The Development of Integrated Bali Tourism Information Portal using Web Scraping and Clustering Methods

To cite this article: I G M Darmawiguna et al 2019 J. Phys.: Conf. Ser. 1165 012010

View the article online for updates and enhancements.
The Development of Integrated Bali Tourism Information Portal using Web Scarping and Clustering Methods

I G M Darmawiguna\textsuperscript{1a}, G A Pradnyana\textsuperscript{1b} and G S Santyadiputra\textsuperscript{1c}

\textsuperscript{1}Informatics Education Department, Faculty of Engineering and Vocational Universitas Pendidikan Ganesha, Bali, Indonesia

E-mail: \textsuperscript{a}mahendra.darmawiguna@undiksha.ac.id, \textsuperscript{b}gede.aditra@undiksha.ac.id, \textsuperscript{c}gsaindras@undiksha.ac.id

Abstract. This study aims to develop integrated tourism information portal prototypes in Bali by using web scrapping and Clustering methods. This study will use the waterfall development model. This study focused on the stages of needs analysis, system design for Tourism information portals in Bali and prototype development at the implementation stage. The initial stage of the web scrapping process is to do web searching in the form of url requests based on the keywords entered. After the data from the web scrapping process is collected, the next step is text processing which consists of several processes, namely parsing or tokenization, stemming in the form of root word search process, and stop word removal or removal of non-essential words. Before the clustering process begins, each word will be given weighting with the TF-IDF method which will form the vector space model as a representation of a text document. Similarity between vector space models will be calculated using the cosine similarity method. After the weighting process, the last step is to cluster the words which then continue to group the websites based on the results of clustering. The results of this study are a website portal that can display tourism information in Bali.

1. Introduction

Tourism is an important economic sector in Indonesia. Tourism is one of the national strategic sectors as a contributor to foreign exchange, an instrument of equitable development and improvement of people's welfare. In 2017, the tourism sector has consistently become a priority program, namely Indonesia's tourism development "Wonderful Indonesia". Tourism is one of the 5 (five) 2017 priority development sectors, namely food, energy, maritime, tourism, industrial areas and Special Economic Zones (KEK), as stated in the RKP 2017. \cite{1}

Bali is one of the islands and provinces in Indonesia that has become a tourist destination for both local and foreign tourists. Bali has various types of tours ranging from nature, culture, to culinary. In 2017, Bali was chosen as the best tourist destination in the world. This award is given by TripAdvisor through Travelers' Choice Award 2017. One indicator of Bali's success as the best tourist destination is promotion. In addition to entering the Wonderful Indonesia video profile segment carried out by the Ministry of Tourism, the local government in Bali is very aggressively promoting. The Badung Regency Regional Tourism Office, Bali, together with the Denpasar City Tourism Promotion Agency held a "sales mission" for the first time in the world history of Indonesian tourism in the City of Manchester, England, which was attended by around 25 top agents. Badung Tourism Office aims to increase by 24\% in 2018. \cite{2}
Information related to Bali as a tourist destination can be accessed on various sites. If the search is done through a search engine such as google.co.id with the keyword "tourism object in Bali" then there will be many pages describing the tourist attraction in Bali. The number of pages that appear after searching in search engines has positive and negative sides. On the positive side, many pages that appear will be able to complement each other about information about tourism objects in Bali. On the negative side, many pages that appear will cause information overload that is the amount of information received by humans so that it is difficult to process it. With the existence of a lot of information, humans are required to be able to sort and combine information obtained from various sources so that it becomes a single unit that is accurate and useful. In addition, many sites are personal blogs from someone whose truth can not be ascertained.

Based on the above problems, it is necessary to develop a media that can collect this information and integrated in a portal. This information portal will integrate information from various sites which are then grouped according to the specified categories. The presentation of information in the portal that will be developed will be made more concise but complete and accurate. The technique used in making this portal is web scraping and clustering methods. Web scraping is the retrieval of semi-structured documents from the internet, generally in the form of web pages in markup languages such as HTML or XHTML and analyzing these documents for specific data from the page. Web scraping has become the most commonly used technique for retrieving data from web pages automatically. Web scraping is related to web indexing which is a universal technique used by almost all search engines. Some studies that implement web scraping include "Implementation of Web Extraction for Hadiths Translated in Indonesian". This study aims to extract, collect, and represent knowledge from the Hadith obtained from several sources on the internet. [3]Another research is "Development of Web Scraping Software for News Sites and Summarizing News". This study aims to create applications that can retrieve news data from various sites. The news data can be summarized which is then accommodated in one place which is then displayed to the reader in the mobile application. [4]Another research related to web scraping is "Application of Web Scraping Techniques on Scientific Article Search Engines". The purpose of this research is to create search engines that implement web scraping to extract scientific journal articles from a number of academic portals both from Indonesia and abroad including the portal Garuda, Indonesian Scientific Journal Database (ISJD), and also Google Scholar. [5]
The next process after doing web scraping is grouping tourism information in Bali by using clustering method. This grouping will be generated automatically. Some studies that implement clustering include "Clustering Indonesian News Articles Using Unsupervised Feature Selection". This study aims to analyze the news by automatically grouping news that has similarities. [6] Another study is "K-Means Based on Neighbors with Latent Semantic Indexing for Documents Clustering." [7] Other research is "Semi-Supervised Text Categorization using Recursive K-Means". This study aims to classify text documents using a semi-supervised learning model. This model works with divide and conquer strategies. [8] Based on the background above, the researchers tried to develop a media that will provide information about tourism in Bali whose information is collected from various sources of the site's pages. The research title is "The Development of Integrated Bali Tourism Information Portal using Web Scraping and Clustering Methods".

2. Method

The research method used in this study is a development research method with a Waterfall model. The stages of the waterfall model are requirements, designs, implementations, verifications, and maintenance, as seen in Figure 2.

![Waterfall Model by Pressman (2015)](image)

Based on the problems that have been studied and described in the background, the system needs analysis is divided into two, namely functional requirement analysis and non-functional requirement analysis. Functional needs of Integrated Information Portal Tourism in Bali Using Web Scraping and Clustering Methods are as follows: (1) Admin login to the system; (2) Admin can enter keywords and the amount of web that will be scrapped to generate a list of google search results (title, web address, and summary of content); (3) Admin can choose information obtained from Google search scrapping results; (4) Admin does the clustering process and assigns categories; (5) The system can scrap google search content; The system can do the pre-processing process of the scrapping results; (6) The system can do pre-processing (stemming, removing stop words, and tokenization); (7) The system can determine the weight the results of tokenization using the TF-IDF algorithm; (8) The system can do data clustering with the K-Means algorithm; (9) The system can display clustering results.

Non-functional needs of Integrated Tourism Information Portal in Bali Using Web Scraping and Clustering Methods are as follows: (1) Users can use the Integrated Tourism Information Portal in Bali properly (Usability); (2) Users can access the Integrated Tourism Information Portal in Bali with any device (Portability); (3) Integrated Tourism Information Portal in Bali has good data security (Reliability) (4) Integrated Tourism Information Portal in Bali has a manual to facilitate users in using the system (Supportability).

The initial stage of the web scrapping process is to do web searching in the form of url requests based on the keywords entered. After the data from the web scrapping process is collected, the next step is text processing which consists of several processes, namely parsing or tokenization, stemming in the form of root word search process, and stopword removal or removal of non-essential words. Before the clustering process begins, each word will be given weighting with the R-TF-IDF method which will form the vector space model as a representation of a text document. Similarity between vector space models will be calculated using the cosine similarity method. After the weighting process, the last step is to cluster the words which then continue to group the websites based on the results of clustering.
3. Results
The research method used in this study is a development research method with a Waterfall model. The stages of the waterfall model are requirements, designs, implementations, verifications, and maintenance.

3.1 Web Scraping and Clustering Design
The flow of web scraping techniques and clustering methods designed in the development of integrated information portal applications can be seen in Figure 3.

![Figure 3. Flow of Web Scraping Techniques and Clustering Methods in Integrated Tourism Information Portal in Bali](image)

3.1.1 Web Scraping
The scraping process begins by determining the initial parameters needed in the search process on the Google search engine. The initial parameter specified is the keyword for searching on google and the number of websites to be retrieved based on the search results.

The process of retrieving data from Google is done using the standard HTML Document Object Model (DOM). From the extraction process with HTML DOM, websites will be obtained based on the keywords entered, complete with the attributes of title, short description, and URL of the website. DOM is an interface that describes XML or HTML documents. DOM is defined as the logical structure of documents and how documents can be accessed and manipulated. DOM is a standard object model for HTML and XML that is platform independent. A web browser does not have to use DOM to display HTML documents. But DOM is needed by JavaScript which will change the appearance of a website dynamically. In other words, DOM is the way JavaScript looks at an HTML page.

3.1.2 Web Clustering
After all the websites that match the keywords have been successfully obtained, the next step is to take the content from each website to do pre-processing stages. Because of the large number of sentences or words of a website, retrieving content is limited by the number of characters. Pre-processing stages begin with the removal of punctuation followed by the stemming process. Stemming is the process of changing the word into a basic word. There are several Indonesian Stemming algorithms, one of them is Sastrawi. Sastrawi stemmer applies an algorithm based on Nazief and Adriani, then enhanced by the CS (Confix Stripping) algorithm, then improved again by the ECS (Enhanced Confix Stripping) algorithm then improved again by Modified ECS algorithm. The next process is to do tokenization. Tokenization is the process of solving words in a sentence. The final stage of pre-processing is to
eliminate stop words. Stop word is a word that is ignored in processing, in this system words that include stop words are stored in a stopwords table in the database. The main characteristics in selecting stop words are usually words that have a high frequency of occurrence such as conjunctions such as "and", "or", "but", "will" and others. The main objective in the application of the stopwords removal process is to reduce the number of words in a document which will affect the speed and performance of the next process.

In this research K-Means algorithm was selected to perform clustering process and to cluster websites, all websites will be represented in the form of vector space models (VSM). VSM is the form of a matrix containing the weight of all words on each website. This weight states the importance or contribution of words to a website and website collection. The importance of a word can be seen from the frequency of appearance on the website. In this study, the weighting process is done by applying the method of R-TF-IDF[9]. Calculation R-TF-IDF is different from the classic tf-idf where in R-TF-IDF also take into account the long normalization issue websites and distribution of occurrence of a word in other websites that affect the value of the weight of the end of a word[9]. Calculation R-TF-IDF is calculated by the following equation:

\[
w_{ij} = \frac{f_{ij}}{\sum_{i} f_{ij}} x \ln \left( 1 + \frac{n_i}{n} \right) x \ln \left( \frac{1}{\sqrt{\sum_{i} f_{ij}}} \right).
\]

\(w_{ij}\) is the weight of the \(i\)-th word in the website \(j\), \(f_{ij}\) is the frequency of the \(i\)-th occurrence of the word in the website \(j\), \(n\) is the total number of websites in the corpus whereas \(n_i\) is the number of websites that contain the word \(i\)-th.

| Website ID | VSM                        |
|------------|----------------------------|
| Web-1      | d:1; a:1,51; e:0,85; f:1,4; g:1,06 |
| Web 2      | b:4; a:1,2; e: 1,2 ; g:1,40 |
| Web 3      | a:1.7; b:1; c:2,3; d:1,5; e:1,7; f:1,8 |
| ...        | ...                        |
| Web n      | e:2; f:1,5; g:3; a:1,2; b:5; d:2,3 |

After successfully forming VSM from each website, the clustering process is performed with the k-means algorithm. The K-Means algorithm works as follows. Given a set of web objects in VSM form and a pre-specified number of clusters \(k\), \(k\) web objects are randomly selected to initialize \(k\) clusters, each one being the centroid of a cluster. The remaining objects are then assigned to the cluster represented by the nearest or most similar centroid. Next, new centroid are re-computed for each cluster and turn all documents are re-assigned based on the new centroid. This step iterates until a converged and fixed solution is reached, where all web objects remain in the same cluster after an update of centroids. The generated clustering solutions are locally optimal for the given data set and the initial seeds. Different choices of initial seed sets can result in very different final partitions. The K-Means algorithm works with distance measures which basically aims to minimize the within-cluster distances. In this research, the distance or similarity between websites can be known by applying the cosine similarity method[10].

The calculation of cosine similarity using the weighting calculation of a web document can be stated in the following formula:

\[
\text{CosSim}(d_i, q_j) = \frac{q_j \cdot d_i}{\|q_j\| \|d_i\|} = \frac{\sum_{j=1}^{n} (q_{ij} \cdot d_{ij})}{\sqrt{\sum_{j=1}^{n} (q_{ij})^2} \sqrt{\sum_{i=1}^{n} (d_{ij})^2}}.
\]
$q_{ij}$ is the weight of the $j$-th word in the website $i$, $d_{ij}$ is the weight of the $j$-th word in the website $i$

Based on the examples in Table 1, the similarities between Web1 and Web2 can be calculated as follows:

$$\text{CosSim}(Q,D) = \frac{(1.51 \times 1.20 + 0.05 \times 1.20 + 1.05 \times 1.40)}{\sqrt{1.51^2 + 0.05^2 + 1.05^2 \times 1.20^2 + 1.40^2}}$$

$$= \frac{4.384}{\sqrt{2.031 \times 2.245}} = \frac{4.384}{4.56} = 0.96$$

The number of clusters are defined by the user, where

3.2 System Design

Based on the problems that have been studied and described in the background, the system needs analysis is divided into two, namely functional requirement analysis and non-functional requirement analysis.

![Figure 4. Use case diagram of Integrated Bali Tourism Information Portal](image)

![Figure 5. Managing Stop Words](image)
3.3 Development

The Integrated Bali Tourism Information Portal system is built using the PHP programming language, using a MySQL database, and the Apache web server. It is shown in figure 5, users can manage stop words lists that are used by the system to pre-processing scrapping text data. Users can search, add, change and delete a word for the stop word list.

Figure 6 is a display of the system interface when the user determines the parameters for web scrapping. There are 2 parameters used in the scrapping process, which are keyword and the number of web pages that will be scrapped.

![Figure 6. Web Scrapping Interface](image)

In figure 7, there will be the results of the scrapping process according to the keywords and number of websites will be stored in the database and displayed to the user. The information displayed is the title, URL and brief description of the website. The next process is to scrap website content according to the URL obtained.

![Figure 7. Results of Web Scrapping](image)

After, scrapping the content of the web, then the next step is pre-processing process and tokenization along with weighting every word.
The result of clustering can be seen in the figure 9.

4. Conclusion
The research method used in this study is a development research method with a Waterfall model. The stages of the waterfall model are requirements, designs, implementations, verifications, and maintenance. Bali tourism integrated information portal that has been developed in this study consists of two main processes, namely web scraping process and clustering process. The process of scraping data from Google is done using the standard HTML Document Object Model (DOM). Before doing web clustering, the data from scraping is done by pre-processing data such as parsing, stemming, stopword removal, and finally the calculation of sentence weight using R-TF-IDF method. Clustering Web is done by applying the K-Means method by measuring web similarity using cosine similarity.
5. References

[1] M. A. Khumaedy, “Tahun 2017 Kita Genjot Sektor Pariwisata,” 6 Januari 2017. [Online]. Available: http://setkab.go.id/tahun-2017-kita-genjot-sektor-pariwisata/.

[2] T. Tresnady, “Bali Promosi Wisata Besar-besaran di Manchester,” 12 November 2017. [Online]. Available: https://www.suara.com/lifestyle/2017/11/12/065244/bali-promosi-wisata-besar-besaran-di-manchester.

[3] Z. Zaira, “Implementasi Ekstraksi Web untuk Hadits yang Diterjemahkan dalam Bahasa Indonesia,” Universitas Indonesia, Jakarta, 2011.

[4] Y. A. Setiawan, “Pembangunan Perangkat Lunak Web Scraping untuk Situs Berita dan Peringkas Berita,” Universitas Atma Jaya Yogyakarta, 2015, 2015.

[5] A. Josi, L. A. Abdillah and Suryayusra, “Penerapan Teknik Web Scraping pada Mesin Pencari Artikel Ilmiah,” Jurnal Sistem Informasi, vol. 5, pp. 159-164, 2014.

[6] D. P. Langgeni, Z. A. Baizal and Y. F. A.W., “Clustering Artikel Berita Berbahasa Indonesia Menggunakan Unsupervised Feature Selection,” in Seminar Nasional Informatika (SENMASIF), Yogyakarta, 2010.

[7] G. A. Pradnyana, I. P. G. H. Suputra and T. A. Wijaya, “K-Means Based on Neighbors with Latent Semantic Indexing for Documents Clustering,” in The First International Conference on Innovative Research Across Disciplines (ICIRAD), Denpasar, 2015.

[8] H. S. Gowda, M. Suhis, D. Guru and L. N. Raju, “Semi-Supervised Text Categorization using Recursive K-Means Clustering,” in International Conference on Recent Trends in Image Processing & Pattern Recognition, Singapore, 2016.

[9] D. Zhu and J. Xiao, “R-tfidf, a Variety of tf-idf Term Weighting Strategy in Document Categorization,” in Seventh International Conference on Semantics, Knowledge, and Grids, Beijing, 2011.

[10] G. A. Pradnyana and A. Djunaidy, “Metode Weighted Maximum Capturing untuk Klasterisasi Dokumen Berbasis Frequent Itemsets,” Jurnal Ilmu Komputer, vol. 6, no. 2, pp. 1-10, 2013.

[11] M. R. Ma’arif, “Integrasi Laman Web Tentang Pariwisata Daerah Istimewa Yogyakarta Memanfaatkan Teknologi Web Scraping dan Text Mining,” Teknomatika, vol. 9, no. 1, pp. 71-80, Juli 2016.

[12] P. Tan, Steinbach and V. Kumar, Introduction to Data Mining, United States: Pearson Education, 2006.

6. Acknowledgements

We express our gratitude for the support provided by Lembaga Penelitian dan Pengabdian Kepada Masyarakat, Universitas Pendidikan Ganesha for funding assistance for this research so that this research can be carried out and completed.