimize the transfer performance of small files. In this paper, we propose the ASS for improving the transfer performance of a large number of small files in Swift. ASS has two parts. In the first stage, the ASS arranges he written objects one by one, and then merges them into large files in chronological order. Those large files are called “volumes”, which are actually stored in Swift. In the second stage, the object-to-volume mapping information (volume id, location) is stored in the key-value store.

The remainders of the paper are organized as follows: Section 2 discusses related works on improving the transfer performance of small files. In Section 3, we described the basic principles of ASS. At the same time, the ASS read algorithm and small file read/write process are introduced. At Section 4, we introduced the experimental environment and analyzed the experimental results. Section 5 concludes the paper.

2. RELATED WORKS

Many people have tried various schemes to improve the small file storage access performance. The index layout strategy can achieve efficient reading of small files by optimizing the physical layout of directory entries, inodes, and data blocks. For example, to reduce the number of IO, C-FFS [15] embeds the inodes in the directory entry and replaces the inodes pointer of the directory entry with inodes. But this strategy has the disadvantage of synchronous recovery operations in a distributed environment. The Cache structure optimization strategy reduces the access time of the storage node by using the external cache CDN and the internal cache, which effectively improves the cache hit ratio. For instance, for efficient access, the Sprite file system uses a stand-alone Cache, and each server node has its own cache space [16]. Lustre leverages the distributed cache space of each client. It uses a collaborative caching strategy that reduces the load on a single server cache [17]. This approach improves file access efficiency. However, the multi-level cache is only effective for hotspot data accessed in the most recent period of time. Due to the small number of hotspot data, it will cause a lot of non-hotspot data access inefficiency.

At present, the combined storage solution is also widely used in the industry. Its main idea is to reduce the amount of metadatas in the metadata server. And it can improve the read/write efficiency of small files by consolidating small files into large data block storage. The consolidation of small files has many different implementations. For example, Hadoop uses its own merging file tool - HAR file archiving. The principle of HAR is to pack multiple small files into one file and then save it to a block. The archive mainly contains metadata and data files [18]. But the merging file tool that comes with the system is often to merge and archive the small files already stored in the system. This can lead to a lot of disk read and write consumption. In fact, it is also possible to merge files on the client before uploading the storage. However, the measure often stores index information locally. When a small file is requested, the system first transmits the entire data block to the client and then reads the offset. This method will result in a large number of invalid data network transmission bandwidth.

Compare with the strategies discussed above, our work differs in two ways: (1) This paper establishes a separate merge engine in Swift object storage. The merge engine combines small files into large files before storing them. It is worth noting that it applies to any small object, such as pictures, documents, etc. (2) For this merge engine, a method of merging files is proposed—ASS.

3. MERGE ENGINE

3.1 An aggregated storage strategy

Swift uses loopback devices and the VFS file system as the underlying storage. In this paper, based on the original Swift framework, a merge engine is added between the object server and the XFS file system. The merge engine uses an aggregate storage strategy. This strategy allows multiple logical files to share the same physical file. It reduces the number of files and metadatas, improves the efficiency of metadatas retrieval and query, and reduces I/O operation delays for file reads. And effectively solved Swift's small file storage problem. The keys to strategy are:

(1) Merge storage: The basic idea of the strategy is to store objects in a volume. Volumes are large files that are stored in Swift actually. This policy stores objects in a volume and separates volumes through Swift's virtual partition, which not only improves the transmission performance of small file, but also ensures Swift's data migration capabilities.

(2) Index storage: The object—volume mapping information (volume id, position) is stored in the key value store (KV server) for cluster maintenance.

The merge engine module includes an object request layer, an object merge layer, a logical map layer, and a physical map layer. When Swift's storage node receives a PUT or GET request from a proxy node, in the original case, Swift Ring uses a Consistent Hashing Algorithm to complete the "object-virtual node-device" mapping. In this paper, since a logical map layer is added, the "object-volume-virtual node-device" mapping is formed. The "volume-virtual node" mapping relationship is a logical mapping, and the "virtual node-device" mapping relationship is a physical mapping. The merge engine module uses ASS, which works as follows.

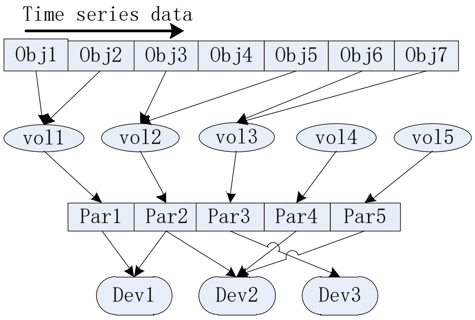


Figure 1. Basic theory of ASS.

In Figure 1, "obj" is the object, "vol" is the volume, and "Par" is Partition. As Figure 1 shows, ASS aggregates files according to the time characteristics of the objects. On the one hand, the solution translates random writes into sequential writes. It reduces the system's garbage collection overhead and data migration overhead. On the other hand, the solution merges and stores the data, which reducing the processing cost of metadatas. Both can effectively improve the transmission performance of small files in Openstack Swift.

3.2 The process of reading and writing files

In this paper, the improvement of Swift framework is embodied in the optimization of reading and writing. The flow of small file read/write operations is shown in the Figure 2.

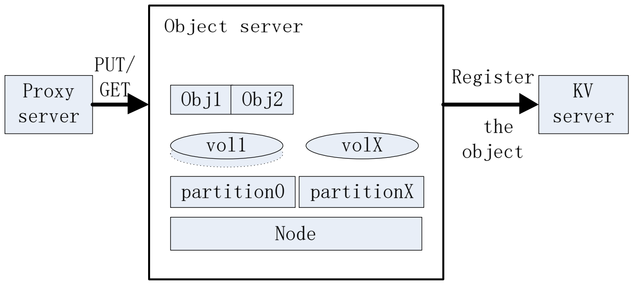


Figure 2. File read-write process.

Write: When the proxy server receives a PUT request from the client, it then forwards the PUT request to the storage nodes. Firstly, storage nodes look for an unlocked volume, or creates a new writable volume and associated lock file (if a new volume is created, it needs to be registered in the KV server). Secondly, storage nodes lock this volume. Storage nodes then appends object information (Object header、Object metadata、Object data) to the end of the volume, just like the shaded part of the figure. The next step is to synchronize the volumes. Finally, storage nodes register objects to the KV server, which is to add new entries to the key-value store.

Read: When the proxy server receives a GET request from the client, it then forwards the GET request to the storage nodes. Firstly, the storage node gets the (volume index、offset in the volume) information of the object from the KV server to locate the volume. The storage node then opens the volume files, gets the offsets, and locates the objects.The reading algorithm of the files is as follows:

```
Filereading(obj Name, obj Size=0)
Currentposition←fileepositon(obj Name)
objheader←header(obj Name)
datasize←datasize(objheader)
```

```
datastartoffset←offset(obj Name)+ dataoffset(objheader)
dataendoffset←datasize+ datastartoffset
if Currentposition>= dataendoffset or obj Size=0
        then call normal Read(c)
if obj Size is normal and obj Size>dataendoffset –
Currentposition
        then Obj Size←dataendoffset- Currentposition
else Obj size←dataendoffset- Currentposition
data←read(filepositon，Obj size)
return data
```

4. EXPERIMENTAL ENVIRONMENT AND RESULTS

4.1 Experimental environment

To verify the effectiveness of the strategy, a small Swift cluster consisting of one proxy node and three storage nodes is built on the virtual machine. The deployment of each service is shown in Table 1.

Table 1. The deployment of each service in Swift cluster

Name	Operating system	Hard-drive sizes	Memory Sizes	Major services
Controller	Centos7	20GB	2GB	Swift client, keystone
Node1	Centos7	20GB	2GB	CARP, HAProxy, Swift storage
Node2	Centos7	20GB	2GB	CARP, HAProxy, Swift storage
Node3	Centos7	20GB	2GB	CARP, HAProxy, Swift storage

4.2 Experimental results

To better test the improved small files storage access performance of the improved Swift framework, many stress testing experiments have been performed on the improved framework. Swift-bench was used as test tool. The experimental test results are as follows.

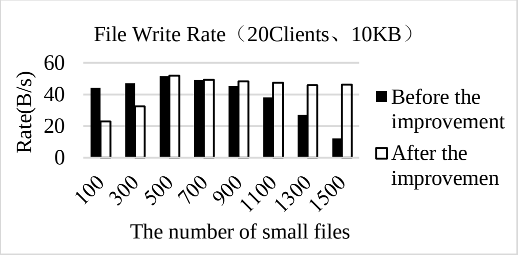


Figure 3. File write rate(20clients、10KB).

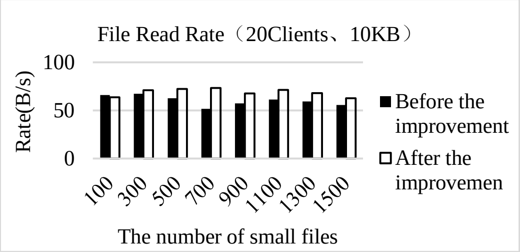


Figure 4. File read rate(20clients、10KB).

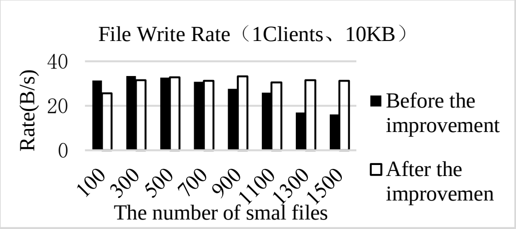


Figure 5. File write rate(1clients、10KB).

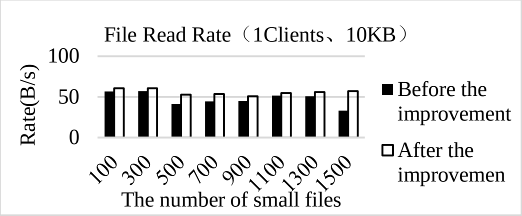


Figure 6. File reading rate(1clients、10KB).

As shown in Figure 3 and Figure 4, in the case of 20 clients writing 10KB small files concurrently, when the number of files is less than 300, the optimized system performance is lower than that of the unoptimized system. However, as the number of files increases, the transmission performance of non-optimized systems gradually decreases, and the performance advantages of optimized systems become more pronounced. In the same situation, the read performance of small files is similar to the former. The scenario where a cluster has only one client is shown in Figure 5 and Figure 6:as the number of files increases, the read/write performance of the optimized cluster is generally greater than that of no optimization. We believe that as the number of files increases, the IO of the system becomes more and more crowded. At this time, the merge strategy can reduce the number of inodes, thereby ensuring the stability of the system performance.

In order to continue to verify the effectiveness of the ASS. In the case of 20 clients, these clients uploads/download 500 small files respectively. At the same time, we record the file access rate in each case as follows. Note that the size of 500 small files is 1KB, 5KB, 10KB···, 100 KB.

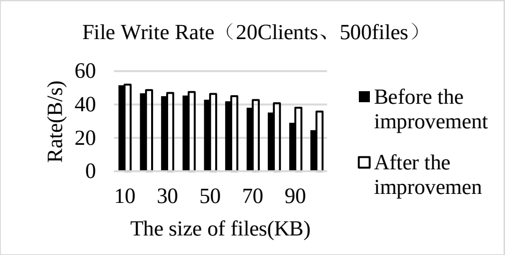


Figure 7. File write rate(20Clients、500files).

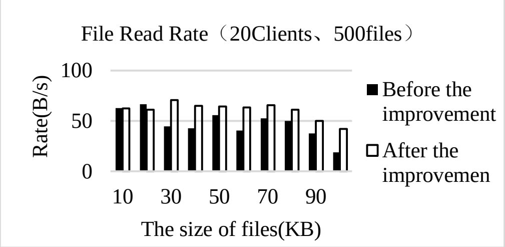


Figure 8. File read rate(20Clients、500files).

As shown in Figure 7, when 20 clients write 500 files at the same time, the improved cluster's small files transfer performance is usually higher than the unimproved cluster. In the same case, the clients read to the cluster. Although the small files transfer performance of the optimized cluster is low when the size of files is less than 20KB, the optimized system performance is more stable overall. And we believe that the improved system improves the storage and access performance of small files.

5. CONCLUSIONS

This paper describes an aggregated storage strategy that is used to improve small file storage performance in Openstack Swift. Based on the original Swift framework, we added a merge engine module between the object server and the XFS file system. This module uses ASS. Then we use ASS to merge small files into volumes. Experiments show that the improved cluster reduces IO congestion and improves the read/write performance of small files.

6. ACKNOWLEDGMENTS

The authors would like to thank the persons who review and give some valuable comments to improve the paper quality. This work was supported by Science and Technology Department of Sichuan Province, Fund of Science and Technology Planning (No. 2018JY0290).

7. REFERENCES

[1] Zwolenski, Matt, and L. Weatherill. "The Digital Universe Rich Data and the Increasing Value of the Internet of Things." Australian Journal of Telecommunications and the Digital Economy 3,2014.

[2] John Gantz, and David Reinsel. "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east." IDC iView: IDC Analyze the Future, 2007:1-16.

[3] J. R Douceur, W. J Bolosky, J. R Lorch, and N. Agrawal. " A five-year study of file-system metadata." ACM Transactions on Storage, 2007: 9-9.

[4] Meyer, T. Dutch, and W. J. Bolosky. "A study of practical deduplication." Usenix Conference on File and Stroage Technologies USENIX Association, 2011:1-1.

[5] A. Chervenak, J. M. Schopf, L. Pearlman, M. H. Su, S. Bharathi, M. D'Arcy, N. Miller, D. Bernholdt and L. Cinquini. "Monitoring the Earth System Grid with MDS4." IEEE International Conference on E-Science and Grid Computing, 2006. E-Science IEEE, 2006:69-69.

[6] D. Beaver, S. Kumar, H. C. Li, J. Sobel, and P. Vajgel. "Finding a needle in Haystack: facebook's photo storage." Usenix Conference on Operating Systems Design and Implementation USENIX Association, 2010:47-60.

[7] Wang, Jing, and Y. Guo. "Scrapy-Based Crawling and User-Behavior Characteristics Analysis on Taobao." International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery IEEE, 2012:44-52.

[8] C. Weddle, M. Charles, J. Qian, A. I. A. Wang, P. Reiher, and G. Kuenning. "PARAID: a gear-shifting power-aware RAID." Usenix Conference on File and Storage Technologies USENIX Association, 2007:30-30.

[9] Sacks, D. "Demystifying Storage Networking DAS, SAN, NAS, NAS Gateways, Fibre Channel, and iSCSI." Ibm Storage Networking, 2001.

[10] S. Ghemawat, H. Gobioff, S. T. Leung. "The Google file system." ACM SIGOPS Operating Systems Review 37, 2003:29-43.

[11] D. Beaver, S. Doug, H. C. Li, J. Sobel, and P. Vajgel. "Finding a needle in Haystack: facebook's photo storage." Usenix Conference on Operating Systems Design and Implementation USENIX Association, 2010:47-60.

[12] G. Decandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, and W. Vogels. "Dynamo: amazon's highly available key-value store." ACM Sigops Symposium on Operating Systems Principles ACM, 2007:205-220.

[13] Lakshman, Avinash, and P. Malik. "Cassandra:a decentralized structured storage system." Acm Sigops Operating Systems Review 44,2010:35-40.

[14] Y. han. "A brief analysis of No SQL database solution Tair. " The electronic commerce,2011:54-61.

[15] L. zhang. Research and implementation of embedded file system based on flash memory. University of Electronic Science and Technology of China, 2005.

[16] Zhong, S, J. Chen, and Y. R. Yang. "Sprite: a simple, cheat-proof, credit-based system for mobile ad-hoc networks." Joint Conference of the IEEE Computer and Communications. IEEE Societies IEEE, 2003:1987-1997.

[17] Nie, Gang, and Q. Xiu-Hua. "Research on Lustre file system based on object-based storage." Information Technology, 2007.

Website:

[18] http://hadoop.apache.org/docs/current/hadoop-archives/
HadoopArchives.html

Semantic Similarity Measures between Terms in the Biomedical Domain within frame work Unified Medical Language System (UMLS)

Abdelhakeem M. B. Abdelrahman
Sudan University of Science and Technology
Collage of Graduate Studies Khartoum, Sudan

Dr. Ahmad Kayed
Department of Computing and Information Technology
Sohar University, Sohar, Oman

Abstract

The techniques and tests are tools used to define how measure the goodness of ontology or its resources. The similarity between biomedical classes/concepts is an important task for the biomedical information extraction and knowledge discovery. However, most of the semantic similarity techniques can be adopted to be used in the biomedical domain (UMLS). Many experiments have been conducted to check the applicability of these measures. In this paper, we investigate to measure semantic similarity between two terms within single ontology or multiple ontologies in ICD-10 “V1.0” as primary source, and compare my results to human experts score by correlation coefficient.

Keywords: Information extraction, biomedical domain, semantic similarity techniques, Unified Medical Language System (UMLS), and Semantic Information Retrieval (SIR).

1. INTRODUCTION

Ontology is test bed of semantic web, capturing knowledge about certain area via providing relevant concept and relation between them. Quality metrics are essential to evaluate the quality. Metrics are based on structure and semantic level. At the present the ontology evaluation is based only on structural metrics, which has not been very appropriate in providing desired results.

Semantic similarity measures are widely used in Natural Language Processing. We show how six existing domain-independent measures can be adapted to the biomedical domain. Semantic similarity techniques are becoming important components in most intelligent knowledge-based and Semantic Information Retrieval (SIR) systems [1]. Measures and tests are provided to define how we can measure the "goodness" of ontology or its resources. Many experiments have been conducted to check the applicability of these measures [4].

General English ontology based structure similarity measures can be adopted to be used into the biomedical domain within UMLS. New approach for measuring semantic similarity between biomedical concepts using multiple ontologies is proposed by Al-Mubaid and Nguyen [2, 3]. They proposed new ontology structure based technique for measuring semantic similarity between single ontology and multiple ontologies in the biomedical domain within the frame work of Unified Medical Subject Language System (UMLS). Their proposed measure based on three features [2]: first Cross modified path length between two concepts. Second, new features of common specificity of concepts in the ontology. Third Local ontology granularity of ontology cluster.

2. BIOMEDICAL DOMAIN ONTOLOGIES

Most of the semantic similarity techniques work in the biomedical domain uses only ontology (e.g. MeSH, SOMED-CT) for computing the similarity between the biomedical terms[9]. However, in this work we use ICD- 10 ontology as primary source to computing the similarity between concepts in biomedical domain.

International Classification of Diseases (ICD): The newest edition (ICD- 10) is divided into 22 chapters: (Infections, Neoplasm, Blood Diseases, Endocrine Diseases, etc.), and denote about 14,000 classes of diseases and related problems. The first character of the ICD code is a letter, and each letter is associated with a particular chapter, except for the letter D, which is used in both Chapter II, Neoplasm, and Chapter III, Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism, and the letter H, which is used in both Chapter VII, Diseases of the eye and adnexa and Chapter VIII, Diseases of the ear and mastoid process. Four chapters (Chapters I, II, XIX and XX) use more than one letter in the first position of their codes. Each chapter contains sufficient three-character categories to cover its content; not all available codes are used, allowing space for future revision and expansion. Chapters I–XVII relate to diseases and other morbid conditions, and Chapter XIX to injuries, poisoning and certain other consequences of external causes. The remaining chapters complete the range of subject matter nowadays included in diagnostic data. Chapter XVIII covers Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified. Chapter XX, External causes of morbidity and mortality, was traditionally used to classify causes of injury and poisoning, but, since the Ninth Revision, has also provided for any recorded external cause of diseases and other morbid conditions. Finally, Chapter XXI, Factors influencing health status and contact with health services, is intended for the classification of data explaining the reason for contact with health-care services of a person not currently sick, or the circumstances in which the patient is receiving care at that particular time or otherwise having some bearing on that person's care [8, 10].

3. SEMANTIC SIMILARITY TECHNIQUES CHALLENGES IN THE BIOMEDICAL DOMAIN

Most of existing semantic similarity techniques that used ontology structure as the primary source can't measure the similarity between terms using single ontology or multiple ontologies in the biomedical domain within frame work Unified Medical Language System (UMLS). However, some of the semantic similarity techniques have been adopted to biomedical domain by incorporating domain information extracted from clinical data or medical ontologies.

4. RELATED WORK

4.1 Rada et al. Proposed semantic distance as a potential measure for semantic similarity between two concepts in MeSH, and implemented the shortest path length measure, called CDist, based on the shortest distance between two concept nodes in the ontology. They evaluated CDist on UMLS Metathesaurus (MeSH, SNOMED, ICD9), and then compared the CDist similarity scores to human expert scores by correlation coefficients.

4.2 Caviedes and cimino. [11] Implemented shortest path based measure, called CDist, based on the shortest distance between two concepts nodes in the ontology. They evaluated CDist on UMLS Metathesaurus (MeSH, SNOMED, ICD9), and then compared the CDist similarity scores to human expert scores by correlation coefficient.

4.3 Pedersen et al.[1] Proposed semantic similarity and relatedness in the biomedicine domain, by applied a corpus-based context vector approach to measure similarity between concepts in

SNOMED-CT. Their context vector approach is ontology-free but requires training text, for which, they used text data from Mayo Clinic corpus of medical notes.

4.4 Wu and Palmer Similarity Measure [11] proposed a new method which define the semantic similarity techniques between concepts C_1 and C_2 as

$$\mathrm{Sim}\,(C_1, C_2) = 2 \times \frac{N_3}{N_1 + N_2 + 2 \times N_3} \qquad (1)$$

Where

N_1 is the length given as the number of nodes in the path from C_1 to C_3 which is the least common super concept of C1 and C2, and

N2 is the length given in the number of nodes on a path from C2 to C3.

N3 represents the global depth of the hierarchy and it serves as the scaling factor.

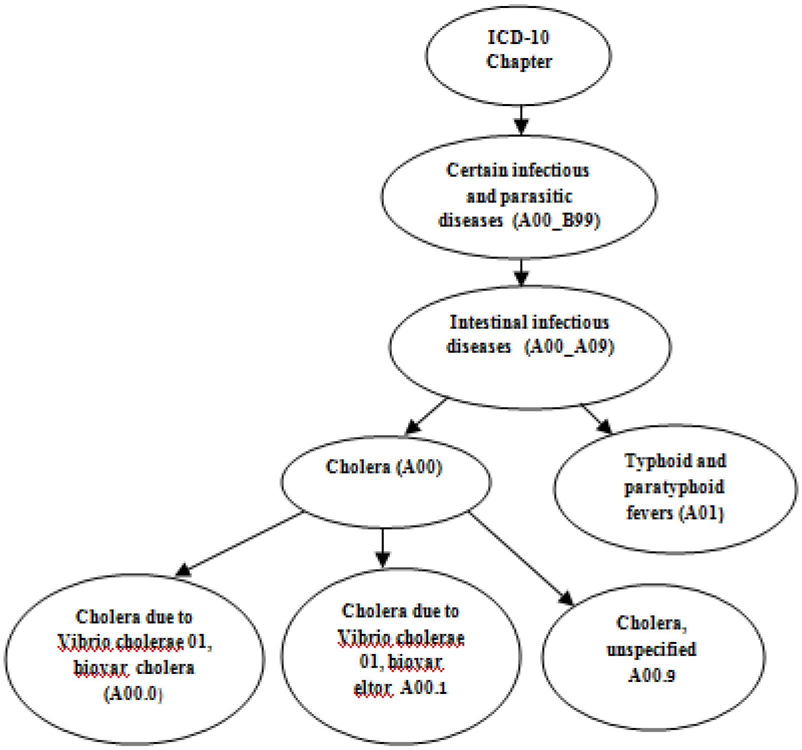

ICD-10 Chapter
Certain infectious and parasitic diseases (A00_B99)
Intestinal infectious diseases (A00_A09)
Cholera (A00)
Typhoid and paratyphoid fevers (A01)
Cholera due to Vibrio cholerae 01, biovar cholera (A00.0)
Cholera due to Vibrio cholerae 01, biovar eltor A00.1
Cholera, unspecified A00.9

Figure 1 fragment of Intestinal infectious diseases

For example from Figure 1: (LCS (A00.1, A00.9) = A00 and LCS(A00 ,A01) = A00_A09) of two concept nodes and N_1, N_2 are the path lengths from each concept node to LCS, respectively. 4.5 Al- Mubaid and Nguyen Similarity technique [5, 11] proposed measure take the depth of their least common subsume (LCS) and the distance of the shortest path between them. The higher similarity arises when the two concepts are in the lower level of the hierarchy. Their similarity measure is:

$$Sim (c_1, c_2) = \log_2 ([L(c_1, c_2) -1] \times [D- depth(L(c_1, c_2)] + 2) \quad (2)$$

Where:

$L(c_1, c_2)$ is the shortest distance between c1 and c_2.

Depth $L(c_1, c_2)$ is depth of $L(c_1, c_2)$ using node counting.

$L(c_1, c_2)$ lowest common subsume of c_1 and c_2.

D is the maximum depth of the taxonomy.

The similarity equal 1, where two concepts nodes are in the same cluster/ontology. The maximum value of this measure occur when one of the concepts is the left most leaf node, and the other concept is the right leaf node in the tree. In the ICD-10 tree let us consider an example in ICD-10 terminology. The category tree is "Intestinal infectious diseases" and is assigned letter A in ICD10 terminology version 2016 at the link (http://apps.who.int/classifications/icd10/browse/2016/en#/A00-A09). This tree looks as follows:

Intestinal infectious diseases [A00-A09]
Cholera [A00]+
Typhoid and paratyphoid fevers [A01]+
Other salmonella infections [A02]+
Shigellosis [A03]+
Viral and other specified intestinal infections [A08]+
Other gastroenteritis and colitis of infectious and unspecified origin [A09]+

The similarity between "Cholera [A00]" and "Typhoid and paratyphoid fevers [A01]" is less similarity than the similarity between "Cholera due to Vibrio cholerae 01, biovar eltor [A00.1]" and "Cholera, unspecified [A00.9]". However, in this measure they take into account the depth

The symbol "+" indicates that the concept can be further expanded into a sub tree (sub-concepts). For example, "Cholera" [A00] can be expanded to be as follows:

Cholera [A00]

Cholera due to Vibrio cholerae 01, biovar cholerae [A00.0]+

Cholera due to Vibrio cholerae 01, biovar eltor [A00.1]+

Cholera, unspecified [A00.9]+

of the LCS of two concepts, in the path length and leacock & chodorwo produce semantic similarity for two pairs [(A00, A01) and (A00.1, A00.9)] in sim (c_1, c_2) measure (Eq 2 in table 1) give high similarity in lower level in the ontology hierarchy ([A00.1, A00.3]).

Table 1: Measures Comparison

Pair of Concepts	**P. L**	**L. C**	**C. K**	**Hisham Al-Mubaid & Nyguan Measure (Eq 2)**
A00 – A01	0.37	2.13	0.91	3.2
A00.1 – A00.9	0.33	2.15	0.91	1.6

The higher numeric similarity result between (A00, A01) means the lower semantic similarity between them.

5. EVALUATION

5.1 Datasets:

There are no standard human rating sets for semantic similarity in biomedical domain. Thus, Hisham Al-Mubaid and Nguyen [3, 11] used dataset from Pedersen et. al [1], which was annotated by 3 physician and 9 medical index experts to evaluate their proposed measure in the biomedical domain.

Table 2 Dataset 1: 30 medical term pairs sorted in the order of the average [1].

Id	*Concept1*	*Concept2*	*Phys*	*Expert*	*Id*	*Concept1*	*Concept2*	*Phys*	*Expert*
4	Renal failure I12.0	Kidney failure I12.0	4.0000	4.0000	**27**	**Acne**	**Syringe**	**2.0000**	**1.0000**
5	Heart I51.5	Myocardium I51.5	3.3333	3.0000	12	Antibiotic (Z88.1)	Allergy (Z88.1)	1.6667	1.2222
1	Stroke I64	Infarct I64	3.0000	2.7778	**13**	**Cortisone**	**Total knee replacement**	**1.6667**	**1.0000**
7	Abortion O03	Miscarriage O03	3.0000	3.3333	**14**	**Pulmonary embolus**	**Myocardial infarction**	**1.6667**	**1.2222**
9	Delusion (F06.2)	Schizophrenia (F06.2)	3.0000	2.2222	16	Pulmonary Fibrosis (E84.0)	Lung Cancer (C34.1)	1.6667	1.4444
11	Congestive heart failure (I50.0)	Pulmonary edema (I50.1)	3.0000	1.4444	**6**	**Cholangiocarcinoma**	**Colonoscopy**	**1.3333**	**1.0000**
8	Metastasis (C77.0)	Adenocarcinoma (C08.9)	2.6667	1.7778	29	Lymphoid hyperplasia (K38.0)	Laryngeal Cancer (C32.0)	1.3333	1.0000
17	Calcification (M61)	Stenosis (H04.5)	2.6667	2.0000	21	Multiple Sclerosis (F06.8)	Psychosis (F06.8)	1.0000	1.0000
10	**Diarrhea**	**Stomach cramps**	**2.3333**	**1.3333**	22	Appendicitis (K35)	Osteoporosis (M80)	1.0000	1.0000
19	Mitral stenosis (I05.0)	Atrial fibrillation (I48)	2.3333	1.3333	23	Rectal polyp (K62.1)	Aorta (I70.0)	1.0000	1.0000
20	Chronic obstructive pulmonary disease (J44.9)	Lung infiltrates (J82)	2.0000	1.8889	24	Xerostomia (K11.7)	Alcoholic cirrhosis (K70.3)	1.0000	1.0000
2	Rheumatoid arthritis (M05.3)	Lupus (L93)	2.0000	1.1111	25	Peptic ulcer disease (K21.0)	Myopia (H52.1)	1.0000	1.0000
3	Brain tumor (G94.8)	Intracranial hemorrhage(I69.2)	2.0000	1.3333	26	Depression (F20.4)	Cellulitis (H60.1)	1.0000	1.0000
15	Carpal tunnel Syndrome (G56.0)	Osteoarthritis (M19.9)	2.0000	1.1111	**28**	**Varicose vein**	**Entire knee meniscus**	**1.0000**	**1.0000**
18	Diabetes mellitus (E10-E14)	Hypertension (I10-I15)	2.0000	1.0000	30	Hyperlipidemia (E78.0)	Metastasis (C77.0)	1.0000	1.0000

5.2 Experiments and Results

Table 2. Test set of 30 medical term pairs sorted in the order of the averaged physicians' scores (taken from Pedersen et. al. 2005 [1]). Al-Mubaid and Nguyen [5, 11] find only 24 out of the 30 concept pairs in ICD-10 using http://apps.who.int/classifications/icd10/browse/2016/en browser version 2010.

Another biomedical dataset was used containing 36 MeSH term pairs [15]. The human scores in this dataset are the average evaluated scores of reliable doctors. UMLSKS browser was used [12] for SNOMED-CTterms, and MeSH Browser [13] for MeSH terms. Table 3, Table 4, Table 5, and Table 6 show Dataset2 along with human scores and scores of Path length, Wu and Palmer's, Leacock and Chodorow's, and Hisham Al-Mubaid & Nguyen techniques calculated using MeSH ontology. The term pairs in bold, in Table 3, Table 4, Table 5, and Table 6, are the ones that contain a term that was not found in MeSH Ontology and they were excluded from experiments.

Table3. Biomedical Dataset 2 (36 pairs) with human similarity scores (Human) and Path length's scores using MeSH ontology.

Id	Concept 1	Concept 2	Human	Path length
1	Anemia	Appendicitis	0.031	8
2	Meningitis	Tricuspid Atresia	0.031	8
.		.	.	.
.		.	.	.
36	Chicken Pox	Varicella	0.968	1

Table 4. Biomedical Dataset 2 (36 pairs) with human similarity scores (Human) and Wu and Palmer's scores using MeSH ontology.

Id	Concept 1	Concept 2	Human	Wu &Palmer
1	Anemia	Appendicitis	0.031	0.364
2	Meningitis	Tricuspid Atresia	0.031	0.364
.		.	.	.
.		.	.	.
36	Chicken Pox	Varicella	0.968	1.000

Table 5. Biomedical Dataset 2 (36 pairs) with human similarity scores (Human) and Leacock and Chodorow's scores using MeSH ontology.

Id	Concept 1	Concept 2	Human	Leacock & Chodorow
1	Anemia	Appendicitis	0.031	1.099
2	Meningitis	Tricuspid Atresia	0.031	1.099
.		.	.	.
36	Chicken Pox	Varicella	0.968	3.178

Table 6. Biomedical Dataset 2 (36 pairs) with human similarity scores (Human) and Hisham Al-Mubaid & Nguyen measure (SemDist) using MeSH ontology.

Id	Concept 1	Concept 2	Human	SemDist
1	Anemia	Appendicitis	0.031	4.263
2	Meningitis	Tricuspid Atresia	0.031	4.263
36	Chicken Pox	Varicella	0.968	0.000

6. CONCLUSIONS AND FUTURE WORK

In this paper we discussed the basics of semantic similarity techniques, the classification of single ontology similarity measures and cross ontologies similarity measures. We prepare a brief introduction of the various semantic similarity measures in biomedical domain. However, from all the above, we can used SemDist as semantic similarity measures in the biomedical domain. In future work, we intend to explore the semantic similarity techniques in the biomedical domain (ICD10, MeSH, and SNOMED-CT) within UMLS frame work. We also prepare implement a web-based user interface for all these semantic similarity techniques and to make it available freely to researchers over the Internet. That will be much helpful for interested researchers in the field of bioinformatics text mining.

7. REFERENCES

[1] Ted Pedersen, et al. " Measures of semantic similarity and relatedness in the biomedical domain ", Journal of Biomedical Informatics 40 (2007) 288–299.

[2] Hisham Al-Mubaid and Hoa A. Nguyen, “A Cluster-Based Approach for Semantic Similarity in the Biomedical Domain” Proceedings of the 28th IEEE, EMBS Annual International Conference New York City, USA, Aug 30-Sept 3, 2006.

[3] Hisham Al-mubaid & Hoa A. Nguyen “Measuring Semantic Similarity between Biomedical concepts within multiple ontologies” IEEE Trans Syst Man Cybern Part C: Appl Rev 2009, 39.

[4] Ahmad Kayed, et al. "Ontology Evaluation: Which Test to Use" 2013 5th International Conference on Computer Science and Information Technology (CSIT), IEEE, pp 45-48, 2013.

[5] Hisham Al-Mubaid and Hoa A. Nguyen, “New Ontology Based Semantic Similarity for the Biomedical Domain”, (2006) p 623 – 628.

[6] S. Anitha Elavarasi, et. al, “A Survey on Semantic Similarity Measure” International Journal of Research in Advent Technology, Vol.2, No.3, March 2014 E-ISSN: 2321-9637.

[7] Nguyen H., Al-Mubaid H. (2006) “New Semantic Similarity Techniques of Concepts applied in the biomedical domain and WordNet.” MS Thesis, University o f Houston Clear Lake, Houston, TX USA, 2006.

[8] World Health Organization, “International statistical classification of diseases and related health problems”. - 10th revision, edition 2010.

[9] Hisham Al-Mubaid and Hoa A. Nguyen, “Using MEDLINE as Standard Corpus for Measuring Semantic Similarity in the Biomedical Domain”, Sixth IEEE Symposium on BionInformatics and BioEngineering (BIBE'06), 2006.

[10] Mirjana Ivanovic& Zoran Budimac, An overview of ontologies and data resources in medical domains, Expert Systems with Applications 41 (2014) 5158–5166.

[11] Montserrat Batet Sanromà, “ontology-based semantic clustering”, PhD Thesis, 2010.

[12] UMLSKS. Available: http://umlsks.nlm.nih.gov

[13] MeSH Browser. Available: http://www.nlm.nih.gov/mesh/MBrowser.html

Evaluating Semantic Similarity between Biomedical Concepts/Classes through Single Ontology

Abdelhakeem M. B. Abdelrahman
Sudan University of Science and Technology
Collage of Graduate Studies Khartoum, Sudan

Dr. Ahmad Kayed
Department of Computing and Information Technology
Sohar University, Sohar, Oman

Abstract Most of the existing semantic similarity measures that use ontology structure as their primary source can measure semantic similarity between concepts/classes using single ontology. The ontology-based semantic similarity techniques such as structure-based semantic similarity techniques (Path Length Measure, Wu and Palmer's Measure, and Leacock and Chodorow's measure), information content-based similarity techniques (Resnik's measure, Lin's measure), and biomedical domain ontology techniques (Al-Mubaid and Nguyen's measure (SimDist)) were evaluated relative to human experts' ratings, and compared on sets of concepts using the ICD-10 "V1.0" terminology within the UMLS. The experimental results validate the efficiency of the SemDist technique in single ontology, and demonstrate that SemDist semantic similarity techniques, compared with the existing techniques, gives the best overall results of correlation with experts' ratings.

Keywords: Biomedical information retrieval, biomedical ontology, semantic similarity measures, Unified Medical Language System (UMLS).

I. INTRODUCTION

Semantic similarity techniques are interested in measuring the semantic similarity, or inversely, semantic distance between two classes/concepts according to a given domain [8]. Semantic Similarity between two terms or sets of documents is defined as the degree of "sameness" between the terms as measured by comparing the information describing their properties [7]. Ontology-based semantic similarity measures are the similarity between two concepts, which is widely used in information retrieval and semantic web service fields [15]. They are can be roughly grouped into two groups as follows: 1) Ontology structure-based measures are those measures that use ontology taxonomy structure (is-a, part of) to calculate the similarity between concepts [5], [16]. In this measure the similarity between concepts is based on the path distance separating the concepts. These measures compute the similarity in terms of the shortest path between two concepts (classes) (group of synonyms) in the taxonomy. Rada et al, [5] proposed their measure as potential measure in the biomedical domain. Their experiments were conducted using MeSH (Medical Subject Headings) biomedical ontology. Wu and Palmer [16] proposed semantic similarity measure of concepts by taking into account the depth of concept nodes only. And 2) Information content-based similarity measures are those measures that use IC of concept derived from corpus statistics to measure the semantic similarity between concepts/classes. However, most of these semantic similarity measures can adapted to be use in biomedical domain. Hisham Al-Mubaid & Nguyen proposed new ontology-based semantic similarity measure that account for the depth of the concept nodes as well as distance (path length) between them. Another recent work on semantic similarity in biomedicine domain by Pedersen, Pakhomov and Patwardhan (2005) [7] in which they proposed a corpus-based context-vector approach to measure similarity between concepts in SNOMED-CT. Our contribution on this paper is compared between these semantic similarity techniques to choose the best measure among different similarity techniques that gives the best correlation and can be used to create our dataset or standard definition to be used to evaluate ontologies in the biomedical domain.

Most ontologies are developed for various purposes and domain [8]. For example, WordNet [8] is a lexical database for general English. In the biomedical domain, the Unified Medical Language System (UMLS) framework [8] includes many biomedical ontologies and terminologies (e.g., ICD-10, SNOMED-CT, MeSH, …etc).

II. BACKGROUND AND RELATED WORK

UMLS The Unified Medical Language System (UMLS) can be considered as an example of terminology which contains many clinical terms and integrates about 100 different vocabularies [1, 2]. It consists of three main knowledge sources: Meta thesaurus (MeSH, SNOMED-CT thesauruses, etc.), Semantic Network, and SPECIALIST Lexicon & Lexical Tools.

MeSH: MeSH, stands for Medical Subject Headings, [2, 3], is one of the source vocabularies used in UMLS. MeSH includes about 15 high-level categories, and each category is divided into subcategories and assigned a letter: A for Anatomy, B for Organisms and C for Diseases, and so on.

SNOMED-CT: SNOMED-CT, stands for Systemized Nomenclature of Medicine Clinical Term [2, 3], was included in UMLS in May 2004. It is a comprehensive clinical terminology, and the current version contains more than 360,000 concepts, 975,000 synonyms and 1,450,000 relationships organized into 18 hierarchies.

The following ontologies can be considered as known ontologies in the medical domain:

NCI Thesaurus (National Cancer Institute Thesaurus): an ontology vocabulary that includes broad coverage of the cancer domain, including cancer related disease, anatomy, genes and drugs.

ICD-10: [4], stand of International Classification Diseases version 10 is one of the most important international medical terminological systems; it was first issued in 1893. Its sixth revision was in 1948, and since this time it has been maintained by the World Health Organization (WHO). The current version is the tenth revision (ICD-10), which was issued in 1992. The initial aim of the ICD was to provide an international classification of death causes in order to produce internationally uniform and thus comparable mortality statistics. The WHO family of international classifications also includes other systems, notably the ICF (International Classification of Functioning, Disabilities and Health) and ICHI (International Classification of Health Inventions). The 22 main sub-categories of ICD-10 include, among others, diseases of the blood and blood-forming organs (D50–D89), endocrine, nutritional and metabolic diseases (E00–E90), mental and behavioral disorders (F00–F99), diseases of the nervous system (G00–G99) and certain infections and parasitic diseases (A00– B99). We present some preliminary observations about ICD-10 and consider the sub-domains I–XVII (codes A00 Q99). Core ontology of ICD-10 must explicate what sub-domains I–XVII address. Six of these domains are classified with respect to systems (nervous system, circulatory system, respiratory system, digestive system, musculo-skeletal system, genito-urinary system), three pertain to special organs (eye, ear, skin), and one domain relates to infectious diseases (A00–B99) and one domain addresses mental and behavioral disorders (F00–F99). Sub-domain level categories Level (i), i = I... XVII may be introduced; their instances are subsumed by the corresponding chapters. The instances of a level category level (i) in ICD-10 exhibit a taxonomic structure. Consider the domain of infections and parasitic diseases (A00–B99) and the associated domain-level category level (I), and includes about 21 high-level categories (taxonomies/sub trees) as shown in *Figure1*. The 2016 release of ICD-10 was used in our experiments.

For example as being described in figure1 below: In our experiment, the similarity is measured using different types of semantic similarity measures. From the evaluation result the best measure will be used in our benchmark dataset to evaluate ontologies in biomedical domain. Figure 1: below describes the biomedical domain type (ICD-10 ontology).

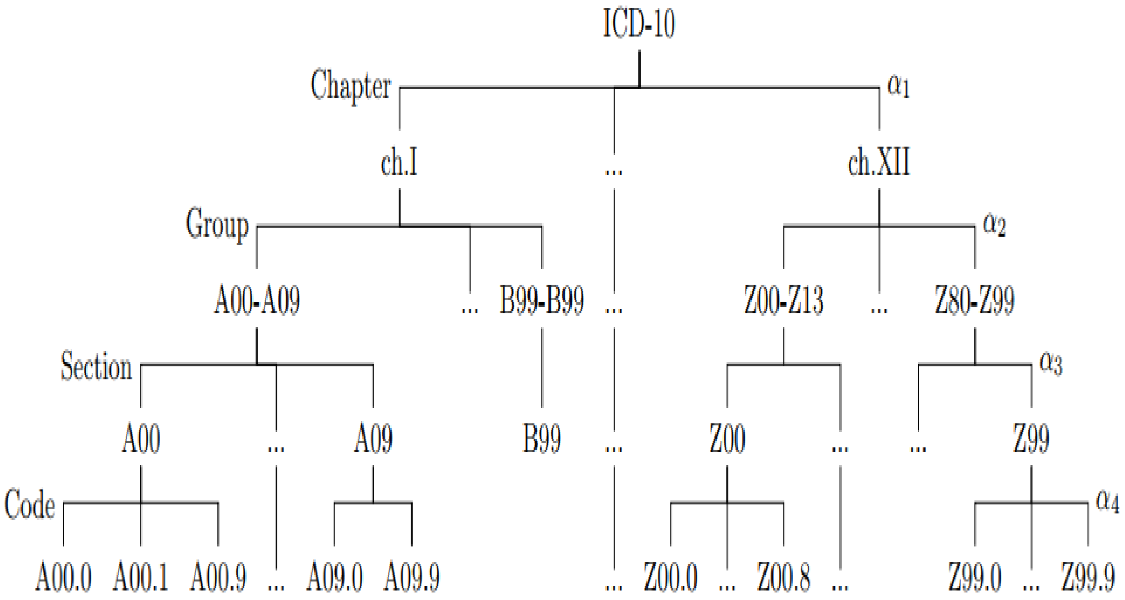


Figure 1: fragment of the ICD-10 taxonomy [17].

RELATED WORK

Rada et al. [5] [10] first proposed a semantic distance measure and applied it into the biomedical domain using MeSH ontology. The semantic distance between two classes is the shortest path length between them.

Caviedes and Cimino [6] [10] implemented the shortest Path length measure, called CDist, based on the shortest distance between two classes' nodes in the ontology. They evaluated their measure *(CDist measure)* on MeSH, SNOMED, ICD9 ontologies based on correlation with human ratings.

Pedersen et al. [7] [10] proposed semantic similarity and relatedness in the biomedicine domain in which they applied a corpus-based context vector approach to measure similarity between concepts in SNOMED-CT. Their context vector approach is ontology free but requires training text, for which, they used text data from Mayo Clinic corpus of medical notes.

Hisham Al-Mubaid & Nguyen [2] [8] proposed measure take the depth of their Least Common Subsume (LCS) and the distance of the shortest path between them. The higher similarity arises when the two concept are in the lower level of the hierarchy. Classes that are more similar with have a lower similarity score than classes that are less similar with this measure.

III. SEMANTIC SIMILARITY MEASURES

Semantic similarity techniques are becoming essential components of most of the information retrieval (IR), information extraction (IE), and other intelligent knowledge-based systems. For example, in IR, similarity measures play a crucial role in determining an optimal match between query terms and the retrieved document in ranking the results such as plagiarism detection [2]. The main semantic similarity measures could be classified into structure-based measures and information content (IC) measures.

1. Structure-based measures: In this measure the similarity between two concepts is based on the path distance separating the concepts. Which include the following types

1.1 Path Length Measure: finds the semantic distance between two concept nodes by finding the shortest path length between them on the ontology.

$$\text{Shortest Path}(C1, \quad C2) \quad = 2 * \text{Max}_{\text{depth}} - \text{len}(c1, c2) \qquad (1)$$

For example, to compute the similarity between "*Hypertensive renal disease with renal failure" (I12.0)* and "*Hypertensive renal disease with renal failure" (I12.0)* the shortest path length between them equal 1 "Using node counting"

Max_Depth of our Taxonomy = 5

So:

Sim (*Hypertensive renal disease with renal failure, Hypertensive renal disease with renal failure*) = 2*5 – 0 = 10 = 100%

.

Sim (*Pure hypercholesterolaemia, Lymph nodes of head, face and neck*) = 2*5 – 2 = 8= 20%

Table 1: Similarity values for two concepts from our taxonomy (Figure 1) using *Path Length* Based Measures (*shortest path*).

id	*Concept1*	*Concept2*	*LCA(c1 c2)*	*Length*	*Similarity*
4	*Hypertensive renal disease with renal failure(I12.0)*	*Hypertensive renal disease with renal failure (I12.0)*	*Hypertensive renal disease with renal failure (I12.0)*	0	100%
11	*Congestive heart failure* (I50.0)	*Left ventricular failure* (I50.1)	*Heart failure I50*	2	80%
. . 30	*Pure hypercholesterolaemia* (E78.0)	*Lymph nodes of head, face and neck (C77.0)*	*ICD10_Chapter*	8	. . 20%

Rada et al. [64] estimates the distance of two concepts C_1 & C_2 as the shortest-path linking them SP(C_1, C_2) and they used biomedical domain to evaluate their work in the information retrieval tasks using shortest path measure.

1.2 Wu and Palmer Measure: in this measure the similarity of concepts is compute by taking into account the depths of concept nodes only. They proposed a measure that has formula as follows:

$$\text{Sim}\ (\text{C1}, \text{C2}) = 2 * \text{depth}(\text{LCS}(\text{C1}, \text{C2}))\ /\ (\text{depth}(\text{C1}) + \text{depth}(\text{C2}))\ \ (2)$$

The score can never be 0 because the depth of the LCS is never 0 (the depth of the root is 1) So the score is 0<Score<=1. When the two classes are the same the score is 1.

$$\text{Sim}\ (\text{C1}, \text{C2}) = \frac{2\text{N}}{\text{N1+N2+2N}} \qquad (3)$$

Where N is the depth of the least common subsume (The least common subsume, LCS (C_1,C_2), of two concept nodes C_1 and C_2 is the lowest node that can be a parent for C_1 and C_2.

From our taxonomy (figure1), we can calculate the similarity between classes C1 and C2 as shown in table2:

Similarity (*Hypertensive renal disease with renal failure*, *Hypertensive renal disease with renal failure*) $= \frac{2*5}{0+0+(2*5)} = 1 = 100\%$

.

Similarity (*Pure hypercholesterolaemia*, *Lymph nodes of head, face and neck*) = $\frac{2*1}{4+4+(2*1)}$ = 0.2 = 20%

Table 2: Similarity values for two concepts from the ICD-10 taxonomy (Figure 1) using Path Length Based Measures (*Wu & Palmer*).

id	Concept1	Concept2	LCS(c1 c2)	Wu & Palmer	Similarity
4	*Hypertensive renal disease with renal failure(I12.0)*	*Hypertensive renal disease with renal failure (I12.0)*	*Hypertensive renal disease with renal failure*	1.00	100%
11	*Congestive heart failure* (I50.0)	*Left ventricular failure* (I50.1)	*Heart failure*	0.80	80%
. . 30	*Pure hypercholesterolaemia* (E78.0)	*Lymph nodes of head, face and neck (C77.0)*	*ICD10_Chapter*	0.20	20%

1.3 Leacock and Chodorow measure:

The similarity between two classes is determined by the shortest path length between two classes node, which connects these two classes in the taxonomy. The similarity is calculated as the negative algorithm of this value. They proposed a measure that has formula as follows:

$$\text{SimL\&C} = -\log\left[\frac{\text{Sp (c1, c2)}}{2(\text{Max\_depth})}\right] \qquad (4)$$

Similarity (Hypertensive renal disease with renal failure, Hypertensive renal disease with renal failure) $= -\log\left(\frac{1}{2(5)}\right) = 1.00$

.

Similarity (Congestive heart failure, Left ventricular failure) $= -\log\left(\frac{3}{2(5)}\right) \quad = 0.52287874528$

.

Similarity (Pure hypercholesterolaemia, Lymph nodes of head, face and neck) $= -\log\left(\frac{9}{2(5)}\right) \quad = 0.045757490560$

Table 3: Similarity values for two concepts from the ICD-10 taxonomy (Figure 1) using Path Leng1th Based Measures (Leacok and Chodorow).

ID	Concept1	Concept2	Length (c1 c2)	Leacok and Chodorow	Sim
4	*Hypertensive renal disease with renal failure(I12.0)*	*Hypertensive renal disease with renal failure (I12.0)*	1	1.00	100%
11	*Congestive heart failure* (I50.0)	*Left ventricular failure* (I50.1)	3	0.52287874528	52%%
. . 30	*Pure hypercholesterolaemia* (E78.0)	*Lymph nodes of head, face and neck (C77.0)*	9	0.0457574905607	. . 5%

2. Information Content-Based Similarity Measure:

2.1 Resnik's Measures

Resnik [9] the similarity between a pair of Classes (C_1 and C_2) is estimated as the amount of taxonomical information they share. In a taxonomy, this information is represented by the Least Common Subsume of both classes (LCS (C_1, C_2)), which is the most specific taxonomical ancestor common to C1 and C2 in a given ontology. Formally:

$$\mathrm{Simres} = -\log(\mathrm{P}\big(\mathrm{LCS}\ (\mathrm{C1}, \mathrm{C2})\big) = \mathrm{IC}\big(\mathrm{LCS}\ (\mathrm{C1}, \mathrm{C2})\big) \qquad (5)$$

$$\mathrm{IC(C)} = \frac{\log(\mathrm{Depth(C)})}{\log(\mathrm{Deep_{max}})} \qquad (6)$$

$$\mathrm{IC}\big(\mathrm{LCS}(\text{Hypertensive renal disease with renal failure,}$$
$$\text{Hypertensive renal disease with renal failure})\big)$$
$$= \mathrm{IC}(\text{Hypertensive renal disease with renal failure})$$

Depth (Hypertensive renal disease with renal failure) = 5 "using node counting"

Deepmax = 5 the maximum depth of ICD10 Ontology.

Then:

$$\mathrm{Simres} = \ \mathrm{IC}(\text{Hypertensive renal disease with renal failure}) = \frac{\log(\mathrm{depth(C)})}{\log(\mathrm{deep_{max}})}$$

$$= \log\frac{(5)}{\log(5)} = 1.00$$

.

$$\mathrm{IC}\big(\mathrm{LCS}(\text{Congestive heart failure,} \qquad \text{Left ventricular failure})\big) = \mathrm{IC}(\text{Heart failure})$$

Depth (Heart failure) = 4 “using node counting”

Deepmax = 5 the maximum depth of ICD10 Ontology.

Then:

$$\text{Simres} = \text{IC(Heart failure)} = \frac{\log(\text{depth(C)})}{\log(\text{deep}_{\text{max}})} = \log\frac{(4)}{\log(5)} = 0.86$$

.

$$\text{IC}\big(\text{LCS(Pure hypercholesterolaemia,} \quad \text{Lymph nodes of head, face and neck)}\big)$$

$$= \text{IC(ICD10\_Chapter)}$$

Depth (ICD10_Chapter) = 1 “using node counting”

Deep$_{\text{max}}$ = 5 the maximum depth of ICD10 Ontology.

Then:

$$\text{Simres} = \text{IC(ICD10\_Chapter)} = \frac{\log(\text{depth(C)})}{\log(\text{deep}_{\text{max}})} = \log\frac{(1)}{\log(5)} = 0.00$$

Table 4: Similarity values for two concepts from the ICD-10 taxonomy (Figure 1) using information content based Measures (Resink).

ID	Concept1	Concept2	LCS(c1 c2)	SimResink	Similarity
4	*Hypertensive renal disease with renal failure(I12.0)*	*Hypertensive renal disease with renal failure (I12.0)*	5	1.00	100%
11	*Congestive heart failure* (I50.0)	*Left ventricular failure* (I50.1)	4	0.86135311614	86%
. . 30	*Pure hypercholesterolaemia* (E78.0)	*Lymph nodes of head, face and neck (C77.0)*	*1*	0.00	0.00%

2.2 Lin’s Measure:

This measure depends on the relation between information content (IC) of the LCS of two classes and the sum of the information content of the individual concepts [9].

$$\text{SimLin (c1, c2)} = \frac{2\times\text{IC(LCS (C1,C2))}}{\text{IC(C1) + IC(C2)}} \qquad (7)$$

From Resink’s Measure:

$$\text{IC(LCS(}\mathit{Hypertensive\ renal\ disease\ with\ renal\ failure,}$$

$$\mathit{Hypertensive\ renal\ disease\ with\ renal\ failure}))$$

$$= \text{IC}(\mathit{Hypertensive\ renal\ disease\ with\ renal\ failure}) = \frac{\log(5)}{\log(5)} = 1.00$$

$$\mathrm{IC}(Hypertensive\ renal\ disease\ with\ renal\ failure) = \frac{\log(\mathrm{depth(C)})}{\log(\mathrm{Deep_{max}})}$$

$$= \log\frac{(5)}{\log(5)} = 1.00$$

Then:

$$\mathrm{SimLin}\ (Hypertensive\ renal\ disease\ with\ renal\ failure,$$

$$Hypertensive\ renal\ disease\ with\ renal\ failure) = \frac{2 \times 1}{1\ +\ 1} = 1.00$$

From Resink's Measure:

$$\mathrm{IC}(\mathrm{LCS}(Congestive\ heart\ failure, \quad Left\ ventricular\ failure))$$

$$= \mathrm{IC}(Heart\ failure) = \frac{\log(4)}{\log(5)} = 0.86$$

$$\mathrm{IC}(Congestive\ heart\ failure) = \frac{\log(\mathrm{depth(C)})}{\log(\mathrm{Deep_{max}})} = \log\frac{(5)}{\log(5)} = 1.00$$

$$\mathrm{IC}(Left\ ventricular\ failure) = \frac{\log(\mathrm{depth(C)})}{\log(\mathrm{Deep_{max}})} = \log\frac{(5)}{\log(5)} = 1.00$$

Then:

$$\mathrm{Sim}Lin(Congestive\ heart\ failure, \quad Left\ ventricular\ failure) = \frac{2 \times 0.86}{1\ +\ 1}$$

$$= 0.86$$

From Resink's Measure:

$$\mathrm{IC}(\mathrm{LCS}(Pure\ hypercholesterolaemia\ , \quad Lymph\ nodes\ of\ head, face\ and\ neck))$$

$$= \mathrm{IC}(\mathrm{ICD10\_Chapter}) = \frac{\log(0)}{\log(5)} = 0.00$$

$$\mathrm{IC}(Lymph\ nodes\ of\ head, face\ and\ neck) = \frac{\log(\mathrm{depth(C)})}{\log(\mathrm{Deep_{max}})} = \log\frac{(5)}{\log(5)} = 1.00$$

$$\mathrm{IC}(Pure\ hypercholesterolaemia) = \frac{\log(\mathrm{depth(C)})}{\log(\mathrm{Deep_{max}})} = \log\frac{(5)}{\log(5)} = 1.00$$

Then:

$$\mathrm{Sim}Lin(ure\ hypercholesterolaemia, \quad Lymph\ nodes\ of\ head, face\ and\ neck)$$

$$= \frac{2 \times 0.00}{1\ +\ 1} = 0.00$$

Table 5: Similarity values for two concepts from the ICD-10 taxonomy (Figure 1) using information content based measures (*Lin*).

ID	Concept1	Concept2	IC(c1)	IC(c2)	IC(LCS(c1,c2))	Sim*Lin*
4	*Hypertensive renal disease with renal failure(I12.0)*	*Hypertensive renal disease with renal failure (I12.0)*	1.00	1.00	1.00	100%
11	*Congestive heart failure* (I50.0)	*Left ventricular failure* (I50.1)	1.00	1.00	0.86	86%
. . 30	*Pure hypercholesterolaemia* (E78.0)	*Lymph nodes of head, face and neck (C77.0)*	1.00	1.00	0.0	0%

3. BIOMEDICAL DOMAIN SIMILARITY MEASURES

3.1 Hisham Al-Mubaid & Nguyen measure [2] [8] proposed measure take the depth of their Least Common Subsume (LCS) and the distance of the shortest path between them. The higher similarity arises when the two concept are in the lower level of the hierarchy. Classes that are more similar with have a lower similarity score than classes that are less similar with this measure. Their similarity measure is:

$$\text{SemDist (C1, C2)} = \log_2([L\ (C1, C2) - 1]^{\alpha}\times[\mathrm{CSpec}(C1, C2)]^{\beta} + k) \qquad (8)$$

$$\mathrm{CSpec}(C1, C2) = D - \text{depth}\left(\mathrm{LCS}(C1, C2)\right) \qquad (9)$$

Where:

$\alpha > 0$ and $\beta > 0$ are contribution factors of two features (Path and CSpec).

Depth (LCS(C1, C2)) is depth of LCS(C1, C2) using node counting.

L(C1, C2) is shortest path length between the two concept nodes.

D is maximum depth of the taxonomy.

K is constant, and CSpec feature is calculated as in (9). We use logarithm function (inverse of exponentiation) for semantic distance (8), which is the inverse of semantic similarity.

To insure the distance is positive and the combination is non-linear, k must be greater or equal to one (k >= l). In this paper, k=l is used in experiments. When two concept nodes have path length of 1 (Path=l) using node counting (i.e., they are in the same node in the ontology), they have a semantic distance (SemDist) equals to zero (i.e. maximum similarity) regardless of common specificity feature.

The maximum value of this measure occurs when one concept is the left-most leaf node, and the other concept is the right-most leaf node in the tree. In ICD10 terminology the maximum value is $\log_2$ ([22-1]*[5-1] + 2) equal 6.4262647547. Therefore, the similarity distance values will be in [1.0000, 6.4262647547] in ICD10 terminology.

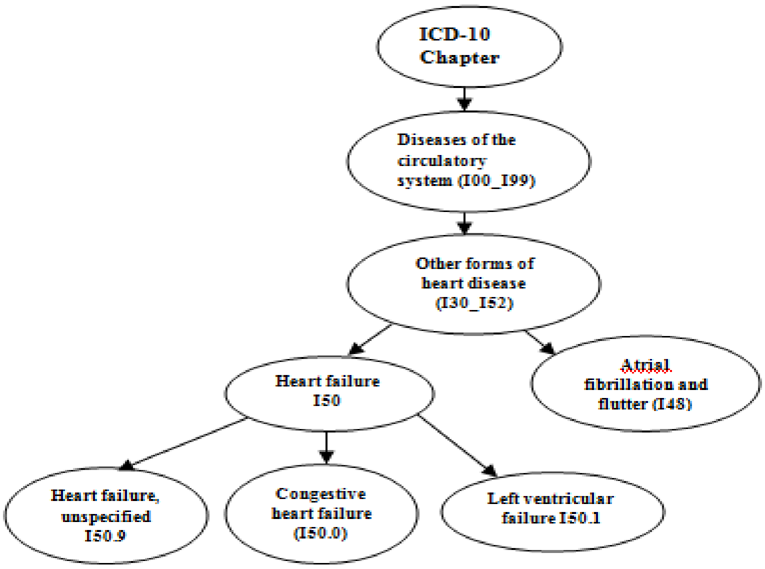

ICD-10 Chapter
Diseases of the circulatory system (I00_I99)
Other forms of heart disease (I30_I52)
Heart failure I50
Atrial fibrillation and flutter (I48)
Heart failure, unspecified I50.9
Congestive heart failure (I50.0)
Left ventricular failure I50.1

Figure 2. Hierarchy tree of eight concepts.

1) The single-cluster path length feature:

From our taxonomy (Figure 2), We can calculate the similarity between classes C1 and C2 as the following:

Path length (*Congestive heart failure, Left ventricular failure*) = 1 “using node counting”

CSpec (*Congestive heart failure, Left ventricular failure*) = D – depth (LCS (*Heart failure*))

= 5 – 4 = 1

So, similarity

Sim (*Congestive heart failure, Left ventricular failure*) = $\log_2([3 - 1]^1 \times [1]^1 + 2) = \log_2(4) = 2$

2) The cross-cluster path length feature:

Let us conceder the example, shown in Figure 3 below. The root is node that connects all the clusters. The path length between two concept nodes (C1 and C2) is computed by adding up the two shortest path lengths from the two nodes to their LCS node (their LCS is the root). For example, in Figure 1, for the two concept nodes *(Heart failure, unspecified, Atrial fibrillation and flutter*), the LCS is the root ICD-10. So, the path length between *Pure hypercholesterolaemia*, and *Lymph nodes of head, face and neck* is calculated as follows:

Path (*Pure hypercholesterolaemia, Lymph nodes of head, face and neck*) = d1 + d2 -1

Where d1 = d (*Pure hypercholesterolaemia*, root) and d2 = d (*Lymph nodes of head, face and neck*, root), where d (*Pure hypercholesterolaemia*, root) is the path length from the root ICD-10 to node *Pure hypercholesterolaemia*, and similarly d (*Lymph nodes of head, face and neck*, root) is the path length from ICD-10 to node *Lymph nodes of head, face and neck*. One is subtracted in the above equation, because the root node is counted twice.

$$\text{Path } (Pure\ hypercholesterolaemia, Lymph\ nodes\ of\ head, face\ and\ neck) = \text{d1} + \frac{2\text{D1} - 1}{2\text{D2} - 1} \times \text{ d2 – 1}$$

$$\text{Path}(Pure\ hypercholesterolaemia,\ \ Lymph\ nodes\ of\ head, face\ and\ neck)\ 5 + \frac{10-1}{10-1} \times\ 5 - 1\ \ = 9$$

$$\text{CSpec (Pure hypercholesterolaemia, Lymph nodes of head, face and neck) = D primary} \ \ - 1\ = \ 5 - 1\ = 4$$

So, similarity

$$\text{SemDist (Pure hypercholesterolaemia, Lymph nodes of head, face and neck)}$$

$$= \log_2(\ [\text{Path - 1}]^{\alpha} \times [\text{CSpec}]^{\beta} + \text{k}) \ \ = \text{Log}_2\ ((9 - 1)\ \times\ (4) + 2) = \log_2 (34) = 5.09$$

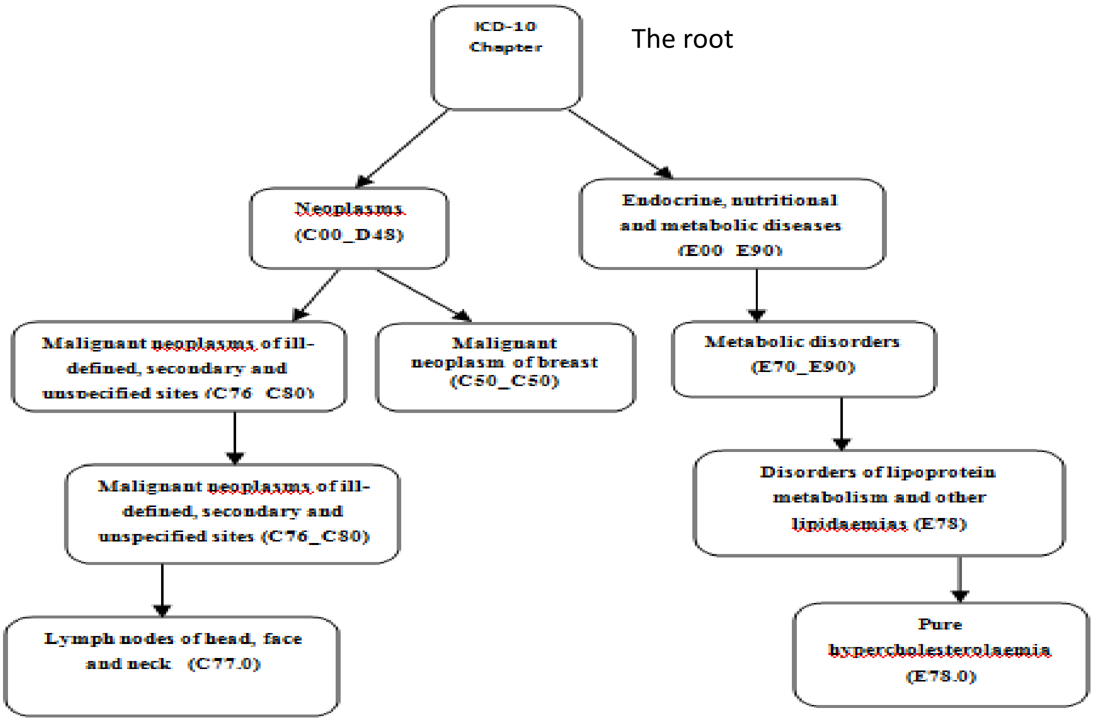

ICD-10 Chapter
The root
Neoplasms (C00_D48)
Endocrine, nutritional and metabolic diseases (E00 E90)
Malignant neoplasms of ill-defined, secondary and unspecified sites (C76 C80)
Malignant neoplasm of breast (C50_C50)
Metabolic disorders (E70_E90)
Malignant neoplasms of ill-defined, secondary and unspecified sites (C76_C80)
Disorders of lipoprotein metabolism and other lipidaemias (E78)
Lymph nodes of head, face and neck (C77.0)
Pure hypercholesterolaemia (E78.0)

Figure 3: Fragment of two clusters in ICD-10 Ontology (C77.0, E78.0).

Table 6: Similarity values for two classes from the ICD-10 taxonomy (Figure 1) using Path Length Based Measures (Al-Mubaid and Nguyen).

ID	Concept1	Concept2	L (c1,c2)	CSPec(c1, c2)	SimDist	Note
4	*Hypertensive renal disease with renal failure(I12.0)*	*Hypertensive renal disease with renal failure (I12.0)*	1	0	1	Same code
11	*Congestive heart failure* (I50.0)	*Left ventricular failure* (I50.1)	3	1	2	Same group
. . 30		*Lymph nodes of head, face and neck (C77.0)*	9	4	5.09	Different chapter

	Pure hypercholesterolaemia (E78.0)					

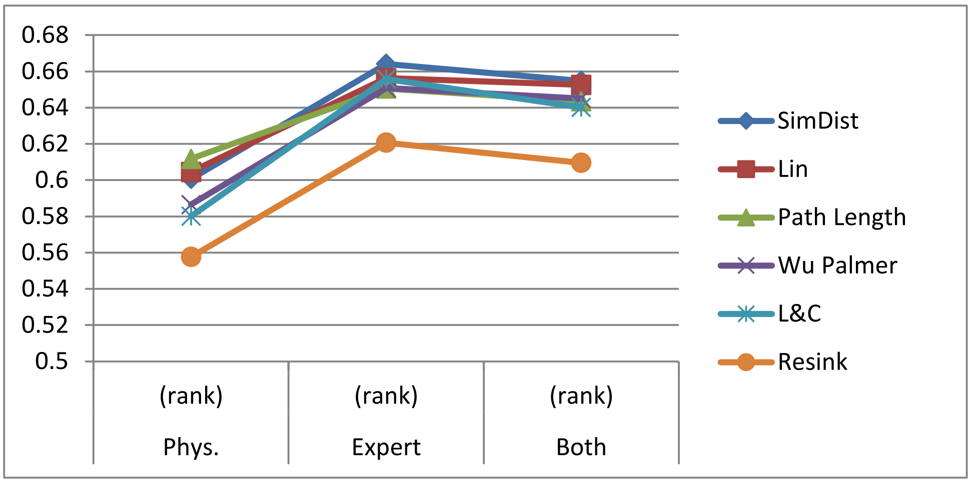
IV. EXPERMENTS AND RESULTS

For experiments, Ontologies of ICD10 were used as information source for the semantic similarity measure and one dataset are used for evaluation. All the measures use node counting for *path lengths* and *depths* of concept nodes. Out of the 30 pairs of Dataset 1 as shown in table 2, only 24 pairs in ICD10 were found. For the six pairs that were not found in ICD10 ontology, average distance/similarity values of the most related concept nodes to each one of them were calculated, so there were 24 pairs in ICD10 ontology in total. The results of absolute correlations with human scores using dataset1, experimented on ICD10 Ontology, are shown in Tables 7 and Figure 4. The experimental results demonstrated that *Al-Mubaid and Nguyen's measure (SimDist)* measure can achieve high correlations with human similarity scores.

Table 7: Absolute correlations with human scores for all measures using ICD10 on Dataset 1

Measure	**Phys. (rank)**	**Expert (rank)**	**Both (rank)**
SimDist	0.6007 (3)	**0.6641 (1)**	**0.6548 (1)**
Lin	0.6045 (2)	0.6563 (2)	0.6526 (2)
Path Length	**0.6118 (1)**	0.6505 (5)	0.6436 (4)
Wu Palmer	0.5865 (4)	0.6508 (4)	0.6451 (3)
L&C	0.5801 (5)	0.6558 (3)	0.6401 (5)
Resink	0.5576 (6)	0.6207 (6)	0.6096 (6)

Figure 4: Results of correlations with human scores for six measures using ICD10 Ontology.



V. EVALUATION:

Dataset:

There are no standard human rating sets of concepts/terms for semantic similarity in the biomedical domain. Thus, to evaluate the six semantic similarity measures, the dataset of 30 concept pairs from Pedersen et al. (2005) [7], (dataset1) which was annotated by 3 physicians and 9 medical index experts. Each pair was annotated on a 4-point scale: “practically synonymous, related, marginally related, and unrelated”.

Table 8 contains the whole pairs of this dataset. The average correlation between physicians is 0.68, and between experts is 0.78. Because the experts are more than the physicians, and the correlation (agreement) between experts (0.78) is higher than the correlation between physicians (0.68), it can be assumed that the experts’ rating scores are more reliable than the physicians’ rating scores.

Only 24 out of the 30 term pairs are found in ICD10 using ICD10 browser version 2010 [11] as some terms cannot be found, 24 pairs was used in the experiments (Pedersen et. al. [7] tested 29 out of the 30 concept pairs as one pair was not found in SNOMED-CT).

The term pairs in **bold**, in Table 8, are the ones that contains a term that was not found in ICD10 Ontology and they were excluded from experiments.

Table7 and figure4 show that the results of correlation with human ratings of physicians, experts, and both (phys. and experts), with the ranks between parentheses. These correlation values (Table7) show that the *SimDist* measure is ranked #1 in correlation relative to experts’ judgments and relative to both (expert and phys. judgments). But relative to physician judgments, the *SimDist* measure is ranked #3. From the applicability point of view, Nguyen and Al-Mubaid Measure (*SimDist*) is the most adequate one, and that can be used in our benchmark dataset. Finally the experiment describe above manually, should be obtained automatically. Hence, we need some software application or tools that can perform all the experiments automatically.

VI. CONCLUSION AND FUTURE WORK

The results discussed in this research has shown that, the SemDist (C1, C2) similarity (proposed by Al-Mubaid and Hoa A. Nguyen) has achieved high matching score by the expert's judgment to measure the similarity between two concepts in biomedical domain. In the future work of this research, we plan to implement a web-based system for all these semantic similarity measures and to make it available to researchers over the Internet.

Table 8 Dataset 1: 30 medical term pairs sorted in the order of the average.

Id	***Concept1***	***Concept2***	***Phys***	***Expert***	***Id***	***Concept1***	***Concept2***	***Phys***	***Expert***
4	Renal failure I12.0	Kidney failure I12.0	4.0000	4.0000	**27**	**Acne**	**Syringe**	**2.0000**	**1.0000**
5	Heart I51.5	Myocardium I51.5	3.3333	3.0000	12	Antibiotic (Z88.1)	Allergy (Z88.1)	1.6667	1.2222
1	Stroke I64	Infarct I64	3.0000	2.7778	**13**	**Cortisone**	**Total knee replacement**	**1.6667**	**1.0000**
7	Abortion O03	Miscarriage O03	3.0000	3.3333	**14**	**Pulmonary embolus**	**Myocardial infarction**	**1.6667**	**1.2222**
9	Delusion (F06.2)	Schizophrenia (F06.2)	3.0000	2.2222	16	Pulmonary Fibrosis (E84.0)	Lung Cancer (C34.1)	1.6667	1.4444
11	Congestive heart failure (I50.0)	Pulmonary edema (I50.1)	3.0000	1.4444	**6**	**Cholangiocarcinoma**	**Colonoscopy**	**1.3333**	**1.0000**
8	Metastasis (C77.0)	Adenocarcinoma (C08.9)	2.6667	1.7778	29	Lymphoid hyperplasia (K38.0)	Laryngeal Cancer (C32.0)	1.3333	1.0000
17	Calcification (M61)	Stenosis (H04.5)	2.6667	2.0000	21	Multiple Sclerosis (F06.8)	Psychosis (F06.8)	1.0000	1.0000
10	**Diarrhea**	**Stomach cramps**	**2.3333**	**1.3333**	22	Appendicitis (K35)	Osteoporosis (M80)	1.0000	1.0000
19	Mitral stenosis (I05.0)	Atrial fibrillation (I48)	2.3333	1.3333	23	Rectal polyp (K62.1)	Aorta (I70.0)	1.0000	1.0000
20	Chronic obstructive pulmonary disease (J44.9)	Lung infiltrates (J82)	2.0000	1.8889	24	Xerostomia (K11.7)	Alcoholic cirrhosis (K70.3)	1.0000	1.0000
2	Rheumatoid arthritis (M05.3)	Lupus (L93)	2.0000	1.1111	25	Peptic ulcer disease (K21.0)	Myopia (H52.1)	1.0000	1.0000
3	Brain tumor (G94.8)	Intracranial hemorrhage(I69.2)	2.0000	1.3333	26	Depression (F20.4)	Cellulitis (H60.1)	1.0000	1.0000
15	Carpal tunnel Syndrome (G56.0)	Osteoarthritis (M19.9)	2.0000	1.1111	**28**	**Varicose vein**	**Entire knee meniscus**	**1.0000**	**1.0000**
18	Diabetes mellitus (E10-E14)	Hypertension (I10-I15)	2.0000	1.0000	30	Hyperlipidemia (E78.0)	Metastasis (C77.0)	1.0000	1.0000

REFERENCES

[1] Hoa A. Nguyen, "New Semantic Similarity Techniques of Concepts Applied in the Biomedical Domain and WordNet" Master Thesis, Dec, 2006.

[2] Hisham Al-mubaid & Hoa A. Nguyen "Cluster-Based Approach for Semantic Similarity in the Biomedical Domain" Proceeding of the 28th IEEE, New York City, Aug 30-Sep 3, 2006.

[3] David Sánchez, "Semantic variance: An intuitive measure for ontology accuracy evaluation", Engineering Applications of Artificial Intelligence 39 (2015) 89–99

[4] World Health Organization., "ICD-10, International Statistical Classification of Diseases and Related Health Problems." 10th Revision. 5th ed. Vol.2 instruction manual (2016).

[5] Rada, et. al. "Development and Application of a Metric on Semantic Net". IEEE Transactions on Systems, Man and Cybernetics, 19,1(1989),17-30.

[6] Caviedes, J. and Cimino, J. "Towards the development of a conceptual distance metric for the UMLS". Journal of Biomedical Informatics 37,77-85, 2004.

[7] Pedersen,T. et al, "Measures of Semantic Similarity and Relatedness in the Medical Domain", University of Minnesota Digital Technology Center Research Report, Journal of Biomedical Informatics April 2006.

[8] Hisham Al-mubaid & Hoa A. Nguyen "Measuring Semantic Similarity between Biomedical concepts within multiple ontologies" IEEE Trans Syst Man Cybern Part C: Appl Rev 2009, 39.

[9] Abdelrahman, A.M.B. and Kayed, A. (2015) A Survey on Semantic Similarity Measures between Concepts in Health Domain. American Journal of Computational Mathematics, 5, 204-214.

[10] Althobaiti, A.F.S. (2017) Comparison of Ontology-Based Semantic-Similarity Measures in the Biomedical Text. Journal of Computer and Communications, 5, 17-27.

[11] http://apps.who.int/classifications/icd10/browse/2010/en.

[12] Hliaoutakis, "Semantic Similarity Measures in MeSH Ontology and their application to Information Retrieval on Medline". Master's thesis, Technical University of Crete, Greek. 2005.

[13] UMLSKS. Available: http://umlsks.nlm.nih.gov

[14] MeSH Browser. Available: http://www.nlm.nih.gov/mesh/MBrowser.html

[15] Kaifeng, et. al. "AN IMPROVED METHOD FOR MEASURING CONCEPT SEMANTIC SIMILARITY COMBINING MULTIPLE METRICS"Proceedings of IEEE IC-BNMT2013.

[16] Wu, Z., and Palmer, M. Verb "semantics and lexical selection" 133-138, 1994.

[17] Roxana Dogaru, et. al, "Searching for Taxonomy-based Similarity Measures for Medical Data"BCI September 2015.

Web Scraping for Estimating new Record from Source Site

Warna Agung Cahyono
Department of Electrical Engineering
University of Brawijaya
Malang, East Java, Indonesia

Wijono
Department of Electrical Engineering
University of Brawijaya
Malang, East Java, Indonesia

Herman Tolle
Department of Information System
University of Brawijaya
Malang, East Java, Indonesia

Abstract: Study in the Competitive field of Intelligent, and studies in the field of Web Scraping, have a symbiotic relationship mutualism. In the information age today, the website serves as a main source. The research focus is on how to get data from websites and how to slow down the intensity of the download. The problem that arises is the website sources are autonomous so that vulnerable changes the structure of the content at any time. The next problem is the system intrusion detection snort installed on the server to detect bot crawler. So the researchers propose the use of the methods of Mining Data Records and the method of Exponential Smoothing so that adaptive to changes in the structure of the content and do a browse or fetch automatically follow the pattern of the occurrences of the news. The results of the tests, with the threshold 0.3 for MDR and similarity threshold score 0.65 for STM, using recall and precision values produce f-measure average 92.6%. While the results of the tests of the exponential estimation smoothing using α = 0.5 produces MAE 18.2 datarecord duplicate. It slowed down to 3.6 datarecord from 21.8 datarecord results schedule download/fetch fix in an average time of occurrence news.

Keywords: Web Scrapping, Exponential Smoothing, MDR, STM

1. INTRODUCTION

The process for collecting any information about competitors from public spaces on the internet with actionable form short or long term strategy is the field of Competitive Intelligent [1]. Companies protect themselves from slander and erroneous rumor of news online. On the other hand the news sites, forums, social media, blogs, etc. can serve as a trigger for customer to customer communication or the customer to the company. And consequently enlarge the area of the market share[2]. Information retrieval techniques from this public site is an early stage that must exist in the system search engine as well as other web mining[3].

Any public sites are autonomous, so that changes to the structure of the HTML tags on the server can occur at any time without being noticed by users. So a method, which may classify datarecord [4] without duplicating by visual senses such as eyes but adaptive to changes in the structure of HTML is required. Method i-robot crawlers [5] picking data record and perform construction of sitemap source site, then do the traverse to the news pages [6]. However, i-robot does not detect more than two of the same news posted on different time. So it needed an additional method for identifying ID post.

From the view point of the server's news sources, if the server detects the intensity of the fetch of the client on the page the same source simultaneously, then it can be categorized under DDoS attack by the system IDS [7]. From the view point of two-way client-server and the problem of duplication of record news on adaptive crawling, the author proposes a method for predicting new posts a news source. so that the intensity of the download page is not often. Then it is not detected as the machine/robot crawlers by the server.

There are several methods of crawling the news with predicting a new news post. First, the novel architecture and algorithm for web page change detection [9], time complexity is small but does not detect changes in the structure of the tags so it is not adaptive and prone to duplication of records.

Second, method Carbon Dating The Web: Estimating the Age of Web Resources, a method where a change of news gathered from several approaches through the header response, RSS XML, backlinks and google's index, then the delta t is calculated from the change in its content [8]. Unfortunately on the server side scripting, date create is dynamic.

Third, A Novel Combine Forecasting Method for Predicting News Update Time, This method uses a combination of Exponential Smoothing and Naive Bayes. The root of the Naive Bayes is exponential smoothing level. The first leaf is a type of news category. Then the second leaf is the number of occurrences of a data set of training [10]. But this method only gets data from RSS.

Fourth, the approach of Mining Data Record [4], a method based on similarity of news segment visually and unsupervised adaptive to change the structure of the HTML source of the news. Although this is not a method to estimate the changes of new content. But the method chosen by the author by combining the exponential smoothing methods. And helpful to know changes some new news in one page. So the datatime series can be taken specifically from each group different news categories in a single page.

Testing on the efficiency of the system are done with the Mean Absolute Error. Testing data on the accuracy of news who successfully learned are done using Recall and Precision to assess the suitability of the record that is being drawn with the original records on the source website[11].

2. THE PROPOSED METHOD

Schedule the fetch/download, which uses exponential smoothing (ES)[12], used to slow the intensity download/fetch followed the appearance of new news site estimate source of its purpose. In Figure 1 below, the system is divided into 4 sub methods namely preparations, MDR, temporary classification, archiving data records, and estimator using ES method.

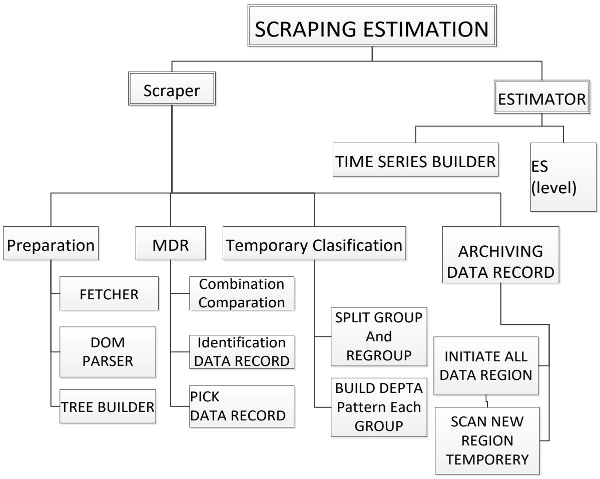


Figure 1. Taxonomy Method Scraping with Estimated New News

2.1 Preparing

Agent response, generated by the download/fetch in the HTML form, in the format parsed into the Document Object Model (DOM) [14]. MDR method uses the DOM data structure as a model for the purposes of comparisons between similar subtree using the edit distance levenstein[16]. To speed up the process of comparison/combination on the MDR method, all nodes except the type element must be removed from the DOM-tree, and each node is modified to contain an array of the results of the combination/comparison for each of the k nodes. In the method of MDR, k is the number of combinations tagtree in a generalize node. In this research the maximum number of combinations is 10 sub tag tree.

2.2 Mining Data Record

Every edit distance results from combinations and comparisons begin n = 1 to k are stored into the data array in each tagnode. At the stage of identification a dataregion, every tagnode checked if the value generated from a combination of levenshtein comparisons under the threshold.

On stage pick datarecord, every child from dataregion sliced into array object datarecord. Each datarecord is formed from one or several combined tagnode also has sub tagtree. At the time of the formation of datarecord also note the download time is saved in each object datarecord.

2.3 Temporary Classification

The entire datarecord in every new dataregion, at this stage, will be divided if between datarecord has similarities Simple Tree Matching (STM) is less than the value of the threshold score[18]. ReGroup algorithm can be seen in Figure 3. The value of the similarity score obtained from a similar number of nodes of the STM results on two tagtree (every tagtree comes from datarecord to compare). To normalize the amount of score values, tagnode in two tagtree calculated then total tagnode divided in half, resulting in an average number of tagnode. A normalized score is the number of matching results STM tagnode divided average number of tagnode. In this research, we uses the value score 0.65.

Although in a dataregion, if there are two datarecord level similarities under the score threshold then it should be split into two different group. And each group has a pattern in the form of tagtree. Tagtree pattern is obtained by aligning [13] the entire tagtree of every member from datarecord group.

2.4 Temporary Classification

After a group has members all datarecord tagtree similar to each other. So any new datarecord in each group is stored into a dataregion which has existed in the form of local files. Before it is stored into a local file datarecord then must be checked against all the href attribute of each anchor tag <a>in datarecord. Researchers use the mongo database to store the array into a single document hrefs. Every document in the database save file location and array contains attribute datarecord hrefs from datarecord. If the combination of hrefs datarecord recently matched with one of the document database, then the datarecord need not be saved. But note download time remains stored in the database for the purposes of the evaluation of the new datarecord post.

Then a combination of hrefs, in each of the daterecord, will be an ID (identity) for every datarecord. Thus, the duplication will not occur when looping download/fetch on the same attribute datarecord hrefs at different time.

The pattern comes from the dataregion which has existed in the local data and the new pattern of a similar group will be carried out using the method of partial tree aligning depta. So the new patterns, aligning results, in the form of tagtree, stored into the pattern in each dataregion.

If there are two equal record posted on different time, all hrefs in datarecord looping will occur. If the entire hrefs on two record match then it can be considered as duplicate records, so it does not need to be saved as archive files. If the entire hrefs don't match the database queries, then save the new datarecord into local files, note also the url, number dataregion, and download time.

2.5 Estimator

This estimation intends to calculate how many new news datarecord that appear during the span of delta T. Thus, emergence of new news time can be estimated. With more, the intensity of the process of repeating browsing on the same content can be slowed.

Forecasting method used is the method of exponential smooting (ES) with parameter α [12]. Input comes from data download time per record. First, calculate how long the average time it takes the appearance of a new datarecord from certain dataregion from one site news source. P is the length of the dataregion. P is calculated from the sum of all datarecord every dataregion which can be displayed on a webpage. Delta T for dataseri derived from the average time the emergence of all datarecord as many as P. for every interval series, count how many datarecord in each interval of time. The format of the input data in the form of a series of new record how many downloaded on any delta t. Then dataseries modelled in the exponential smooting. The SES model used is a single/level [15] :

$$S_t = \alpha X_t + (1-\alpha)S_{t-1} \qquad (1)$$

$$\hat{X}_t(m) = S_t \qquad (2)$$

3. ESTIMATION OF DATA EXTRACTION

Referring to figure 1, this method is generally broken down into two major parts, namely web data capture and record appearance estimator news. Web data retrieval starting from fetcher, MDR, a temporary classification until datarecord archiving. While the estimated new news starting from the initiation of length P datarecord and average download time (ΔT) every dataregion, then formation of dataseries, then estimate the amount of n records that appear in the next interval. If ΔT is the average download time is full of P one data region. Then, the difference of time Δti is posting the datarecord i subtracted to the time posting datarecord i-1. A provision on the interval to the i as follows:

$$\Delta ti = \frac{T}{n_i} \qquad (3)$$

$$next\ fetch = lastfetch + \Delta ti * (P - 2) \qquad (4)$$

nextfetch is the estimated time of next fetching, which type of data is datetime. While lastfetch is the latest time fetch/download. i is interval, which is the time t exist in there. P is the length (maximum amount) of datarecord which is able to display at web page at certain dataregion.

3.1 Data Used

Sampling data web is the latest news from datarecord dataregion which has the pattern combination of generalized-node 1n [4]. Estimator is the sample data dataseries formed from records stored per dataregion from each site. With the initiation of shown in table 1 below.

Table 1. Characteristics of latest news datarecord at 27 Juni 2018 12:48 until 28 Juni 2018 04:57 UTC

URL	NoRegion/ Length P/ amount Dataregion	average ΔT (minute)
http://www.cnnindonesia.com/	1/6/6	52.6
http://www.tribunnews.com/	6/49/11	84.2
http://www.metrotvnews.com/?	14/26/21	92.3
http://www.merdeka.com	3/40/6	170.6
http://www.tempo.co	5/7/6	23.7
http://www.kompasiana.com	4/95/9	568
http://www.harianjogja.com	5/20/5	152.2
http://www.dream.co.id	6/7/7	565.5
http://www.antaranews.com	4/33/11	98.98
http://www.detik.com	5/38/6	62.7
http://www.rmol.co	4/39/8	244.2
http://www.inilah.com	8/28/14	185.4

3.2 Data Flow System Diagram

Plot this data describes the workings of the system, starting from setting URL parameter, T, scoreT, and α. URL parameters used as target fetch web scraping system. T is the levenshtein edit distance threshold on MDR. While scoreT is a normalized similarity score results of STM[18]. The parameter scoreT is used on a temporary classification and archiving, for comparisons of datarecord in the form of tagtree.

In Figure 4 below, the parameter α is the parameter for the level (St) in the exponential smoothing. Estimation of the next Download is shown in equation (3) and (4). Whereas temporary classification and archiving datarecord implemented in dataRecord archiving algorithms, SaveCheckDuplicate, and Regroup in Figure 2, 4 and 5. As shown below:

```
SaveCheckDuplicate(NoRg, DRCi, DRGLs)

FileLocation=queryDB(DRCi.hrefs)
if FileLocation is not exist
 NewFileLocation=SaveScrap(DRCi,
 DRCi.hrefs)
 SaveDB(NoRg, NewFileLocation, DRGLs)
End If
```

Figure. 2 Algorithm : Duplication check before hrefs are saved.

In Figure 2, NoRg is index number from the dataregion. Each dataregion which have been at the local file is always given the ID number in the region. QueryDB is query on mongo database to check whether the combination of hrefs in datarecord has existed with the return value in the form of file storage location datarecord. DRCi datarecord is no. i have just extracted. DRGLs is all the dataregion which has been in the local. SimpanDB is a function/method stores the address of the file locations on the database, on the DRCi dataregion number to NoRg from DataRegion DRGLs.

```
ReGroup(DRGi)

DRG2s=init collection dataregion
while(DRGi ≠ 0)
 DRC=pop up datarecord from DRGi
 for(i=1; i <= DRG2s.size; i++)
  if STM(DRG2s[i].pattern,DRC > =
  scoreT) then exit loop for;
  end if
 end for
 $T_s$=DRC
 if i <= DRG2s.size then
  $T_s$=PartialTreeAlign(DRG2s[i].pola, DRC)
  If DRC unable aligned then
   i= DRG2s.size+1
  End if
 End if
 DRG2s[i].pattern=$T_s$
 DRG2s[i].add(DRC)
End while
return DRG2s
```

Figure. 3 ReGroup Algorithm for temporary classification .

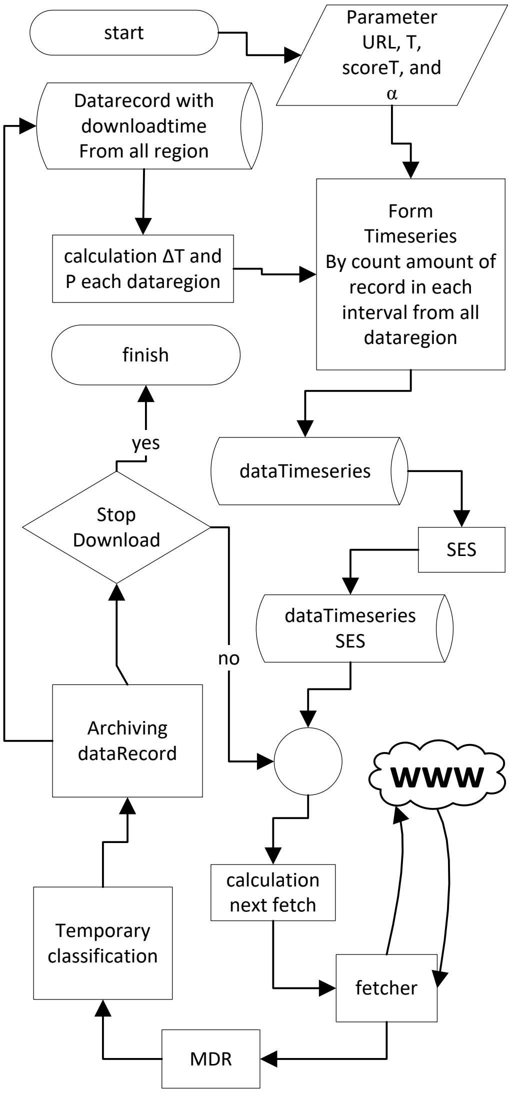

start
Parameter URL, T, scoreT, and α
Datarecord with downloadtime From all region
Form Timeseries By count amount of record in each interval from all dataregion
calculation ΔT and P each dataregion
finish
yes
dataTimeseries
Stop Download
SES
dataTimeseries SES
no
Archiving dataRecord
www
calculation next fetch
Temporary classification
fetcher
MDR

Figure. 4 Diagram of a system of Scraping with the fetch being estimated.

```
ArchivingDataRecord(DRGs,DRGLs)

1.  while(DRGs ≠ 0)
2.     DRGi=pop up dataregion from DRGs
3.     DRG2s=ReGroup(DRGi)
4.     while(DRG2s ≠ 0)
5.       DRG2i=popup dataregion from DRG2s
6.       polaDRGi=take from DRG2i.patern
7.       for(i=1;i<=DRGLs.size;i++)
8.          if STM(polaDRGi,DRGLs[i].patern) > =
       scoreT
9.               then    NoRg=i
10.      End For
11.      if NoRg not yet set then NoRg= DRGLs.size+1
12.        while(DRG2i  ≠ 0)
13.          DRCi=popup datarecord from DRG2i
14.          SaveCheckDuplicate(NoRg, DRCi, DRLGs)
15.        End while
16.    End while
17. End while
```

Fig. 5 Algorithm of Data Record Archiving.

In Figure 3, DRGi is a temporary Dataregion with the number i. DRGi generated by MDR after fetch/download just happened. On line 1 Figure 3, DRG2s is an empty dataregion. DRC on line 3 is temporary datarecord originating from dataregion DRGi. Line 4 will be skipped if DRG2s still empty. On line 5, compare the pattern of the dataregion tagtree belongs to each DRG2s with datarecord tagtree DRC.

In figure 5, polaDRGi at line 6 is tagtree, tagtree pattern which is belong to each the dataregion i's DRG2. tagtree Patterns on everytime dataregion used for pattern matching to determine the suitability of the dataregion on certain dataregion. STM on line 8 is the SimpleTreeMatching algorithm[18].

3.3 Accuracy Experiment

This research uses 12 URL test. Testing at this stage of this web data capture must be done before the stage of forecasting. This is to tell how valid the data used in web data retrieval system, including this estimation. Testing is done using web data recall and precision[11]. In Figure 4, there is a section of the MDR, temporary classification, archiving datarecord which respectively using threshold editdistance not more than 0.3 for levenshtein and threshold score similarity ternormalisasi not less than 0.65 for STM. While the selected dataregion is a dataregion which latest news is shown.

Table 2. Testing : TP(True Positive), Recall, precision, dan f-measure for 12 news sites at 2 August 2017 12:10 until 7 Agustus 2017 05:16 UTC

Sites	TP	Recall (%)	f-measure
www.inilah.com	28	87.5	93.3%
www.cnnindonesia.com	18	85.7	92.3%
www.tribunnews.com	48	92.3	96%
www.metrotvnews.com	26	96.3	98.1%
www.merdeka.com	29	96.7	98.3%
www.dream.co.id	13	76,5	86.7%
www.tempo.co	31	100	100%
www.kompasiana.com	100	100	100%
www.antaranews.com	50	98	99%
www.harianjogja.com	17	95.2	97.6%
www.detik.com	36	94.7	97.3%
www.rmol.co	29	87.9	93.5%
Average		92.6	98.8

The precision column is not shown in table 2 due to all urls have value 100%. 12 sites in table 2 have a value of 100% precision so there are no data received incorrect/false. The average recall values was 92.6%, meaning that there is a still unread datarecord in small amounts, usually the last datarecord in the dataregion. But the data still can be used for datatime series because no data was wrong. In addition, the latest data on the web view will shift. So in the end, will still be readable datarecord by scraping system.

The next testing phase is to test the efficiency. The second test is done using MAE for measuring efficiency of download traffic. Efficiency is done by comparing the value of MAE download fix and value estimation-based download MAE. Download the fix, with time, done in the time span ΔT – (P-2) * Δti, Δti value obtained from equation (3). While the range of download time that ter-estimation is calculated based on the difference in time nextfetch subtracted by lastftech in units of minutes. The estimation is done with a value of α 0.5 can be seen in the following table.

Table 3. The average test results of MAE on schedule scraping, miss and duplication datarecord refers to the fix average history.

Sites	MAE Duplication
www.inilah.com	11.96
www.cnnindonesia.com	5.88
www.tribunnews.com	37.1
www.metrotvnews.com	12.7
www.merdeka.com	14.4
www.dream.co.id	6.9
www.tempo.co	6.8
www.kompasiana.com	79.2
www.antaranews.com	16.5
www.harianjogja.com	9.9
www.detik.com	37.1
www.rmol.co	23.8
Average	21.8

At the second trial was performed on the parameter α 0.5 which means balanced between the influence of the actual data and the average history results estimation. Testing with parameters α = 0.5 on 12 URL address is performed on every dataregion, since each URL address has more than one data region. The results of testing with α = 0.5 can be seen in the following table:

Table 4. The average test results of MAE on schedule the scraping miss and duplication datarecord refers to ES estimation.

Sites	MAE Duplication
www.inilah.com	9.5
www.cnnindonesia.com	5.99
www.tribunnews.com	12.1
www.metrotvnews.com	7.8
www.merdeka.com	7.3
www.dream.co.id	6.96
www.tempo.co	6.5
www.kompasiana.com	89.1
www.antaranews.com	10.95
www.harianjogja.com	8.3
www.detik.com	37.3
www.rmol.co	16.5
Average	18.2

Test estimation based on fix average time dataregion and test-based estimation of exponential smooting with parameters α = 0.5 can lower MAE of 3.6.

4. CONCLUSION

In this paper, an approach web scraping with estimated download time is used to avoid the intensity of downloads are too often. Test data shows that there is no negative data that is read by the system. Test of scraping has done using levenshtein edit distance threshold parameters of 0.3 and similarity threshold parameter STM of 0.65, which shows the f-measure 98.8% with 100% precision. So none datarecord which will not be stored in the database. The next test on the efficiency of the system with the SES 0.5 alpha parameters required. Test results indicate MAE decrease download time of 3.6. It showed a reduced number of duplicate datarecord who found a result too often download/fetch. In other words schedule fetch became slower, so the vacuum of download time can reduce the network traffic load. And reduce the impact of the system IDS from the server site news.

5. FUTURE WORK

Minus 2 on formula 4 is used to match the records of the database. then the next job is how to get rid of minus 2. So we need a method to recognize the paging url of the datarecord in the dataregion.

6. REFERENCES

[1] W. Y. C. Thompson S.H. Teo, "Accessing The Impact Of Using The Internet For Competitive Intelligence," Information And Management, 2001.

[2] D. J. F. W. Glynn Mangold, "Social media: The new hybrid element of the promotion mix," Business Horizons, pp. 257-365, 2009.

[3] S. P. Yugandhara Patil, "Review of Web Crawlers with Specification and Working," IJARCCE, pp. 220-223, 2016.

[4] R. G. Y. Z. Bing Liu, "Mining Data Records in Web Pages," SIGKDD, 2003.

[5] J.-M. Y. W. L. Y. W. L. Z. Rui Cai, "iRobot: An Intelligent Crawler for Web Forums," WWW 2008, 2008.

[6] B. S. P. P. Namrata H.S Bamrah, "Web Forum Crawling Techniques," International Journal of Computer Applications, vol. 8, pp. 36-41, 2014.

[7] A. A. N. V. Dusan Stevanovic, "Feature evaluation for web crawler detection with data mining techniques," Expert Systems with Applications, 2012.

[8] N. Salah Eldeen, "Carbon Dating The Web: Estimating the Age of Web Resources," International World Wide Web Conference Committee (IW3C2), pp. 1075-1082, 2013.

[9] D. K. S. Deepak Kumar Ganeshiya, "A novel architecture and algorithm for web page change detection," International Advance Computing Conference (IACC), 2013.

[10] X. Z. W. Z. Y. W. Memeng Wang, "A Novel Combine Forecasting Method for Predicting News Update Time," Fourth International Symposium on Information Science and Engineering, pp. 227-231 , 2012.

[11] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness &

Correlation," Journal of Machine Learning Technologies, p. 37–63, 2011.

[12] T. M. J. A. Cooray, Applied Time Series: Analysis and Forecasting, Oxford: Alpha Science International Limited, 2008.

[13] Y. Zhai and B. Liu, "Web data extraction based on partial tree alignment," Chiba, Japan, 2005.

[14] Philippe Le Hégaret; Ray Whitmer; Lauren Wood, "Document Object Model (DOM)," 19 January 2005. [Online]. Available: https://www.w3.org/DOM.

[15] Gardner and E. S., "Exponential smoothing: The state of the art—Part II," ELSEVIER, International Journal of Forecasting, p. 637–666, 2006.

[16] V. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," Soviet Physics Doklady, p. 707–710, 1966.

[17] A. Papana, "Short-Term Time Series Prediction For A logistics Outsourcing Company," in Outsourcing Management for Supply Chain Operations and Logistics Service, Thessaloniki, Bussines Science Reference, 2013, pp. 150-160.

[18] W. Yang, "Identifying syntactic differences between two programs," Software—Practice & Experience, pp. 739 - 755, 1991.

Analysis of Comparison of Fuzzy Knn, C4.5 Algorithm, and Naïve Bayes Classification Method for Diabetes Mellitus Diagnosis

Putri Elfa Mas`udia
Departement of Electrical Engineering
State Polytechnic of Malang
Malang, Indonesia

Ridwan Rismanto
Departement of Technology Information
State Polytechnic of Malang
Malang, Indonesia

Abdullah Mas`ud
Departement of Electrical Engineering
State Polytechnic of Malang
Malang, Indonesia

Abstract: Early detection of diabetes mellitus (DM) can prevent or inhibit complication. There are several laboratory test that must be done to detect DM. The result of this laboratory test then converted into data training. Data training used in this study generated from UCI Pima Database with 6 attributes that were used to classify positive or negative diabetes. There are various classification methods that are commonly used, and in this study three of them were compared, which were fuzzy KNN, C4.5 algorithm and Naïve Bayes Classifier (NBC) with one identical case. The objective of this study was to create software to classify DM using tested methods and compared the three methods based on accuracy, precision, and recall. The results showed that the best method was Fuzzy KNN with average and maximum accuracy reached 96% and 98%, respectively. In second place, NBC method had respective average and maximum accuracy of 87.5% and 90%. Lastly, C4.5 algorithm had average and maximum accuracy of 79.5% and 86%, respectively.

Keywords : Fuzzy KNN, C4.5 Algorithm, Naïve Bayes Classifier, Diabetes Mellitus

1. INTRODUCTION

Diabetes mellitus (DM) is a disease marked by high level of blood sugar caused by impaired insulin secretion, insulin disruption, or both.DM is a heterogeneous group marked by increase on glucose level in the blood or hyperglycemia [1].

There are various classification methods, such as K-nearest neighbor (KNN), fuzzy KNN (F-KNN), decision treemethod using C4.5 algorithm, Naïve Bayes classifier (NBC) method, and many other methods. In previous studies, one of this methods was used to classify a problem without analyzing which classification method produce the best result. Yanita Selly conducted a study in 2013 to compare KNN and F-KNN methods. The result showed that F-KNN method is better than KNN method, as accuracy of F-KNN reached 98% while KNN only had 96% accuracy [12].

The result was then further analyzed in this study, where F-KNN, decision tree method using C4.5 algorithm, and Naïve Bayes classifier (NBC) method were compared. The results of these three methods were analyzed to obtain the best classification method.

A. *Research Objective*

1. To apply fuzzy KNN method, decision tree method using C4.5 algorithm, and Naïve Bayes classifier (NBC) methodin diagnosing DM.
2. To create a software to compare the three methods based on accuracy, time, precision, and recall.

B. *Related Results from Previous Studies*

Yanita Selly dkk compared DM classification using K-nearest neighbor (KNN) and fuzzyKNN methods. KNN is a classification method that perform strict prediction on tested data based on k nearest neighbor. Meanwhile, F-KNN predicts tested data based on membership value of tested data in each class, and then class data with highest membership was selected as resulting predicted class. The study results showed that F-KNN method is better than KNN method, as accuracy of F-KNN reached 98% while KNN only had 96% accuracy [12].

Other study conducted by Parida Purnana regarding detection of Type II DM using Naïve Bayesbased on particle swarm optimization. In the study, particle swarm optimizationwas used to improve accuracy in detecting DM. The study result showed that this method had 98.16% accuracy and 0.99 AUC, thus it can be classified as ‘excellent classification’ [9].

Larissa dkk conducted study regarding classification of client using C4.5 algorithm as creditingbasis. This study classify clients of a bank, so that when a problem occurs, the bank could easily obtain rules from the resulting decision tree. With decision tree method using C4.5 algorithm, process of gathering information was faster and more optimal with larger number of data, therefore the error in decision making could be minimized [4].

2. SYSTEM PLANNING

The steps of this research were:

1. Studying literatures regarding fuzzy KNN, C4.5 algorithm, and Naïve Bayes classifier methods.
2. Studying dataset from Indian Pima Diabetes that were used as trainingdata
3. Designing software to perform classification in accordance with tested methods.
4. Applying fuzzy KNN, C4.5 algorithm, and Naïve Bayes classifier methodsto diagnose DM.

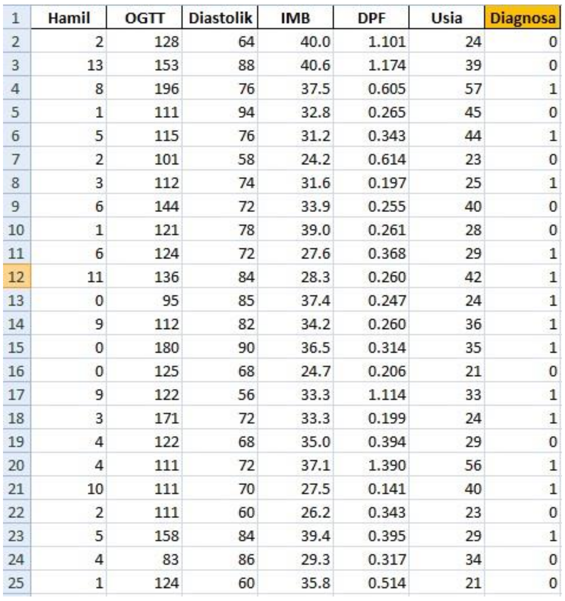
5. Testing and analyzing the results of each method and calculating the accuracy.

2.1 Data Preprocessing Method

This study used dataset that were then classified as training data and testingdata. These data were obtained from UCI machine learning repository database: Indian Pima Diabetes in http://archieve.ics.uci.edu. There are 768 clinical data in Indian Pima database but not all attributes are completely available.

768 clinical data obtained from Indian Pima Database were preprocessed, which means that insignificant data was deleted to maximize classification result. Missing valueor data with incomplete attributes was treated using rules from [LES-12], since the classification result is highly influential to training data.

From 8 parameter of data, parameters of TSFT and INS were deleted since the missing value was very large. The preprocessed data are displayed in Figure 1.



1	Hamil	OGTT	Diastolik	IMB	DPF	Usia	Diagnosa
2	2	128	64	40.0	1.101	24	0
3	13	153	88	40.6	1.174	39	0
4	8	196	76	37.5	0.605	57	1
5	1	111	94	32.8	0.265	45	0
6	5	115	76	31.2	0.343	44	1
7	2	101	58	24.2	0.614	23	0
8	3	112	74	31.6	0.197	25	1
9	6	144	72	33.9	0.255	40	0
10	1	121	78	39.0	0.261	28	0
11	6	124	72	27.6	0.368	29	1
12	11	136	84	28.3	0.260	42	1
13	0	95	85	37.4	0.247	24	1
14	9	112	82	34.2	0.260	36	1
15	0	180	90	36.5	0.314	35	1
16	0	125	68	24.7	0.206	21	0
17	9	122	56	33.3	1.114	33	1
18	3	171	72	33.3	0.199	24	1
19	4	122	68	35.0	0.394	29	0
20	4	111	72	37.1	1.390	56	1
21	10	111	70	27.5	0.141	40	1
22	2	111	60	26.2	0.343	23	0
23	5	158	84	39.4	0.395	29	1
24	4	83	86	29.3	0.317	34	0
25	1	124	60	35.8	0.514	21	0

Figure 1 Preprocessed Data

Used training data had six parameters, which were hamil, ogtt, diastolik, IMB, DPF and Usia. These parameters were used in classification process. The value of each parameter was used to determine diabetes diagnosis, where value of ‘1’ means positive diabetes and ‘0’ means negative diabetes

2.2 System Description

This study compared three classification methods, which were fuzzy KNN, C4.5 algorithm, and Naïve Bayes classifier. Classified data were generated from Indian Pimadatabase. This study designed a software for classification process and the results were used to determine the best method. Process of the study is displayed in Figure 2.

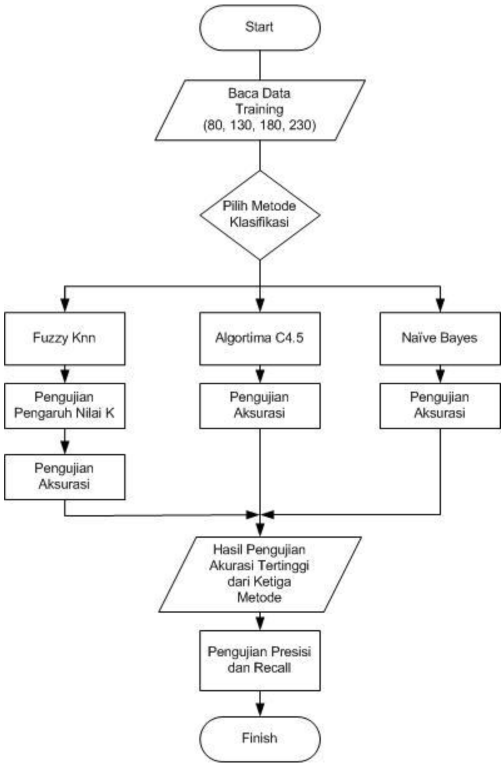

Start
Baca Data Training (80, 130, 180, 230)
Pilih Metode Klasifikasi
Fuzzy Knn
Algortima C4.5
Naïve Bayes
Pengujian Pengaruh Nilai K
Pengujian Aksurasi
Pengujian Aksurasi
Pengujian Aksurasi
Hasil Pengujian Akurasi Tertinggi dari Ketiga Metode
Pengujian Presisi dan Recall
Finish

Figure 2. Flowchart of the Study

2.3 Classification using Fuzzy KNN Method

In general, classification using fuzzy KNN method was conducted following these steps:

1. Input normalized training data.
2. Determine k value as initial parameter.
3. Determine weight exponent (m), this study used m = 2.
4. Calculate distance between new record data and each record training data using Euclidian distance.
5. Calculate membership value of each class, class with the highest membership value then used to determine new target.
6. Output was the result of the class with the highest membership value.

2.4 Classification using Fuzzy KNN Method

In Anyanwu journal, Podgorelec explains that C4.5 algorithm is a development of ID3 algorithm that is used in generating decision tree. C4.5 algorithmis not limited to binary number and is able to generate decision tree with multiple variables. Attributes in C4.5 algorithm generate one branch for every attribute branch in default [7]. Steps of classification process using decision tree C4.5 in general can be expressed as flowchart that is displayed in Figure 3.

Flowchart of decision tree C4.5 algorithm can be explained as:

1. Training data was required for classification process
2. Since data hamil, Ogtt, Diastolik, IMB, DPF and Usia were numerical data, early classification was done to minimize branch for further selection
3. Training data was required for classification process

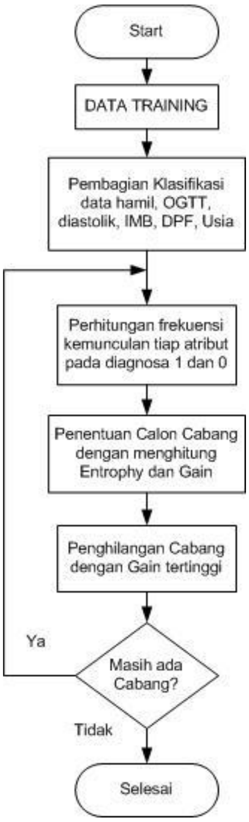

Start
DATA TRAINING
Pembagian Klasifikasi data hamil, OGTT, diastolik, IMB, DPF, Usia
Perhitungan frekuensi kemunculan tiap atribut pada diagnosa 1 dan 0
Penentuan Calon Cabang dengan menghitung Entrophy dan Gain
Penghilangan Cabang dengan Gain tertinggi
Ya
Masih ada Cabang?
Tidak
Selesai

Figure 3. Flowchart of C4.5 Algorithm

4. Frequency of occurrence of each data in positive and negative diabetes diagnosis was calculated.
5. Branch was determined by calculating entropy and gain according to aforementioned formula.
6. If initial branch/root had been determined, then the second branch was determined by removing parameter of the obtained branch. This process was repeated until there was no branch candidate.
7. If there was not any branch candidate, then the process was finished and decision tree had been generated.

2.4 Classification Using Naïve Bayes

Han (2006) explains that NBC uses Bayesian algorithm to calculate total probability. In NBC, probability of one word will be classified as one category(posterior probability),and it is based on the highest previous probability (prior probability).Naïve Bayes works by calculating the number of occurrence of specific attribute in particular category.

In Naïve Bayes with non-numerical data, probability of occurrence of specific category can be directly calculated, then it is multiplied with every attribute. However, with numerical data, this cannot be done as the data is continuous. For numerical data, the probability is calculated using Gaussian equation. Flowchart of Naïve Bayes is shown in Figure 4.

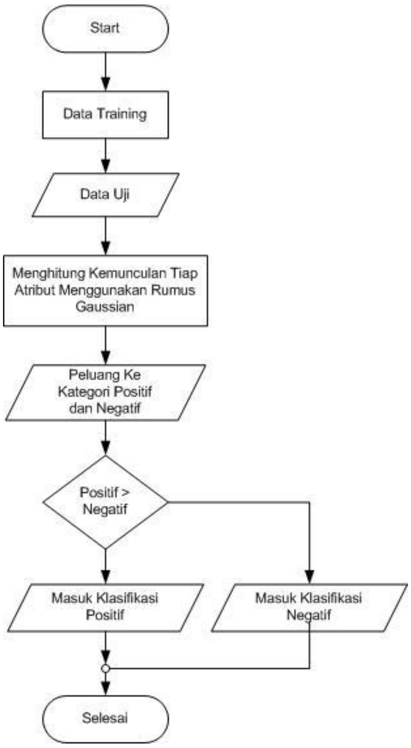
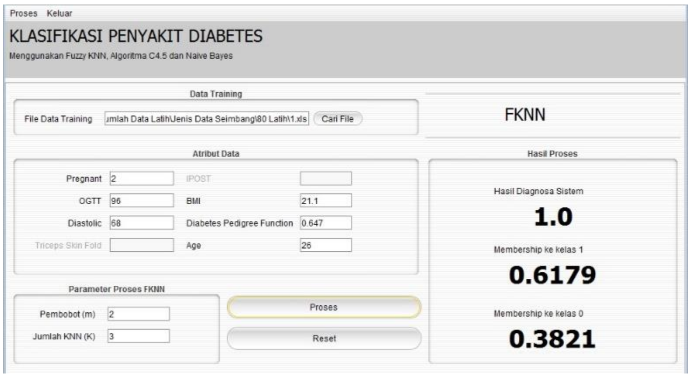

Start
Data Training
Data Uji
Menghitung Kemunculan Tiap Atribut Menggunakan Rumus Gaussian
Peluang Ke Kategori Positif dan Negatif
Positif > Negatif
Masuk Klasifikasi Positif
Masuk Klasifikasi Negatif
Selesai

Figure 4. Flowchart of Naïve Bayes Classifier

3. RESULT AND DISCUSSION

3.1 Classification Data Using Fuzzy Knn

When fuzzy KNN method was selected as feature process, then the diagnosis could be conducted in individual or collected data. Figure 5 shows classification using individual testing data.



KLASIFIKASI PENYAKIT DIABETES
Menggunakan Fuzzy KNN, Algoritma C4.5 dan Naive Bayes

Data Training
File Data Training: ımlah Data Latih\Jenis Data Seimbang\80 Latih\1.xls

FKNN

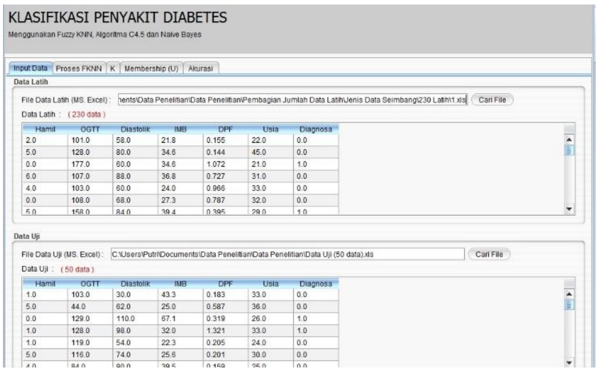
Atribut Data
Pregnant: 2
IPOST:
OGTT: 96
BMI: 21.1
Diastolic: 68
Diabetes Pedigree Function: 0.647
Triceps Skin Fold:
Age: 26

Parameter Proses FKNN
Pembobot (m): 2
Jumlah KNN (K): 3

Hasil Proses
Hasil Diagnosa Sistem
1.0
Membership ke kelas 1
0.6179
Membership ke kelas 0
0.3821

Figure 5.Display of Fuzzy KNN Classification using Individual Testing Data

There were several features in the classification using testing data collection: 1) input data, which was for entering the training data and testing data collection; 2) process FKNN, which was for entering the k and m parameter values and for clarification. Display of the clarification of the testing data collection is shown in Figure 6.



KLASIFIKASI PENYAKIT DIABETES

Menggunakan Fuzzy KNN, Algoritma C4.5 dan Naive Bayes

Data Latih

File Data Latih (MS. Excel) : nents\Data Penelitian\Data Penelitian\Pembagian Jumlah Data Latih\Jenis Data Seimbang\230 Latih\1.xls

Data Latih : (230 data)

Hamil	OGTT	Diastolik	IMB	DPF	Usia	Diagnosa
2.0	101.0	58.0	21.8	0.155	22.0	0.0
5.0	128.0	80.0	34.6	0.144	45.0	0.0
0.0	177.0	60.0	34.6	1.072	21.0	1.0
6.0	107.0	88.0	36.8	0.727	31.0	0.0
4.0	103.0	60.0	24.0	0.966	33.0	0.0
0.0	108.0	68.0	27.3	0.787	32.0	0.0
5.0	158.0	84.0	39.4	0.395	29.0	1.0

Data Uji

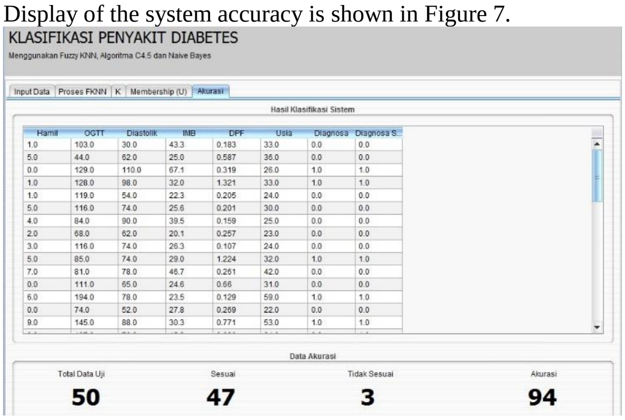
File Data Uji (MS. Excel) : C:\Users\Putri\Documents\Data Penelitian\Data Penelitian\Data Uji (50 data).xls

Data Uji : (50 data)

Hamil	OGTT	Diastolik	IMB	DPF	Usia	Diagnosa
1.0	103.0	30.0	43.3	0.183	33.0	0.0
5.0	44.0	62.0	25.0	0.587	36.0	0.0
0.0	129.0	110.0	67.1	0.319	26.0	1.0
1.0	128.0	98.0	32.0	1.321	33.0	1.0
1.0	119.0	54.0	22.3	0.205	24.0	0.0
5.0	116.0	74.0	25.6	0.201	30.0	0.0
4.0	84.0	90.0	39.5	0.159	25.0	0.0

Figure 6 Display of the clarification of the testing data collection Fuzzy KNN

Feature ‘k’ was used to observed results in accordance with the number of k and feature ‘membership’ was used to observe the membership value, while feature ‘accuracy’ was used to calculate the system accuracy whether the results fit the previous theories. Display of the system accuracy is shown in Figure 7.



KLASIFIKASI PENYAKIT DIABETES

Menggunakan Fuzzy KNN, Algoritma C4.5 dan Naive Bayes

Hasil Klasifikasi Sistem

Hamil	OGTT	Diastolik	IMB	DPF	Usia	Diagnosa	Diagnosa S...
1.0	103.0	30.0	43.3	0.183	33.0	0.0	0.0
5.0	44.0	62.0	25.0	0.587	36.0	0.0	0.0
0.0	129.0	110.0	67.1	0.319	26.0	1.0	1.0
1.0	128.0	98.0	32.0	1.321	33.0	1.0	1.0
1.0	119.0	54.0	22.3	0.205	24.0	0.0	0.0
5.0	116.0	74.0	25.6	0.201	30.0	0.0	0.0
4.0	84.0	90.0	39.5	0.159	25.0	0.0	0.0
2.0	68.0	62.0	20.1	0.257	23.0	0.0	0.0
3.0	116.0	74.0	26.3	0.107	24.0	0.0	0.0
5.0	85.0	74.0	29.0	1.224	32.0	1.0	1.0
7.0	81.0	78.0	46.7	0.261	42.0	0.0	0.0
0.0	111.0	65.0	24.6	0.66	31.0	0.0	0.0
6.0	194.0	78.0	23.5	0.129	59.0	1.0	1.0
0.0	74.0	52.0	27.8	0.269	22.0	0.0	0.0
9.0	145.0	88.0	30.3	0.771	53.0	1.0	1.0

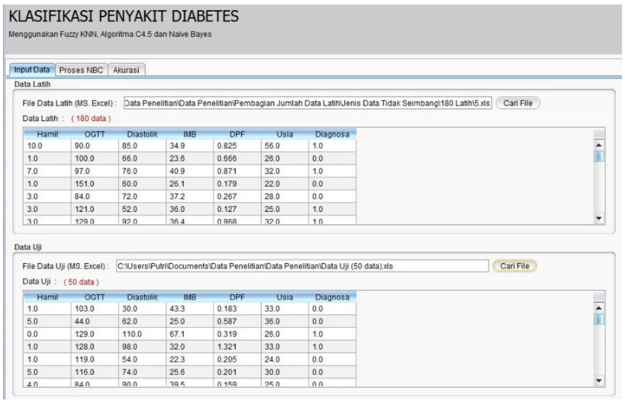
Data Akurasi

Total Data Uji	Sesuai	Tidak Sesuai	Akurasi
50	47	3	94

Figure 7 Display of Feature ‘System Accuracy FKNN’

3.2 Classification Data Using Naïve Bayes

There were several features in the classification using testing data collection: 1) input data, which was for entering the training data and testing data collection; 2) process NBC, which was for clarification process. Display of the clarification of the testing data collection is shown in Figure 8.



KLASIFIKASI PENYAKIT DIABETES

Menggunakan Fuzzy KNN, Algoritma C4.5 dan Naive Bayes

Data Latih

File Data Latih (MS. Excel) : Data Penelitian\Data Penelitian\Pembagian Jumlah Data Latih\Jenis Data Tidak Seimbang\180 Latih\5.xls

Data Latih : (180 data)

Hamil	OGTT	Diastolik	IMB	DPF	Usia	Diagnosa
10.0	90.0	85.0	34.9	0.825	56.0	1.0
1.0	100.0	66.0	23.6	0.666	26.0	0.0
7.0	97.0	76.0	40.9	0.871	32.0	1.0
1.0	151.0	60.0	26.1	0.179	22.0	0.0
3.0	84.0	72.0	37.2	0.267	28.0	0.0
3.0	121.0	52.0	36.0	0.127	25.0	1.0
3.0	129.0	92.0	36.4	0.968	32.0	1.0

Data Uji

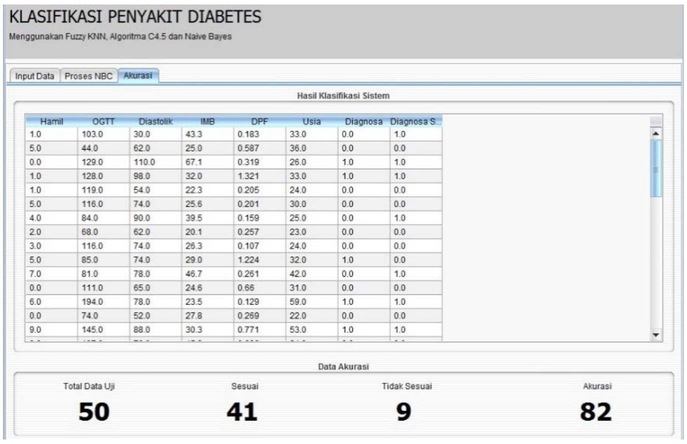
File Data Uji (MS. Excel) : C:\Users\Putri\Documents\Data Penelitian\Data Penelitian\Data Uji (50 data).xls

Data Uji : (50 data)

Hamil	OGTT	Diastolik	IMB	DPF	Usia	Diagnosa
1.0	103.0	30.0	43.3	0.183	33.0	0.0
5.0	44.0	62.0	25.0	0.587	36.0	0.0
0.0	129.0	110.0	67.1	0.319	26.0	1.0
1.0	128.0	98.0	32.0	1.321	33.0	1.0
1.0	119.0	54.0	22.3	0.205	24.0	0.0
5.0	116.0	74.0	25.6	0.201	30.0	0.0
4.0	84.0	90.0	39.5	0.159	25.0	0.0

Figure 9 Input Data in Naïve Bayes

Feature ‘Proses NBC’ was used for clarification, while the feature ‘accuracy’ was used to calculate the level of system accuracy. The display of the system accuracy is shown in Figure 10.



KLASIFIKASI PENYAKIT DIABETES

Menggunakan Fuzzy KNN, Algoritma C4.5 dan Naive Bayes

Hasil Klasifikasi Sistem

Hamil	OGTT	Diastolik	IMB	DPF	Usia	Diagnosa	Diagnosa S...
1.0	103.0	30.0	43.3	0.183	33.0	0.0	1.0
5.0	44.0	62.0	25.0	0.587	36.0	0.0	0.0
0.0	129.0	110.0	67.1	0.319	26.0	1.0	1.0
1.0	128.0	98.0	32.0	1.321	33.0	1.0	1.0
1.0	119.0	54.0	22.3	0.205	24.0	0.0	0.0
5.0	116.0	74.0	25.6	0.201	30.0	0.0	0.0
4.0	84.0	90.0	39.5	0.159	25.0	0.0	1.0
2.0	68.0	62.0	20.1	0.257	23.0	0.0	0.0
3.0	116.0	74.0	26.3	0.107	24.0	0.0	0.0
5.0	85.0	74.0	29.0	1.224	32.0	1.0	0.0
7.0	81.0	78.0	46.7	0.261	42.0	0.0	1.0
0.0	111.0	65.0	24.6	0.66	31.0	0.0	0.0
6.0	194.0	78.0	23.5	0.129	59.0	1.0	1.0
0.0	74.0	52.0	27.8	0.269	22.0	0.0	0.0
9.0	145.0	88.0	30.3	0.771	53.0	1.0	1.0

Data Akurasi

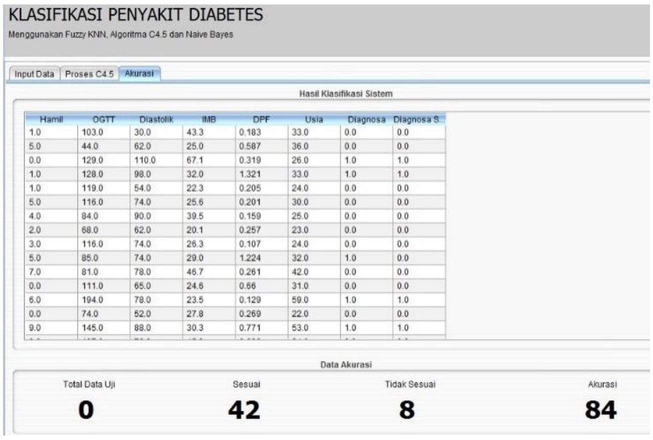
Total Data Uji	Sesuai	Tidak Sesuai	Akurasi
50	41	9	82

Figure 10 Display of the System Accuracy of Naïve Bayes

3.3 Classification Data Using Decision Tree C4.5

There were several features in the classification using testing data collection: 1) input data, which was for entering the training data and testing data collection; 2) process C4.5, which was for clarification process. Display of the clarification of the testing data collection is shown in Figure 10.

Feature ‘accuracy’ was used to observe the level of accuracy of the algorithm C4.5 classification by the system and the results were compared with the previous theories. The display of the results of the accuracy of the algorithm C4.5 is shown in Figure 11.



KLASIFIKASI PENYAKIT DIABETES

Menggunakan Fuzzy KNN, Algoritma C4.5 dan Naive Bayes

Hasil Klasifikasi Sistem

Hamil	OGTT	Diastolik	IMB	DPF	Usia	Diagnosa	Diagnosa S...
1.0	103.0	30.0	43.3	0.183	33.0	0.0	0.0
5.0	44.0	62.0	25.0	0.587	36.0	0.0	0.0
0.0	129.0	110.0	67.1	0.319	26.0	1.0	1.0
1.0	128.0	98.0	32.0	1.321	33.0	1.0	1.0
1.0	119.0	54.0	22.3	0.205	24.0	0.0	0.0
5.0	116.0	74.0	25.6	0.201	30.0	0.0	0.0
4.0	84.0	90.0	39.5	0.159	25.0	0.0	0.0
2.0	68.0	62.0	20.1	0.257	23.0	0.0	0.0
3.0	116.0	74.0	26.3	0.107	24.0	0.0	0.0
5.0	85.0	74.0	29.0	1.224	32.0	1.0	0.0
7.0	81.0	78.0	46.7	0.261	42.0	0.0	0.0
0.0	111.0	65.0	24.6	0.66	31.0	0.0	0.0
6.0	194.0	78.0	23.5	0.129	59.0	1.0	1.0
0.0	74.0	52.0	27.8	0.269	22.0	0.0	0.0
9.0	145.0	88.0	30.3	0.771	53.0	1.0	1.0

Data Akurasi

Total Data Uji	Sesuai	Tidak Sesuai	Akurasi
0	42	8	84

Figure 11 Results of the Accuracy of the Algorithm C4.5

3.4 System Testing Method

The methods for system testing were:

1. Training Data Testing, the test was done with equal amount of the testing data, which was 50, but with various training data: 80, 120, 160, 200 data training.
2. Results of the weight exponent (m) in Fuzzy KNN. This was because m determined how much the distance weight between each neighbor to the membership value.

3. Duration test for the clarification process between fuzzy KNN, C4.5 algorithm and Naïve Bayes classifier
4. Accuracy test between fuzzy KNN, C4.5 algorithmandNaïve Bayes classifier. This test used accuracy formula.
5. Precision test amongfuzzy KNN, C4.5 algorithm and Naïve Bayes classifier
6. Recall test fuzzy KNN, C4.5 algorithm and Naïve Bayes classifier

3.5 Testing and Analysis Results in FKNN

The results of the test using fuzzy KNN method on the balanced and unbalanced training data are displayed in Table 1 and Table 2.

Table 1. Result of Fuzzy KNN Test On Balance Training Data

K	System Accuracy (%)				
	80 training data	130 training data	180 training data	230 training data	Average
2	70	84	86	88	82
4	80	86	92	94	88
6	92	92	92	94	92.5
8	90	90	94	94	92
10	90	96	96	96	94.5
12	92	96	98	98	96

Table 2. Result of Fuzzy KNN Test On Unbalance Training Data

K	System Accuracy (%)				
	80 training data	130 training data	180 training data	230 training data	Average
2	76	82	82	86	81.5
4	84	86	86	94	87.5
6	84	86	86	94	87.5
8	84	94	96	96	92.5
10	88	94	94	96	93
12	92	96	98	98	96

The results shows that the more training data, the higher system accuracy. This means that as the number of training data increases, the number of record with distance near the predicted data class also increases, which in turn improves the accuracy.

The test results of balanced training data show that accuracy tended to increase, except for k=8 where it slightly decreased. Meanwhile, for 180 and 230 training data, all system accuracy increased from k=2 to k=12.

Test results of unbalanced training data show that for 80, 130, and 230 training data, all system accuracy increased from k=2 to k=12. Meanwhile for 180 training data, system accuracy slightly decreased on k=10.

Test results of both balanced and unbalanced training data show that the number of training data is directly proportional to system accuracy. The slight decrease in several tests was insignificant and system accuracy was tended to be stable.

3.6 Testing and Analysis Results in C4.5

This test was aimed to observe which type of training data generated the best results. Each training data was tested five times to obtain the best results.

The test results of balanced training data shows that the best data was obtained from the second experiment with respective average and maximum accuracy of 78% and 84% for 130 training data. Meanwhile, the test results of unbalanced training data shows that the best data was obtained from the third experiment with respective average and maximum accuracy of 79.5% and 86% for 80 training data. All test results are displayed in Table 3

Table 3a. Test Result of C4.5 Algorithm Test on Balance Training Data

Training Data	System Accuracy (%)					
	Balanced Training Data					
	1	2	3	4	5	Avg
80	70	80	80	88	80	79.6
130	72	84	78	78	80	78.4
180	74	74	74	68	62	70.4
230	74	74	74	74	74	74
Avg	72.5	78	76.5	77	74	**75.6**

Table 3b. Test Result of C4.5 Algorithm Test on Unbalance Training Data

Training Data	System Accuracy (%)					
	Unbalanced Training Data					
	1	2	3	4	5	Avg
80	80	86	86	80	82	82.8
130	82	84	84	80	80	82
180	62	62	74	74	74	69.2
230	74	74	74	74	74	74
Avg	74.5	76.5	79.5	77	77.5	**77**

Each training data was tested five times to observe the results change, then average of system accuracy was taken. The results showed that the number of training data is inversely proportional to system accuracy. This probably caused by the increasing number of training data makes it more difficult to generate decision tree.

The test of this study was done on two type training data. Balanced training data means that the diagnosis was evenly distributed on positive and negative results. Meanwhile, in unbalanced training data, the diagnosis was random, which means that there was no record of the number of positive and negative results. In this test, average of system accuracy from two types of training data was calculated. The average results are shown in Table 4.

Table 4. Average Accuracy of Training Data Type

Number of Training Data	Balanced Training Data (%)	Unbalanced TrainingData (%)
80	79.6	82.8
130	78.4	82
180	70.4	69.2
230	74	74
Average	**75.6**	**77**

The results show that unbalanced training data had better accuracy with 77% compared with balanced training data (75.6%).

3. 7 Testing and Analysis Results in NBC

This test was aimed to observe which type of training data generated the best results. Each training data was tested five times to obtain the best results. The results can be seen at Table 5.

Table 5a. Test Result of Classification Using Naïve Bayes

Training Data	System Accuracy (%)					
	Balanced Training Data					
	1	2	3	4	5	Average
80	90	80	86	84	80	84
130	86	88	88	84	82	85.6
180	86	86	90	82	86	86
230	86	86	86	86	86	86
Average	87	85	87.5	84	83.5	**85.4**

Table 5b. Test Result of Classification Using Naïve Bayes

Training Data	System Accuracy (%)					
	Unbalanced Training Data					
	1	2	3	4	5	Average
80	84	86	80	86	86	84.4
130	84	84	84	88	86	85.7
180	86	86	86	90	84	86.4
230	86	86	86	86	86	86
Average	85	85.5	84	87.5	85.5	**85.5**

The test results of balanced training data shows that the best data was obtained from the third experiment with respective average and maximum accuracy of 87.5% and 90% for 180 training data. Meanwhile, the test results of unbalanced training data shows that the best data was obtained from the fourth experiment with respective average and maximum accuracy of 87.5% and 90% for 180 training data. The results show that the number of training data is directly proportional to system accuracy.

The test of this study was done on two type training data. Balanced training data means that the diagnosis was evenly distributed on positive and negative results. Meanwhile, in unbalanced training data, the diagnosis was random, which means that there was no record of the number of positive and negative results. In this test, average of system accuracy from two types of training data was calculated. The average results are shown in Table 6.

Table 6. Average Accuracy of Training Data Type

Number of Training Data	Balanced Training Data (%)	Unbalanced Training Data (%)
80	84	84.4
130	85.6	85.7
180	86	86.4
230	86	86
Average	**85.4**	**85.5**

As can be seen from Table 6, there was no significant difference between both data types with the accuracy difference only 0.01%, with unbalanced training data had slightly higher accuracy than balanced training data.

3. 8 Testing and Analysis Results in Naïve Bayes

In the previous tests, accuracy of fuzzy KNN, C4.5 algorithm, and Naïve Bayes had been tested in detail. From the test result, the best accuracy of each method was compared with other methods without considering training data type. The comparison result of accuracy of all methods is displayed in Table 7.

Table 7. Comparison of Accuracy of The Three Methods

Training Data	Accuracy (%)		
	Fuzzy KNN	C4.5 Algorithm	Naïve Bayes
80	92	86	86
130	96	84	88
180	98	74	90
230	98	74	86
Average	**96**	**79.5**	**87.5**

The result shows that fuzzy KNN method had the highest accuracy with 96%.

4. CONCLUSSION

From the results, it can be concluded that:

1. The system was able to classify DM diagnosis using fuzzy KNN, C4.5 algorithm, or Naïve Bayes with average accuracy of all methods was 88.5%.
2. In classification using fuzzy KNN method, the highest accuracy was obtained in 180 training data and k=12, with accuracy of 98% and average accuracy of all training data of 96%. The results show that the more training data, the higher system accuracy. This means that as the number of training data increases, the number of record with distance near the predicted data class also increases, which in turn improves the accuracy.
3. In classification using C4.5 algorithm, the highest accuracy was obtained in 80 training data, with accuracy of 86% and average accuracy of all training data of 79.5%.
4. In classification using Naïve Bayes method, the highest accuracy was obtained in 180 training data, with accuracy of 90% and average accuracy of all training data of 87.5%. The results show that the more training data, the higher system accuracy.
5. Based on the results of accuracy test, the best classification method was fuzzy KNN with average accuracy of 96%, followed by Naïve Bayes method with 87.5%, and lastly C4.5 algorithm with average accuracy of 79.5%.
6. Based on the results of precision and recall test, the best classification method was fuzzy KNNwith precision and recall of 0.94 and 1, respectively. This result shows that the accuracy is directly proportional to precision and recall.

5. REFERENCES

[1] Brunner and Suddarth. 2002. *Buku Ajar Keperawatan Medikal Bedah*, edisi 8 volume 2. Jakarta : EGC.

[2] Keller, James. 1985. *A Fuzzy K-Nearest Neighbor*. IEEE vol. SMC-15, No. 4

[3] Kusrini, 2007. *Design and implementation of building decision tree using C4.5 algorithm.* Proceedings of SEAMS-GMU Conference 2007.

[4] Larrisa Navia Rani, 2015. *Klasifikasi Nasabah Menggunakan Algoritma C4.5 Sebagai Dasar Pemberian Kredit*. Jurnal Kom Tek Info Fakultas Ilmu Komputer Volume 2 No. 2. ISSN : 2356-0010

[5] Li D, Deogun JS, Wang K (2007) Gene Function Classification Using Fuzzy K-Nearest Neighbor Approach.

[6] Manning, D. Cristopher, Prabakhar Raghavan dan Hinrich Schutze. 2009. *An Introduction to Information Retrieval*. Cambridge University Press.

[7] Maimon, O. dan Last, M. 2000. *Knowledge Discovery and Data Mining, The Info-Fuzzy Network (IFN) Methodology*. Dordrecht: Kluwer Academic.

[8] Mistra. 2005. 3 *Jurus Melawan Diabetes Mellitus*. Jakarta : Puspa Swara.

[9] Purnana, Parida. Dan Supriyatno, Catur.2013 *Deteksi Penyakit DiabetesType II denganNaive Bayes Berbasis Particle Swarm Optimization.* Jurnal Teknologi Informasi Volume 9 No.2 ISSN 1414-9999.

[10] Sunjana, 2010, *Aplikasi Mining Data Mahasiswa dengan Metode Klasifikasi Decision Tree*, Seminar Nasional Aplikasi Teknologi Informasi, Vol 7 pp. 24-29.

[11] Tandra, Hans. 2009. *Osteoporosis Mengenal, Mengatasi, dan Mencegah Tulang Keropos*. Jakarta: Gramedia Pustaka Utama.

[12] Yanita, Selly, ridho, ahmad. & lailil. 2013. *Perbandingan K-Nearest Neighbor dan Fuzzy K-Nearest Neighbor pada Diagnosis Penyakit Diabetes Melitus.*Jurnal Doro Volume 2 no.10

[13] Han, J. and Kamber, M., 2006. Data Mining: Concepts and Techniques, University of Illinois at Urbana-Champaign