ISSN: 1978-0524

Recommendation system for web article based on association rules and topic modelling

Guntur Budi Herwanto[a,1,*] Annisa Maulida Ningtyas[b,2]

[a] *Department of Computer Science and Electronics, Universitas Gadjah Mada, Yogyakarta, Indonesia*
[b] *Departement of Health Information and Services, Universitas Gadjah Mada, Yogyakarta, Indonesia*
[1] *gunturbudi@ugm.ac.id*;* [2] *annisamaulidaningtyas@ugm.ac.id*

ABSTRACT

The World Wide Web is now the primary source for information discovery. A user visits websites that provide information and browse on the particular information in accordance with their topic interest. Through the navigational process, visitors often had to jump over the menu to find the right content. Recommendation system can help the visitors to find the right content immediately. In this study, we propose a two-level recommendation system, based on association rule and topic similarity. We generate association rule by applying Apriori algorithm. The dataset for association rule mining is a session of topics that made by combining the result of sessionization and topic modeling. On the other hand, the topic similarity made by comparing the topic proportion of web article. This topic proportion inferred from the Latent Dirichlet Allocation (LDA). The results show that in our dataset there are not many interesting topic relations in one session. This result can be resolved, by utilizing the second level of recommendation by looking into the article that has the similar topic.

Keywords:
Website
Recommendation
Topic Modelling,
Latent Dirichlet Allocation
Association Rule

I. Introduction

Currently, the world wide web becomes a knowledge base for numerous information around the world. A lot of industry starts to utilizing the benefit of the world wide web, including tourism industry. Gretzel [1] said the internet is the primary source of information in the domain of tourism. With the in-creasing information of the tourism destination on the Internet, the traveler is no longer dependent on travel agents [2]. The tourists prefer to seek information over the internet, even its an itinerary or individual reviews of each place. However, the internet can be overwhelming for the novice traveler due to various sources of information. The information that comes out is often not quite what they want. Recommendation system can be a tool to resolve the issue [3] and also provide useful information to help the user to make his choice. To be able to make a good recommendation, the system should be able to identify the user interests based on other users who have the same preferences [4]. The analysis of the browsing pattern of users can provide valuable information to the website owner. Such analysis can be done by applying data mining technique into web data.

Web mining is referred to as the application of data mining technique to the web data [5]. Web mining classified into three categories, namely the web usage mining, we content mining, and web structure mining [6]. Web usage mining is a process of picking up information from the user how to use websites [7]. The objective is to capture, model, and analyze the behavioral patterns and profiles of users interacting with a website [8]. One way to achieve this goal is by using association rule mining. Web usage association rule mining has long been a traditional data mining method for automatic extraction of potentially interesting information about the behavior of the website visitors from the web usage log files [9]. The most commonly used algorithms for association mining are Apriori [10]. While this algorithm can discover meaningful association pattern, the problem of too much execution time and generating too many items due to URL variation exists [11]. This issue can be resolved by utilizing content information from the web content mining process.

Web content mining is a process of picking up information from texts, images, and other contents [7]. Mostly, web content can be seen as a text data or documents. Several analysis such as topic modeling can reveal the theme from the documents [12]. Thematic information obtained from the content may also increase the understanding of the pattern generated from the web usage mining [13].

DOI: 10.26555/jifo.v12i2.a12629

In this study, we propose two-level recommendation, using association rule mining and topic modelling. We also aimed to prevent the twin problems that exists in Apriori by reducing the variations of itemset by utilizing topic modeling. Itemset that usually formed as a set of URL changed to become a set of the topic. In addition to reducing generated association rule, this combination can make the rule more meaningful because of thematic information that contained in the rule[14].

This paper is organized into 5 sections. Section 1 describes the background, current technique, and the case study. Section 2 describes the related work of web mining. Section 3 de-scribes the proposed system of this study. Section 4 show the result and research findings. Finally, Section 5 summarizes the conclusion.

II. Related Work

The growth of users and contents on the World Wide Web can be an enormous potential for some website owner, to seek for interesting user behavior in their industry niche including tourism [1]. User behavior can be seen by analyzing the usage data obtained from the web server. Examination of user actions in interacting with a website can offer insights causing to customization and personalization of a user's web practice [15].

Web Usage Mining can provide online recommendation effectively[16] . sunil [16] propose an architecture for online recommendation in Web Usage Mining System. The author presents the architecture of online recommendation in Web usage mining (OLRWMS) for improving the accuracy of clas-sification by the interaction between classifications, evaluation, and the current user activates and user profile in the online phase of this architecture. Another recommendation system proposed by Destyaputri [17]used three-level recommendation system based on association rule discovery, news articles in the same category, and similarity between news articles. By combining collaborative filtering approach and content-based filtering, experiment results show that the technique produces reliable news recommendation.

Association rule mining especially Apriori algorithm have been studied to uncover potential user browsing behaviors and creating recommendation [7] [18] [11]. Rawat [7] customizes Apriori Algorithm to become custom-built apriori. The goal is to find effective pattern analysis. The author found that analyzing web logs can not only provide an interesting pattern but also help in creating an adaptive website. More trying to uncover the disadvantages of the apriori algorithm by com-paring with another algorithm, the results appear that apriori has more execution time than the other algorithm. Lazcoretta analyzes the process of discovering association rules in this kind of big repositories and of transforming them into user-adapted recommendations by the two-step modified Apriori technique. The results show that their approach can provide better recommendation services by analyzing the behavior of a single user by all other users of web-based information systems.

Web log data is the primary source for analyzing user behavior in the web usage mining. Such models can be extended by adding web content as a source for analyzing user behavior especially for the website that has an extensive content like news portal. The combination of both sources can make a significant improvement on the recommendation of news articles [19]. Semantic analysis [17] or Topic Modeling [20] often be used to analyze the semantic meaning of the content, so it can be combined with log data to enrich the sense of the pattern.

In this research, we aim to combine the association rule mining with topic modeling. We use apriori as the algorithm for association rule mining, and enrich the association rule with Latent Dirichlet Allocation (LDA) [12]. Apriori allows us to discover the association among user clicks. Meanwhile, LDA makes it possible to identify several topics under what the user clicks. By combining these two, we can produce a good recommendation for the future users.

III. Method

Our recommendation system built consisting offline and online phase. Offline phase aims to generate the knowledge base that will be used as the basis for recommendations on the online phase. There are two levels of the knowledge base that will be utilized as the basis for the recommendation. The first one is the association rules and the second is the similarity of content based on topic modeling.

Web content and web server logs are the primary sources for this recommendation system. These two sources are used to generate the first knowledge base which is association rules. Firstly, web server logs will be transformed into URL sessions. Then, web contents are processed in topic modeling to find the topic model. The result from both of these sources combined to generate the session topic. Each session topics is an item which becomes the input to the association rule mining. The output of the association rule mining is becoming the first knowledge base. The second knowledge base generates purely from the topic similarity of the document inferred by the result of topic modeling. The workflow of these offline process can be seen in Figure 1.

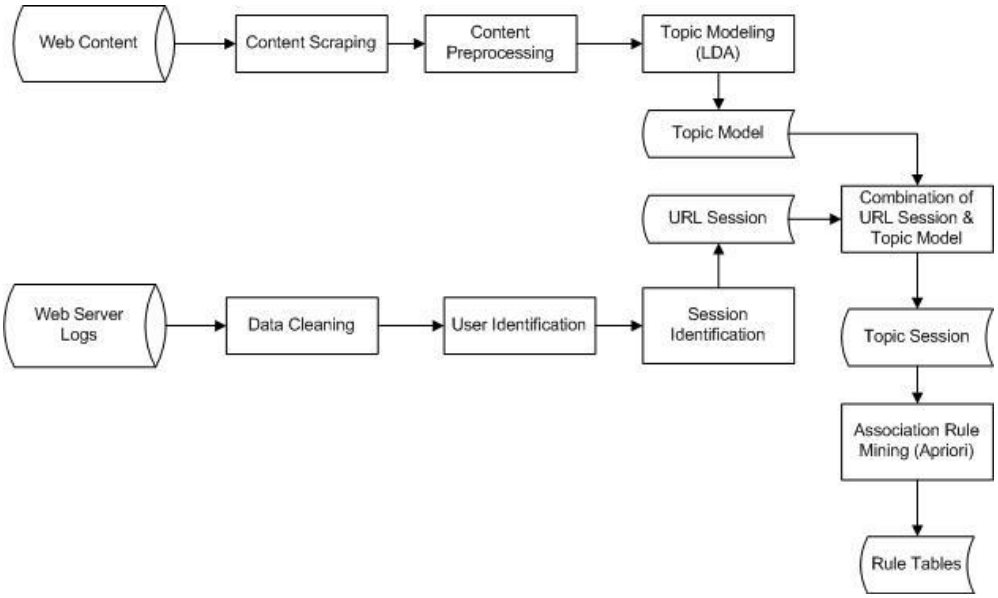

Fig. 1. Offline Phase

The online phase then uses these knowledge base as the reference for the recommendation. When the user started to navigate the website, a set of topic item are shaped. As-sociation rules have the priority for the recommendation. A rule that has the highest support and confidence will become the recommendation. If there are no rules that match, then the similar document based on the topic distribution will be selected as the recommendation. The workflow of these online process can be seen in Figure 2.

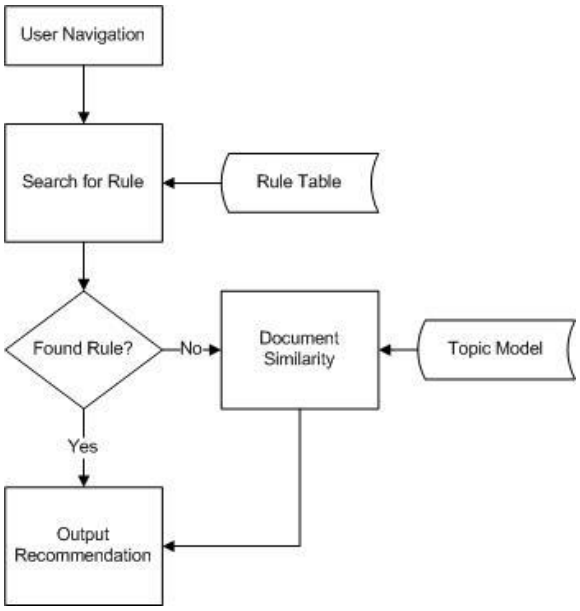

Fig 2. Online Phase

A. Web Log Data Preprocessing

The interaction between the users and the website is recorded on a web server logs. This log file has several relevant fields that were used for analysis such as the identity of the user and also the information on the accessed pages. Before these log files are ready for analysis, data cleaning is needed. There are four steps involved in the cleaning process. The first is eliminating request with the error status code (4xx and 5xx). The second step is to remove the request on the page that is not related to user activity such requests on images, videos, styles (CSS), and scripts (javascript). The next step is to eliminate requests on the page that are not directly related to content such as contact page, about, and sitemap. The last step is to eliminate the request made by robots, crawlers, or spiders.

After the logs data are cleaned, user and session need to be identified. A user can be recognized by the identical of IP address and user agent (Chitraa and Thanamani, 2011). Once the user is found, we need to identify the session. The simplest way can be achieved by looking at the timeout. Most commercial services using 30 minutes as the default timeout (Cooley et al, 1999). The output of this process is in the form of a URL session length n, and can be represented in (1):

$$S_s = (u_s^1 u_s^2 \dots u_s^L) \tag{1}$$

where the session S_s consist of URL u_s^L that belongs to session s with the length of L.

B. Topic Modelling

Topic modeling is a statistical method for discovering patterns and themes in the corpus of document [2] [12]. In this study, we used Latent Dirichlet Allocation (LDA) that Introduced by Blei [12]. LDA allows us to identify topics in web content.

In this research, web content acquired by the web scraping technique recursively to the entire address of the website. This content needs to be transformed into LDA corpus. This transformation consists of tokenization, stopword removal and forming into the bag of words. The output of LDA is a model that contains the topic with the probability of the word. This model can be used to find the document topic probability vector that represents the distribution of topics from each of the document. Document topic probability vector can be described in (2):

$$DT_i = P_{u_d}^{T_0}, P_{u_d}^{T_1}, \dots., P_{u_d}^{T_k} \tag{2}$$

Where DT_i is document topic probability vector, $P_{u_d}^{T_k}$ is the affinity of topic k on document i. This vector can be used to see the similarity between documents based on topic distribution, and used as a knowledge base for the second recommendation.

C. Session and Topic Model Combination

The output of the sessionization process is a series of URL. A URL included in the analysis contains an article. In gudegnet, there are more than 9.000 articles with its own topics. In the previous process, we have a document topic probability vector, so that we can get the topic with the highest probability. This one topic then substitutes with the URL in session to become a session topic. Session topics can be represented as follows.

$$ST_s = T_s^1 T_s^2 \dots T_s^L \tag{3}$$

Where the session ST_s consist of topic T_s^L that belong to session s with the length of L.

D. Association Rule Mining

Session topic from (3) contains a list of topic that accessed together in a session. Therefore it can show the interest degree of a user on some topic [2]. To uncover the interesting relationship between

the topics we used Association Rule Mining. Association Rule Mining is a fundamental data mining task. Association rule can be used to find all co-occurrence relationships and Web usage patterns [8]. We used the best known association rule mining algorithm which is the Apriori algorithm proposed in [10].

This relationship can be expressed into association rules. Given $I = i_1, i_2, \dots, i_m$ be a set of items, where each I is session topic (ST_s) and i is an item or topics (T_s^L). Let $T = t_1, t_2, \dots, t_n$ be a set of transactions or a set of session topics where each session topic is a set of topics such that $t_i \subseteq I$. Support and confidence are used as the metrics for the rule. The main purpose of association rule mining is to discover all association rules in T that have support and confidence greater than or equal to the minimum support and minimum confidence given by user [8]. An association rule can be represented as, $X \rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap I = \emptyset$.

E. Recommendation System

In the online phase, recommendation system generated based on two level of the knowledge base, association rules and topic similarity.

1) Recommendation Based on Association Rules: An example of association rules is :
Culinary → Accommodation [$support = 20\%, confidence = 80\%$]

The rule says that 20% visitors read culinary article and accomodation article in one session, and those who read the culinary article also read accomodation article 80% of the time. The left part of the rule is known as the antecedent, and the right part known as the consequent. These co-occurence relationship used as a recommendation for the web articles. During user surfing sessions, the system will match the sessions with the antecedent, and if such rule is found then the system will recommends the consequent item.

2) Recommendation Based on Topic Similarity: It is possible that during the browsing of a user, no session is match the association rule. Therefore, it needs a knowledge base that assure recommendations. We used document similarity based on document-topic vector generated in (3). The intuition behind LDA is that document is a mixture of multiple topics [12] with different probabilities. Given that probabilites vector, we can measure the similarity of documents by using cosine similarity.

IV. Results

We have collected a month of user request from gudegnet web server log. It was contained 15.795.173 requests, and by doing data cleaning, we can reduce it into 212.694 requests. Then we perform user identification and sessionization and we got 148.666 session with the average length is 1,4 URL per session. From the result, we can see that the bounce rate in gudegnet is very high. We decided to remove the session that contains only have one URL. After this elimination, we got 30.032 sessions with 3.1 URL per session. From the content side, we scraped 12.908 articles from gudegnet, with 10.791 in Bahasa Indonesia, and 4.893 in English. On each language, we performed topic modeling so we got 2 topic model, with each of the model contain 10 topics. Then, we make document topic probability for each article based on particular language. The example of topic model in bahasa Indonesia can be seen in Table I.

We combine URL session and topic, to become the session topic. In the session, it is highly possible that a user can access the same topic throughout the navigation. The objectives of this association rule analysis are to seek the different topic that frequently access together. We make a distinct topic selection in every session and eliminate the item set that only has 1 topic. This process resulted in 7.619 session topic. This would become the input for association rule generation with Apriori.

We perform apriori with minimum support 0,01 and minimum confidence 0,6. With these parameters, there are 144 frequent itemset and 11 rule. The sample result of the frequent itemset can be seen in the following. The number means the topic number base on Table 1.

$$[3, 1, 6]\ , support\ :\ 0.025$$

$$[10, 1, 6]\ , support\ :\ 0.029$$

$$[1, 6]\ , support\ :\ 0.098$$

$$[10, 6]\ , support\ :\ 0.103$$

Based on the frequent itemset, the association rules can be generated. The sample of association rule generated can be seen in the following.

$$[1, 8, 3]\ \rightarrow\ [6]\ , confidence\ :\ 0.698$$

$$[10, 8, 6]\ \rightarrow\ [1]\ , confidence\ :\ 0.699$$

$$[1, 7, 3]\ \rightarrow\ [6]\ , confidence\ :\ 0.734$$

$$[10, 1, 8]\ \rightarrow\ [6]\ , confidence\ :\ 0.744$$

$$[10, 1, 3]\ \rightarrow\ [6]\ , confidence\ :\ 0.772$$

When the user navigate into article with the highest probability of topic is topic 10, the system will give recommendation several article on topic 6. We can see the support for this rule is very low in 1% but the confidence is pretty high in 77%.

Table 1. Topic Model

Topic	Words
1	0.022*desa + 0.014**pasar* + 0.014**wisata* + 0.008**restaurant* +0.008*rw + 0.007**menikmati* + 0.006**sekar* + 0.006**kerajinan* + 0.006**pengunjung* + 0.005**lokasi*
2	0.015**indonesia* + 0.014**mandiri* + 0.013**sleman* +0.008**lingkungan* + 0.007**kantor* + 0.007**buku* + 0.007**bantul* + 0.007**bank* + 0.006**bni* + 0.006**perusahaan*
3	0.039**atm* + 0.025**seni* + 0.019**indonesia* + 0.018**tunai* + 0.016**bca* + 0.013**pameran* + 0.011**tersedia* + 0.011**gallery* + 0.010**art* + 0.008**karya*
4	0.020**buka* + 0.019**gudeg* + 0.018**ayam* + 0.017**warung* + 0.016**menu* + 0.015**nasi* + 0.014**goreng* + 0.012**restoran* + 0.011**soto* + 0.010**pariwisata*
5	0.022**candi* + 0.020*villa + 0.018*hotel + 0.012*fasilitas + 0.010*uang + 0.009**terletak* + 0.009**restoran* + 0.009*bangunan + 0.009*museum + 0.006*parkir
6	0.034**tour* + 0.030*travel + 0.019**wisata* + 0.015**pantai* + 0.014**paket* + 0.014**borobudur* + 0.012**mobil* + 0.010**parangtritis* + 0.010**jam* + 0.010**harga*
7	0.036**rt* + 0.033**smp* + 0.018**negeri* + 0.009**bantul* + 0.008**harjo* + 0.006**jepang* + 0.006**cv* + 0.006**umbul* + 0.005**mangkubumi* + 0.005**giwangan*
8	0.020**pendidikan* + 0.017**rumah* + 0.015**pelayanan* + 0.012*tk + 0.011**masyarakat* + 0.011**sakit* + 0.011**daerah* + 0.011**kesehatan* + 0.009*circle + 0.009**dokter*
9	0.015**upacara* + 0.014**masjid* + 0.011**dusun* + 0.009**desa* + 0.008**gunungan* + 0.007**jawa* + 0.007**gunung* + 0.007**sultan* +0.006**kyai* + 0.006**makam*
10	0.079**informasi* + 0.054**kota* + 0.051**terbaru* + 0.029**istimewa* + 0.027**kontak* + 0.027**perkembangan* + 0.026**gudang* + 0.026**detail* + 0.026**simak* + 0.026**tertera*

The second level of recommendation is based on the topic similarity between articles. We infer model from Table I into all web articles to create a document topic (2). The example of document topic probability can be seen in Table 2.

Table 2. Document Topic

Document	Topic Probability
Pantai Drini	Topic 1*0.0532096046423 + Topic 2*0.0457563449665 + Topic 3*0.144193425948 + Topic 4*0.198742440219 + Topic 5*0 + Topic 6*0.195690258319 + Topic 7*0 + Topic 8*0 + Topic 9*0.359529963764 + Topic 10*0
DA Transport	Topic 1*0 + Topic 2*0 + Topic 3*0 + Topic 4*0 + Topic 5*0.044931817055 + Topic 6*0.71084746703 + Topic 7*0.239733171757 + Topic 8*0 + Topic 9*0 + Topic 10*0
Gulai Kepala Ikan Pak Untung	Topic 1*0.144676810114 + Topic 2*0 + Topic 3*0 + Topic 4*0.605434609718 + Topic 5*0 + Topic 6*0 + Topic 7*0.122715187125 + Topic 8*0.10850285919 + Topic 9*0 + Topic 10*0

We can get topic similarity by comparing the probability of each document. The result of the top 3 similar item, based on articles in Table II, can be seen in Table 3.

Table 3. Document Similarity

Document	Similar Document : Similarity
Pantai Drini	Pantai Ngobaran : 0.93888807 Goa Rancang Kencono : 0.91741419 Pantai Sadeng : 0.90404648
DA Transport	Windu Rent Car : 0.99549359 AB Yogya Transport : 0.99484509 Nusa Santana Prima Tour dan Travel : 0.9756009
Gulai Kepala Ikan Pak Untung	Warung Bu Ageng : 0.97529632 Lusidus Vegetarian : 0.97240818 Keripik Belut : 0.97150409

When the user navigate into ”Pantai Drini” , the system will give top 3 recommendation based on the similarity. The recommendation is ”Pantai Ngobaran” ,”Goa Rancang Kencono” , and ”Pantai Sadeng” with the similarity over 90%.

V. Conclusion

In this study, we proposed two level of recommendation based on association rules and topic similarity. We used Apriori as the association rule mining algorithm and LDA as the topic modeling algorithm. We implement this technique specifically for a website that has much content such as city directory that we used for this study. From the result, we can see that there are only 11 association rules generated with the confidence below 80%. The result says that there are not many associations between different topics. On the other hand, we can get much better results on the second level which is the topic similarity. The top three recommended documents have over 90% similarity.

Acknowledgement

The authors would like to thank PT Citraweb Indonesia owner of gudegnet for providing the data and their support.

References

[1] U. Gretzel, “Intelligent systems in tourism,” Ann. Tour. Res., vol. 38, no. 3, pp. 757–779, Jul. 2011 [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0160738311000776

[2] O. Arbelaitz, I. Gurrutxaga, A. Lojo, J. Muguerza, J. M. Pérez, and I. Perona, “Web usage and content mining to extract knowledge for modelling the users of the Bidasoa Turismo website and to adapt it,”

Expert Syst. Appl., vol. 40, no. 18, pp. 7478–7491, Dec. 2013 [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0957417413005198

[3] D. Buhalis and R. Law, "Progress in information technology and tourism management: 20 years on and 10 years after the Internet—The state of eTourism research," Tour. Manag., vol. 29, no. 4, pp. 609–623, Aug. 2008 [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0261517708000162

[4] B. Pan and D. R. Fesenmaier, "Travel information search on the internet: a preliminary analysis.," 2003, pp. 242–251.

[5] J. Borges and M. Levene, "Data Mining of User Navigation Patterns," pp. 92–112, 2007.

[6] A. S. Lalani, "Data Mining of Web Access Logs,MS Thesis," pp. 1–70, 2003.

[7] S. S. Rawat and L. Rajamani, "Discovering Potential User Browsing Behaviors Using Custom-Built Apriori Algorithm," Int. J. Comput. Sci. Inf. Technol., vol. 2, no. 4, pp. 28–37, 2010.

[8] B. Liu, Web Data Mining. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011 [Online]. Available: http://link.springer.com/10.1007/978-3-642-19460-3

[9] M. Dimitrijevic and Z. Bosnjak, "Pruning statistically insignificant association rules in the presence of high-confidence rules in web usage data," Procedia Comput. Sci., vol. 35, no. C, pp. 271–280, 2014 [Online]. Available: http://dx.doi.org/10.1016/j.procs.2014.08.107

[10] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," in Proceedings of the 20th International Conference on Very Large Data Bases, 1994, pp. 487–499 [Online]. Available: http://dl.acm.org/citation.cfm?id=645920.672836

[11] E. Lazcorreta, F. Botella, and A. Fernández-Caballero, "Towards Personalized Recommendation by Two-step Modified Apriori Data Mining Algorithm," Expert Syst. Appl., vol. 35, no. 3, pp. 1422–1429, 2008 [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2007.08.048

[12] D. M. Blei, "www.cs.princeton.edu/~blei/papers/Blei2012.pdf," Cs.Princeton.Edu, pp. 77–84 [Online]. Available: http://www.cs.princeton.edu/~blei/papers/Blei2012.pdf%5Cnpapers2://publication/uuid/228B8D31-9CA0-447B-A144-3D6B0EA97493

[13] O. Arbelaitz, I. Gurrutxaga, A. Lojo, J. Muguerza, J. M. Pérez, and I. Perona, "Enhancing a Web Usage Mining based Tourism Website Adaptation with Content Information."

[14] M. Zanker, M. Fuchs, W. Höpken, M. Tuta, and N. Müller, "Evaluating Recommender Systems in Tourism — A Case Study from Austria," 2008, pp. 24–34.

[15] J. Vellingiri and S. C. Pandian, "A Survey on Web Usage Mining," Glob. J. Comput. Sci. Technol., vol. 11, no. 4, pp. 67–72, 2011.

[16] Sunil and M. Doja, Recommender System Based on Web Usage Mining for Personalized E-learning Platforms, vol. 5. 2017.

[17] D. M, A. G. M. Erwin, and N. D, "2013, News Recommendation in Indonesian Language Based on User Click Behaviour."

[18] N. More and N. P. More, "Recommendation of Books Using Improved Apriori Algorithm," Int. J. Innov. Res. Sci. Technol, vol. 1, no. 4, pp. 80–82, 2014.

[19] H. . Husin, "News Recommendation Based on Web Usage and Web Content Mining," in ICDE Workshops, 2013.

[20] N. Dave, K. Potts, V. Dinh, and H. U. Asuncion, "Combining association mining with topic modeling to discover more file relationships," Int. J. Adv. Softw., vol. 7, no. 3, pp. 3–4, 2014.