Audience Prism: Segmentation and Early Classification of Visitors Based on Reading Interests

Lilly Kumari
IIT Roorkee
Uttarakhand, India
lillyupt@iitr.ac.in

Sunny Dhamnani
IIT Kharagpur
West Bengal, India
sunnyd@cse.iitkgp.ernet.in

Akshat Bhatnagar
IIT Kanpur
Uttar Pradesh, India
akshatb@iitk.ac.in

Atanu R. Sinha
Leeds School of Business
University of Colorado
Boulder, Colorado, USA
atanu.sinha@colorado.edu

Ritwik Sinha
Adobe Research, Bangalore
Karnataka, India
ritwik.sinha@adobe.com

ABSTRACT

The largest Media and Entertainment (M&E) web portals today cater to more than 100 Million unique visitors every month. In Customer Relationship Management, customer segmentation plays an important role, with the goal of targeting different products for different segments. Marketers segment their customers based on customer attributes. In the non-subscription based media business, the customer is analogous to the visitor, the product to the content, and a purchase to consumption. Knowing which segment an audience member belongs to, enables better engagement. In this work, we address the problems: 1) How can we segment audience members of an M&E web property based on their media consumption interests? 2) When a new visitor arrives, how can we classify them into one of the above defined segments (without having to wait for consumption history)? We apply our proposed solution to a real world data-set and show that we can achieve coherent clusters and can predict cluster membership with a high level of accuracy. We also build a tool that the editors can find valuable towards understanding their audience.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Clustering, Information filtering, Selection process*

Keywords

Digital Media Analytics, Clustering, Segmentation, Prediction, Classification

CODS '16, March 13-16, 2016, Pune, India
© 2016 ACM. ISBN 978-1-4503-4217-9/16/03…$15.00
DOI: http://dx.doi.org/10.1145/2888451.2888459

1. INTRODUCTION

According to the Pew Research Center [10], major Media and Entertainment (M&E) websites today serve large audiences. The largest of these cater to more than 100 Million unique visitors every month. The editors of these organizations today are limited in their ability to understand and take appropriate actions to engage the large volume of traffic they receive. Organizations would like to have a deeper understanding of their audience's interests, with the goal of catering to their needs and wants. Better understanding can lead to better personalization. The larger the number of highly engaged visitors to a media outlet's web properties (websites or apps), the better is their ability to demand higher rates from advertisers.

In this work we aim to achieve two goals. The first is to segment visitors: How can we segment the audience of an M&E web property based on their media consumption interests? Second, predicting segment membership: When a new visitor arrives, how can we classify them into one of the above defined segments, without having to wait for consumption history? In the first part of the problem, we are interested in finding coherent groups of audience members of a Media site, where the grouping is performed on audience interest. While such post-facto analysis may be of interest to an editor, the real power of our approach can be realized by classifying a new visitor into one of these segments. Hence, in the second part, we want to address the challenge of a "cold start"; that is, when an audience member visits the website for the first time, the website does not have any clue about the visitor and hence may not know the best way to engage this visitor.

The advantage of solving this problem are manyfold, we describe these next. The information we generate may be used for personalizing the visitor's view of the website. Also, past content missed by the visitor may be recommended to her. A further application is to provide a personalized summary of content for the visitor in the form of an email digest. Targeting specific user segments for genres of media content which have limited readership, can also be achieved with this approach. Finally, our approach can help by providing advertisers the ability to understand audience groups when they advertise on certain content. While we believe these are natural applications of our work, we have not implemented them in this paper.

The paper is organized as follows. In section 2, we describe related work in this direction. Next, in section 3, we describe the methodological details of our approach. We describe our data-set and exploratory analysis of the data in section 4. The results of our approach are detailed in section 5. We finally conclude the paper in section 6. In the rest of the paper, we will use visitor and audience interchangeably, and use publisher and M&E site interchangeably.

2. RELATED WORK

Analysis of social media data, and models of information propagation on social media has received significant attention. Specifically, the problem of inferring latent interests of users on such networks has received attention. For instance, in some studies, a user's interests have been inferred from her neighbours [8, 18]. However, these attempts have exploited the inherent structure of OSNs, where one has the explicit "likes" of the members, in addition to the social graph. A more general approach to implicit affinity detection has been studied [16], but this work still relies on the existence of a social graph. None of these works directly translate to the setting of M&E web properties, where social graphs are not observable, and explicitly stated affinities are unavailable.

Recommender systems [7] provide another approach to studying these problems, but they are plagued by the cold-start problem, where statements cannot be made about a visitor until they have provided enough information about their interests by consuming content. Also, while a recommender system may have success in recommending the right set of articles, it is directly not applicable for the editor of an M&E site to extract insightful understanding of audiences.

The publisher ecosystem has three players, the content creators (publishers and M&E sites), the advertisers and the audience members. The academic study of this ecosystem has focused a lot on the relationship between the advertiser and the publisher [5]. Additionally, the relationship between the advertiser and the audience member has been studied; in particular, the question of estimating the value of individual channels and ads have been studied [15, 14, 19]. But, unlike in OSNs, little work has been published in the study of the relationship between the publisher and the audience members.

On the other hand, in the publishing industry, the leaders in data driven decisioning are trying to achieve some of this on their own. For example, The Huffington Post [13] performs significant testing of their content, as well as layouts. The New York Times [4] is trying to group articles based on its social media consumption, providing an advertiser the ability to buy a package of stories that are then trending on Twitter. However, the lack of a systematic study about the information which is contained in the relationship between the publisher and the audience provides a motivation for our work.

3. METHODS

In this section, we start with a description of the overall flow of our approach. We then dive into the technical details of our approach.

3.1 Workflow

There are two parts in our work. The first is segmenting visitors based on their interests in different content and the second is predicting segment membership, with as little information about the visitor as possible. Below are the steps performed in these two parts in more detail.

3.1.1 Segmenting Visitors

The first part of our work is an attempt to answer the following question: "How can we segment the audience members of an M&E web property based on their media consumption interests?". We start with click-stream data from the publisher's web property. For all unique articles that are consumed, we collect the topic distribution. A user's topic preference vector is computed by aggregating the topic distributions of all the articles consumed by this individual. Next, we cluster the individuals into groups using Spherical K-Means. We use Silhouette analysis to determine the optimal number of clusters in the data. Finally, to visualize the validity of our clustering approach, we perform non-parametric dimension reduction on this data and visualize it in lower dimensions (2 or 3). Technical details of each of these steps is provided in Section 3.2.1.

3.1.2 Predicting segment membership

In the second part of our paper, we address the question: "When a new visitor arrives, how can we rapidly classify that visitor into one of the above defined segments?". There are two aspects to answering this question. The first part requires building an offline model on historic data. This is tuned to predict with high accuracy, given a visitor's attributes, the segment membership of the visitor. In the second part, applied on streaming data, when a visitor arrives on the web property we predict the visitor's segment membership using the offline model built earlier. This information may be used for taking personalized action. The technical details of our approach are detailed in Section 3.2.2.

3.2 Approach

In this section, we describe the technical details of our approach.

3.2.1 Segmenting visitors

We start with the click-stream information of all visitors to the web property of an M&E portal. Let us say that the i^{th} visitor has seen k_i articles $A_{i1}, A_{i2}, ..., A_{ik_i}$. Let us further say that the j^{th} article has a topic distribution of $(T_{j1}, ..., T_{jl})'$, where T_{jm} is the level of the m^{th} topic expressed in the j^{th} article. The set of possible topics is prespecified to belong to a set of size l. This information may be captured using multiple tools that provide supervised article categorization using an existing knowledge base [11]. Some examples of such services include Alchemy API[1], Open Calais[2], TextWise[3], and uClassify[4]. Combining the topic distributions of all the articles consumed by a visitor, we get this visitor's topic preference,

$$\tilde{T^i} = (T^i_1, ..., T^i_l)' = \frac{1}{k}\sum_{j=1}^{k}(T_{j1}, \cdots, T_{jl})' \quad (1)$$

[1] http://www.alchemyapi.com/

[2] http://new.opencalais.com/opencalais-api/

[3] http://textwise.com/api/categorization/

[4] https://www.uclassify.com/

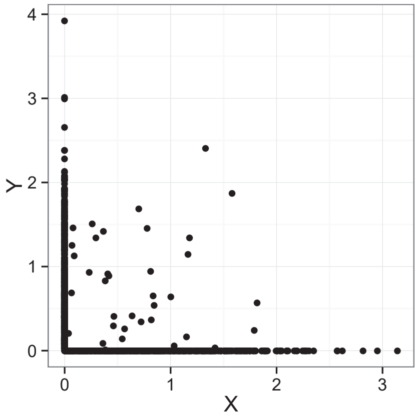

Figure 1: Simulated data to describe the applicability of Spherical K-Means to sparse high dimensional data. Note that in higher dimensions, a lot of the data will lie on lower dimensional sub-spaces because one visitor is unlikely to have read articles on many different topics.

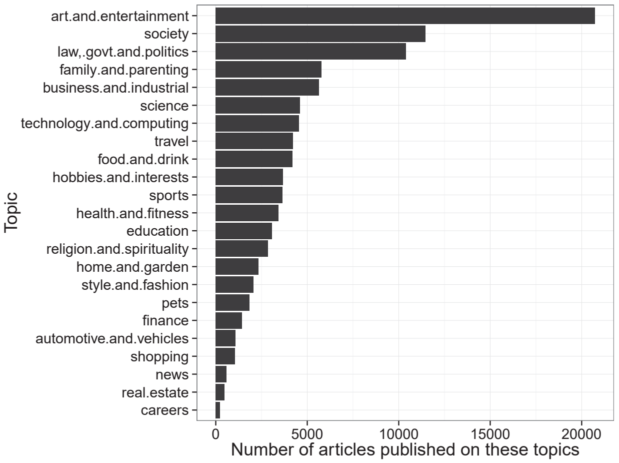

Figure 2: The distribution of articles which contain the different topics. Note that a single article may contain multiple topics, thus, the sum of over all topics can be greater than the total number of articles available in the data.

Thus, each visitor may be expressed as a vector in the l dimensional space, assuming there are l distinct topics. Next, we define the cosine distance between two individuals as one minus the cosine of the angles between their respective topic vectors. That is,

$$d(\tilde{T^i},\tilde{T^j}) = 1 - \frac{\sum\limits_{m=1}^{l} T_m^i \times T_m^j}{\sqrt{\sum\limits_{m=1}^{l} {T_m^i}^2}\sqrt{\sum\limits_{m=1}^{l} {T_m^j}^2}} \qquad (2)$$

Using the above distance function between two visitors, we can run the Spherical K-Means algorithm for cluster determination [3], where the centre of the cluster is set in such a manner that the angle between the points within a cluster is made uniform and minimal. Spherical K-Means is well suited for clustering of document corpora with sparseness in high dimensions [20]. Because of the fact that a single visitor is unlikely to have consumed articles from many different topics, individuals are unlikely to have non-zero values in many of the dimensions. In Figure 1, we simulated some data to describe the characteristics that justify choosing Spherical K-Means. As can be seen in Figure 1, in higher dimensions, a lot of the data may lie in lower dimensional sub-spaces. Using Euclidean distance, as in K-Means [2] is not correct given the nature of this data. Spherical K-Means on the other hand is appropriate for such data, given that it creates angular clusters as opposed to circular clusters. (In our data, out of the 23 topics enlisted in Section 4, visitors show significant sparseness in their topic preferences with the statistical mode of the distribution at 5, please refer to Figure 3 in Section 4).

The Spherical K-Means algorithm takes K (the number of clusters) as an input. The algorithm can be run for a range of values of K. Once the clusters are computed, we are still left with deciding what value of K leads to the most coherent clusters. This decision can be made based on a number of measures of cluster purity [1]. We decided to use the Silhouette coefficient [12], which contrasts the average within cluster distance with the average distance for a point from the nearest cluster. Silhouette analysis has a computational complexity of $O(n^2)$. To overcome the challenge of scalability, we computed the Silhouette coefficient based on a random sample of points from each cluster. With a sample of size 5000 from each cluster, this estimated Silhouette coefficient (which lies between 0 and 1) has a standard deviation bounded by 1/5000.

The average of $\tilde{T^i}s$ for all individuals i who belong to one cluster provides a measure of the topics that are most expressed within that cluster. In other words, these are the topics that are of most interest to visitors who fall in this segment.

Once clustering has been performed, it is natural to be interested in visualizing the clusters to see if they are coherent. The standard approach to dimension reduction for visualizing cluster coherence has been to apply Principal Component Analysis approaches to this data. However, since we are not using Euclidean distance, such an approach is not appropriate to our situation. To overcome this challenge, we apply the t-SNE [17] approach to reduce the dimensionality of our data. This approach performs a nonlinear feature space dimension reduction for data representation in lower dimensions. One has the option of selecting whether to display the data in 2 or 3 dimensions.

3.2.2 *Predicting Segment Membership*

While segmenting visitors based on their topic preferences is interesting, it is not sufficient to take any action. The reason is that the consumption needs to have already happened; it is a post-facto analysis that limits its use to an M&E web site. To overcome this challenge, we aim to predict segment membership of a visitor based on their first visit to the web property.

Let us say that based on historic data, segments of visitors have been built. Assume that these segments are $(C_1, ..., C_K)$. Then, each visitor falls in one of K segments. The challenge lies in being able to guess the segment before the visitor build her consumption history. This may be achieved by

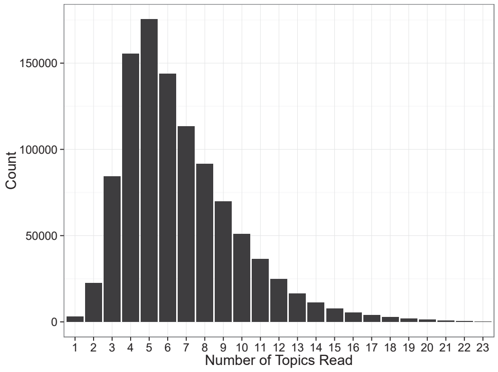

Figure 3: The number of topics read by each visitor. A vast majority of visitors have read articles which contain ten or fewer topics.

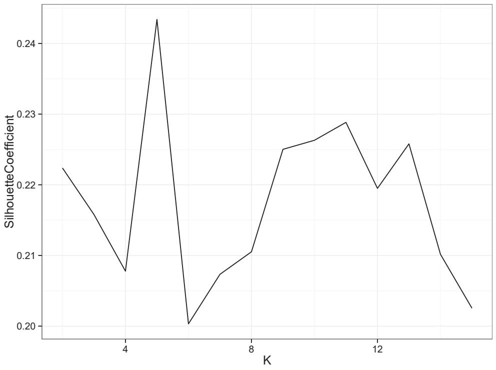

Figure 4: Silhouette Plot for Spherical K-Means to determine the right number of clusters in the data. The optimal choice is where this plot peaks, which leads to 5 clusters.

building a predictive model.

Given a set of attributes of a visitor, denoted by the vector $\tilde{x_i}$, we are interested in guessing y_i, the segment the i^{th} visitor belongs to (one of $C_1, ..., C_K$). This requires building the classifier $P(y_i = C_r|\tilde{x_i})$, where r $\in$ {1, 2, ..., K}. We explored different options, and settled on a multi-class Random Forest classifier [6] to estimate these probabilities. The class for which this predicted probability is the highest is the predicted segment of the visitor. We evaluated the quality of the fit by exploring the confusion matrix, and the precision and recall in each class.

As mentioned earlier, the vector $\tilde{x_i}$ contains information about a collection of attributes of the i^{th} visitor. In our data the following attributes were available and used. The first type of attribute was what are collected in typical web analytics tools[5] and the second class of features was the topic distribution of the first content consumed by the visitor. When a visitor arrives on the M&E web property and we observe the vector $\tilde{x_i}$, we can immediately make a prediction about the segment this individual belongs to based on the predictive model above. Thus, we know an individual's interests without any delay.

4. DATA DESCRIPTION

Our invention is implemented on data from a web magazine that covers current affairs, commentary, essays, fiction and so on. This data spans over one month of web activity (March of 2015). The data is filtered to extract only those readers who have read at least 2 articles during the month. This led to $1,023,399$ unique readers whom we group into segments based on their readership. The URLs for content read by these visitors added up to $41,049$ unique articles. The contents of these articles were collected and these were passed to Alchemy Taxonomy API which gives confidence scores corresponding to 23 high level topics[6] . Figure 2 shows the number of articles which contain each of the 23 topics. It is to be noted that a single article can span multiple topics.

The first goal of our work is to segment visitors based on their topic vectors. The crucial aspect that determined the choice of our analysis method is presented in Figure 3. From the figure, it is clear that most visitors have not been exposed to more that 10 different topics. Also, the statistical mode of this distribution is at 5. Given this data, it is easy to see that a large proportion of our data lies in a lower dimensional sub-space of the original 23 dimensional space. This provides further evidence to choose Spherical K-Means over K-Means, as expecting circular clusters in this data is not realistic.

5. RESULTS

In this section, we summarize the results of our approach applied to real data.

5.1 Segmenting Visitors

The first step of our work requires us to build Spherical K-Means for different values of K. We explored the range of values from 2 to 15 for K. This lead to the Silhouette plot in Figure 4. The point at which this plot peaks is where the clusters have the highest within cluster similarity and the smallest between cluster similarity; hence this is an optimal choice for the number of clusters. Thus, in the rest of the work, we decided to use 5 as the number of clusters for our analysis, as suggested by Figure 4.

Once, we identify the number of clusters and build the segments based on this choice, we can visualize the distribution of topics across the segments. Figure 5 provides this information. As can be seen, for this data-set the dominating topics in cluster 2 are "Law, Government and Politics" and "Society", while the dominating theme in cluster 5 are "Family and Parenting", "Sports" and "Technology and Computing". Cluster 4 is dominated by "Family and Parenting" and "Health and Fitness". Clusters 1, and 3, have only one dominating cluster each; these are "Art and Entertainment" and "Food and Drink", respectively.

After creation of the clusters, we explored the quality of

[5] "Visit Num", "Geo City", "Geo Country", "Geo DMA", "TZ Offset", "Browser", "OS", "OS Version", "Device", "Login Status", "First Hit Time GMT", "New vs Repeat"

[6] The 23 topics returned by Alchemy API are: "art and entertainment", "automotive and vehicles", "business and industrial", "careers", "education", "family and parenting", "finance", "food and drink", "health and fitness", "hobbies and interests", "home and garden", "law, govt and politics", "news", "pets", "real estate", "religion and spirituality", "science", "shopping", "society", "sports", "style and fashion", "technology and computing", and "travel".

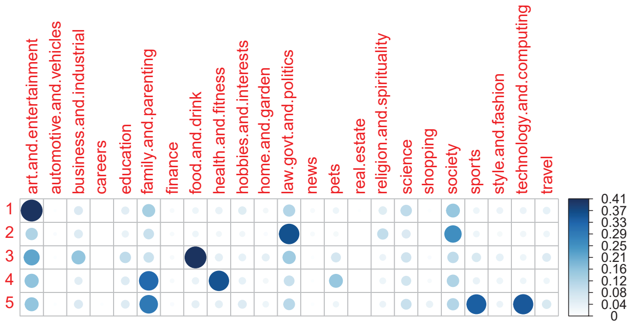

Figure 5: Distribution of topics across segments. The prevalence of a topic in a cluster is denoted by the size of the bubble as well as the intensity of the colour.

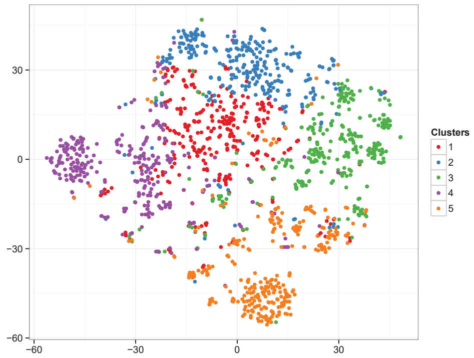

Figure 6: Representation of our data, reduced to two-dimensions using t-SNE. The color of the plot denotes the cluster assigned by the Spherical K-Means algorithm.

the fit by first performing t-SNE an approach to non-linear dimension reduction. While we can perform dimension reduction to any size, we experimented with reducing our data to 2 and 3 dimensions. This data (reduced to 2 dimensions) is plotted in Figure 6 for a sample of 1,500 points (with equal representation from each of the 5 clusters). As we see, our clusters have good separation. There are only a few instances where points of one color lie in a neighbourhood dominated by points of a different color.

Table 1: Confusion Matrix of our Classifier based on user attributes, and topic distribution of first article consumed (all numbers in %).

	Predicted Segment				
Observed Segment	1	2	3	4	5
1	12.26	3.07	1.57	1.54	1.56
2	3.57	12.43	1.42	1.39	1.20
3	4.96	3.16	7.93	1.93	2.02
4	4.66	3.30	1.86	8.53	1.65
5	4.86	3.59	1.99	1.95	7.61

Table 2: Confusion Matrix of our Classifier based on user attributes, and topic distribution of first *two* articles consumed (all numbers in %).

	Predicted Segment				
Observed Segment	1	2	3	4	5
1	13.53	2.57	1.54	1.28	1.09
2	2.90	14.01	1.17	0.95	0.97
3	3.35	2.20	12.31	1.13	1.02
4	3.19	2.28	1.53	12.09	0.91
5	3.55	2.36	1.67	1.28	11.14

5.2 Predicting Segment Membership

Once the segments are identified, we are interested in predicting a visitor's segment membership with the lowest possible delay from the time the individual comes to the web property. We build a Random Forest classifier [2] with 5 classes (for the 5 segments above) to achieve this goal. As features, we first attempt to use the user's attributes [7] (excluding their readership information). This lead to a classifier with very limited accuracy. To overcome this challenge, we next added the topic distribution of the first article consumed by the visitor, that is another 23 features. To ensure that we are not inflating the results of this analysis, we did two modifications to our analysis. First, we only took those visitors who read at least 6 articles on the web site. Next, we ensured that the segment of the customer was determined from all but the first article for this visitor.

With these variables, we were able to achieve a classifier that achieved significantly better results than random classification. Table 1 provides the confusion matrix for this prediction model. As we can see, about half (49%) of the predictions lie along the diagonal. At random, this proportion would be 20%. Thus, our classifier has a lift of 2.5 times over the random strategy. While this is better than not using the topics of the first article, we wondered if this can be further improved.

A natural next question is to ask if including the topic of the second article improves the accuracy further. We performed this analysis and the results are presented in Table 2. As we can see, we have an accuracy of 63%, which leads to a lift of 3.15 over the random targeting strategy. Table 3 provides the segment specific precision and recall for the different classes, for this classifier. Thus our approach to early prediction of visitor segment can be accomplished with a high level of confidence.

6. CONCLUSIONS

As the size of audience consuming digital media continues to grow, it has become important for digital media publishers to have a more personal, individual-level understanding of their audience for better audience engagement and retention. Current techniques provide only an aggregate-level understanding of media audience, based on visitor history while neglecting many of the attributes of every individual visitor to a website [9].

[7] "Visit Num", "Geo City", "Geo Country", "Geo DMA", "TZ Offset", "Browser", "OS", "OS Version", "Device", "Login Status", "First Hit Time GMT", "New vs Repeat"

Table 3: Classifier with user attributes and topics for first two articles. Segment specific precision and recall (all numbers in %).

Segment	Precision	Recall
1	51	68
2	60	70
3	68	62
4	72	60
5	74	56

In our work, we segment visitors to an M&E web property based on their media consumption interests. We applied the Spherical K-Means algorithm as it is more applicable to the characteristics of such data. We have shown that coherent clusters can be built. We further classify a new visitor to a website into one of these segments. We have shown that we are able to achieve a 2.5 times improvement over random targeting with user attributes and the topics of the first article read. In addition, with information about the first two articles consumed by a visitor, we can achieve a lift of 3.15 times over random targeting. This addresses the cold-start problem that many classification and recommendation algorithms face. We also built a web based tool that captures our technology of audience segmentation and cluster prediction for a new visitor. We hope that such a tool will help the M&E website understand its audiences better.

The information of visitor segment can be used for a variety of purposes like personalization of a visitor's view of the website, recommendation of past content, genre-wise targeting of specific user segments, personalized content summarization, and so on. Further research may be directed at understanding these applications better.

7. REFERENCES

[1] M. Charrad, N. Ghazzali, V. Boiteau, and A. Niknafs. Nbclust: an r package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, 61(6):1–36, 2014.

[2] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

[3] K. Hornik, I. Feinerer, M. Kober, and C. Buchta. Spherical k-means clustering. *Journal of Statistical Software*, 50(10):1–22, 2012.

[4] M. Ingram. The nyt is doing something smart by using twitter trends to target ads. http://bit.ly/1Of5xxx, 2015-10-04.

[5] H. Katz. *The media handbook: A complete guide to advertising media selection, planning, research, and buying*. Routledge, 2014.

[6] A. Liaw and M. Wiener. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[7] G. Linden, B. Smith, and J. York. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, 2003.

[8] J. Mcauley and J. Leskovec. Discovering social circles in ego networks. In *ACM Transactions on Knowledge Discovery from Data (TKDD)*, volume 8, February 2014.

[9] B. Mobasher, H. Dai, T. Luo, Y. Sun, and J. Zhu. Integrating web usage and content mining for more effective personalization. In *Electronic commerce and web technologies*, pages 165–176. Springer, 2000.

[10] K. Olmstead and E. Shearer. Digital news - audience: Fact sheet. http://pewrsr.ch/1DKhEJx, 2015-10-04.

[11] D. Quercia, H. Askham, and J. Crowcroft. Tweetlda: supervised topic classification and link prediction in twitter. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 247–250. ACM, 2012.

[12] P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[13] D. Segal. Arianna huffington's improbable, insatiable content machine. http://nyti.ms/1Ne8b6y, 2015-10-04.

[14] R. Sinha, S. Mehta, T. Bohra, and A. Krishnan. Improving marketing interactions by mining sequences. In *Web Information Systems Engineering, WISE 2015*. Springer, 2015.

[15] R. Sinha, S. Saini, and N. Anadhavelu. Estimating the incremental effects of interactions for marketing attribution. In *Behavior, Economic and Social Computing (BESC), 2014 International Conference on*, pages 1–6. IEEE, 2014.

[16] M. Smith. *Implicit affinity networks*. PhD thesis, Brigham Young University, 2007.

[17] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

[18] Z. Wen and C.-Y. Lin. On the quality of inferring interests from social neighbors. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010.

[19] M. M. Yadagiri, S. K. Saini, and R. Sinha. A non-parametric approach to the multi-channel attribution problem. In *Web Information Systems Engineering–WISE 2015*, pages 338–352. Springer, 2015.

[20] S. Zhong. Efficient online spherical k-means clustering. In *Proceedings of 2005 IEEE International Joint Conference on Neural Networks*, volume 5, pages 3180–3185. IEEE, 2005.