Referral Traffic Analysis: A Case Study of the Iranian Students' News Agency (ISNA)

Roya Hassanian-Esfahani*
Research Institute for ICT, ACECR, Tehran, Iran
r.hassanian@ictrc.ac.ir
Mohammad-Javad Kargar
Department of Computer Engineering, University of Science and Culture, Tehran, Iran
kargar@usc.ac.ir

Received: 27/June/2015 Revised: 23/Dec/2015 Accepted: 31/Jan/2016

Abstract

Web traffic analysis is a well-known e-marketing activity. Today most of the news agencies have entered the web providing a variety of online services to their customers. The number of online news consumers is also increasing dramatically all over the world. A news website usually benefits from different acquisition channels including organic search services, paid search services, referral links, direct hits, links from online social media, and e-mails. This article presents the results of an empirical study of analyzing referral traffic of a news website through data mining techniques. Main methods include correlation analysis, outlier detection, clustering, and model performance evaluation. The results decline any significant relationship between the amount of referral traffic coming from a referrer website and the website's popularity state. Furthermore, the referrer websites of the study fit into three clusters applying K-means Squared Euclidean Distance clustering algorithm. Performance evaluations assure the significance of the model. Also, among detected clusters, the most populated one has labeled as "Automatic News Aggregator Websites" by the experts. The findings of the study help to have a better understanding of the different referring behaviors, which form around 15% of the overall traffic of Iranian Students' News Agency (ISNA) website. They are also helpful to develop more efficient online marketing plans, business alliances, and corporate strategies.

Keywords: Referral Traffic; Data Mining; K-means Clustering; Online News.

1. Introduction

News is a rich and message-containing context which conveys answers to some fundamental questions including what, who, how, when, where, why, for whom and so on [1-4]. It also covers a conceptual hierarchy of story, event, and topic [5]. The first electronic newspaper "News Report" was established in the University of Illinois in 1974 followed by the Columbus Dispatch starting its online service in 1980. In the late 1990s thousands of online newspapers came into existence [6].

News websites are making millions of news articles available to the public through different web services these days. Most of the professional news agencies have their own web-based services; many non-professional news agencies are also participating in the process of producing and distributing online news.

Website users come from different channels including organic search services, paid search services, referral links, direct hits, links from online social media, e-mails and so on [7]. Website owners (webmasters) can understand how to address their users' needs by paying attention to their feedbacks and analyzing their behaviors via web usage mining, which refers to activities done to understand users' behaviors by tracking the users' footprints on the web. For example, a company can realize the behavior of its visitors by tracking its referral audience, who come from website A to website B, and is called the referral audience of website B. Referral audience are acquired through referral customer acquisition, a significant marketing strategy in all industries, especially online ones, and a certain part of modern Customer Engagement (CE) Cycle indicated in Fig. 1.

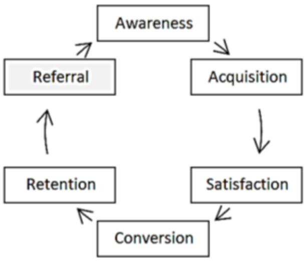

Fig. 1. Customer Engagement (CE) cycle [8]

To measure the number of acquired referral audience, one needs to measure referral traffic, a portion of website's overall traffic. In this paper we model the referral traffic of a news website via analyzing referral acquisitions and the popularity status of the referrers through data mining techniques. This approach of websites usage mining, at the best of our knowledge, has not been addressed in previous works and is used for the first time in this study.

The paper is organized as follows. In Section 2, we present a review on previous works and in Section 3

* Corresponding Author

research methodology is described. Pre-processing activities including cleaning, filtering, validation, outlier detection and normalization are described in Section 4 while Section 5 provides processing activities consisting relationship between investigation and clustering. Descriptive statics of data as well as modeling details are presented in Section 6 and conclusion and proposals for further studies are provided in Section 7.

2. Related Works

Empirical studies on online users' behavior have been performed utilizing diverse approaches ranging from descriptive social analysis and characterization to statistical analysis. Applying data mining techniques on online user-activity logs has led to valuable field literature; however, mining a website's referral traffic is a deprived category.

A common approach in web usage mining is web data clustering, which falls into sessions-based and linked-based approaches [9]. In the sessions-based approach, the one utilized in this study, the system log files of users' behaviors are retrieved, cleaned and analyzed. There are three sources of system log files including those stored on server side, on the client side and on proxy servers [10].

By investigating different time series for modeling page views, Omidvar et al. [11] propose application of regression techniques. Their research independent variables include *visitors' sources* with three options, *connection speed* with six options and *visitors type* with two options. And dependent ones include *page views of visitors' source* with three options, *page views of visitors' speed* with six options and *page views of visitors' type* with two options. The results of their study outline interesting findings. For example, search engine visitors cannot be described by single regression, while referral visitors are well-defined by linear regression. A more detailed report of their work is presented in [12].

Xu and Liu [13] cluster web users by a combination of vector analysis and K-means clustering algorithm. Their solution is independent of sessions. In another study, Xie and Phoha [14] propose a belief function based on Dempster-Shafer's theory of combining evidence to cluster web users into different profiles. Another study on clustering web usage sessions based on access patterns is performed by Fu et al. [15]. They propose a generalization based clustering method which uses an attributed-oriented induction method to reduce the dimensionality of data.

The same issue is addressed based on a-priori association rules discovery and usage-based clustering algorithm in [16]. They define two page types consisting content purpose and navigational purpose based on the amount of time spent by users. A summary of the key related works is provided in Table 1.

Table 1. Summary of key related works

Approach	Technique	Ref.
Analyzing the effectiveness of a website's analytic variables on each other	Linear regression	[11, 12]
Clustering web users	Combination of vector analysis and K-means clustering of log data	[13]
Web users profiles clustering	Belief function	[14]
Clustering web usage sessions based on access patterns	Hierarchical clustering	[15]
Web users' profile clustering	A-priori association rules	[16]

3. Research Method

The research was set to address two questions:

- Is there any significant relationship between referral traffic coming from referrer websites and their popularity status?
- How can online referral traffic be modeled through data mining techniques?

Research steps were designed according to [17] as shown in Fig. 2. It starts from data gathering and goes through data preprocessing and processing steps. In the last step the defined model is interpreted by field experts.

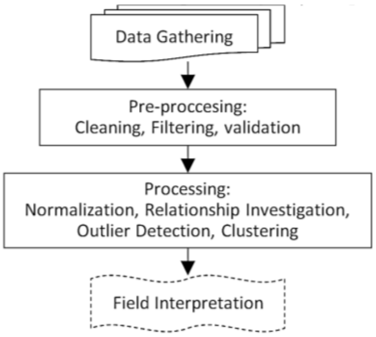
Data Gathering
Pre-proccesing:
Cleaning, Filtering, validation
Processing:
Normalization, Relationship Investigation,
Outlier Detection, Clustering
Field Interpretation

Fig. 2. The research steps

Considering the timing of the study, after running a pilot test on data of a week, the study was conducted on data of 1 month from 2015.01.17 to 2015.02.16.

3.1 Sample Population and Research Variables

The main website of Iranian Students' News Agency (ISNA) was selected for the study. ISNA is a non-governmental Iranian pure online news agency, launched in 1999. The agency provides 44 different online news services and in a month it has an average of 14,717,950 audiences, of whom around 15% are referrals.

There are two research variables in the study: *Referral Traffic* and *Referrer's Alexa Rank*.

- *Referral Traffic*

Referral traffic is defined as "the segment of traffic that arrives at a website through a link on another website" [18]. There are different traffic metrics with slight differences including "hit", "session", "page view" and "page visit". A "session" is a delimited number of a users' explicit web requests across one or more web servers [19], and since the research does not follow click streams inside the website and only referral acquisitions are counted, the "session" was opted in this study. Therefore, referral traffic in this study is the number of visit sessions made on the main ISNA domain from referrer websites.

Referral traffic can be studied through HTTP protocol, which conveys a field named "HTTP header field". Checking it, the new webpage can understand where on the web the user's request has been sent.

Most of the times, referral traffic happens when there is a link to the second webpage on the first webpage. The user hits the link and the browser sends a request to the server of the second webpage having the HTTP header field embedded. To gather this variable, Google Analytics (GA), one of the Google's free web services used in many web mastering and monitoring tasks, was used. The validity of using GA data as input of data mining processes has been confirmed through experimental studies in [20, 21].

Although GA is one of the best and the most widely used website statistics service [22], the most source of probable inaccuracy in the research results could be originated from inaccuracies in GA statistics. It is a known problem which is discussed in several previous works. It uses users' IP to detect their locations. This may have a great impact on the validity of location-based reports in some areas. For this study, since surfing the web through proxy servers is a common manner for Iranian Internet users, to prevent any ambiguities, none of the GA's location-based information and services were used. All other scripts were coded in Python 2.7 as a strong tool for web mining and data scraping with lots of well-defined libraries including BeautifulSoup, urllib, urllib2 and lxml.

- *Referrer's Alexa Rank*

A global scoring system was needed to estimate the popularity state of referrer websites in the study. Two free public available scoring systems were Google PageRank and Alexa Rank. Google PageRank consisted of 10 score levels from 0 to 10, which did not make enough discrimination for our purpose. However, Alexa Rank assigning scores from 1 -the rank of Google.com itself- to multi-millions, was opted for this study. The other reason for choosing Alexa Rank over Google PageRank was that there were legal automatic mass extraction methods available for Alexa Rank [23], while automatic Google PageRank extracting was officially illegal and involved hacking techniques [24].

Alexa is a well-known American company in global ranking of the websites. It performs ranking based on traffic data collected from more than 30 million websites [25]. The validity of this score as a suitable general popularity metric has been proved in several previous studies such as [26, 27].

4. Pre-Processing

The frequency of total referral traffic by days is provided in Fig. 3. The chart depicts a periodic pattern during the days of the week. The most referral acquisitions happen on Sundays and the least ones on Thursdays and Fridays. This pattern is in line with the ISNA's overall page visit pattern that is fewer during Iranian weekend days.

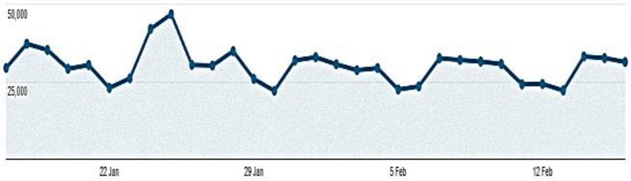

Fig. 3. The number of referral acquisitions by days (Source: GA)

After collecting data, the first step of pre-processing in any data mining project is data cleaning. In this study, to clean the data, first an integration phase was done to transform the format of collected referral URLs into the plain format needed for the next processes. The noisy data of URLs along with repeated data were then detected and removed. With regard to the research scope, sub domains were also ignored and their values were added to their related domains. A validation test was performed to omit all broken and problematic links from the list. Finally, a clean list of 6,572 referrers in a month was reached. Among them there were some nationally access blocked links which had to be recognized to be treated differently in the next steps. The statistics of collected referral URLs for one month are presented in Fig.4.

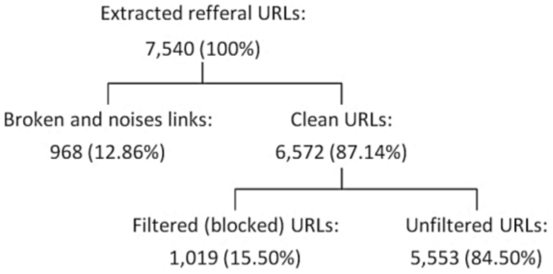

Fig. 4. Statistics of collected referral URLs

The examination shows that 1,019 sessions– around 15.50%– of the referral requests are originated from nationally access blocked domains, among which Facebook.com is the greatest. This is in line with statistics from [28], which reports Facebook.com as the number one social referral traffic source for news websites in 2013. It is a remarkable result at the same time as ISNA does not have any official account in blocked online social media.

4.1 Outlier Detection

Outliers are data that appear far away from others and diverge from the overall pattern in the dataset. The possible negative effects of outliers on data analysis make their detection an essential part in analyzing and modeling processes. Moreover, outlier detection can give a better insight concerning the dataset.

Fig. 5 depicts variables' box plots, a graphical tool to understand the behavior of data toward its distribution.

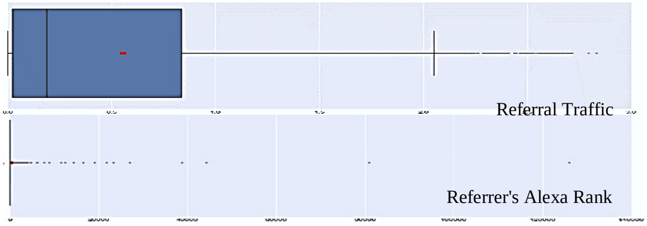

Fig. 5. Variables' box plots

As the outlier detection process strongly depends on data and research context, the Referrer's Alexa Rank outlier detection had to be rejected in the interpreting phase since being in the extremes (the least or the most) does not reflect any significant information, except the weakness or the strength of the referrer website from popularity point of view.

Referral Traffic box plot indicates four unusual behaviors shown as four diverged nodes. They are far from the others and can be interpreted as outliers. These nodes from right to left are presented in Table 2.

Table 2. Referral Traffic Outlier detection results

Website	Referral Traffic	Referrer's Alexa Rank
www.parseek.com	126039	14562
www.khabarfarsi.com	80915	2687z
www.khabarpu.com	44339	3043
www.trakhtorlink.com	38802	64690

Checking result with the field experts reveals that the first three websites are automatic news aggregator web services and the last one (trakhtorlink.com) is a sport specific news service. The surprising point which can be concluded from Table 2 is that none of the referral traffic outliers have significant status in Alexa Rank criterion.

4.2 Normalization

As many distance-based and density-based clustering algorithms are isotropic in all directions, applying them on unequal ranges of variances is equivalent to putting more weights on some variables. In this regard normalization should be considered essential in pre-processing phase to improve the accuracy of clustering model.

One of the most popular normalization methods is Min-max normalization which applies a linear transformation on the original data [29] and gives the same importance to all the variables. In this study, normalization shifted the data ranges as follows:

Referral Traffic: (1, 126,039) → *(0, 1)*
Referrer's Alexa Rank: (1, 28,335,166) → *(0, 1)*

5. Processing

Processing methods involve relationship investigation and data modeling (clustering) which are described in this section.

5.1 Relationship Investigation

The first step in investigating the relationship between two variables is the examination of their scatter plot. Scatter plot is one of the most effective graphical tools for determining if there are any relationship, patterns, or trends between two numerical attributes [30]. As Fig. 6 depicts, reviewing the overall pattern of scatter plotted data does not show any meaningful relationship between the two research variables.

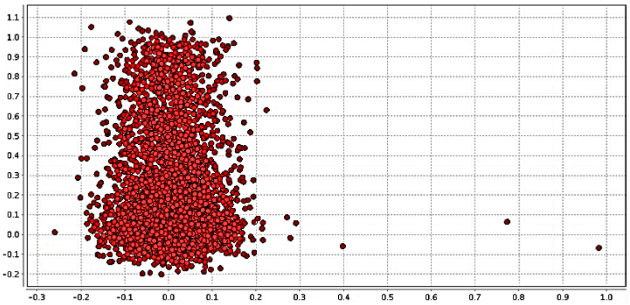

Fig. 6. Scatter plotted data with added jitter

Correlation analysis is a statistical technique to investigate whether two pairs of quantifiable variables are related. Correlation results between the two research variables are reported in Table 3.

Table 3. Correlation analysis results

	Referral Traffic	Referrer's Alexa Rank
Referral Traffic	1	-0.03479
Referrer's Alexa Rank	-0.03479	1

As it is shown in Table 3, the correlation coefficient is near to zero that confirms there is no significant relationship between the two research variables.

In other words, the relationship investigation on research data demonstrates that a website's Alexa Rank is not related to its Referring Traffic for the understudy website. This may be interpreted as the generality of Alexa ranking system. Or, the gap between popularity trends and news seeking manners.

5.2 Clustering

Clustering refers to a variety of methods aiming to group data in such a way that data in the same groups – called clusters- share more similarities (upon some factors) than to those in other groups. Clustering models fall into several approaches including connectivity-based, centroid-based, distribution-based and density-based. There is no predefined rule to choose a model, but selecting an appropriate one depends on the dataset and the aims of the study.

In this study after applying several clustering algorithms, evaluation results showed that distance-based algorithms were best fitted to the dataset. Therefore, based on experimental analysis and in line with previous works, we opted for the K-means Squared Euclidean Distance clustering algorithm.

K-means clustering is one of the most popular unsupervised clustering algorithms. It stores K centroids and then forms clusters. Each point is considered to belong to a particular cluster if it is closer to that cluster's centroid than any others.

6. Findings

6.1 Descriptive Statistics

Descriptive statistics of the two research variables are provided in Table 4.

Table 4. Statistics report of the dataset

Item	Referral Traffic	Referrer's Alexa Rank
Mean	122.2039	5100651
Standard Error	30.3745	96454.65
Median	3	1401456
Mode	1	716577
Standard Deviation	2263.053	7186357
Sample Variance	5121409	5.16E+13
Kurtosis	1915.734	1.398512
Skewness	40.33055	1.571618
Range	123169	28335165
Minimum	1	1
Maximum	123170	28335166
Sum	678354	2.83E+10
Confidence Level (95.0%)	59.54591	189088.9

Great Standard Error (SE) and Standard Deviation (SD) values show inaccuracy of dataset mean due to their diverse distribution. Values of skewness above plus one for both variables is a sign of non-normality, which is confirmed through variables boxplots in Fig. 5, too.

In 86 detected referring Top-level Domains (TLDs), the com with 52% and ir with 41% are the most frequent referrers. Top ten most frequent referring TLDs are provided in Fig. 7.

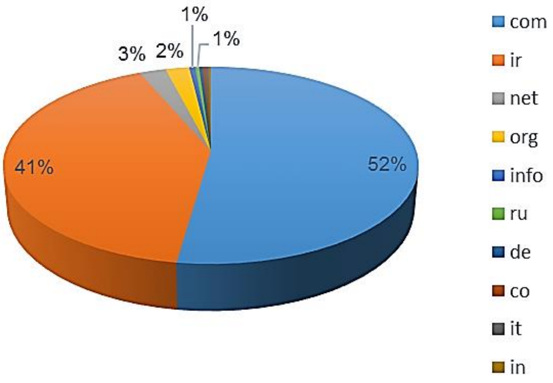

Fig. 7. Top ten most frequent referring TLDs

Since com is the most popular TLD in the world, it’s the most frequent in our dataset, too. And as ISNA is an Iranian news website, its second most referral traffic comes from ir TLDs which is the national Iranian TLD code. Other national TLD codes in the top ten list are ru (Russia), de (Germany), co (Colombia), it (Italy) and in (India).

From the Referrals Traffic point of view, com and ir are again the biggest referrers to ISNA website. Top ten biggest referrer TLDs are shown in Fig. 8.

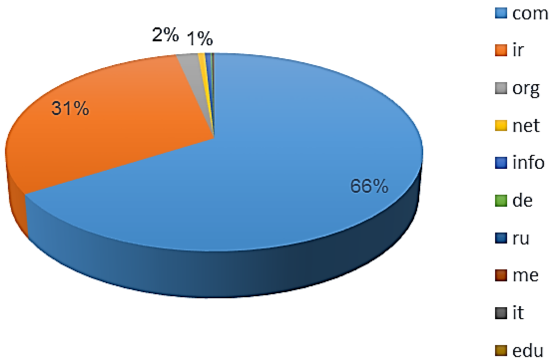

Fig. 8. Top ten biggest referrer TLDs

The two Figures depict that com TLD, which covers 52% of referrers, brings the 66% of referral traffic to the ISNA. And ir TLD, which covers 41% of referrers, brings 31% of Referral Traffic to the ISNA.

There exist some other TLDs such as me (Montenegro) which are referring a great monthly traffic to ISNA; however, they are not listed in the most frequent referrer TLDs.

As around 15% of ISNA referral traffic requests are initiated from nationally access blocked websites, the dispersion of TLDs reported in Figures 7 and 8 can be interpreted to the users equipped with VPNs or other filtering bypassing tools.

6.2 Modeling

For model selection, three main clustering algorithms were applied on data with the following overall results (Table 5).

Table 5. Applied clustering models and overall achieved results

Approach	Model	Result
Density-based	DBSCAN	The model was incapable of detecting distinct clusters and resulted in only one big cluster.
Connectivity-based	Hierarchical	The model resulted in 10019 clusters, which did not convey enough information for the purpose of our modeling. It also took a great amount of time in comparison with other algorithms.
Centroid-based	K-means	The model led to predefined number of clusters with different performances due to parameter settings.

As Table 5 shows, one of the most common clustering algorithms DBSCAN failed to detect more than one cluster. It may be because of the condense density of the dataset in two-dimension space, Although preprocessing activities proved nonuniformity distribution of variables. Hierarchical modeling also let to tremendous number of clusters in several hierarchies. Consequently, K-means clustering was opted. The next task was to determine the right value for K, the number of clusters.

For model evaluation against different values of K, as an unsupervised technique was used, two indices of the average within centroid distance and Davies Bouldin Index (DBI) were examined. It should be mentioned that values are multiplied by -1 for the optimization propose, however it does not effect on the selection process.

Results from the average within centroid distance calculation for different number of clusters in K-means clustering model from 2 to 10 are provided in Fig. 9. It is calculated by averaging the distance between the centroid and all examples of a cluster. The average within centroid distance is a measure of clusters compactness and reduces as the clusters become more compact.

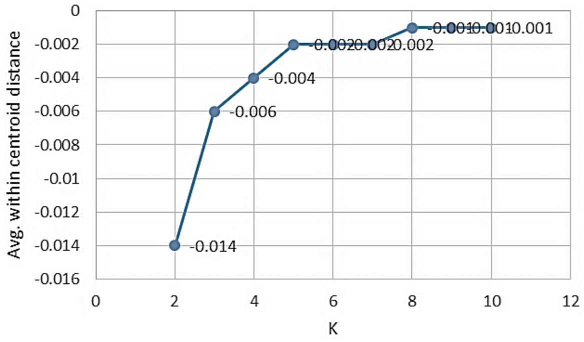

Fig. 9. The Average within centroid distance for different values of K (the number of clusters)

Fig. 9 reveals that the quality of modeling practice is in its best for Ks equal to 2, 3 and 4. It also decreases as K value increases.

DBI is another clustering quality criterion: the lower the value of DBI, the better the quality of clustering. Fig. 10 shows DBI values regarding to K variation from 2 to 10.

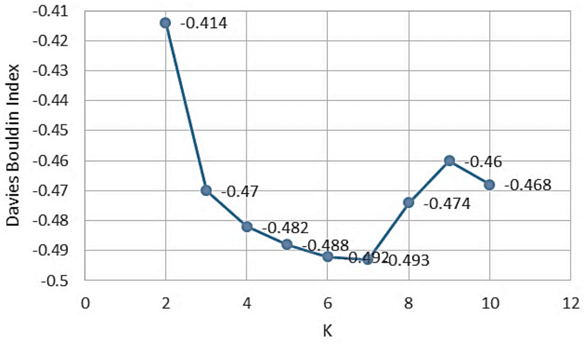

Fig. 10. DBI calculated for different values of K (the number of clusters)

The figure shows that DBI decreases dramatically when K goes from 2 to 3. It then decreases smoothly as K moves from 3 to 7. The curve rises when K goes from 7 to 9.

As a trade-off between both quality indices, the best clustering model could be selected on K=3, where the average within centroid distance and DBI are both presenting a relatively good quality simultaneously.

Finally, the performance evaluation results of clustering model for K=3 are provided in Table 6.

Table 6. Performance evaluation results for K=3

Criterion	Result
Avg. within centroid distance	-0.006
Avg. within centroid distance_cluster_0	-0.006
Avg. within centroid distance_cluster_1	-0.011
Avg. within centroid distance_cluster_2	-0.003
DBI	-0.470

Table 6 shows that the most compact cluster is cluster_2, follows by cluster_0.

Visual clustering results are provided in Figure 11. The vertical axis represents referrer's Alexa Ranks, while the horizontal axis is for the referral traffic during one month.

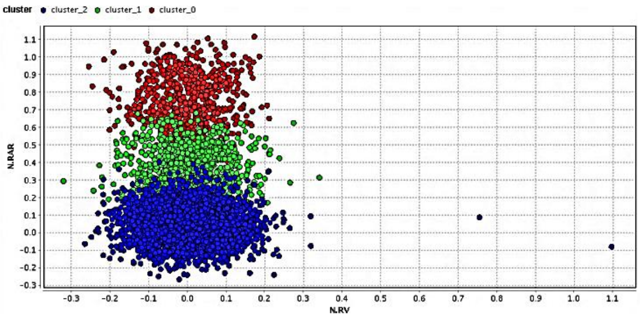

Fig. 11. Detected clusters (with added jitter) for K-means Squared Euclidean Distance algorithm (K=3)

Cluster modeling leads to three separate clusters with the characteristics mentioned in Table 7.

Table 7. Generated cluster modeling

Name of Cluster	Color	No. of Items
Cluster_0	Red	670
Cluster_1	Green	996
Cluster_2	Blue	3,885

The Interpretation of results by field experts is one of the most important steps in data mining projects. This task connects the results of a research to the real life situations. Some notable points about research results are provided below.

Referring to the dataset, it can be observed that most of the cluster_0 members are ir TLDs or com TLDs which are hosted in Iran.

Moreover, members of cluster_2 are mostly automatic news aggregator websites. Automatic news aggregators, also called feed aggregators, are types of software that collect news articles and other textual materials automatically and repeatedly from different online sources. They usually set the data in some predefined categories and deliver them to the end users in the form of comprehensive lists of news. The number of sources varies from hundreds to thousands. The different nature of such websites makes them important to detect them effectively in the process of traffic analysis. From this point of view, identifying them as a distinct cluster is a promising achievement of the clustering process.

In Alexa Rank system, a website with smaller rank has more traffic and generally seems most popular. Therefore, as the Alexa Rank increases, the popularity of the website decreases as it earns less monthly traffic. In this study, the most valuable referrer group would be cluster_2, which in the same time is the biggest cluster. In the same vein, the less important group of referrer is one with greater Alexa Ranks, cluster_0, which is also the smallest cluster with 670 members.

7. Conclusion and Further Studies

This research was the first step to study online news content reposting status and its penetration on the web. In this study, the goal was to understand the referral audience traffic for the website. Two metric were needed. The first metric, referral traffic, was opted to investigate the number of referrals to collect the number of referral sessions. The second metric, the Alexa Rank, was chosen to estimate the extent of importance or popularity of the referrer websites.

Between several common clustering models the K-means algorithm led to better results. And by examining several values, K=3 led to the best initial value for clustering.

Some new knowledge was achieved by utilizing data mining techniques. These findings lead to a better understanding about the origin and the nature of a website’s referral traffic. The findings of the study can be used for future online marketing activities, information quality improvements, and the development of better and more efficient business alliances and strategies.

Analyzing the reposting status of referrer websites and investigating the relation between reposting and referring manners are proposed as the next steps of this research.

Acknowledgements

We hereby extend our sincere appreciation to ISNA for assisting us during the process of accessing the data of the study.

References

[1] B. Kovach and T. Rosenstiel, The Elements of Journalism: What Newspeople Should Know and the Public Should Expect, Completely Updated and Revised: Random House Digital, Inc., 2007.

[2] J. B. Singer, "Five Ws and an H: Digital challenges in newspaper newsrooms and boardrooms," The International Journal on Media Management, vol. 10, pp. 122-129, 2008.

[3] D. E. Appelt, "Introduction to information extraction," Ai Communications, vol. 12, pp. 161-172, 1999.

[4] M. Karlsson and J. Strömbäck, "Freezing the flow of online news: Exploring approaches to the study of the liquidity of online news," Journalism Studies, vol. 11, pp. 2-19, 2010.

[5] X. Dai and Y. Sun, "Event identification within news topics," in IEEE International Conference onIntelligent Computing and Integrated Systems (ICISS), 2010, pp. 498-502.

[6] D. Shedden. (2014). New Media Timeline (1969-2010). Available: http://www.poynter.org/latest-news/business-news/tracker/28803/new-media-timeline/

[7] R. Watkins. (2014, 2/18/2015). Understanding Google Analytics: What is Acquisition? Available: http://southernweb.com/2014/01/understanding-google-analytics-acquisition/

[8] K. Ertell, "The Missing Links in the Customer Engagement Cycle," 2010.

[9] A. Vakali, J. Pokorný, and T. Dalamagas, "An overview of web data clustering practices," in Current Trends in Database Technology-EDBT 2004 Workshops, 2005, pp. 597-606.

[10] R. Iváncsy and I. Vajk, "Frequent pattern mining in web log data," Acta Polytechnica Hungarica, vol. 3, pp. 77-90, 2006.

[11] M. A. Omidvar, V. R. Mirabi, and N. Shokry, "Time Series modeling of visitors’ type on web analytics."

[12] M. A. Omidvar, V. R. Mirabi, and N. Shokry, "Analyzing the impact of visitors on page views with Google analytics," arXiv preprint arXiv:1102.0735, 2011.

[13] l. H. Xu and H. Liu, "Web User Clustering Analysis based on K-Means Algorithm," International Conference on Information, Networking and Automation (ICINA), vol. 2, pp. 6-9, 2010.

[14] Y. Xie and V. V. Phoha, "Web user clustering from access log using belief function," in Proceedings of the 1st international conference on Knowledge capture, 2001, pp. 202-208.

[15] Y. Fu, K. Sandhu, and M.-Y. Shih, "A generalization-based approach to clustering of web usage sessions," in Web Usage Analysis and User Profiling, ed: Springer, 2000, pp. 21-38.

[16] B. Mobasher, R. Cooley, and J. Srivastava, "Creating adaptive web sites through usage-based clustering of URLs," in Knowledge and Data Engineering Exchange, 1999.(KDEX'99) Proceedings. 1999 Workshop on, 1999, pp. 19-25.

[17] M. K. Jiawei Han, "Data mining: concepts and techniques," in TheMorgan Kaufmann Series in DataManagement Systems, ed: Elsevier, 2006, pp. 1-55860. p7.

[18] B. J. Davies and N. F. Glasser, "Analysis of www. AntarcticGlaciers. org as a tool for online science communication," Journal of Glaciology, vol. 60, pp. 399-406, 2014.

[19] Google Support Page. (2/19/2015). How a session is defined in Analytics. Available: https://support.google.com/analytics/answer/2731565?hl=en

[20] T. Raudenbusch, "Can Google Analytics be a reasonable alternative to collect data for Web Usage Mining?," E-Business Applications, 2013.

[21] P. Sawant and R. Kulkarni, "A Knowledge Based Methodology To Understand The User Browsing Behavior For Quality Measurement Of The Websites Using Web Usage Mining," International Journal Of Engineering And Computer Science, vol. 2, pp. 1522-1538, 2013.

[22] W3Techs. (2/13/2015). Usage of traffic analysis tools for websites. Available: http://w3techs.com/technologies/overview/traffic_analysis/all

[23] Alexa. (2015, 5/29/2015). Alexa Internet Terms of Use Agreement. Available: http://www.alexa.com/help/terms

[24] Google. (2014, 5/29/2015). Google Terms of Service, Part:Using our Services. Available: http://www.google.com/intl/en/policies/terms/

[25] Alexa. (2/19/2015). About Us. Available: http://www.alexa.com/about

[26] A. Olteanu, S. Peshterliev, X. Liu, and K. Aberer, "Web credibility: Features exploration and credibility prediction," in Advances in Information Retrieval, ed: Springer, 2013, pp. 557-568.

[27] A. Thakur, A. Sangal, and H. Bindra, "Quantitative measurement and comparison of effects of various search engine optimization parameters on Alexa Traffic Rank," International Journal of Computer Applications, vol. 26, pp. 15-23, 2011.

[28] J. Kosur. (2013, 7/2/2015). Facebook rules social media referral traffic for mainstream media. Available: http://socialnewsdaily.com/15911/facebook-rules-social-media-referral-traffic-for-mainstream-media/

[29] N. Karthikeyani Visalakshi and K. Thangavel, "Impact of Normalization in Distributed K-Means Clustering," International Journal of Soft Computing, vol. 4, pp. 168-172, 2009.

[30] M. K. Jiawei Han, "Data mining: concepts and techniques," in TheMorgan Kaufmann Series in DataManagement Systems, ed: Elsevier, 2006, pp. 1-55860. p60.

Roya Hassanian-Esfahani received her B.Sc in Electronic Engineering and M.Sc in Information Technology Engineering from Shiraz University, Iran. As a current Ph.D Candidate in Research Institute for Information and Communication Technology, ACECR, Tehran, Iran, her focus is on news mining. Hassanian-esfahani's dominant research interests include web information quality assessments, data mining and web mining.

Mohammad Javad Kargar received B.Sc and M.Sc degree in computer engineering from IAU- University of Science and Research, and Ph.D Degree in Information Technology and Multimedia system from University Putra Malaysia in 2008. He is currently an assistant professor at the department of computer engineering in University of Science and Culture, Tehran, Iran. He is founder of the International Conference on Web Research Conference series. His research interests include web and Information Quality, Web and Data Mining and Distributed systems