Sugarcane Variety Identification Using Dynamic Weighted Directed Acyclic Graph Similarity

Adi Heru Utomo
Department of Information Technology
State Polytechnic of Jember
Jember, Indonesia, 68101
adiheruutomo@polije.ac.id

Riyanarto Sarno, R.V. Hari Ginardi
Department of Informatics Engineering
Faculty of Information Technology
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia, 60111
{riyanarto, hari}@if.its.ac.id

***Abstract*—Dynamic wDAG Similarity algorithm can be applied to sugarcane annotation. At first, we have to make a wDAG structure of many different varieties of sugarcane. We also have to make wDAG of sugarcane that will be annotated. Then, we have to calculate the similarity between wDAG types of sugarcane that will be annotated and wDAG of all the existing types of sugarcane. This similarity calculation results will present sequence similarities ranging from the most similar to the most distant from sugarcane varieties were annotated. This Dynamic wDAG Similarity algorithm has difference compared with the previous wDAG Similarity algorithm. WDAG used in this research has the node labeled , arc labeled and arc weighted, where the weight of the arc can be changed dynamically. This research fixes the previous studies of static wDAG, in which the weight values on the arc of wDAG can not be changed. On Dynamic wDAG, the weight on each arc is based on the fuzzy calculations that show the tendency of sugarcane varieties were annotated. And the fuzzy value is calculated based on agronomic traits of sugarcane to be annotated. Leaf node is the part of wDAG that will be compared first. The similarity calculation result between the two wDAG is affected by data on a leaf node to be compared and the weights of the arcs. The result shows that this method gained the average of Precision of 96%, the average of Recall of 88.5%, and the average of Accuracy of 96%.**

***Keywords*—Dynamic wDAG, Sugarcane classification, Sugarcane variety identification, wDAG similarity.**

I. INTRODUCTION

Knowledge of sugarcane variety annotation is a difficult thing to do. The data on this knowledge is a complex and heterogeneous data. Currently, annotations of sugarcane varieties can be done manually by expert sugarcane. This is because information about the sugarcane plant is descriptive qualitative semantics. It causes the annotation process is difficult for the common people as well as computer programs.

We did a previous research on sugarcane annotations using forward chaining expert system. In the research, the traits of each type of sugarcane should be known in advance before created the application program. In this research, the program will be made in order to search for files that contain characteristics of sugarcane for each type of sugarcane on the internet. File that contains the characteristics of each type of sugarcane will be converted into a data structure. In this study, the data structure used is Weighted Directed Acyclic Graph (wDAG).

In this research, we will create an application of Sugarcane Annotations using Dynamic Weighted Directed Acyclic Graph (DwDAG) similarity algorithm. In this method the sugarcane metadata compiled into a wDAG that node labeled, arc labeled and arc weighted, in which the weight of the arc can be changed dynamically. WDAG has been selected for representing the traits of the sugarcane crop because in wDAG a child node can have more than one parent node. This is useful when representing some of the traits of sugarcane. For example, cortex has two parents, node, and internode.

This research fixes the previous study of wDAG in which the value of the weight on the arc can not be changed [1], [2]. This study was conducted to replicate the annotation process is done manually by experts in which experts have a tendency in determining the outcome of the annotations. For example, an expert in sugarcane crop will more quickly identify the type of sugarcane varieties just by looking at the main traits of the sugarcane crop without having to see all the existing sugarcane crop traits. By this method, the cane annotation can be done more quickly.

The logic can be applied in the algorithm of annotation using wDAG. When using Static wDAG similarity, where the value of the weight on the arc can not be changed, then to find the most similar sugarcane wDAGs in the database to wDAG of sugarcane crop to be annotated, it has to calculate the similarity with all of the sugarcane wDAGs in the database sequentially. By using the Dynamic wDAG, the similarity calculation can be done by determining the priority of comparison in the order of sugarcane varieties tendency.

II. METHODS

The Dynamic wDAG similarity algorithm is as follows. First, a wDAG database for every variety of sugarcane will be created. Then, the agronomic traits and wDAG containing morphological traits of the sugarcane to be annotated will be included. Weighting for each arc of the wDAG is based on the fuzzy calculation of agronomic traits that are included. To calculate the similarity of the two wDAG, Jing's wDAG similarity algorithm is used. In this algorithm, the leaf nodes are the part of wDAG that will be compared first. Next will be compared all of nodes and arc on the upper level until it

978-1-5090-1381-4/16/$31.00 ©2016 IEEE

reaches the root. The method proposed in this research is shown in Fig. 1 .

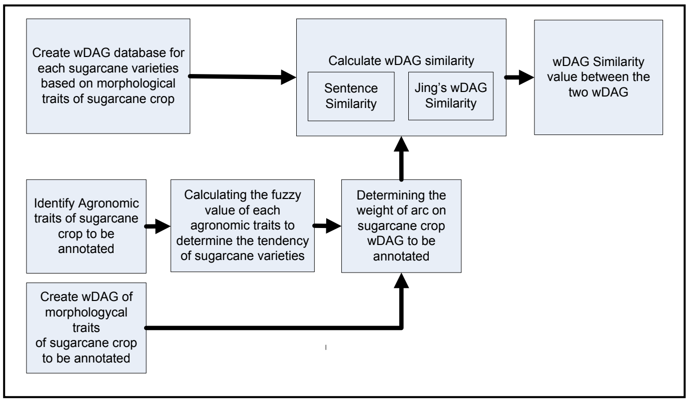
Create wDAG database for each sugarcane varieties based on morphological traits of sugarcane crop
Calculate wDAG similarity
Sentence Similarity
Jing's wDAG Similarity
wDAG Similarity value between the two wDAG
Identify Agronomic traits of sugarcane crop to be annotated
Calculating the fuzzy value of each agronomic traits to determine the tendency of sugarcane varieties
Determining the weight of arc on sugarcane crop wDAG to be annotated
Create wDAG of morphologycal traits of sugarcane crop to be annotated

Fig. 1. Diagram of The Proposed Method

Dynamic wDAG Similarity

The characteristics of sugarcane varieties that will be annotated represented in a wDAG which has a node labeled, arc labeled, and arc weighted. An example of wDAG representation can be seen in Fig. 2.

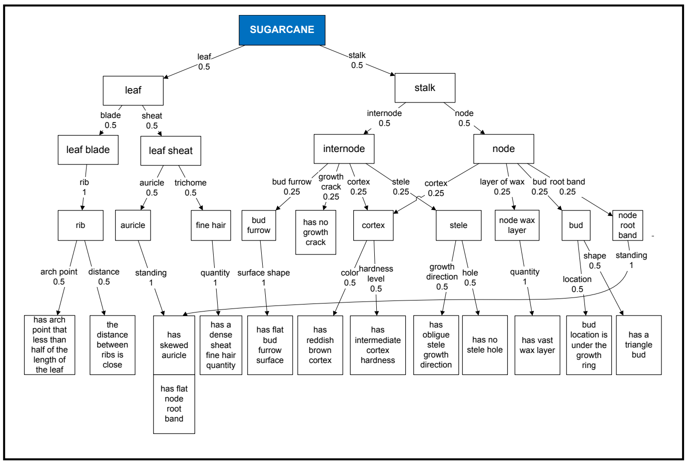

Fig. 2. Representation of the Sugarcane Plant Morphology characteristics in a wDAG

In Fig. 2, shown the representation of morphological traits of a sugarcane crop in a wDAG. In the representation shown that sugarcane has the following characteristics: 1. has arch point that less than half of the length of the leaf; 2. the distance between ribs is close; 3. has skewed auricle; 4. has flat node root band; 5. has a dense sheat fine hair quantity; 6. has flat bud furrow surface; 7. has no growth crack; 8. has reddish brown cortex; 9.has intermediate cortex hardness; 10. has oblique stele growth direction; 11. has no stele hole; 12. has vast wax layer; 13. bud location is under the growth ring; 14. has a triangle bud.

In wDAG above, the weight value is normalized. The sum of weights of all branches in asub wDAG is 1, and the value of each branch is 1 divided by the number of branches on a sub wDAG.

The first step to do is look for the tendency of sugarcane varieties based on agronomic traits of the sugarcane to be annotated. The agronomic traits of the sugarcane to be annotated are germination rate is slow, stalk diameter is medium, the amount of flower is medium, final ripening, coir level is medium, first appeared shoots is late.

By using a forward chaining expert system rule base [3], based on agronomic traits mentioned above, we can see the tendency of the type of sugarcane that will be annotated.

Structure of Forward Chaining Expert System Rule Base

In this case, the forward chaining is a search strategy that initiated the process of collection of data or facts. From these data can be searched to a conclusion that a solution of the problems faced. Inference engine looking to the rules in the knowledge base that premise is in accordance with the data, and then from that principle has obtained a conclusion. Forward chaining starts the search process with the data so that this strategy also called data driven.

TABLE I. BASE VARIABLE LIST

Variabel	*Initial*	*Variabel*	*Initial*
germination is slow	A	final ripening	L
germination is medium	B	coir level is slightly	M
germination is fast	C	coir level is medium	N
stalk diameter is small	D	coir level is plenty	O
stalk diameter is medium	E	first appeared shoots is early	P
stalk diameter is thick	F	first appeared shoots is medium	Q
flowering is slightly	G	first appeared shoots is late	R
flowering is medium	H	stem density is tightly	S
flowering is plenty/sporadic	I	stem density is medium	T
early ripening	J	stem density is distantly	U
medium ripening	K		

In this research, the rule was made to classify the data automatically [4]. Some of the rules used are described in Table 2 as follows:

TABLE II. CONCLUSION VARIABLE QUEUE

Number	*Rule List*
R-1	IF A AND (E OR F) AND (G OR H OR I) AND (K OR L) AND N THEN “BULULAWANG”
R-2	IF B AND H AND E AND T THEN “PS 862”
R-3	IF B AND S AND E AND I AND (K OR L) THEN “PS 864”
R-4	IF B AND T AND E AND H AND J THEN “PS 881”
R-5	IF C AND T AND E AND (G OR H OR I) AND (J OR K) THEN “VMC 76-16”

Based on forward chaining expert system rule base, obtained a tendency that the sugarcane variety to be annotated is a Bululawang.

Each variety of sugarcane has the typical traits. The typical trait of Bululawang is that the cane has a triangle bud. The typical traits of PS 881 are that the can has a wide and shorter leaf, short auricle, has a round bud but not outstanding, and has fine hair in their leaf sheat. And the typical traits of PS 864 is that the cane has half curved leaf, short auricle, and has a standout round bud.

Since the main characteristic of the Bululawang lies in the bud, so that the weight of arc named bud on wDAG above can be transformed into a greater weight [5], as shown in Fig. 3.

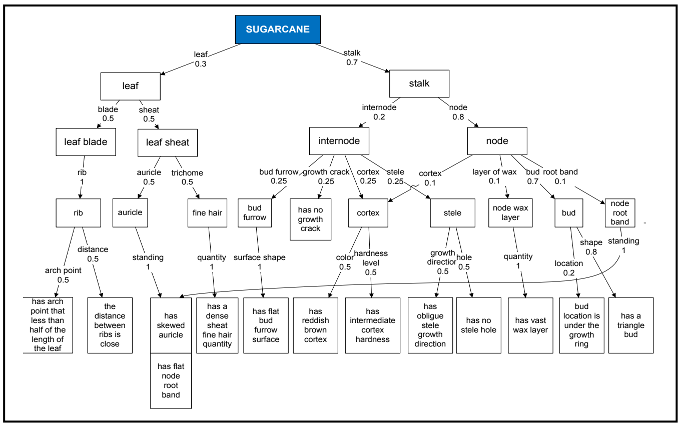

Fig. 3. Representation of the Sugarcane Plant Morphology characteristics in a wDAG with Tendency

Furthermore, the similarity between the wDAG and all wDAG stored in the database will be calculated by priority compared by the wDAG of Bululawang sugarcane variety first.

Similarity Calculation

The algorithm calculating the similarity between the two wDAG contained in papers [2]. Fig. 4 shows a wDAG representing morphological characteristics of a Bululawang sugarcane variety to be compared to wDAG that has a tendency that have a kind of Bululawang sugarcane variety as shown in Fig. 3.

wDAGsim (g , g ') is used to calculate the similarity of each pair wDAG from bottom to top. The similarity of each pair of sub wDAG at the top level is calculated based on the similarity of sub wDAG at the level below it. The weight of arcs is also considered. Weight values were averaged using the arithmetic average $(w_{i} + + w_{i}^{'})/2$. The average value is multiplied by the wDAG similarity and done recursively. First, the wDAG similarity is obtained based on the similarity of leaf nodes. Leaf node similarity is based on the sentence similarity semantically calculated using a Semantic Textual Similarity Systems (UMBC EBIQUITY CORE) algorithm. The statistical method is based on distributional similarity and Latent Semantic Analysis (LSA).

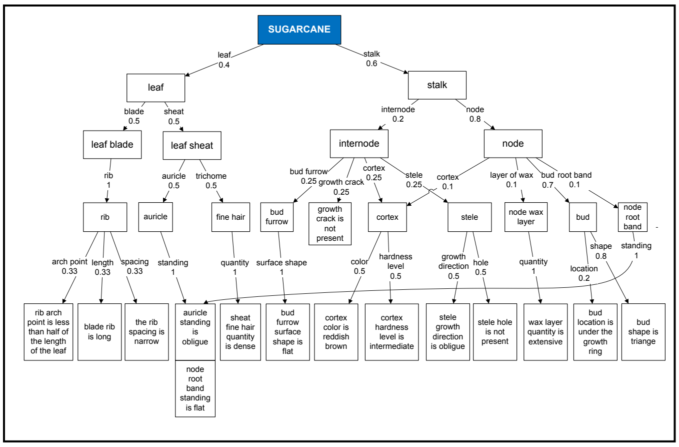

Fig. 4. wDAG Representation of the Bululawang Sugarcane Variety

LSA Word Similarity Model

LSA word similarity model used is the improvement of research on LSA word similarity previously conducted in 2013 and the research on SemEval semantic text similarity tasks performed in 2014 [6], [7]. LSA relies on the fact that semantically similar words (e.g., cat and kitten or nurse and doctor) are more likely to occur near one another in the text. Thus evidence for word similarity can be computed from a statistical analysis of a large text corpus. Raw word co-occurrence statistics extracted from a portion of a Stanford WebBase dataset [8].

Stanford POS tagger is used on the corpus to perform Part of speech tagging and lemmatization [9]. A sliding window of fixed size over the entire corpus was used to count Word/term co-occurrences. Two co-occurrence models were generated using window sizes ±1 and ±4. The more precise context which is better for comparing words of the same part of speech while the larger one is more suitable for computing the semantic similarity between words of different syntactic categories was provided by the smaller window.

A predefined vocabulary of 22,000 common English open-class words and noun phrases, extended with about 2,000 verb phrases from WordNet was used as a base of the word co-occurrence models. When words/phrases are POS tagged, the final dimensions of the word/phrase co-occurrence matrices are 29,000×29,000. After transforming the raw word/phrase co-occurrence counts into their log frequencies and select the 300 largest singular values, singular value decomposition was applied to the word/phrase co-occurrence matrices [10]. Then, the LSA similarity between two words/phrases is defined as the cosine similarity of their corresponding LSA vectors generated by the SVD transformation.

The simple align-and-penalize algorithm was used to compute the semantic similarity of two text sequences, [6]. In this research, we obtain the sentence similarity between the two sentences of the two leaf node using online tool on the site http://swoogle.umbc.edu/SimService/phrase_similarity.html.

WDAG Similarity Calculation

WDAG similarity calculation formulated as follows.

$$DAGsim(g, g') = \begin{cases} 0 & \textit{the root node labels of } g \textit{ and } g' \textit{ are not identical} \\ 1 & g \textit{ and } g' \textit{ are leaf node} \\ \Sigma \begin{cases} wDAGsim(g_i, g'_j).\frac{(w_i + w'_j)}{2} & g_i \textit{ and } g'_j \textit{ are not missing} \\ wDAGsim(g_i, \varepsilon).\frac{(w_i + 0)}{2} & g_i \textit{ is missing in } g' \\ wDAGsim(\varepsilon, g'_j).\frac{(0 + w'_j)}{2} & g'_j \textit{ is missing in } g \end{cases} \\ \sum_{j=1}^{breadth\_of\_g'} wDAGsim(\varepsilon, g'_j).\frac{(0 + w'_j)}{2} & \textit{only } g \textit{ is a leaf node} \\ \sum_{i=1}^{breadth\_of\_g} wDAGsim(g_i, \varepsilon).\frac{(w_i + 0)}{2} & \textit{only } g' \textit{ is a leaf node} \end{cases} \quad (1)$$

where,

wDAGsim(g, g’): similarity of two wDAGs g and g'

wDAGsim(g_i, g'_j): intermediate similarity of the i-th and j-th sub-wDAGs of the wDAGs g and g', respectively

w_i and w'_j: arc weights of the i-th and j-th child of the root node of wDAG g and g', respectively.

$\mathcal{E}$: an empty wDAG

i: increase from 1 to g

j: increase from 1 to g'.

In the example above, the behavior of the algorithm can be described as follows. ‘Rib’ wDAG similarity will be calculated first. To calculate the ‘Rib’ wDAG similarity, leaf nodes similarity below ‘Rib’ wDAG are required. The result of the calculation of sentence similarity of the leaf node is as follows. *SentenceSim*(rib arch point is less than half of the length of the leaf, has arch point that less than half of the length of the leaf) = 0.94908255, *SentenceSim*(blade rib is long, $\mathcal{E}$) = 0, *SentenceSim*(the rib spacing is narrow, the distance between ribs is close) = 0.43991673. Each of that similarity value has to be multiplied by the average weight of each arc of rib, ie 0.94908255 * (0.3333+0.5)/2, 0 * (0.3333+0)/2, and 0.43991673 * (0.3333+0.5)/2. ‘Rib’ wDAG similarity result obtained is 0.578726550012. Next will be calculated the similarity of sub wDAG above it, ie 'leaf blade' wDAG. The similarity value of wDAG 'leaf blade' is the value of 'rib' wDAG similarity multiplied by the average weight of the ‘leaf blade’ arc, namely 0.578726550012 * (1+1) / 2, so that the results obtained of 'leaf blade' wDAG similarity is 0.578726550012. In the same way, will be calculated for all sub wDAG similarity, and finally obtained wDAG similarity value of the both wDAG above is 0.770804284

III. PERFORMANCE EVALUATION

The evaluation of sugarcane variety identification using Dynamic Weighted Directed Acyclic Graph similarity may use the confusion matrix and the ROC curve/AUC (Area Under the Curve).

Confusion Matrix

Confusion Matrix is a method for evaluation using a matrix table as shown in table 3 [11]. The table show that the dataset is composed of two classes, one class is presumed as positive and the other negative [12]. In the next stage, the confusion matrix is generating value accuracy, precision, and recall.

TABLE III. CONFUSION MATRIX MODEL

Correct Identification	**Identified as**	
	+	-
+	True positive	False Negative
-	False Positive	True Negative

True positive is a case where the identification is predicted correctly as positive. A true negative is a case where the identification is predicted correctly as negative. A false positive is a case where the identification is predicted incorrectly as positive, while a false negative is a case where the identification is predicted incorrectly as negative.

Precision (P) is the number of true positives (T_p) over the number of true positives plus the number of false positives (F_p).

$$P = \frac{Tp}{Tp + Fp} \quad (2)$$

Recall (R) is the number of true positives (T_p) over the number of true positives plus the number of false negatives (Fn).

$$R = \frac{Tp}{Tp + Fn} \quad (3)$$

Accuracy (A) is the number of true positives (T_p) plus the number of true negatives (T_n) over the total number of existing identification.

$$A = \frac{Tp + Tn}{Tp + Tn + Fp + Fn} \quad (4)$$

Identification test performed five times for each sugarcane varieties. There are five wDAGs of five sugarcane varieties that will be annotated, each of which will be compared with five wDAGs of five varieties of sugarcane that the varieties are already known.

The testing result for each variety of sugarcane can be seen in Table 4 below.

TABLE IV. THE TESTING RESULT OF THE SUGARCANE IDENTIFICATION

Sugarcane Variety	**Precision**	**Recall**	**Accuracy**
Bululawang	1	0.8	0.96
PS 881	1	1	1
PS 882	1	0.625	0.88
PS 864	1	1	1
VMC 76-16	0.8	1	0.96

From the table above, it can be seen that the average of Precision is 0.96 or 96%, the average of Recall is 0.885 or 88.5%, and the average of Accuracy is 0.96 or 96%.

ROC Curve

ROC curve shows the identification accuracy and compares visually. ROC express confusion matrix. ROC is a two-dimensional graph with false positives as a horizontal line and a true positive as a vertical line. The calculation result is visualized with the ROC curve (Receiver Operating Characteristic) or AUC (Area Under the Curve). ROC has a diagnostics value level, that is [13]:
a. Accuracy value 0.90 – 1.00 = excellent identification
b. Accuracy value 0.80 – 0.90 = good identification
c. Accuracy value 0.70 – 0.80 = fair identification
d. Accuracy value 0.60 – 0.70 = poor identification
e. Accuracy value 0.50 – 0.60 = failure

The ROC processing result of sugarcane variety identification using Dynamic Weighted Directed Acyclic Graph similarity is 0.96 with excellent identification value level.

IV. Conclusion

In this paper has shown that the forward chaining expert system rule base is effectively implemented to create a dynamic wDAG. The experimental result shows that the precision and accuracy of sugarcane variety identification using Dynamic Weighted Directed Acyclic Graph similarity has excellent level identification value.

References

[1] R. Sarno, K. Ghozali, B. A. Nugroho, and A. Hijriani, “Semantic Matchmaking using Weighted Directed Acyclic Graph,” in *International Seminar on Applied Technology, Science, and Arts*, 2011, pp. 329–334.

[2] A. Kunaefi, T. Nagai, H. Nakano, and R. Sarno, “Semantic Web Service Discovery Using Weighted Directed Acyclic Graph,” *J. Kursor*, 2013.

[3] A. Abraham, “Rule-Based Expert Systems,” *Handb. Meas. Syst. Des.*, 2005.

[4] A. Arwan, B. Priyambadha, R. Sarno, M. Sidiq, and H. Kristianto, “Ontology and semantic matching for diabetic food recommendations,” in *2013 International Conference on Information Technology and Electrical Engineering: “Intelligent and Green Technologies for Sustainable Development”, ICITEE 2013*.

[5] Y. Anistyasari and R. Sarno, “Weighted ontology for subject search in Learning Content Management System,” in *2011 International Conference on Electrical Engineering and Informatics, ICEEI*, 2011, pp. 1–4.

[6] L. Han, A. Kashyap, T. Finin, J. Mayfield, and J. Weese, “UMBC_EBIQUITY-CORE: Semantic Textual Similarity Systems,” *Proc. 2nd Jt. Conf. Lex. Comput. Semant.*, vol. 1, pp. 44–52, 2013.

[7] A. Kashyap, L. Han, R. Yus, J. Sleeman, T. Satyapanich, S. Gandhi, and T. Finin, “Meerkat Mafia: Multilingual and Cross-Level Semantic Textual Similarity Systems,” *Proc. 8th Int. Work. Semant. Eval. (SemEval 2014)*, no. SemEval, pp. 416–423, 2014.

[8] J. Cho, H. Garcia-Molina, T. Haveliwala, W. Lam, A. Paepcke, S. Raghavan, and G. Wesley, “Stanford WebBase components and applications,” *ACM Trans. Internet Technol.*, vol. 6, no. 2, pp. 153–186, 2006.

[9] P. Tagger and C. D. Manning, “Enriching the Knowledge Sources Used in a Maximum Entropy,” in *EMNLP ’00 Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13*, 2000, pp. 63–70.

[10] P. Li, C. Burgess, and K. Lund, “The acquisition of word meaning through global lexical co-occurrences,” *Proc. Thirtieth Annu. Child Lang. Res. Forum*, pp. 166–178, 2000.

[11] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. San Fransisco: Morgan Kauffman, 2006.

[12] M. Bramer, *Principles of Data Mining*. London: Springer, 2007.

[13] F. Gorunescu, *Gorunescu - Data Mining Concept Model Technique*. Berlin: Springer, 2011.