8/10/2017

Analysis of Online Users' Sequential Click Stream Data — Evidence from MSNBC

INSY 5377 Team Project Report

Instructor: Jie Zhang

Abishek Ganesh, Akanksha Shrivastava, Caston Fernandes, Jie Zhang, Padmavathi Karunaiananda Sekar

GROUP NAME: J-PAAC

TABLE OF CONTENTS

INTRODUCTION

MSNBC (formerly stylized as msnbc) is an American news cable and satellite television network that provides news coverage and political commentary from NBC News on current events. MSNBC and its website were created in 1996 by Microsoft and General Electric's NBC unit, which is now the Comcast-owned NBCUniversal. Microsoft invested $220 million for a 50% share of the cable network, while MSNBC and Microsoft would share the cost of a $200 million newsroom in Redmond, Washington for msnbc.com. NBC supplied the channel space and hence its name is a combination of "MSN" and "NBC". MSNBC was created as an alternative to CNN. The general news site was rebranded as NBCNews.com and a new msnbc.com was created as the online home of the cable news channel.

The network's goal of attracting a younger, tech-savvy audience in 1996 failed to materialize. In 1999, MSNBC began a partnership with The Washington Post that permitted more integrated coverage on the website. The msnbc.com website, a separate company, remained relatively successful, becoming the most-used online news site in 1997, 1998, and 1999.

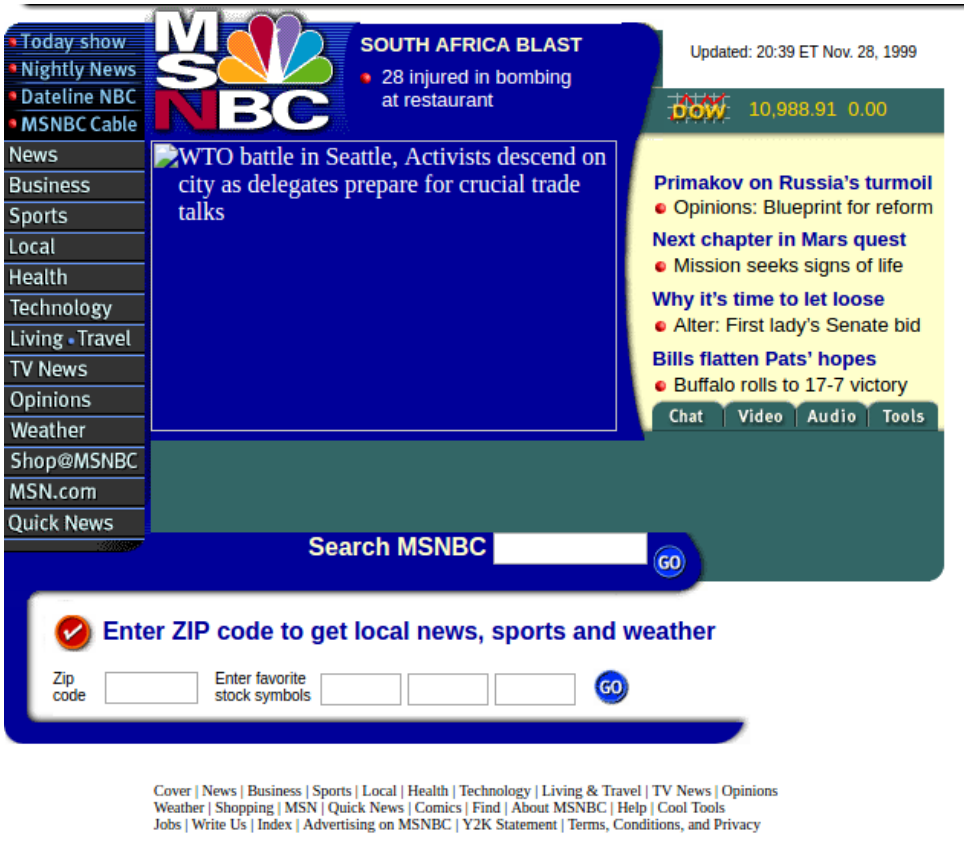

Fig 1: msnbc.com in 1999

SCOPE

Clickstream data is composed of thousands, millions or even billions of hits that tell the story of how, who, and what visitors are doing on your website. On a Web site, clickstream analysis (also called clickstream analytics) is the process of collecting, analyzing, and reporting aggregate data about which pages a website visitor visits -- and in what order. The path the visitor takes through a website is called the clickstream.

There are two levels of clickstream analysis, traffic analytics and e-commerce analytics. Traffic analytics operates at the server level and tracks how many pages are served to the user, how long it takes each page to load, how often the user hits the browser's back or stop button and how much data is transmitted before the user moves on. E-commerce based analysis uses clickstream data to determine the effectiveness of the site as a channel-to-market.

Clickstream allows you to automate the collection of events like tracking every login, button click or hover, purchase, abandonment of purchase or any other action etc. Analysis of this data provides the insights needed to deliver a better customer experience. It also allows your product and marketing teams identify opportunities and optimize their website or application to more impactfully reach our users. It can be used for improving/personalizing search or making responsive recommendations.

Because an extremely large volume of data can be gathered through clickstream analysis, many e-businesses rely on big data analytics and related tools such as Hadoop to help interpret the data and generate reports for specific areas of interest. Clickstream analysis is considered most effective when used in conjunction with other traditional market evaluation resources.

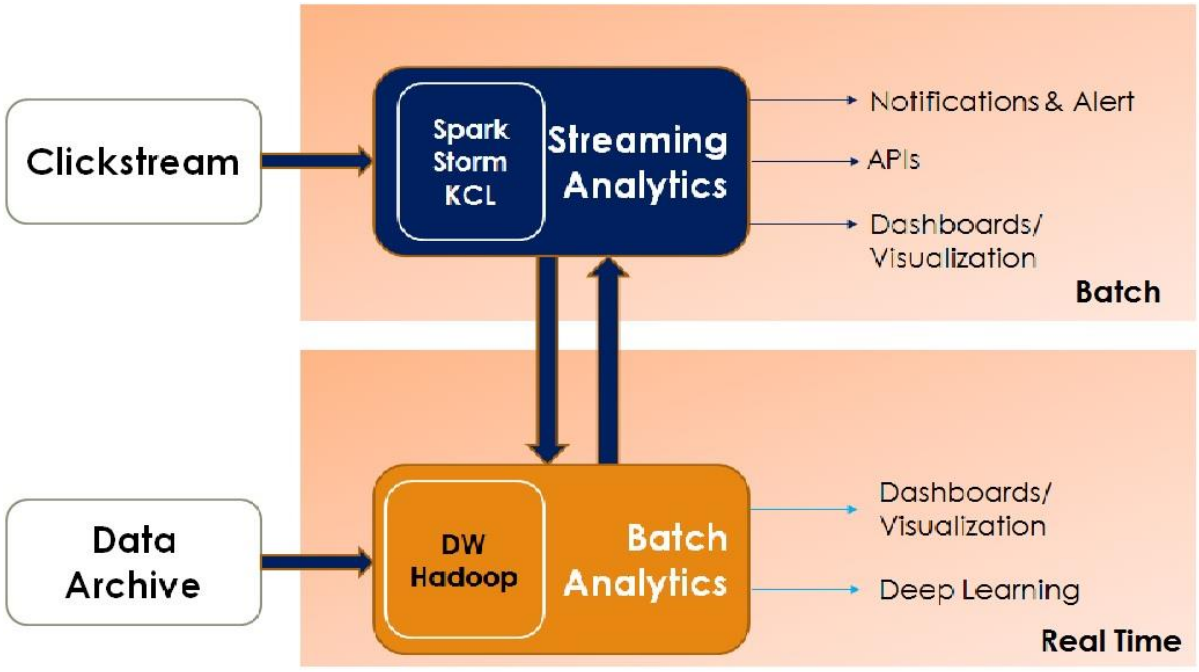

Fig 2: Clickstream Data Analytics

In this study, we focus on analyzing the clickstream data that comes from Internet Information Server (IIS) logs for msnbc.com and news-related portions of msn.com for the entire day of September 28, 1999 PST. Using this data, we analyze the detailed path/interaction sequences that users take on msnbc.com website to mine insights from the digital analytics data that helps to create true competitive advantage. We also do some Social Network Analysis on this data to better understand the interaction sequences on the website.

RESEARCH QUESTIONS

The dataset we have taken into consideration is the msnbc.com's clickstream data for one day. It contains sequence of page categories viewed by many users in 24 hours. Based on this data, we focus to answer the below research questions.

1. Which is the most popular page category?
2. Which is the most common page category the user visits first?
3. Which is the most common page category that the users end their visit on this website?
4. What is the maximum & minimum number of page views for each page category?
5. Which page category should be least considered for advertising based on minimum page views?
6. Which page category has the maximum bounce rate and exit rate?
7. What are the most common page category triads?
8. Can we build a recommendation system that recommends the next page category to view based on the previous two-page category visits?

The above questions would enable msnbc.com to manage and optimize their content on the webpages to enhance their business and boost marketing strategies.

We also perform Social Network Analysis on the page categories to calculate the centrality measures and page rank associated with each page category. By analyzing the website as a graph structure, we draw insights about the website and figure out how to optimize the website to serve users better.

DATA DESCRIPTION

The data comes from Internet Information Server (IIS) logs for msnbc.com and news-related portions of msn.com for the entire day of September 28, 1999 PST. Each sequence in the dataset corresponds to page views of a user during that 24-hour period. Each event in the sequence corresponds to a user's request for a page. Requests are recorded at the level of page category. The categories are "frontpage", "news", "tech", "local", "opinion", "on-air", "misc", "weather", "health", "living", "business", "sports", "summary", "bbs" (bulletin board service), "travel", "msn-news", and "msn-sports".

Each category is associated with an integer starting with "1". For example, "frontpage" is associated with 1, "news" with 2, "tech" with 3 and so on. For the sake of convention, we will address the page category as a page henceforth. Some relevant information about the dataset is summarized below:

- Number of users: 989818
- Average number of visit per user: 5.7
- Dataset Characteristics: Sequential
- Attribute Characteristics: Categorical

```
% Different categories found in input file:

frontpage news tech local opinion on-air misc weather msn-news health living business msn-sports sports summary bbs travel

% Sequences:

1 1
2
3 2 2 4 2 2 2 3 3
5
1
6
1 1
6
6 7 7 7 6 6 8 8 8 8
6 9 4 4 4 10 3 10 5 10 4 4 4
1 1 1 11 1 1 1
12 12
1 1
```

Fig 3: Dataset

The above figure represents the clickstream data. Each line in the sequence indicates the page category viewed by a user before exiting from the website. For example, a user viewed tech category page (3) and then moved to news category page (2) and so on. Another user viewed the frontpage (1) and clicked on another article in the frontpage (1) before exiting from the website.

INSIGHTS & DISCUSSION

This section consists of data analysis performed on the clickstream data. The summary statistics of the dataset is shown in the below table.

count	989818
mean	4.747129
std	21.25739
min	1
25%	1
50%	2
75%	5
max	14795

Tab 1: Summary Statistics for the dataset

LANDING PAGE

Landing page is a web page that appears in response to clicking on a search engine optimized search result or an online advertisement. The below plot shows the frequency of landing pages.

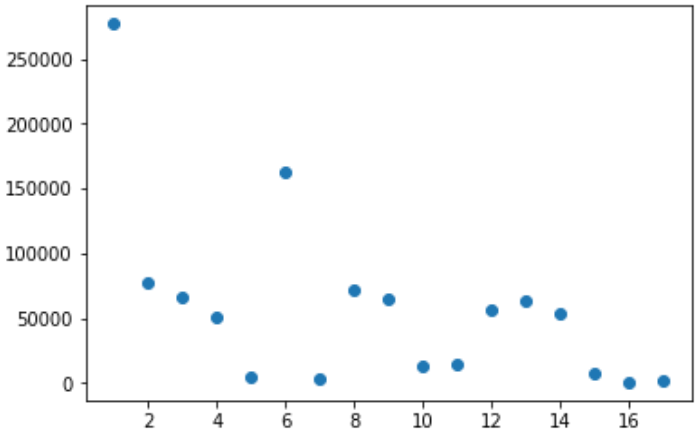

Fig 5: Landing Page frequency

From the above figure, we can see that "Frontpage" (1) is the most common landing page and "msn-news" (16) is the least common landing page.

EXIT PAGE

The Exit Page is the last page accessed during a visit. The below plot shows the frequency of exit pages.

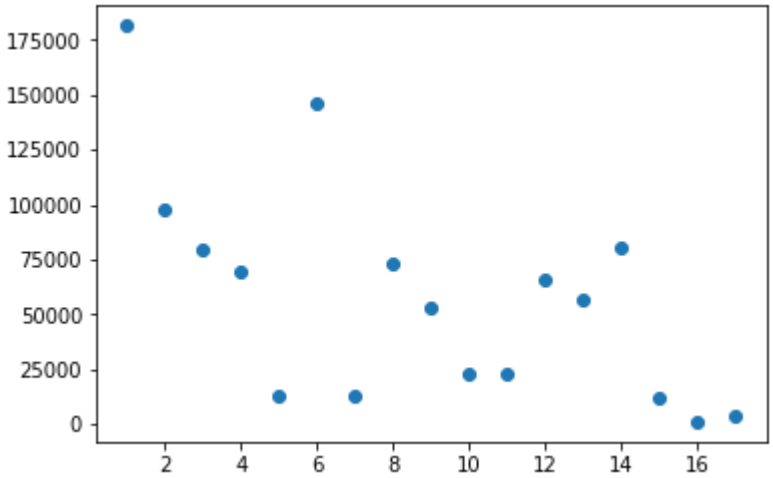

Fig 6: Exit Page frequency

From the above figure, we can see that "Frontpage" (1) is the most common exit page and "msn-news" (16) is the least common exit page.

MINIMUM & MAXIMUM PAGEVIEWS

The below plot shows the number of pageviews for pages from different categories.

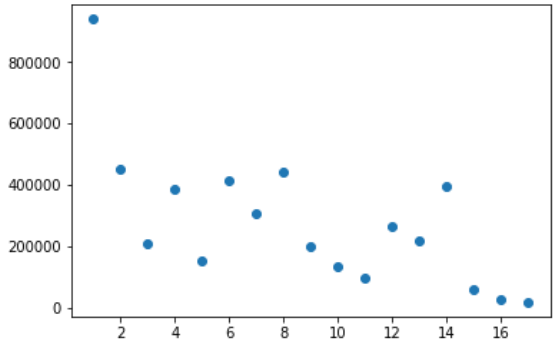

Fig 7: Minimum & Maximum pageviews

From the above figure and table shown below, we can see that "Frontpage" (1) has maximum pageviews and "msn-sports" (17) is minimum pageviews.

Pages	Exit Rate	Bounce Rate	Page Counts
1	0.579604127	0.20748731	940469
10	0.456882583	0.453948397	131760
11	0.390384916	0.494220665	96817
12	0.586345525	0.449596419	264899
13	0.730831211	0.290651891	216125
14	0.677340563	0.377803464	395880
15	0.402054795	0.360444625	56576
16	0.500480307	0.136150235	25249
17	0.309013265	0.465871438	16972
2	0.557340575	0.465633021	452387
3	0.654844688	0.632328635	207479
4	0.571324115	0.386050157	386217
5	0.505302757	0.22243256	151409
6	0.671235969	0.562954923	414928
7	0.156270959	0.043576047	305615
8	0.764691732	0.235855098	439398
9	0.583810094	0.305953001	196614

Tab 2: Web Analytics Metrics

BOUNCE RATE

It represents the percentage of visitors who enter the page and then leave ("bounce") rather than continuing to view other pages within the same site. Tab 2 shows the bounce rate of different pages. From the above table, we see that the bounce rate is high for "tech" (3) page.

EXIT RATE

It is the percentage of visitors to a site who actively click away to a different site from a specific page, after possibly having visited any other pages on the site. The below figure shows the exit rate of different pages. From table 2, we see that the exit rate is highest for "weather" (8) page.

SOCIAL NETWORK ANALYSIS

We perform Social Network Analysis on the dataset by treating each page category as a node and identifying the relationship that exists between the various pages. Some of the metrics we generated are summarized and discussed below:

Pages	Degree Centrality	Degree	Betweenness Centrality	Eigen Vector Centrality	Closeness Centrality	Page Rank
bbs	0.125	2	0	2.12E-11	0	0.0625
business	0.1875	3	0	0.195002308	0.3125	0.039283731
frontpage	0.6875	11	0.345833333	0.652732065	0.568181818	0.196422265
health	0.1875	3	0	0.195002308	0.3125	0.039283731
living	0.1875	3	0	0.195002308	0.3125	0.039283731
local	0.25	4	0	0.308287043	0.328947368	0.058923811
misc	0.3125	5	0.004166667	0.379201213	0.347222222	0.078564727
msn-news	0.125	2	0	2.12E-11	0	0.0625
msn-sports	0.1875	3	0	0.063966294	0.231481481	0.039302321
news	0.1875	3	0	0.195002308	0.3125	0.039283731
on-air	0.25	4	0	0.308287043	0.328947368	0.058923811
opinion	0.125	2	0	2.12E-11	0	0.0625
sports	0.25	4	0.075	0.214112826	0.347222222	0.058944407
summary	0.125	2	0	2.12E-11	0	0.0625
tech	0.1875	3	0	0.195002308	0.3125	0.039283731
travel	0	0	0	0	0	1.46E-42
weather	0.125	2	0	2.12E-11	0	0.0625

Tab 3: Summary of Centrality Measures

DEGREE

Degree of a node is the number of connections it has with other nodes. This measure indicates how well connected a node is with the rest of the nodes in the network. From Tab 3, we can see that "Frontpage" has the highest degree and "travel" has the lowest degree.

DEGREE CENTRALITY

It is defined as the number of links incident upon a node (i.e., the number of ties that a node has). Higher value for degree centrality indicates that a node is popular in the network. From Tab 3, we can see that "Frontpage" has the highest degree centrality and "travel" has the lowest degree centrality.

BETWEENNESS CENTRALITY

Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. Nodes that have higher Betweenness centrality tend to be powerful players in the network for transmission of information. From Tab 3, "frontpage" has the highest betweenness centrality and most of the other nodes (with value as 0 for betweenness centrality) are not considered powerful in the network.

CLOSENESS CENTRALITY

Closeness centrality (or closeness) of a node is the average length of the shortest path between the node and all other nodes in the graph. This measure indicates how close a node is with respect to all other nodes in the network. From Tab 3, "frontpage" has the highest closeness centrality and many other nodes are having low closeness centrality.

EIGENVECTOR CENTRALITY

It is a measure of the influence of a node in a network. It assigns relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. From Tab 3, "frontpage" has the highest eigenvector centrality and "travel" has the lowest eigenvector centrality.

PAGE RANK

PageRank is an algorithm used by Google Search to rank websites in their search engine results. PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites. In Tab 3, we see that "frontpage" has the highest page rank for msnbc.com while "travel" has the lowest page rank.

NETWORK ANALYSIS

The below graphs show the distribution of the dataset.

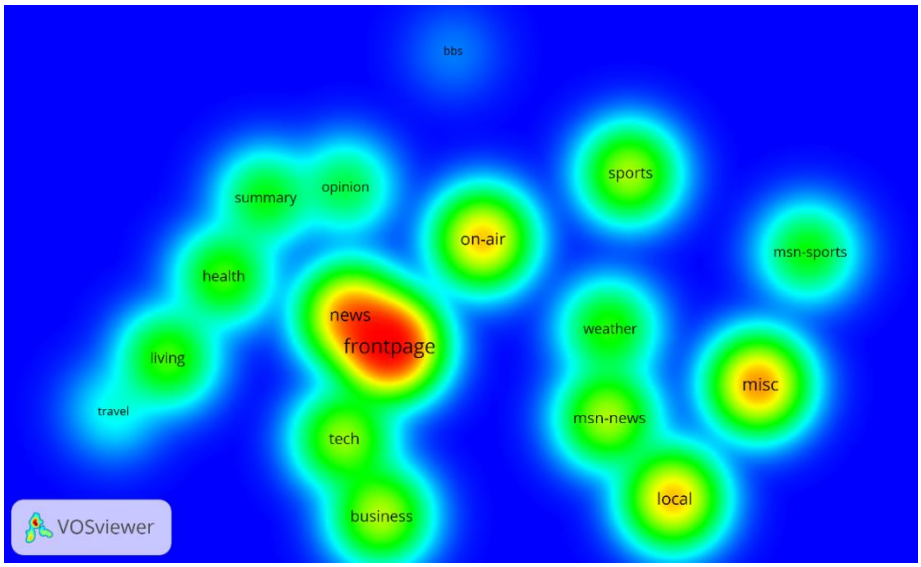

Fig 8: Density graph of the Dataset

From the above graph, we find certain pages are extremely cohesive. This is because of the high frequency of page visit from one category to another. For example, travel, summary, health, living and travel are highly cohesive because users prefer to move from one page to another in the above categories. Similarly, sports, misc, msn-sports, and bbs are not cohesive with any other page categories.

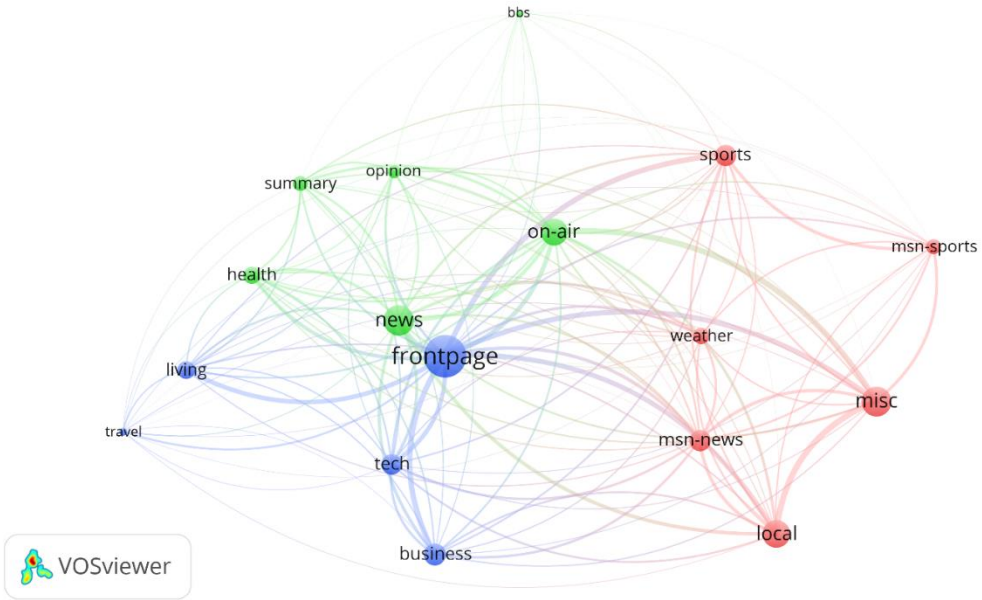

Fig 9: Network Link for the dataset

From the above figure, we notice that there are three distinct clusters. The lines connecting different nodes indicate the sequence of visit from one category to another.

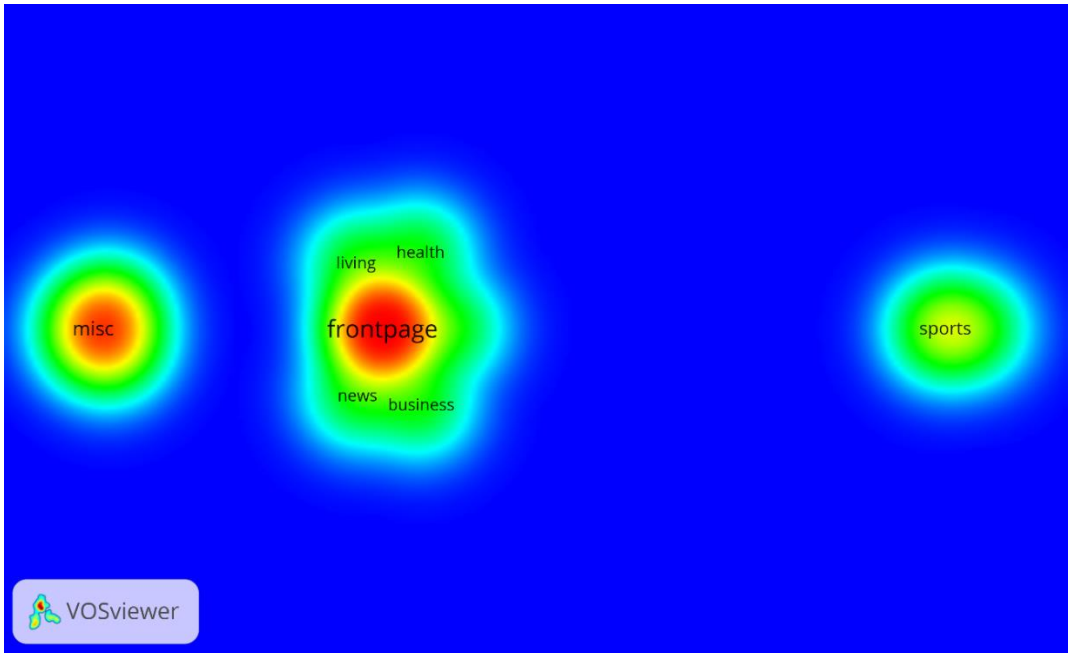

Fig 10: Density graph showing connected nodes after applying threshold

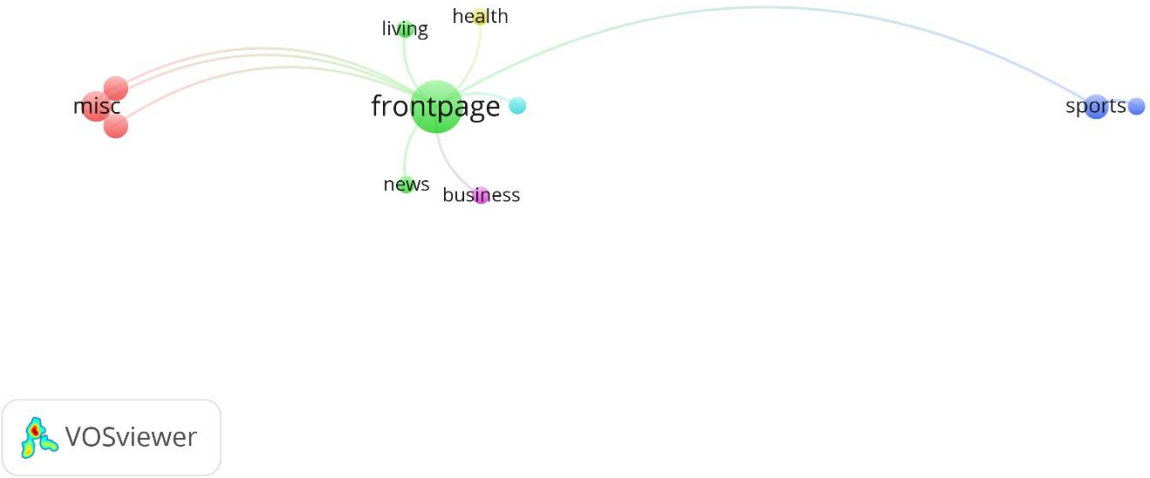

Fig 11: Network link showing connected nodes after applying threshold

In the above graphs, we apply a threshold of mean of the co-sequence (i.e. whenever two pages appear one after the other). This is done to identify the most important nodes in the network. The above graphs show that frontpage, living, health, news and business are commonly viewed categories while misc and sports are two different entities not cohesive with any other categories. The network link graph shows the commonly navigated page categories.

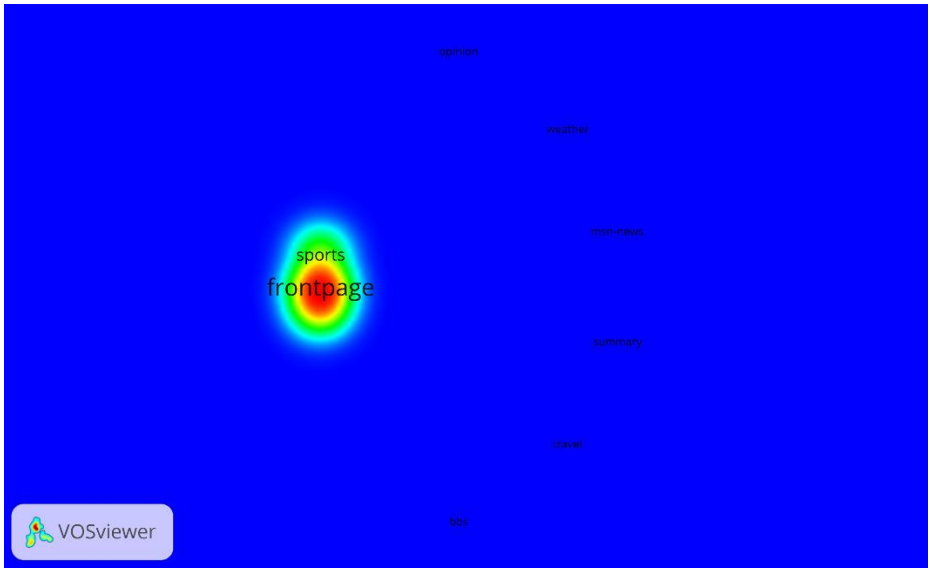

Fig 12: Density graph showing position of all nodes after applying threshold

Fig 13: Network link showing position of all nodes after applying threshold

From figure 12 and 13, we see a distinct core-periphery structure that central notes have sequence of visits. While the other nodes that are peripheral do not have many connections from the core nodes.

RECOMMENDER SYSTEM

After analyzing the centrality measures, we try to recommend page categories to visit based on previous pages visited. We do this to understand human behavior with respect to preference based on similar users.

From Tab 4, we see that most users (1328) prefer to navigate from "news" (2) to "opinion" (5) and choose to read more content from "opinion" (5) page. Thus, we can recommend page 5 to a new user who visits page 2 and then page 5. Similarly, many users (234 users) prefer to navigate from "on-air" (6) to "opinion" (5) and then navigate to "weather" (8). Thus, for a new user, we can recommend the sequence of pages to visit as 6—5—8.

Recommendation for Seq of 2, 5		Recommendation for Seq of 6, 5	
('2', '5', '5')	1328	**('6', '5', '5')**	1172
('2', '5', '2')	469	**('6', '5', '6')**	323
('2', '5', '6')	172	**('6', '5', '8')**	234
('2', '5', '1')	160	**('6', '5', '2')**	164
('2', '5', '3')	62	**('6', '5', '1')**	113
('2', '5', '15')	62	**('6', '5', '11')**	67
('2', '5', '12')	53	**('6', '5', '15')**	65
('2', '5', '8')	49	**('6', '5', '10')**	61
('2', '5', '11')	48	**('6', '5', '16')**	42
('2', '5', '10')	47	**('6', '5', '7')**	40
('2', '5', '14')	35	**('6', '5', '3')**	35
('2', '5', '4')	33	**('6', '5', '12')**	32
('2', '5', '7')	20	**('6', '5', '14')**	23
('2', '5', '9')	17	**('6', '5', '4')**	21
('2', '5', '17')	7	**('6', '5', '9')**	13
('2', '5', '16')	6	**('6', '5', '17')**	6
('2', '5', '13')	4	**('6', '5', '13')**	2

Tab 4: Recommendation for sequence of pages

LESSONS LEARNT

This study paves way for understanding clickstream data and how to perform Web analytics on the data. We get a thorough understanding of concepts like landing page, entry page, exit page, bounce etc. and standard metrics like bounce rate, exit rate and pageviews. Based on the sequence of page visits, we recommend the next page to visit based on user preference.

By analyzing the data as a network, we have calculated centrality measures such as degree, degree centrality, betweenness centrality, eigenvector centrality and closeness centrality. It helps us gain knowledge about interpretation of centrality measures in our dataset. We also understand how page rank works and incorporated it for our dataset to help us identify the most important page category.

We also used VOSviewer to identify the network structure – if it is clumpy or has a core-periphery structure. By plotting density graphs, we analyzed the most cohesive groups of page categories. The network link figures helped us understand the different clusters that exists in our dataset.

CHALLENGES FACED

This study is very interesting to understand how clickstream data can be used to draw meaningful insights. However, there are a few challenges that could not be addressed due to lack of sufficient information.

The dataset consists of only categorical data i.e. the sequence of page visits. It lacks more information about clickstream which if present would have enabled for more advanced analysis. If the dataset had information about the users and their activity on the website, it would've helped perform better web analytics. Hence, our analysis turned out to be quite limited in scope.

The data sample is restricted to just one day. It also lacks information about the exact time of the day when the data was collected. This inhibits our approach to provide better recommendation based on time.

While analyzing the dataset as a network, to find the structure of the network, we must pick a threshold value. This value was picked based on mean instead of the minimum, maximum, standard deviation, or other parameters. The choice of picking mean was only based on the assumption that the data would be properly distributed. The validity of such approach remains unknown.

CONCLUSION

By analyzing msnbc.com's clickstream data, we have identified that the "frontpage" is the most popular page. We can organize marketing content in the first five pages with maximum page views to generate more revenue. We also see that the pages "tech" and "on-air" have high bounce rate. To fix this, we would suggest the business owner to make the page content more engaging for users. We also identified pages like "sports" and "misc" that are not central compared to other nodes in the network. To fix this, we can recommend these pages to visitors when they are viewing more popular pages so that we can improve the pageview of the entire website.

REFERENCES

[1] https://en.wikipedia.org/wiki/MSNBC

[2] https://en.wikipedia.org/wiki/History_of_MSNBC:_1996%E2%80%932007

[3] http://searchcrm.techtarget.com/definition/clickstream-analysis

[4] https://www.quora.com/What-are-the-applications-of-clickstream-analysis

[5] http://www.blastam.com/blog/move-into-limitless-world-of-clickstream-data-analysis

[6] http://archive.ics.uci.edu/ml/datasets/msnbc.com+anonymous+web+data

[7] https://en.wikipedia.org/wiki/Landing_page

[8] https://piwik.org/faq/general/faq_87/

[9] https://en.wikipedia.org/wiki/Centrality#Degree_centrality

[10] https://en.wikipedia.org/wiki/PageRank

[11] http://www.vosviewer.com/getting-started