Investigating Cluster Stability when Analyzing Transaction Logs

Daniel Grech
University of Sheffield
Sheffield
United Kingdom
dgre090@gmail.com

Paul Clough*
University of Sheffield
Sheffield
United Kingdom
p.d.clough@sheffield.ac.uk

ABSTRACT

Data-driven approaches have become increasingly popular as a means for analyzing transaction logs from web search engines and digital libraries, for example using cluster analysis to identify common patterns of search and navigation behavior. However, steps must be taken to ensure that results are reliable and repeatable. Although clustering patterns of user interaction behavior has been previously explored, one aspect that has received less attention is *cluster stability* that can be used to aid cluster validation. In this paper we compute stability based on the Jaccard coefficient to investigate the cluster stability when using different subsets of transaction log data from WorldCat.org. Results provide insights into different types of search behaviors and highlight that clusters of varying degrees of stability will result from the clustering process. However, we show that additional investigation beyond the results of cluster stability is required to fully validate the resulting clusters.

1. INTRODUCTION

With the increased availability of user-system interaction data has come reliance on the use of data-driven techniques for mining and analyzing data. One particular area that has received considerable interest in recent years is search or transaction log mining [1]. Valuable insights can be gained from analyzing the traces people leave when they search for and navigate digital information. These transaction logs provide a unique resource to drive the next generation of digital services and applications [2]. The use of unsupervised learning techniques, such as *clustering*, have been widely used for various tasks in transaction log analysis. This includes identifying user interests from query logs, grouping query refinements according to users' information needs, providing query suggestions, identifying changes in user intent and identifying tasks [1, 2].

In this paper we investigate categorizing users' general search and navigational patterns using user-system interaction data derived from WorldCat.org transaction logs, together with cluster analysis. Clustering is often used as a data-driven method to explore and group common patterns of interaction [3, 4, 5]. However, the use of cluster analysis raises many questions, such as what is the optimal set of clusters and how 'good' are the resulting clusters? One approach to validate the results of clustering is to assess *cluster stability* [6]. The stability of clusters could be affected by a variety of factors, including sample size, selected features, algorithm, parameter settings and distance/similarity metric. The idea behind cluster stability is that the optimal clustering of a data set is the clustering that is most stable. Despite the importance of cluster stability, however, there is little empirical work undertaken in the area of transaction log analysis.

We use a simple method based on the Jaccard coefficient and subsets of data to determine cluster stability. Results highlight that clusters of varying stability are produced when using cluster analysis. However, results also show the limitations with relying solely on cluster stability to validate clustering outputs and the need for manual inspection. Our work is similar to [7], but instead of using cluster stability to determine the optimal number of clusters we use it to help with validating the results of clustering. The following research questions are considered: *[RQ1] How stable are the clusters produced from applying cluster analysis to a sample of Worldcat.org transaction logs?* and *[RQ2] How can cluster stability be used to validate the clustering of transaction logs?* The remainder of the paper is structured as follows. Section 2 describes related work; Section 3 describes our experimental setup; Sections 4 and 5 present and discuss results; finally Section 6 concludes the paper and provides avenues for further research.

2. RELATED WORK

2.1 Mining transaction logs

This paper investigates validity when clustering transaction logs into groups that signify distinct patterns of user-system interaction. Examples of past work that have analyzed search patterns of transaction logs using cluster analysis include Chen & Cooper [4] who applied hierarchical agglomerative clustering to detect distinct patterns of user behavior for a library catalog system. They manually derived 47 variables that could be extracted from the transaction

*This is the primary contact author.

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

logs of an online library catalog. They used cluster analysis to find groups of similar sessions and came up with six clusters of general usage patterns. Wolfram et al. [3] used similar a similar approach to identify distinctive session characteristics from three web search transaction logs. Stenmark [5] used self-organizing maps (SOMs) to identify clusters of user behavior in intranet search logs. Weber & Jaimes [8] base their study of usage patterns on Broder's taxonomy of user intent, but perform automated analyses of users' sessions using features derived from activities within the session and also from external demographic information. Using k-means clustering they show that different user profiles have distinct patterns of search behavior. Jones & Klinker [9] also use supervised learning to automatically segment sessions into higher level *missions* and lower level *goals*. In most cases the stability of clustering is not investigated or reported. Heer et al. [7] is one of the exceptions in which they investigate user-system interaction activity using cluster stability to determine the optimal number of clusters.

2.2 Cluster stability

Clustering methods will generate clusterings for almost any dataset, even if the data is homogeneous by nature. Therefore, performing some form of cluster validation to ensure that the clusters produced are meaningful is an important step in the process [7]. A key concept in the field of cluster validation is the notion of *cluster stability*. The idea behind cluster stability is that the optimal clustering of a dataset is the clustering that is most stable. Hennig [6] captures the intuition behind cluster stability by stating that clusters which are meaningful and valid "shouldn't disappear easily if the data set is changed in a non-essential way". This viewpoint is shared by Von Luxburg [10] who states that a clustering structure on a data set is stable if when it is applied to "several data sets from the same underlying model or of the same data generating process", it outputs fairly similar results. Von Luxburg insists that, in the field of cluster stability, the way that clusters look is not important; all that matters is the clusters can be constructed in a stable manner.

Ben-David et al. [11] describe the possible scenarios that can lead to unstable clusters and state that such clusters are usually the result of one of the following phenomena:

Multiple global optima. If the global optimizer of the clustering objective function is not unique then this will always lead to unstable clusters.

Small sample size. If the sample size is not large enough to ensure that the cluster structure is well-pronounced, then instability will be observed.

Algorithmic instability. If the clustering algorithm can converge to very different solutions by ending up at different local optima then instability will be present. Such instability would not exist if an algorithm which always terminates at the global optimum existed.

Geometric instability. It is possible that the mechanism behind stability based model selection is not consistent with the geometric model of the underlying distribution leading to low score for stability.

All of the definitions emphasize that clusters which are meaningful should be reproducible using different datasets from the same underlying distribution or under slightly different algorithmic conditions. In this study we identify different 'types' of sessions within a large transaction log using clustering and validate results by quantifying and analyzing the stability of the resulting clusters using the Jaccard Coefficient (see Section 3.3).

3. METHODOLOGY

3.1 Dataset

The dataset used to investigate cluster stability is a search log from WorldCat.org, the world's largest bibliographic data base, with more than 300 million bibliographic records and over 2 billion holdings from more than 70,000 libraries[1]. Log data for two months of WorldCat.org (October 2012 and April 2013) were used (74,711,963 entries in total). Preparation of the logs included filtering out non-human traffic, such as web search engine crawlers, together with segmenting the logs into *sessions* (with sessions consisting of more than 100 queries removed). A simple and efficient time-based method using a 30 minute cut-off period was used to segment sessions, resulting in 25,395,469 user sessions[2]. This paper is not concerned with defining sessions or broader units of interaction, such as tasks or missions [9]; rather, we focus on validating the outputs of cluster analysis on transaction logs where session boundaries have already been defined.

The log data contains many types of user-system interactions, such as the user issuing a query, selecting to view an item from the search results, viewing other pages from the WorldCat interface, navigating links (e.g., related items or further information), logging into a user account, etc. The use of different features from the log are being investigated to characterize patterns of search and navigation; however, in this paper we focus on a small subset of features and the issue of cluster validation. The average duration of a session is 41 secs (σ=81 secs) with 55% of sessions originating from the US. The mean number of actions (queries and viewed items) is 3.07 per session, with a typical session consisting of 2 item views (μ=2.2, σ=2.62). 42% of sessions consist of single query searches with the mean number of queries being 3.3 queries per session (σ=4.1). Sessions with viewed items only are typically referrals from external sites, such as web search engines, and subsequently do not contain query actions that occur within WorldCat.

3.2 Cluster analysis

To cluster sessions from WorldCat.org we first extract the following descriptive features to represent sessions (based on [3]): (i) duration of session; (ii) number of queries used to search for items (issued within WorldCat); (iii) average query length; (iv) number of viewed item pages (excluding clicks on other links within WorldCat, such as viewing help and login pages); and (v) number of different subjects viewed. The final feature represents the diversity in subject of the items (e.g., books or DVDs) that users view. This is based on subject information provided about each bibliographic resource provided by OCLC. In future work we plan to investigate a wider range of features to capture richer

[1] http://www.oclc.org/worldcat/catalog.en.html

[2] Various methods can be used to segment transaction logs into sessions, but the 30 minute cut-off heuristic was shown to be adequate for this task (see [12] for further details).

and more complete interaction patterns. Prior to clustering, the scores for the features must be normalized. In this study the feature values were scale normalized using Principal Component Analysis (PCA), which we also used to conduct preliminary exploration of the dataset. Five principal components emerged from the dataset with the first two accounting for 55% of the variability. The clustering algorithm was subsequently applied to the data projected onto the principal components.

Various clustering algorithms have been used in prior work. In our work we use the DBSCAN density-based clustering algorithm [13] as its execution time is almost linear and therefore highly suited to larger datasets, such as transaction logs. Density-based clustering algorithms represent the data in a spatial manner and aim to find regions of high density separated from low density regions. Such an approach has the advantages of being able to identify irregular shaped clusters, requiring only one dataset iteration and does not require the number of clusters to be defined prior to initialization. Two parameters must be set for DBSCAN: ϵ and $MinPts$. In DBSCAN a cluster is defined as a set of densely-connected points (controlled by ϵ) which maximize density-reachability and must contain at least $MinPts$ points. Parameter values of $\epsilon = 0.4$ and $MinPts = 200$ were chosen through empirical investigation.

3.3 Computing stability

We used Hennig's approach for assessing cluster stability that uses the Jaccard coefficient as a measure of similarity between two sets based on set membership [6]. The approach assesses stability by re-sampling the original dataset with the assumption that points drawn from the same underlying distribution should give rise to more or less the same clusterings. The procedure used is as follows:

1. Cluster the entire dataset. This is the 'best' clustering of the dataset as it includes all data points.

2. Re-sample new datasets from the original one and cluster again.

3. For every cluster in the original clustering find the most similar cluster (i.e., that with the highest similarity score) in the new clustering using the Jaccard coefficient and record its value.

4. Compute the cluster stability for every cluster in the original clustering as the mean of the similarity scores over the re-sampled datasets.

In our study cluster stability was calculated by applying the clustering process 100 times on samples (using sampling without replacement) of the original dataset, each containing 10,000 sessions.

4. RESULTS

DBSCAN produced a group of 10 clusters from the log data with around 20% classified as 'noise' – points too far away from any of the produced clusters to be considered for inclusion and discarded from further analyses. Table 1 summarizes the clusters and shows mean values for the original features, as well as stability scores. The clusters can be mapped to attributes through the use of session identifiers. Clusters 1 and 2 account for 54% of the sessions with stability scores of 0.87 and 0.85 respectively.

The most stable clusters are clusters 4 and 5 that account for 19% of sessions. These four clusters have stability scores over 0.80 that suggests they are unlikely to have arisen from the clustering process by chance. Clusters 8, 9 and 10 are the least stable and suggest noise or that they should have been combined with other clusters. The estimated stability is moderately correlated with the size of the cluster ($r = 0.67$, $p < 0.05$) indicating that larger clusters *generally* tend to be more stable than smaller ones.

Cluster 1 is similar to clusters 3, 6, 9 and 10. In fact, these clusters all share the same mean values (0.00) for number of queries and average query length. However, they vary in session duration and the magnitude of items viewed. Sessions that belong to cluster 3 vary from those in cluster 1 by being longer (around 5 minutes in length) and containing more item views which tend to span across two subjects on average. Clusters 6, 9 and 10 are also similar but vary in duration, number of items viewed and number of subjects viewed. These typically reflect sessions in which users are referred to WorldCat.org and therefore have searched in external websites and do not query in WorldCat (i.e., number of queries = 0).

Cluster 2 is the second largest cluster comprising around 25% of the data points not classified as noise. On average, sessions in this cluster have a duration of around 2 mins, contain 1-2 queries (of around 3 words) and involve viewing 1-2 items which generally belong to the same subject.

Clust Num	Size (%)	Duration (secs)	Num of Queries	Query Len	Item Count	Subj Count	Stab score
1	(30%)	123.98	0.00	0.00	2.15	1.00	0.87
2	(24%)	117.31	1.61	3.31	1.51	1.00	0.85
3	(12%)	298.15	0.00	0.00	3.07	2.00	0.76
4	(10%)	468.49	0.00	0.00	0.00	0.00	0.96
5	(9%)	131.18	2.42	2.79	0.00	0.00	0.91
6	(3%)	508.63	0.00	0.00	4.50	3.00	0.66
7	(8%)	211.15	1.83	2.95	2.74	2.00	0.68
8	(2%)	225.28	1.87	2.37	3.64	3.00	0.45
9	(1%)	340.57	0.00	0.00	4.70	4.00	0.44
10	(1%)	296.07	0.00	0.00	7.63	4.00	0.20

Table 1: Cluster summary: size, mean values for original features and stability scores

Clusters 5, 7 and 8 are similar to cluster 2 in the sense they contain sessions involving a couple of queries and viewing of items which tend to belong to different subjects. The average query length is fairly constant across all sessions in these clusters. Sessions in cluster 5 contain the most queries on average (μ=2.42), although do not lead to the user viewing items. On the other hand, sessions in cluster 7 contain less queries on average (μ=1.87), but users in such sessions tend to view on average 3-4 items from 2 or more subjects. Similarly, for users in cluster 8, the average number of queries is 1.87.

The distinguishing factor, however, between sessions in clusters 7 and 8 seem to be that users who are part of the sessions in cluster 8 tend to view a larger number of items which usually span across an average of 3 different subjects. Initial inspection would suggest that cluster 4 seems to be unique in comparison with other clusters. Sessions in this cluster tend to last around 8 minutes but do not contain any queries or the viewing of page items. Reasons for this could include: errors in the sessionization process, interaction behavior not captured by the current feature set or people using Worldcat as a service within external sites.

5. DISCUSSION

With regards to RQ1 cluster stability scores range from 0.20 to 0.96. DBSCAN successfully identifies different types of patterns of user-system interaction that can be interpreted in light of how users interact with WorldCat. However, before drawing inferences from the resulting clusters it is essential to validate the results to reduce the possibility that the clusters were identified by chance and do not actually reflect differences in the underlying data. In relation to DBSCAN unstable clusters represent data points that should either have formed part of another cluster or should have been classified as noise. From results presented in Section 4, the indications are that the most unstable clusters (clusters 8, 9 and 10) should probably have formed part of other more stable clusters. One possible reason for this could be the fact that the ϵ parameter of DBSCAN is a global parameter and cannot be adjusted per-cluster.

With respect to RQ2 cluster stability scores can be used help determine the optimum number of clusters and evaluate the "goodness" of the resulting clusters [7]. Hennig [6] states that large stability values do not necessarily indicate that the underlying clusters are valid. However, he also emphasizes that small stability values are always informative, indicating that the underlying clusters are either meaningless in relation to the true underlying model, or that instabilities exist in the clusters or the clustering methods used. In the case of the results in Table 4 the most stable cluster (cluster 4) is markedly different from other clusters and is likely indicative of users who are using Worldcat via external services, the actions of which are not captured in the current feature set. Since these sessions do not reflect user activity within Worldcat.org one might argue they should be filtered out along with robot traffic. There may also be other reasons for cluster 4, but this does suggest that despite stable clusters typically being meaningful and valid [6] more in-depth analyses must be carried out to better interpret the clusters and gain a complete and accurate picture of user behavior [7, 14]. When utilizing data-driven approaches then applying methods for validating results is important. However, cluster stability alone is not enough to fully validate the results of clustering.

6. CONCLUSIONS

This paper has investigated cluster stability for identifying groups of sessions based on features indicative of user-system interaction. The DBSCAN clustering algorithm is used to form clusters that are then validated using a simple approach for assessing cluster stability based on comparing clusters from samples of the dataset with the original clustering using the Jaccard coefficient. As one might expect the results of the clustering contained a mix of stable and unstable clusters. There is clearly instability when clustering that calls for the need to model varying parameters to arrive at a stable set of clusters. Future work will investigate stability with respect to other criteria, such as varying feature sets, sample sizes, parameter settings and alternative log data. In addition, we plan to investigate alternative methods for computing stability.

7. ACKNOWLEDGEMENTS

Work partially supported by the Google Faculty Award "Developing a taxonomy of search sessions". We also thank OCLC for providing anonymised transaction logs.

8. REFERENCES

[1] Silvestri, F.: Mining query logs: Turning search usage data into knowledge. Found. Trends Inf. Retr. **4** (2010) 1–174

[2] Agosti, M., Crivellari, F., Di Nunzio, G.M.: Web log analysis: A review of a decade of studies about information acquisition, inspection and interpretation of user interaction. Data Min. Knowl. Discov. **24** (2012) 663–696

[3] Wolfram, D., Wang, P., Zhang, J.: Identifying web search session patterns using cluster analysis: A comparison of three search environments. Journal of the American Society for Information Science and Technology **60** (2009) 896–910

[4] Chen, H.M., Cooper, M.D.: Using clustering techniques to detect usage patterns in a web-based information system. Journal of the American Society for Information Science and Technology **52** (2001) 888–904

[5] Stenmark, D.: Identifying clusters of user behavior in intranet search engine log files. Journal of the American Society for Information Science and Technology **59** (2008) 2232–2243

[6] Hennig, C.: Cluster-wise assessment of cluster stability. Computational Statistics and Data Analysis (2007) 258–271

[7] Heer, J., Chi, E.H., Chi, H.: Mining the structure of user activity using cluster stability. In: in Proceedings of the Workshop on Web Analytics, SIAM Conference on Data Mining (Arlington VA, ACM Press (2002)

[8] Weber, I., Jaimes, A.: Who uses web search for what: And how. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. WSDM '11, New York, NY, USA, ACM (2011) 15–24

[9] Jones, R., Klinkner, K.L.: Beyond the session timeout: Automatic hierarchical segmentation of search topics in query logs. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management. CIKM '08, New York, NY, USA, ACM (2008) 699–708

[10] Von Luxburg, U.: Clustering stability: An overview. Now Publishers Inc (2010)

[11] Ben-David, S., Von Luxburg, U.: Relating clustering stability to properties of cluster boundaries. COLT **2008** (2008) 379–390

[12] Wakeling, S., Clough, P.: Determining the Optimal Session Interval for Transaction Log Analysis of an Online Library Catalogue. In: Proceedings of the 38th European Conference on IR Research (ECIR'16). Springer (2016) 703–708

[13] Ester, M., peter Kriegel, H., S, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise, AAAI Press (1996) 226–231

[14] Grimes, C., Tang, D., Russell, D.M.: Query Logs Alone are not Enough. In: Proceedings of the WWW 2007 Workshop on Query Logs Analysis: Social and Technological Challenges. (2007)