Journal of Computer Science

Original Research Paper

An Automatic Website Menu Comparison among Indonesia's University Websites for Designing Labeling System of an Indonesia University Website

I Kadek Aditya Cahaya Putra, Dana Sulistiyo Kusumo,
Anisa Herdiani, lndra Lukman Sardi and Shinta Yulia Puspitasari

School of Computing, Telkom University, Bandung, Indonesia

Article history
Received: 19-07-2018
Revised: 21-12-2018
Accepted: 21-02-2019

Corresponding Author:
Kusumo, School of Computing, Telkom University, Bandung, Indonesia
Email: danakusumo@telkomuniversity.ac.id

Abstract: Label on websites represents the contents of information. To avoid delivery of incorrect information, website developers need to design a good labeling system, thus the labels can provide the information that the website owners want to convey to their users. One of the approaches to design a labeling system on a website is by comparing and studying the labeling system used on competitor websites. In this research, it was compared the labels of website menus, used by 11 university websites in Indonesia, by calculating the similarities of the label using Levenshtein distance. The result of label comparison was analyzed, which then can be used as designing source of labeling system for building and improving labeling systems of university website menus in Indonesia.

Keywords: Labeling System, Information Architecture, Website Menu

Introduction

Website design must consider the design and structure of information (Djonov, 2007; Larson and Czerwinski, 1998) because bad organization information of website can make users difficult to find information (Gullikson *et al*., 1999). Therefore, in developing a website, it is necessary to organize information to make information easy to be found by users (Rosenfeld and Morville, 1998). Website information architecture is the design of the structure and organization of information on a website thus users can easily find information on a website (Rosenfeld and Morville, 1998). One of the key components of web information architecture is labeling system.

A label on a website is used to represent information contained on a website (Rosenfeld and Morville, 1998). Good labeling system is required by a website owner to deliver information needed by website users and to facilitate the users in finding information (Rosenfeld and Morville, 1998). There are several ways to design a labeling system; one of them is by comparing and studying labeling systems used on similar and relevant websites (Rosenfeld and Morville, 1998).

In this research, it was compared 11 labeling systems of university website menus in Indonesia. The expected results are rankings, frequency and similarity degree of compared labels of website menus, which can be used as label design sources for website label menu. It was evaluated the results of labels comparison between university websites with the labels of indexed obtained on the National Higher Education Standard documents, issued by the Directorate General of Higher Education and the documents of the Higher Education Institution Accreditation Standards. Finally, to evaluate the comparison results, the comparison label results was reviewed by an expert, which then it can also be used to design the labeling system for university website in Indonesia.

The recommendations on the labeling system for university websites in Indonesia can help the content developers to use common patterns of website label menu integrating with specific information according to objectives of website owners. Using the common patterns can benefit websites because the users are accustomed to browsing and finding information on similar websites.

Literature Review

Information Architecture

Information architecture is defined as a discipline to organize, label, navigate and search for information on a website to help the user easily to find information (Rosenfeld and Morville, 1998). In this research, it only focused on a labeling system design. A labeling system selects term used that can exactly represent information and concepts in a website. Labels are one of the clearest ways to show users how organizations and navigation systems are on a website. For example, in the organization system, labels consist of Faculty, Academic and Admission, meanwhile in the navigation system, labels used are Home, Search and Contact Us. Labels that represent information must minimize the error of information to be conveyed to users. If it is found questions or confusions over labels, there should be clarification and explanation of labels. Labels should educate users about new concepts and help them quickly identify needed information.

In labeling system, there are two types of labels: Text and iconic (Rosenfeld and Morville, 1998). Text labels can be divided into four types: Contextual link, header, navigation system and index term. First, label as contextual link is a hypertext-linked label contained in a document or piece of information. Second, label as headings is a label that is often used as a title that describes the piece or the whole information. Label as heading can be visualized into the hierarchy by using numbering, font size, font type and another form. Third, label as navigation system is a label that represents the options on the navigation system, which users can move from one information space to other information spaces. Lastly, label as index terms is a label in the form of keywords and subject headings that represent content for browsing and searching purposes. In this research, it focused specifically on the label as navigation.

Designing the labeling system needs sources for labeling. There are three methods that can be used as sources of labeling system (Rosenfeld and Morville, 1998).

Using Our Own Website

Our website may already have labeling system by default. At least the previously used labels have been based on some decisions during the creation of the site, so it is possible that not all the labels need to be replaced. Instead, it can be used as the starting point for developing a complete labeling system by considering the decisions made while creating the original system.

Comparing with Competitor Websites

Searching other websites as references of labeling system. When conducting this step, it is possible to find similar labeling patterns of competitor websites. These patterns may not be the industry standard but can be used as the label option for websites.

Referring to the Thesaurus and Controlled Vocabulary

Thesaurus and controlled vocabulary are useful resources created by professionals in the library field. This vocabulary is often available to the public and has been designed for widespread use. This vocabulary is useful for filling the labeling system used to index the content.

Web Content Analysis

Comparing labels between websites to find the similarity and dissimilarity of labels are related to website content analysis (Herring, 2010; McMillan, 2000). Website content analysis is the use of content analysis on the website (Herring, 2010; McMillan, 2000). There are five processes for conducting content analysis for the website: Research questions/hypotheses formulation, sample selection, category definition, coding scheme, coding data analysis and interpretation (Herring, 2010; McMillan, 2000). The labels comparison can be seen as the process of defining categories of coding units. Also, the context unit in the category definition process is the labeling system of university websites in Indonesia.

Web Information Extraction

Information extraction transforms unstructured text into information in a structured form (Cowie and Lehnert, 1996). An example of unstructured information is web page content. The first process of web information extraction is web crawling to extract data on web page. Web crawling is a crawling technique commonly used by search engines to search for useful information from sets of web page sources (Olston and Najork, 2010). Web crawlers are used to make copies of visited web pages. The copies are then used for further processing such as indexing and extracting information.

A special term for web page extraction is called web scraping. Web scraping is an extraction technique used to obtain structured data from web pages (Vargiu and Urru, 2012; Ferrara *et al*., 2014). There are two common techniques of web scraping. First, automated techniques using machine learning and second, manual techniques by defining the template for each page to be extracted (Vargiu and Urru, 2012; Ferrara *et al*., 2014). The web template scraping technique is conducted by defining the template using the XPath expression. Scrapy is an application framework for searching and extracting data from web pages (Myers and McGuffee, 2015). Scrapy extracts data from a web page by defining the template for each page to be extracted (Myers and McGuffee, 2015). The advantage of web scraping template technique is the high accuracy of extracting results, because the template definition is as accurate as possible (Vargiu and Urru, 2012; Ferrara *et al*., 2014) (Myers and McGuffee, 2015).

Levenshtein Distance

Levenshtein Distance (LD) is the measure of syntactic similarity between two strings, the source string (s) and the target string (t) (Ristad and Yianilos, 1998; Haldar and Mukhopadhyay, 2011). The LD value is derived from the number of deletion, insert, or substitution operations required to convert string s to t. The greater the Levenshtein distance, the more dissimilarity between compared strings. Mathematically, the Levensthein distance between 2 strings a and b is expressed in Equation 1:

$$LD_{a,b}(i,j)=\begin{cases}0 & if\ i=j=0,\\ \min\begin{cases}LD_{a,b}(i-1,j)+1\\ LD_{a,b}(i,j-1)+1\\ LD_{a,b}(i-1,j-1)+1_{a_i\neq b_j}\end{cases} & else,\end{cases} \quad (1)$$

Where:
a = The first string
b = Second string
i = First string length
j = The length of the second string

For example, the LD value between "Verify invoice" and "Verification invoice" is 7 because of substitution "*y*" and "*i*" and the addition of "cation".

Vector Space Model

Vector Space Model (VSM) uses vectors to represent terms in documents (Berry *et al.*, 1999). In VSM, a document collection can be represented as a term-document matrix (or term matrix frequency) (Berry *et al.*, 1999). Each cell in the matrix corresponds to the weight counted from the occurrence of a term in a document (Berry *et al.*, 1999).

Text Preprocessing

Text Preprocessing is a process of transforming unstructured data into structured data according to user need for further mining processes (sentiment analysis, summary, clustering of documents, etc.) (Uysal and Gunal, 2014). The steps of the text preprocessing as follows (Uysal and Gunal, 2014):

Parsing

Document parsing splits document structures into separated components.

Lexical Analysis

It is also popular as tokenization. Tokenization is the process of cropping each input string into tokens. In principle, this process is to separate each word that compiles a document.

Stop Word Removal

In this process, stop words are removed because they are common words and not represented information in a document. The database of the stop words collection can be used to remove these kinds of words found in the tokenization results.

Phrase Detection

In this step, input data capture is not only word tokens, but also two or more words into phrases.

Stemming

Stemming is the process of converting a word into its root word.

Research Methodology

In this research, a system was built to collect, to pre-process and to compare the 11 university website menus labels. Then the comparison results were analyzed and evaluated for proposing labeling system for university websites in Indonesia. The research methodology can be seen in Fig. 1.

Data Collection

First, it was defined university websites which were used as inputs of our proposed system. In this research, the researchers crawled 11 menus labels of university websites by using scrappy. The 11 websites were ten best university websites in Indonesia ranked by webometrics on 5th September 2016 at 4:22 PM and also one additional university website, which is the working place of one of the experts in this study. The Indonesian well-known names of these universities are UGM, UI, ITB, IPB, UB, UNPAD, UNDIP, UNAIR, UNUD, UNSYIAH and TEL-U. The aim of this research was to propose the automatic comparison of university website labels; thus the 11 university websites were sufficient to compare automatically and to analyze the comparison results.

From these 11 websites, it was only the navigation labels in website menus was crawled. In addition, the documents of the Institutional Accreditation of University (IAU) and the National Standard of Higher Education (NSHE) were used to create index terms to evaluate the results of the label comparison of websites.

Data Preprocessing

This step preprocessed data of 11 website labels and the two documents. For the website labels, there were three processes. First was removing redundant labels, second was labeled error checking using Indonesian stemmer and third was the manual fixing of label writing errors. Preprocessing documents were aimed to parse the documents creating index terms.

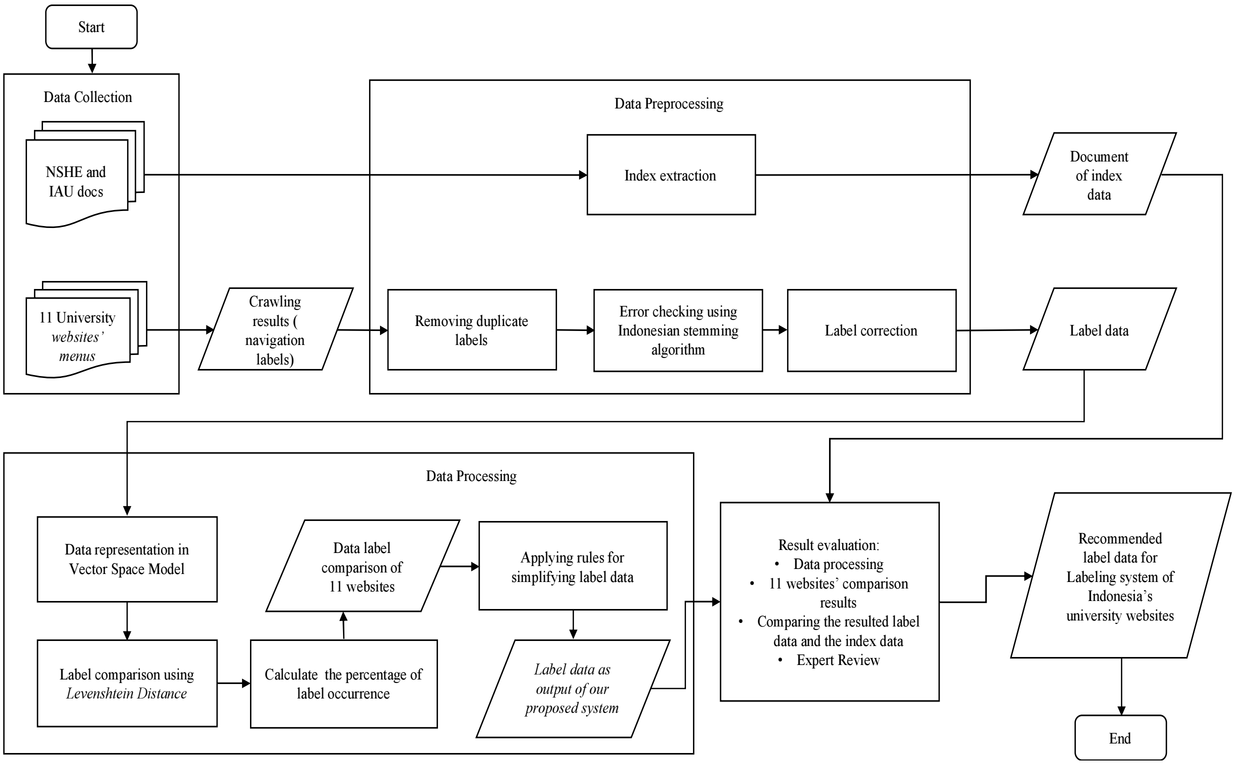
Start
Data Collection
NSHE and IAU docs
11 University *websites' menus*
Crawling results (navigation labels)
Data Preprocessing
Index extraction
Removing duplicate labels
Error checking using Indonesian stemming algorithm
Label correction
Document of index data
Label data
Data Processing
Data representation in Vector Space Model
Label comparison using *Levenshtein Distance*
Calculate the percentage of label occurrence
Data label comparison of 11 websites
Applying rules for simplifying label data
Label data as output of our proposed system
Result evaluation:
• Data processing
• 11 websites' comparison results
• Comparing the resulted label data and the index data
• Expert Review
Recommended label data for Labeling system of Indonesia's university websites
End

Fig. 1: Research methodology consisting of data collecting, data preprocessing, data processing and result evaluation

Data Processing

Data processing compared all university websites' labels to yield percentage label appearance on each website. These results can inform the similarity and ranking of each label in the 11 university websites. Data processing began with combining all label data from 11 university websites then formed a label-document matrix using Vector Space Model (VSM) representation. Each document in the matrix represents all labels in each university. Levenshtein Distance was used to compare between all labels and labels in each document in the matrix. The results were the percentage of occurrence on each label in each university website and on all websites.

Based on the results of the percentage of occurrence of each label on all websites, it was created rules that could be used for simplifying and evaluating labels between two labels compared having the same first word. The aim was to ensure the final label having higher occurrence among other labels with the same first word, for example, Fakultas (English: Faculty) vs Fakultas Teknik (Faculty of Engineering). The first has a higher occurrence than the latter then the second is not considered as the output of the system. The rules are as follows.

For labels consist of more than one word then matching the first word with other labels.

If there were other labels having similar name for their first word and they appeared more than one time, then the proposed system created a new label named with the first word and excluded the other labels that the first words had the same name.

The new label then was re-compared to calculate its similarity for both labels on each website and on all websites (average similarity).

There were the percentage of new labels for each website and all websites.

The average similarity of the new labels then was compared to the previous labels that were excluded.

If the new label had lower similarity value, then the excluded labels would be put back in the proposed system.

The new labels would be saved for further analysis in the comparison with the IAU and the NSHE documents and expert review.

The results of the label comparison process and the implementation of the above rules were suggested as a labeling system of university website in Indonesia.

Data Analysis

The aim of data analysis was to evaluate the results of the proposed system. First, it was evaluated the results of the label corrections on the data preprocessing and the results of the label comparisons. Then, it was compared the results of website label comparison and the label indexes of the IAU and the NSHE documents. The comparison of the results with these two documents was aimed to check the suitability of the label comparison results with the two documents representing the national standard governing university operation in Indonesia.

The Levenshtein distance was used to count the occurrence of each label in the IAU and the NSHE documents. Lastly, the interview and discussion of the label comparison results were conducted with four website content experts. Two of them are directly responsible for managing two university website contents and the others are experienced website developers. The aim of the expert reviews was to complement the previous approach uncovering important labels that did not exist in the indexes of the IAU and the NSHE documents. The reviewers informed the degree importance of the comparison results of labels.

Results and Discussion

In this section, it discusses and analyzes data preprocessing, the comparison results of 11 university website labels, the comparison between the results of the label comparison and the indexes of the IAU and the NSHE documents and expert review on the results of the label comparison.

Analysis of Data Preprocessing

The preprocessing results of data labels on each university website are presented in Table 1 listed from the highest ranking of the Webometric to the lowest.

It is shown in Table 1 that there are many labels removed because of duplicated values as the results of crawling processes.

Based on Table 1, it can be seen that on each university website exist incorrect words after being checked by Indonesian stemming algorithm (Adriani *et al*., 2007). But the errors were not only due to writing errors, but also the use of abbreviated words (such as common terms used in each university), English words and other words that could not be detected using the Indonesian stemming algorithm. After it was corrected all wrong words, these labels became the inputs of the label comparison process among 11 university websites.

Analysis of the Results of 11 University Website Labels' Comparison

There were two scenarios for comparing the labels. The first scenario was the comparison of labels before and after the correction of the word label. The example of the results can be seen in Fig. 2 and 3, respectively. The second scenario was to apply the rules (see Data Processing section).

Table 1: Preprocessing results (after label correction)

	Labels			
University	Crawled	Preprocessed	Correct	Wrong
UGM	693	33	29	4
UI	7132	123	92	31
ITB	2759	89	65	24
IPB	5372	68	51	17
UB	4311	199	77	122
UNPAD	320	32	15	17
UNDIP	227	224	122	102
UNAIR	7921	92	68	24
UNUD	13860	126	42	84
UNSYIAH	740	37	18	19
TEL-U	14632	211	95	116

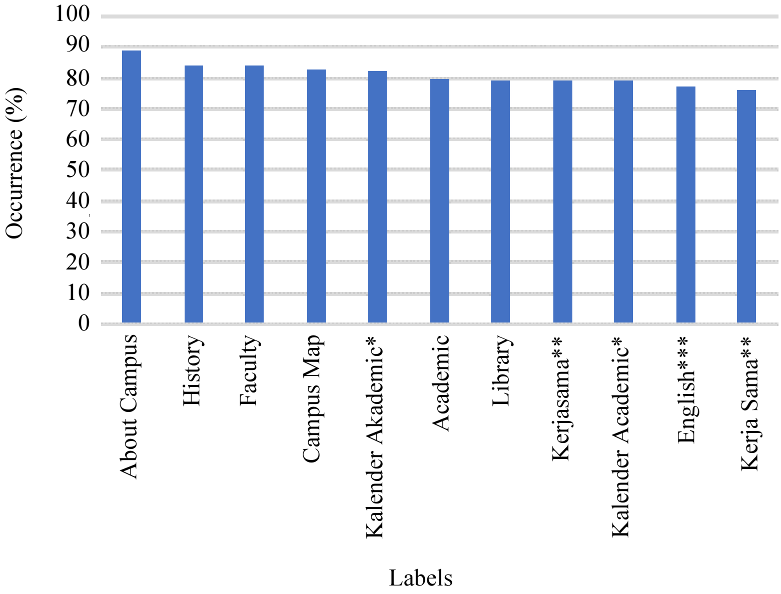

Fig. 2: A Sample of 11 labels with highest percentage occurrence before word correction
Note: All labels are translated into English, except: * has English word 'Academic Calendar', ** has English word 'Cooperation' and *** was not translated into English

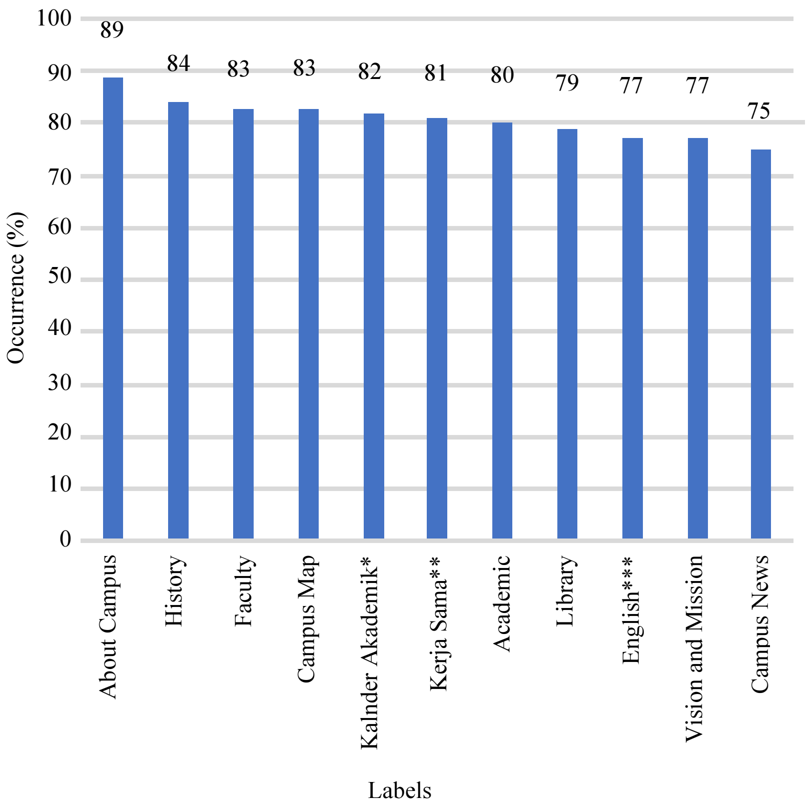

Fig. 3: A Sample of 11 labels with highest percentage occurrence after word correction
Note: All labels are translated into English, except: * has English word 'Academic Calendar', ** has English word 'Cooperation' and *** was not translated into English

In Fig. 2., it was found several labels that are similar, - but it was due the writing errors considered different by our proposed system, such as the labels of ‘Kalender Akademik’ and ‘Kalendar Akademik’ (in Indonesian and the correct words are ‘Kalender Akademik’; the meaning in English is Academic Calendar) and ‘Kerjasama’ and ‘Kerja Sama’ (in Indonesian and the correct words are ‘Kerja Sama’; the meaning in English is Cooperation). These words had the high percentage of occurrence. Because this research used the Levenshtein distance which calculated the value of similarity based on the letters contained on each label being compared, the correction of the wrong word was required on label preprocessing before comparison to increase the percentage occurrence of correct labels.

The label corrections increased the percentage of correct label occurrence (see the example in Fig. 3). In addition, after label correction, the amount of label data decreased from 865 labels to 854 labels (see Table 2). For example, the correct label of cooperation in Indonesia is ‘Kerja Sama’. After word correction process, correcting ‘Kerjasama’ into ‘Kerja Sama’ (both refer to the same meaning ‘Cooperation’ in English), the percentage increased form below 80% into above 80%. Previously there were six websites using the word ‘Kerjasama’ instead of ‘Kerja Sama’.

In the second scenario, to obtain the label having higher occurrence indicating a common label among university websites, it was analyzed the results of the implementation of the rules from previous data processing evaluation between two labels having the same first word (see Data Processing section). In Fig. 3, the label named ‘Fakultas’ (English: Faculty) has the percentage of occurrence above 80%, but in 854 labels (as the results of label correction) there are other labels with the first word beginning with ‘Fakultas’, for example ‘Fakultas Teknik’ (English: Faculty of Engineering). There were other labels having the same condition with the *Faculty* case. Therefore, the rules number 1 to 4 was implemented to reduce the number of similar labels and gained total 456 labels (Table 2).

However, in 456 labels, it was found that there were some labels that previously had high percentage of occurrences removed after applying the rules number 1 to 4. There were some labels such as 'Tentang Kampus' (English: About Campus), Kalender Akademik (English: Academic Calendar) and 'Peta Kampus' (English: Campus Map) and other similar labels after deleting words following the first word becoming 'Tentang' (in English: About), 'Kalender' (in English: Calendar) and 'Peta' (in English: Map), respectively, having lower percentage of occurrence. Therefore, the rules number 5 to 7 was implemented to gain higher percentage of occurrence comparing the original and modified labels. The higher percentage of occurrence of labels indicates important labels in the proposed system. Total labels from this process are 706 and become the output of website labels' comparison for the next processes in our research. The number of labels after word correction (854 labels) becomes 706 after implementing the rules (Table 2).

Analysis of the Results of the Labels' Comparison with the Institutional Accreditation of University (IAU) and the National Standard of Higher Education (NSHE) Index Terms

The comparison result of label data is then be compared with index list obtained from the IAU (BNAHE, 2015) and NSHE (MRTHE, 2015) documents. The purpose of comparison is to evaluate the labels generated by the important points contained in the policies related to higher education in Indonesia. A sample result of the label comparison with the indexes of the document of the IAU and NSHE can be seen in Fig. 4.

Based on Fig. 4 there are several examples of labels that have 100% similarity with the IAU (BNAHE, 2015) and NSHE (MRTHE, 2015) documents such as 'Administration', 'Academic', 'Accreditation' and other labels. Based on Fig 4, it is also seen that there is one label that has high percentage of the comparison results between labels that are 80%, also it has 100% resemblance to the existing indexes in documents (BNAHE, 2015) and (MRTHE, 2015) as in the label 'Academic'. It is also found one label that have percentage less than 60% but has 100% resemblance to the existing indexes in documents (BNAHE, 2015) and (MRTHE, 2015) such as the 'Administration' label. Based on the comparison of the labels with the document index, the label 'Administration' is one of the important labels even though it has low percentage. Therefore, the use of the national higher education documents can help to consider what labels need to be displayed on the university website to conform to the higher education standards in Indonesia.

Table 2: Number of labels generated from comparison processes

Label comparison processes	Total of Labels
After word correction	854
Applying the rules-step 1 - 4	456
Applying the rules-step 5 - 7	706

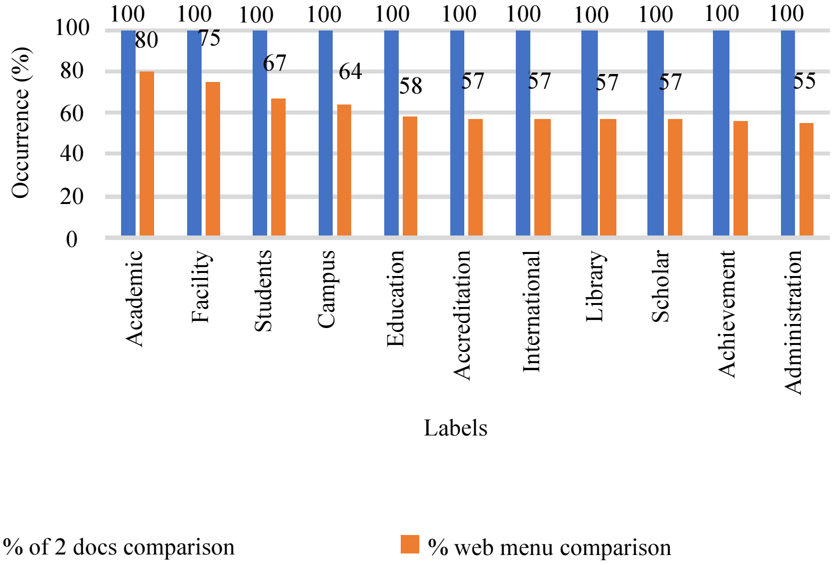

Fig. 4: A Sample of labels comparison between the IAU and NSHE documents and the web menus **Note:** All labels are translated into English

Analysis of the Results of the Labels' Comparison with Experts

To evaluate actual selections of labels used in the university website, it was involved four website developer experts in this research. Two of them are specifically in charge of managing contents in the university website. The evaluation was conducted by showing the program and the results to the experts. It is also asked several questions concerning the labels and our system performance.

Based on the evaluation, only 25 labels out of 706 labels (Table 2) were not used in their university websites. The examples of non-selected labels are included 'LPSE Campus (in English: Campus' Electronic Procurement Service)', 'Student Blog', 'Webmail', 'Student Email' and 'Staff Blog'. Those labels relate to specific internal businesses operations (such as students, staffs, employees, etc.), thus not all universities use the labels. Other unselected labels are not used in their websites but are used in sub-domains of the website.

There are some labels having small percentage based on the comparison results, but it is important to be used in their university websites. For example, the occurrence percentage of 'Helpdesk' label is 35% while one expert considered it as important label. This shows that some universities may not have the same point of view about the importance of certain labels representing entities in the universities.

Meanwhile regarding to the system contributions, all experts agreed that the system helped them for comparing website labels. The process was faster than manual comparison and it enabled them to compare as many websites as they want. However, it still needed to be evaluated which labels were relevant to their business processed and could be used on their university websites.

Conclusion

The label comparison results yield the percentage of label occurrence on each website and overall university website to build the labeling system of university websites in Indonesia. With proper data preprocessing of the label data, the percentage of occurrences to some labels is better compared without data preprocessing. By the rules implementation, to simplify the comparison of label data with the same first word, it can reduce the number of labels of comparison results. The labels comparison with indexes in the Institutional Accreditation of University (IAU) and the National Standard of Higher Education (NSHE) documents is helpful to consider the labels on the university website to conform to the higher education standards in Indonesia.

Based on experts' judgments, the proposed system can be used as recommendation of labeling systems and reduce resource needed to compare the labels of university websites. Hopefully, web content developers can also use the proposed system to design labeling system for university website in Indonesia, - by selecting top comparison results and common labels and by analyzing other labels according to the needs of their internal business processes.

Related to web content analysis, the contribution of this research is to automatically generate coding units as labels representing the content of information. The quantitative analysis (Herring, 2010) as expected is measured by occurrence of labels. High frequency of occurrence indicates common labels used by different websites. Because the processes of label comparison and generation are automatic, it can reduce the bias in a coding process and processing time compared to manual coding (McMillan, 2000). This proposed system is not cover limited (McMillan, 2000) and selected site hierarchy but it can cover complete hierarchy of compared websites. In addition, this proposed system can also dynamically compare websites based on users' inputs.

Preprocessing label data using the Indonesia stemming algorithm can be used to check writing errors for Indonesian language labels but cannot be used for English language labels and special terms used on each campus. Further work needs to add features on data labels preprocessing to handle English labels used as well as specific terms. We also suggest the implementation of semantic similarity matching to complement existing syntactic similarity matching, which aims to uncover compared labels having semantic similarity values.

Acknowledgment

The authors would like to thank Monte, Rama, Jefry and Arief for developing http://webscompare.com/, the website's version of this proposed system with more generic and automatic functionalities. The authors also express its gratefulness to Yahdi for his help during this research.

Funding Information

This research is supported by Telkom University using University's Excellent Grant scheme Batch II-2017.

Author's Contributions

I Kadek Aditya Cahaya Putra: Conducted the experiment and wrote the research report.

Dana Sulistiyo Kusumo, Anisa Herdiani and lndra Lukman Sardi: Designed this research, wrote and revised this paper.

Ethics

There are no ethical issues of this paper because all the data crawled from open and public websites.

References

Adriani, M., J. Asian, B. Nazief, S.M.M. Tahaghoghi and H.E. Williams, 2007. Stemming Indonesian: A confix-stripping approach. ACM Trans. Asian Lang. Inform. Process., 6: 1-33. DOI: 10.1145/1316457.1316459

Berry, M., Z. Drmac and E. Jessup, 1999. Matrices, vector spaces and information retrieval. SIAM Rev., 41: 335-362. DOI: 10.1137/S0036144598347035

BNAHE, 2015. Institutional Accreditation of University. Board of National Accreditation for Higher Education. https://banpt.or.id/instrumen/APT.rar

Cowie, J. and W. Lehnert, 1996. Information extraction. Commun. ACM, 39: 80-91. DOI: 10.1145/234173.234209

Djonov, E., 2007. Website hierarchy and the interaction between content organization, webpage and navigation design: A systemic functional hypermedia discourse analysis perspective. Informat. Design J., 15: 144-162. DOI: 10.1075/idj.15.2.07djo

Ferrara, E., P. De Meo, G. Fiumara and R. Baumgartner, 2014. Web data extraction, applications and techniques: A survey. Knowledge-Based Syst., 70: 301-323. DOI: 10.1016/j.knosys.2014.07.007

Gullikson, S., R. Blades, M. Bragdon, S. McKibbon and M. Sparling *et al*., 1999. The impact of information architecture on academic web site usability. Electronic Library, 17: 293-304.

Haldar, R. and D. Mukhopadhyay, 2011. Levenshtein distance technique in dictionary lookup methods: An improved approach. ArXiv:1101.1232.

Herring, S.C., 2010. Web Content Analysis: Expanding the Paradigm. In: International Handbook of Internet Research Hunsinger, J., L. Klastrup and M. Allen (Eds.), Springer Netherlands, pp: 233-249. DOI: 10.1007/978-1-4020-9789-8_14

Larson, K. and M. Czerwinski, 1998. Web page design: Implications of memory, structure and scent for information retrieval. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Apr. 18-23, ACM, USA Los Angeles, California, USA, pp: 25-32. DOI: 10.1145/274644.274649

McMillan, S.J., 2000. The microscope and the moving target: The challenge of applying content analysis to the world wide web. Journalism Mass Commun. Quarterly, 77: 80-98. DOI: 10.1177/107769900007700107

MRTHE, 2015. National standard of higher education number 44. Ministry of RTHE.

Myers, D. and J.W. McGuffee, 2015. Choosing scrapy. J. Comput. Sci. Coll., 31: 83-89.

Olston, C. and M. Najork, 2010. Web crawling. Foundat. Trends Inform. Retrieval, 4: 175-246. DOI: 10.1561/1500000017

Ristad, E.S. and P.N. Yianilos, 1998. Learning string-edit distance. IEEE Trans. Patt. Anal. Mach. Intell., 20: 522-532. DOI: 10.1109/34.682181

Rosenfeld, L. and P. Morville, 1998. Information architecture for the world wide web. O'Reilly and Associates, Inc., Sebastopol, CA, USA.

Uysal, A.K. and S. Gunal, 2014. The impact of preprocessing on text classification. Inform. Process. Manage., 50: 104-112. DOI: 10.1016/j.ipm.2013.08.006

Vargiu, E. and M. Urru, 2012. Exploiting web scraping in a collaborative filtering- based approach to web advertising. Artificial Intell. Res., 2: 44-44. DOI: 10.5430/air.v2n1p44