Label Comparison Analysis Using Hybrid Similarity for Labeling Systems of Indonesian University Websites (English Menu Version)

I Gusti Bagus Ady Sutrisna ¹, Dana S. Kusumo ² and Ibnu Asror ³
¹,²,³ Telekom University, Bandung, Indonesia

Email: ¹igustibagus.as@gmail.com, ²danakusumo@telkomuniversity.ac.id, ³iasror@telkomuniversity.ac.id

Abstract. Labeling systems are an important component in designing a university website. In order to avoid misinformation, website developers need to design a website labeling system that can represent information that a website owner wants to deliver to the users of her website. One way of designing a labeling system is by comparing and studying the labeling system used on competitor's websites. Syntactic similarity is usually used for comparing the websites' labels. However, it has a limitation that can only calculate similarities based on strings only, so that the possibility of two compared labels having the same meaning will be considered as different labels. To provide the meaning of the label, the comparison using semantic similarity is used. However, semantic similarity has limitations such as not being able to process two words or more and possibly having no path exists between two word senses. Therefore, this research proposes a hybrid similarity where the highest result of syntactic and semantic similarity comparison between two labels is combined and able to cover the limitation of two methods. The hybrid similarity shows the results of possibility of higher score between combined syntactic and semantic similarity score with suggestion on the list label (based on semantic result). The expert suggest output of hybrid list label to help developer create and reviewing existing label, and then participant who answers the questionnaires 82.16% agree with label that show the acceptance of the label. The list label based on hybrid score can help and assisting the web developer to create new labeling systems for a new website. It is also help developer to improving an existing website’s labeling systems. Keywords: university website, labeling systems, syntactic similarity, semantic similarity, hybrid similarity, Levenstein Distance, Wu and Palmer.

1. Introduction
Labeling systems on the website is a word that can be used correctly to represent a group of information or concept on the website [1]. A bad labeling system can produce the label is not representative and differentiate each other, still uses jargon elements, does not give a good impression, and confusing label make a bad feedback from user [1]. Therefore, labeling systems are very important to provide appropriate information to the user.
Labeling system in university website is used to provide university information. Labeling systems help not only internal website users (faculty member, staff, and students), but also external website users is also need information about university information such as location, faculty, and others [2]. The selection of labels should be relevant and represent the information content of a website. Mistakes in the website's label selection causes website users unable to get the information they needed and desired. This causes a loss to the website owner [2].

One of the methods to create or improve labeling systems is by studying and comparing the competitor websites [1]. The Labeling Systems can be developed from a comparison between the targeted website and other websites in the same domain. Comparing method is used because there is no standard in making or improving Website University. By comparing each other, it will be easier to find pattern in labeling system. This pattern may not be a standard, but it can help at least inform label choices [1].

The labeling system is only based on textual label, so that other structural or hierarchical systems were not included in this study (textual label only). The type of labeling systems studied is the navigation label system because every university website consistently has the navigation label system (top menu, main menu, and footer menu) [1]. The university website used as sample comparison are university websites in Indonesia that get the top 10 ranking in Webometrics Indonesia and Telkom university website [3]. The Webometrics is used because Webometrics is one of the tools or systems to measure or assess the progress of all the best universities in the world through the university website [4].

In the previous research, the comparison of labeling systems was done for different purposes [5] and manually [6]. Kadek [7] made the first step to comparing labeling systems automatically with syntactic similarity with levenshtein distance. For the dataset, it used Indonesia university website with 706 labels. However, the result of syntactic similarity has a limitation that can only calculate similarities based on strings only, although the possibility of having the same meaning will be considered as a different label [8]. So to provide meaning of the label, the comparison using semantic similarity is used with English language. English language is used because semantic similarity based on WordNet (complete dictionary for English language) [9]. Semantic similarity using Wu and Palmer method is used because it has the advantage based on the shortest path concept of WordNet that only uses "is a relation" of the two concepts to provide meaning of the label [9]. However, the semantic similarity has the limitations of not being able to process two words or more and possibly having no path exists between two word senses. Therefore, this research proposes a hybrid similarity where the score is combined and able to cover the limitation of each other. The result from hybrid similarity is labeling systems based on score of syntactic and semantic similarity combined.

Regarding to the background stated above, we identify the research question as follows:

1. How to apply hybrid similarity (syntactic and semantic) methods for comparison of labeling systems?
2. Can result of comparing labeling systems used for creating and designing labeling systems for Indonesian Website University (English menu version)?

The hypothesis of this research is output of list common label based on the hybrid score can be used as guidance for Indonesian university website with English menu version. The list common label based on hybrid score can help and assisting the web developer to create new labeling systems for a new website. It is also help developer to improving an existing website’s labeling systems.

As delimitation this research limit the labeling system is based on textual label, so other systems such as Organization Systems, Navigation Systems, and Search Systems were not included in this study. Only use navigation labels to be extracted consists of top menu, main menu, and footer menu. The language used is the English language obtained from the extraction process on Indonesian university website with English menu version (there are also other languages even though the website is in English version).
2. Related Work

2.1. Labeling Systems
The definition of labeling systems on the website is how a word can be used correctly to represent a group of information or concept on the website [1]. Label is used to representing information based on the labeling systems [1]. A label is used to represent larger pieces of information on Website, for example "Contact Us" representing content about contact names, addresses, phone, fax, and email information. Hence, the purpose of the label is to communicate information efficiently [1].

2.2. Extraction
To extract labeling system from website, web crawling is used for the process. Web crawling is a search technique commonly used by search engines to find useful information from a collection of web page sources [10]. Web crawlers are used to make copies of visited web pages. The copy is used for further processing such as indexing and extracting the information.

2.3. Preprocessing
According Preprocessing of labeling systems used text processing to transforming unstructured data into structured [12]. Preprocessing is used because semantic similarity only process one word with English language only [9]. In this research the steps are used on the process is removing redundant label, stopwords removal, Indonesian language and term removal, tokenization, checking label on WordNet, and stemming.

The result of preprocessing of labeling system is only one word label with English. The semantic similarity is used in process [9]. Two words or more cannot be processed on semantic because the WordNet can only process one word with English language [9].

2.4. Comparing with Syntactic Similarity Methodology
Levenshtein distance is a measure of similarity between two strings, the source string (s) and the target string (t) [8]. The score is derived from the number of deletions, inserts, or substitutions required converting (s) to (t). The higher gap from levenshtein distance, the string that is compared becomes less similar. Then, to measure Levenshtein Distance between 2 strings a, b is expressed in equation 1.

\[
\operatorname{lev}_{a,b}(i,j) = \begin{cases} 
\max(i,j) & \text{if } \min(i,j) = 0, \\
\min \left\{ \begin{array}{l}
\operatorname{lev}_{a,b}(i-1,j) + 1 \\
\operatorname{lev}_{a,b}(i,j-1) + 1 \\
\operatorname{lev}_{a,b}(i-1,j-1) + 1_{a_i \neq b_j}
\end{array} \right. & \text{otherwise}, 
\end{cases}
\]

**Equation 1.** Levenshtein Distance Calculation Formula

By:
- a = the first string
- b = second string
- i = first string length
- j = length of the second string

By using the score and then can calculate the similarity score between two strings with the syntactic similarity method. Then to calculate the score of similarity using the formula:

\[
sim = 1 - \left( \frac{\text{Dis}}{\text{MaxLength}} \right) \times 100\%
\]

**Equation 2.** Similarity Calculation Formula

Sim = similarity / similarity score
Dis = distance levenshtein (using the previous formula)
MaxLength = longest string score
2.5. Comparing with Semantic Similarity

Wu and Palmer defined as the similarity of two concepts based on the depth of Least Common Subsumer and the shortest path (Semantic Similarity) [9]. The calculation process performed by Wu and Palmer is to find the shortest path of each concept on the WordNet, and then each formed path is combined to search for the LCS (Least Common Subsumer). Searching LCS by finding the sense that often arises from the two paths connected based on WordNet. Equation 3 is the formula used to calculate WUP [6] that has been modified by WS4J.

\[
\text{sim}(c_1, c_2) = \frac{2 \times \text{depth}(lcs(c_1, c_2))}{\min(\text{depth}(c_1)) + \min(\text{depth}(c_2))} \times 100\%
\]

**Equation 3.** Wu and Palmer Calculation Formula

In Equation 2.3, there is the term depth. It is measured from the root to the specified node. Concept of two labels has been found on WordNet and calculated by semantic similarity using Wu and Palmer. The role of WordNet in this research can be the source of corpus to make the semantics of the term, by relying on the concept of taxonomy [9].

3. Research Methodology

3.1. Research Design

In the research design system the process system is applied to get the labeling systems and then reviewed by four expert and 60 research participants. The whole research process flowchart is shown in Figure 1.

![Figure 1. The Whole Research Process Flowchart](image-url)

3.2. Extraction Process

Extraction of labeling systems is used to get the data that needed on this research. The universities used as sample comparison are universities in Indonesia that get the top 10 ranking in Webometrics Indonesia and Telkom university website, so that there are 11 university websites used for comparison. The selection of university websites using Webometrics is because Webometrics is one of the tools or systems to measure or assess the progress of all the best universities in the world through the university website [4]. Total 11 of university website are used because it can be representing for labeling systems based on Webometrics [4].
Extraction process is done through web scraping method. This method serves to obtain labeling systems by defining the template manually to the page you want to search for navigation. Each website will have different web scraping code that is because the difference in HTML (Hyper Text Markup Language) structure. In each web page scraping will be defined HTML tags that contain navigation labels on every web [3].

3.3. Preprocessing Process
1. The function of stopwords removal is to remove the stopwords such as “of”, “the”, and etc. The stopwords is not used because it is focusing on English language only. Another language such as Indonesian language and term also removed. In this research, Labeling Systems is focusing on English language only.
2. The purpose tokenization is to cut 2 or more words into single word [12]. The system sees labels that have 2 or more words. For examples with “vision and mission” label data, then tokenization perform data processing into 3 labels namely “vision”, “and”, and “mission”. The new label of tokenization is entered in the array so it adds from the previous label.
3. WordNet checking for every word generated from tokenization [12]. Checking system will output data label data containing one word English language only. There are types of label data contained in the data label. The one word English language pass on checking in the WordNet and the other (non-one word English language) not pass.
4. Process of stemming is to get the root word from the English word [12]. Stemming will produce output in the form of each word that make up the label will turn into a root word. Examples of “faculties” and “faculty”, then it will turn into the basic word “faculty”. Some of the basic word case is different from the original word example of “media” to “medium”, and so forth. websites using Webometrics is because Webometrics is one of the tools or systems to measure or assess the progress of all the best universities in the world through the university website [4].

3.4. Comparing Process
1. Form a matrix representation M * N to measure the similarity score of labels on each university website. And then form a combination of labels from each label on the university website and then eliminate label if there is the same label.
2. The rows (M1,M2,...,Mn) in the matrix will be filled in by a combination of all the labels from each university website while the columns (N1,N2,...,N11) in the matrix will be filled with 11 label files from each university website. Initialize values on a matrix with a value of 0.
3. Calculating the similarity of one label on a row with the labels on 11 label files will use the similarity method. After comparing one label with each label at the university, the highest score obtained and then to be selected.
4. From getting the highest label score from each university website, it will be calculated according to the highest average score.

3.5. Expert and User Review
The importance of expert review is used to know how the actual selection of the use of labels on the website that occurred in the field so that requires knowledge from expert review. The results labeling systems then re-analysis to see which labels are better to use as the labeling system on the university website. Then the results of the analysis will be presented along with the combined score from hybrid of the label on each web. Labeling systems based on expert review is used to inform common labels for labeling system on the university website.

The user review is analysis of labeling systems based on the research. The importance of user review used to find out the labels that are often accessed by user when browsing or looking for information in the website of Indonesian universities with English menu version. There are 60 users who answer the questionnaire from student, internal campus, and external campus. The user is divided
into 3 categories and has many characteristics such as the age and how often they browsing using website. The results of the label list in the re-analysis to see which labels are often accessed by user.

4. Experiment and Analysis

4.1. Extraction Result

The purpose of extracting is to get all data from labeling system. Total label extracted from stopwords removal and Indonesian language and term is about 755 labels and used for process syntactic similarity. The label is extracted from each website for this research.

Preprocessing is used for semantic similarity. Process of preprocessing method is tokenization to split words from 2 words or more, word checking with WordNet, and application of stemming to get the word WordNet base. Total label preprocessing is about 390 labels.

4.2. Comparison Result

Extraction of labeling systems that has total 755 label then processed using syntactic similarity on 11 university websites. The expected result using syntactic similarity is to know the characteristic and the use of labeling systems (syntactic similarity based). Each label on the combined label dataset is compared to each label owned by each university website by using levenshtein distance to find what label on each website has the highest similarity score with the existing label. After find each of highest score, and then label ranked based on average score. From the table its show when label have similar string and then the score is 100%, but when any different string compared its score below 100%.

The result of preprocessing of labeling systems that has total 390 labels then processed using comparison labels on 11 university websites. The expected result using semantic similarity is to know the characteristic and use of labeling systems (semantic similarity based). Each label on the combined label dataset is compared to each label owned by each university website by using Wu and Palmer to find what label on each website has the highest similarity. After find each of highest score, and then label ranked based on average score.

Based on score from the result of Syntactic and Semantic Similarity that has total 755 and 390 labels and then it is processed using comparison labels on 11 university websites. The expected result using hybrid similarity is to know the characteristic and the use of labeling systems (hybrid similarity based). Each label on the combined label dataset is compared to each label owned by each university website based on highest score based on Levensthein Distance and Wu and Palmer. After find each of highest score, and then label ranked based on combined score. The example from hybrid similarity results (top 5) can be seen in Figure 2.

**Figure 2.** The example from hybrid similarity results (top 5)

On the result of average from using hybrid, the score based on two types single only for the highest score. On the process there is the average highest average for the list label. However, on the semantic process there is some of suggestion that has same meaning cannot be place on university website such as “family” or related label. To know the combined highest score,
there are ranking based on combining syntactic and semantic score not only average. The combined score only focusing on “Semantic & Syntactic” Score from the highest to lowest. The common label from hybrid similarity score can be seen in figure 3.

In the Figure 3 is a sample of hybrid similarity comparison results. The highest score most common label are “faculty” and “program”. The secondary is “academic”, “International”, and “library” based on list labels output with hybrid scores. If the score gets smaller from highest score most common label then the label becomes uncommon and is rarely used on the university website. In the case of hybrids, checking between words is based on the similarity of strings between words and the meanings of each of the words being compared. Therefore, the result data obtained in the process that occurs is a combination of the number of occurrences or syntactic similarity processes between strings in the comparison and semantics allow for changes in the string or change the word based to the results of comparison between meanings. Syntactic can compare either one or two words or more while semantics only apply to one-word cases whereas two or more words do not apply. The results obtained in hybrid comparisons are the result of syntactic similarity, semantic similarity results, and hybrid similarity results that allow having a different string types. The process on the combined can be used as list label that has highest semantic and syntactic score to choose label. For the syntactic based on same string and semantic based on same string or meaning suggestion for case of the label. The hybrid score based on below 27.23% and considered is list label for uncommon list. The uncommon and considered label from hybrid score can be seen in figure 4.

Based on 755, 690 having scores below 27.27%. That is mean can be categorized to uncommon scores. Score 18.18%, score 9.09% is uncommon label from the list. The considered score based on two words or more and syntactic and semantic only score. The syntactic compared all the label data
based on string similarity. The label that has a highest score to another one means highest similar to another label. Label that have 100 score can be change to themselves or to another label which means they have similarities. In the process semantic, it is possibly no path exists between two word senses on WordNet. The score used in hybrid similarity is from syntactic, semantic, and syntactic and semantic combined. The considered score semantic with no path exists between two word senses can be replacement using syntactic score with one word or two or more words. The result of combined and considered score helps developer choose the label data to create or improving the website.

Based on the comparison analysis, the data can be used as a labeling system in hybrid development. Based on rankings, the data that has the highest syntactic score and has semantic score, or has a combined score label can be selected through the string changing list of the label.

4.3. Expert and User Review Result
Based on interviews with the four expert, the firstly they choose label from meeting with academic of university and benchmarking with other website. The standard labeling system for the website is done by benchmarking the websites of other universities. In the benchmarking process is observed on what labels are often used on university websites. However, because benchmarking is still done manually so the numbers of websites that can be benchmark only 3-4 or 2 campus and cannot be done on the entire label. They only use for menu design and content of the website. By using the system that has been built, the number of websites and the number of labels that are compared can be more and the benchmarking process can be done faster. Any comparison on the website has a syntactic, semantic, or hybrid score so that it can choose based to the score displayed from the expert review. The Semantic is better than syntactic, because semantic is closer to its human language or meaning than syntactic. Hybrid is better because it is not fixed on one term that has similarity.

So based on the results of interviews with experts can be said that the system has been built and the resulting label results of comparison can used to inform common labels for labeling system on the university website. The data then sorted back in accordance with the needs of each of the website maker or standard website maker in each university.

The comparison labels and the built-in system are reviewed by user who use the website, to the respondent is choosing from list label based on hybrid score. So the label from system can giving the respondent from how often accessed the labels. For label selection criteria based on a hybrid similarity score from the results from common to uncommon labels. System has been built can help the process of making labeling standards on the university website.

Based on user review from 9 July 2018 to 16 July 2018 there are 60 users who answer the questionnaire from student, internal campus, and external campus. The user is divided into 4 categories and has many characteristics such as the age and how often they browsing using website. The participant based on range age 18 to 36 and has many kind of profession.

The answer from questionnaire has much different kind such as agree with the all label, did not agree, and semi agrees from label that they choose. In the question there are divided with appearance label from 11 website. The user chooses labels that are often accessed and give the feedback to the system. If user is agree with all labels on questionnaire they choose agree answer with all labels chosen. But if user is agree only few of labels then they write for choosing such as example only academic, and research based on user they choose. The last are disagree if user did not agree with all labels then they choose disagree.
5. Conclusion

Experiment shows the process of comparison with syntactic, semantic and hybrid method creates labeling systems with score. After comparing, the list label with different characteristic will be obtained and going through process analyzing. Hybrid is used because the possibility of higher score between combined syntactic and semantic with suggestion on the labeling systems. Hybrid similarity is able to cover the limitation of syntactic and semantic similarity. The highest score most common label are “faculty” and “program” and the secondary is “academic”, “International”, and “library” based on list labels output with hybrid scores. If the score gets smaller from highest score most common label then the label becomes uncommon and is rarely used on the university website. Based on the results of interviews with four expert review and 60 user review acceptance and then the result can be used to inform common labels for labeling system on the university website. The four expert suggest hybrid for the list label to help developer create labeling systems and improving existing labeling systems, and then user 82.16% agree based on Figure 4.6 with label that shown on questionnaire (58.16% agree and 24% semi agree) that show the acceptance of the label.

For a recommendation, added new method to handle English language with two words or more in process of semantic similarity. In the future work, semantic similarity using domain thesaurus to prevent label with specific domain meaning such as home (in website domain) and other terms.

References

[1] Lou Rosenfeld, Peter Morville, “Information Architecture for the World Wide Web Third Edition, O'Reilly,” 2015
[2] Shelley Gullikson, Ruth Blades, Marc Bragdon, Shelley McKibbon, Marnie Sparling, Elaine Toms, “The Impact of Information Architecture on Academic Web Site Usability,” 1999
[3] Jsoup HTML parser in java. [Online], Available: smlcodes.com. [Accessed January 2018]
[4] Mike Thelwall, Liwen Vaughan, and Lennart Bjorneborn, “Webometrics,” 2003
[5] Bing Liu, Kaidi Zhao, Lan Yi, “Visualizing web site comparisons,” 2001
[6] Ping Zhang, Gisela Von Dran, Paul Blake, Veerapong Pipithsuknsut, “A Comparison of the Most Important Website Features in Different Domains: An Empirical Study of User Perceptions,” Proceedings of Americas Conference on Information Systems (AMCIS’2000), Long Beach, CA. August 10-13, 2000. pp. 1367-1372, 2000
[7] I Kadek Aditya Cahaya Putra, Dana S. Kusumo, Anisa Herdiani, Indra Lukman Sardi, Shinta Yulia Puspitasari, “An Automatic Website Menu Comparison among Indonesia’s University Websites for Designing Labeling System of An Indonesia University Website,” Journal of Computer Science, Accepted for Publication in 2018
[8] Rishin Haldar, Debajyoti Mukhopadhyay, “Levenshtein Distance Technique in Dictionary Lookup Methods: An Improved Approach,” arXiv:1101.1232, no. Ld, pp. 286–293, 2011
[9] Lingling Meng, Runqing Huang, Junzhong Gu, “A Review of Semantic Similarity Measures in WordNet,” International Journal of Hybrid Information Technology Vol. 6, No. 1, January, 2013
[10] Robertus Theodore, “Model Ekstraksi Data Pada web,” Pros. Konf. Nas. Inform., p. 175, 2013
[11] Wael Hassan Gomaa, Aly Fahmy,”A Survey of Text Similarity Approaches, International Journal of Computer Applications” (0975 – 8887), April 2013
[12] Muhammad Ali Fauzi, “Text Pre-Processing.” 2016. [Online]. Available: malifauzi.lecture.ub.ac.id/files/2016/02/Text-Pre-Processing-v2.pptx. [Accessed Jan 2018]