*Sindh Univ. Res. Jour. (Sci. Ser.) Vol. 48 **(4D)** 27-30 (2016)*

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Discovering Users Navigation of Online Newspaper using Association Rules

H. S. HUSIN

Institute of Information Technology (UniKL MIIT), University Kuala Lumpur – Malaysian

Received 29rd August 2016 and Revised 28th October 2016

Abstract: We analyse behaviour of non-subscribed users browsing an online newspaper. We collect Web server logs from a Malaysian online newspaper for 28 days in April 2012. Association rules are used to understand user behaviour based on their visits recorded on the Web server logs. We found that National section is mostly accessed, together with Sports and Entertainment. Further analysis on article pages showed that users tend to read the same genre of articles. Discussion on findings are compared with previous work are also presented.

Keywords: Association rules; Web usage mining; Web server log; Online newspaper; User behaviour

1. INTRODUCTION

Association rules can discover the correlations between pages that are requested together during the server session. These rules can suggest the potential relationship between pages that are regularly accessed together, although they may not be directly connected (Eirinaki andVazirgiannis, 2003). We can discover hidden information from large Web log data collection and frequent patterns of Web pages visited by user can be used to obtain useful information about user's navigation behaviour (Iváncsy andVajk, 2006) . We can use association rules to predict the users' next requests from Web log data. (Géry andHaddad, 2003). By and large, association rules are used to identify the pages that are being accessed together is a session. Association rules can also be a basis to predict future requests by users. In our work, we use the association rules to identify the pages that are frequently accessed together in a session.

2. METHODOLOGY

In this study, we collect web server logs from a daily Malaysian newspaper, *Berita Harian* for four weeks in the month of April 2012. The April data set contains 53 225 310 records containing logs from 1 April 2012 till 28 April 2012.

We follow the Web usage mining pre-processing techniques suggested by (Cooley, Mobasher, andSrivastava, 1999). In the case of Web browsing, association rules can capture relationships among the different URLs based on the users' navigation trails. After we identify the user and session, we create the transaction file for data mining using association rules to find out the relationship between the different pages that users accessed. The Apriori algorithm finds groups of URLs frequently occurring together in many transactions (Srikant and Agrawal, 1996).

There are three pre-processing steps taken to ensure that the logs are ready for analysis stage. The first task is to do data cleaning, followed by user identification and lastly session identification (Cooley *et al*., 1999). The items that are removed from the data set are unsuccessful requests, images files, automated requests, RSS requests, Facebook requests and other requests that are part of the page such as the gallery and video files.

User identification in Web server logs are based on a combination of the IP address and the user agent string that is type of browser and operating system. If a user has the same IP address, but different type of user agent, it is assumed that it denotes a different user (Cooley *et al*., 1999).

Session identification is intended to divide the page accesses of each user into individual sessions. The literature suggests using a timeout. One study established a timeout of 25.5 minutes based on empirical data (Catledge andPitkow, 1995). Another method suggested a session-duration-based set a session to a 30-minute threshold (Cooley *et al*., 1999). For our studies, we set the timeout for five minutes and take all sessions that have at least two pages. This is based from a previous study that assumed a page session to be over if the user remained on a single page for five minutes or more and the next page viewed is then assumed to be start a new page study (Huntington andJ amali, 2008).

In the context of finding Web pages that are visited together, it is similar as to find associations among items in transaction databases or to find items that are

bought together in a basket. A set of session is defined as $S = \{S_1, ..., S_n\}$ where n is the number of sessions in S and a set of URLs or page request is a number of URLs in S defined as $U = \{U_1, ..., U_n\}$.

When the data is cleaned, we create the transaction file listing all the pages that are accessed by users. Next, we generate the association rules using the package *arulez* for Apriori association rule in R. In reference to transactions on the Web, association rules capture the relationships between the pages. For example, *{National, World} ⇒ {Sports} [support=0.01, confidence=0.5]* represents the relationship of users that access National and World pages whom also accessing the page Sports with confidence of 50%.

The support value represents that item set *{National, World, Sports}* was present in 1% of user sessions. There are two parts of session analysis for this work. The first part is to investigate what section pages are being read together in the same session throughout April 2012. For this purpose, we used two support values of 0.01 and 0.1 and confidence of 0.5.

The translation for each section is as follows

- Nasional - National
- Sukan - Sports
- Dunia - World
- Hip - Entertainment
- Ekonomi - Economy/Business
- Pendidikan - Education
- Agama - Religion
- Rencana - Features
- Ratu - Women
- Skuad - Squad
- Mukadepan - Frontpage
- Sastera - Literature
- Surat - Letters
- Mutakhir - Latest
- Wilayah – Regional

To analyse the pair of section pages that are accessed together, we choose sessions containing requests to section pages of at least two requests in a session. To analyse the type of articles that users like to read, we select Monday 2 April 2012. The article title is abbreviated to nine characters in lower case. The URL title is abbreviated to simplify the process to generate the graphs and to make the graphs more readable. We chose sessions that have at least two requests for the article page.

3. FINDINGS

The results and findings are divided to two main parts; the first one discussing the section pages for the 28 days in April 2012 and the other part on the article pages that were requested by users on Monday 2 April 2012

3.1 Analysis of frequently accessed section pages for April 2012

Frequent item sets are set of transactions or items that appear many times in a database or a basket. In our context, the frequent item set represent the frequent access pattern to the Web pages that are requested by users. There are two steps in association rule mining; which are to find all frequent item sets and to generate high confidence rule (Agrawal *et al*., 1996). The *Apriori* principle states that if an item set is frequent, then all of its subset must also be frequent (Borges andLevene, 2007). For example, if *{National, Sports, World}* is a frequent item set, any transaction that contains *{National, Sports, World}* must also contain its subsets; *{National, Sports}, {National, World}, {Sports, World}, {National}, {Sports},* and {*World}.*

On the other hand, if an item set such as *{Literature, Letters}* is infrequent, then all of its supersets are infrequent too. This technique to trim the rules based on the support measure is known as support-based pruning (Tan, Steinbach, andKumar, 2005). In **(Fig-1)** for the support of 0.01, there are more section pages appear, as compared when the support threshold is increased to 0.1 in **(Fig-2)**. Nevertheless, the top three section pages are National, Sports and World.

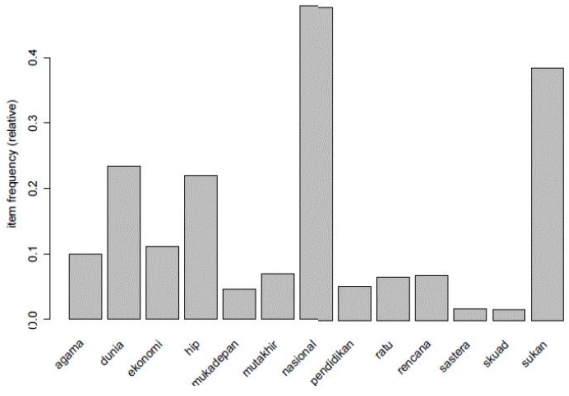

Fig-1 Item frequency plot for section pages in April 2012 with support = 0.01

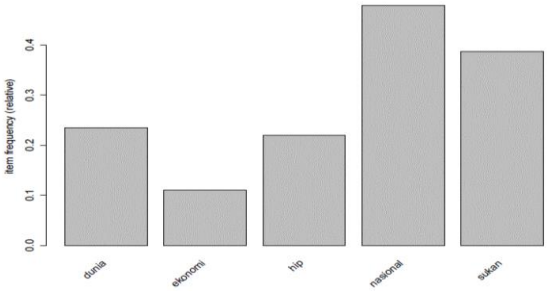

Fig-2 Item frequency plot for section pages in April 2012 with support = 0.1

Table-1 lists the frequent item sets for section pages with the support of 0.1. The most frequent pair is National and Sports, followed by National and World.

Table-1: Frequent itemset for sections page in April 2012 with support 0.01

No	itemset	support
1	{nasional, sukan}	0.14704075
2	{dunia, nasional}	0.13602367
3	{dunia, sukan}	0.10058465
4	{hip, nasional}	0.09721369
5	{dunia, nasional, sukan}	0.0689919
6	{ekonomi, nasional}	0.06583036
7	{dunia, ekonomi}	0.05852661
8	{dunia, hip}	0.05735028
9	{hip, sukan}	0.05718349
10	{agama, nasional}	0.05543656

From the association rules in **(Table 2)** with support of 0.01 and confidence 0.1, we can identify the less popular sections and their correspondence section that are read together

Table-2: Association rules for section pages in April 2012 with smaller support value of 0.01

No	itemset	support
1	{sastera} ⇒{pendidikan}	0.01111365
2	{sastera} ⇒{rencana}	0.01106975
3	{pendidikan} ⇒{rencana}	0.01990976
4	{ratu} ⇒{rencana}	0.02281546
5	{pendidikan} ⇒{ratu}	0.01562582
6	{ratu} ⇒{agama}	0.02695893
7	{rencana} ⇒{agama}	0.02611619
8	{pendidikan} ⇒{agama}	0.01947083
9	{agama} ⇒{ekonomi}	0.03511421
10	{rencana} ⇒{ekonomi}	0.02333339

Based on the association rule mining, we discovered that the section Literature and Education have very high confidence of being read together, despite a very low count of support. This suggests that although not many users read Literature and Education sections, but when they do, they will read these two sections together.

3.2 Analysis of frequently accessed article page on Monday 2 April 2014

We analysed the article pages that are frequently accessed based on the support threshold of 0.01 and confidence of 0.5. The frequent itemsets for article titles accessed on Monday 2 April 2012 is shown in **Table 3.** We accessed the archives of *Berita Harian* to find out the articles are in which section. From the table, the articles from Sports of *{balotelli, robinhoga}* has the highest support, followed by articles *{serahteru, wanitadak}* from National sections. The subsequent itemsets are mostly from National sections. We looked into other days and in general, the itemsets for articles are of the same section.

Table-3: Frequent item sets for article page on Monday

No	itemset	support
1	{balotelli, robinhoga}	0.0183
2	{serahteru, wanitadak}	0.0146
3	{safeesali, thailands}	0.0134
4	{jpnkedahb, wanitadak}	0.0133
5	{duaperagu, pintusemp, wanitadak}	0.0133
6	{chinaunte, serbuanre}	0.0120
7	{pintusemp, serahteru}	0.0111
8	{wanitadak, wwfgesake}	0.0110
9	{duaperagu, gapenasyo}	0.0109
10	{duaperagu, jpnkedahb}	0.0109

The association rules for the articles shown in **(Table-4)** is arranged according to the lift value. All the frequent item sets for the articles have very high lift of more than 1, with the top six association rules are for articles from the National sections.

Table-4: Association rules for article page on Monday 2 April 2012

No	Association rules	support	lift
1	{jpnkedahb} ⇒ {serahteru}	0.0106	24.82
2	{gapenasyo} ⇒ {chinatari}	0.0102	15.33
3	{wwfgesake} ⇒ {duaperagu}	0.0103	11.95
4	{chinatari, pintusemp}⇒ {duaperagu}	0.0107	11.42
5	{gapenasyo} ⇒ {pintusemp}	0.0109	11.26
6	{gapenasyo} ⇒ {duaperagu}	0.0109	10.98
7	{robinhoga} ⇒ {balotelli}	0.0183	9.92
8	{serahteru} ⇒ {pintusemp}	0.0111	9.53
9	{serahteru} ⇒ {duaperagu}	0.0108	9.02
10	{jpnkedahb} ⇒ {duaperagu}	0.0109	8.97

Although the first two rules do not correspond with the top two frequent itemsets for the articles, but the rules show that users are reading news articles of the same sections. **(Table-5)** shows the article label, the corresponding title and sections for each articles that are frequent.

Table-5: Article titles and labels for Monday 2 April 2012

Article label	Section	Article title in Malay/English
serbuanre	Sports	Serbuanredwarriors/ The Red Warrior attacts
safeesali	National	Safeesalibercerai/ Safee Sali divorce
wanitadak	National	Wanitadakwadiserangpengacaratv / Woman attacks TV host
chinatarik	National	Chinatariklaranganimport / China detracts import ban
pintusemp	National	Pintusempadansesakribuanpulang / The border is packed with thousands coming home
balotelli	Sports	Balotellibuatmanciniresah / Balotelli makes Mancini anxious
robinhoga	Sports	Robinhogagalbantumilanraihtigamata / Robin Hog fails to collect 3 points
serahteru	National	Serahteruskepejabatkl / Submit to KL office
jpnkedahb	National	Jpnkedahbanturusnadapatmykad / JPN Kedah helps to get MyKad
gapenasyo	National	GAPENAsyorsasterasubjekwajib / GAPENA suggest Literature as compulsory subject
wwfgesake	National	WWFgesakerajaankurangkanlesenpukattunda / WWF urgest government to control trawlers licence

4. CONCLUSIONS

From the derived association rules, we found that users are always looking for local news and they want to keep updated with current affairs. This finding also shows that the users want to know the current issues that happen locally as well as news around the globe as suggested by (Rathmann, 2002).

Apart from the high requested section pages, we also investigate the section pages that are less popular to see the frequent pairs that are accessed together. From the association rule mining, we have discovered that section Literature, Education and Features have high support and confidence. This may be due to the fact that the newspaper has weekly inserts that are targeted for education, particularly for primary and secondary schools. The findings for sections that are associated strongly are different from previous literature. In (Batista, 2001), they found that there are strong associations between Politics and Society, Politics and International News and between Society and International News. Previous study on New York Times revealed that international news dominate the online usage, as compared to local news (Nicholas and dHuntington, 2000).

The difference in the findings may be due to the fact that *Berita Harian* is considered as the national dailies, a go-to source of news for local and current news that happen in Malaysia. As compared to The New York Times, that may be recognized as reliable. Users are inclined to target on specific contents of their linking when they read the website.

ACKNOWLEDGMENTS

We thank *Berita Harian* for allowing us to use the web server logs and Bakhtiar Abdul Hamid for assistance in data collection.

REFERENCES:

Agrawal, R., H. Mannila, R. Srikant, H. Toivonen, A. I. Verkamo, others. (1996). Fast Discovery of Association Rules. Advances in {K}nowledge {D}iscovery and {D}ata {M}ining, 12(1), 307–328.

Batista, P., M. J. Silva, (2001). Mining on-line newspaper web access logs. In 12th International Meeting of the Euro Working Group on Decision Support Systems EWG-DSS.

Borges, J., M. Levene, (2007). Evaluating variable-length markov chain models for analysis of user web navigation sessions. Knowledge and Data Engineering, IEEE Transactions on, 19(4), 441–452.

Catledge, L. D., J. E. Pitkow, (1995). Characterizing browsing strategies in the World-Wide Web. Computer Networks and ISDN Systems, 27(6), 1065–1073.

Cooley, R., J. Srivastava, (1999). Data preparation for mining world wide web browsing patterns. Knowledge and Information Systems, 1(1),5–32.

Eirinaki, M., M. Vazirgiannis, (2003). Web mining for web personalization. ACM Transactions on Internet Technology (TOIT), 3(1), 1–27.

Géry, M., H. Haddad, (2003). Evaluation of web usage mining approaches for user's next request prediction. In Proceedings of the 5th ACM international workshop on Web information and data management 74–81.

Huntington, P. D. N., H. R. Jamali, (2008). Website usage metrics: A re-assessment of session data. Information Processing andManagement, 44, 358–372.

Iváncsy, R., I. Vajk, (2006). Frequent pattern mining in web log data. Acta Polytechnica Hungarica, 3(1), 77–90. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.101.4559&rep=rep1&type=pdf

Nicholas, D., P. Huntington, (2000). Evaluating the use of newspaper web sites logs. International Journal on Media Management, 2(2), 78–88.

Purcell, K., L. Rainie, T. Rosenstiel, K. Olmstead, (2010). Understanding the participatory news consumer: How internet and cell phone users have turned news into a experience. Pew Research Center, 1–63.

Rathmann, T. A., (2002). Supplement or substitution? {T}he relationship between reading a local print newspaper and the use of its online version. Communications, 27(4), 485–498.

Tan, P. N., V. Kumar, (2005). Introduction to Data Mining, (First Edition). Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Tewksbury, D., (2006). What do Americans really want to know? Tracking the behavior of news readers on the Internet. Journal of Communication, 53(4), 694–710.