Recommendations to Improve Content Viewership Yield for Skyscanner

(Mid Term Report)

Team SkyTrek:

Aseem PRABHAT

Jedaiah TAN Jia Le

NGUYEN Viet Huy

ANLY482

Project Update

Following the proposal review with the Professor Kam, the team shared their concerns in streamlining the project scope with the client. The following table demonstrates the final assessment of the revised project objectives.

<table>
<tr><th>Past Objective</th><th>Mid Term Assessment</th></tr>
<tr><td>Identify the different web content factors that affect content performance in order to differentiate between high and low performing content (regression)
What are the common attributes that lead to some content pieces drawing in most of the traffic?
Is there an ideal standard format for news articles that best caters to the needs of the users?</td><td>Objectives could be streamlined to keep the focus on identifying high performing article attributes.</td></tr>
<tr><td>What are the most popular content themes that resonate with users in each of the given markets namely Singapore, Malaysia and Thailand?</td><td>Client’s primary objective is to explore insights pertaining to the Singapore news site</td></tr>
<tr><td>Should the strategic focus be on generating more articles to drive traffic or focus on fewer quality articles?</td><td>Client highlighted on going constraints in releasing more articles and preferred exploring the effectiveness of paid and unpaid articles.
Client also expressed greater interest in increasing organic growth (unpaid searches)</td></tr>
<tr><td>Facilitate the content planning process by way of an interactive dashboard</td><td rowspan="2">[To be dropped]
Team presented advising professor’s concerns on project scope size and obtained client's approval to streamline project focus on the above mentioned objectives</td></tr>
<tr><td>What is the role of seasonality and annual trends in the online readership pattern of Skyscanner users?</td></tr>
<tr><td>How can the content planning process be streamlined in order to create maximum impact with minimal resources?</td><td>[To be dropped]
Objective involves a time consuming study of the team’s current content planning process which might not be feasible given the other high priority objective from the client</td></tr>
</table>

The client further confirmed that this project's primary objective is to identify how to increase returns of investments given her current efforts (limited resources). Further studying the information available and details of this project, the team finalized on the following project objectives:

1. To increase organic growth by identifying key article attributes that draw high levels of traffic and interest
2. To identify cost-effective paid sources
3. To identify and validate the most effective content themes for the Singapore market. Following which, the analysis process can be replicated for the Malaysia and Thailand market

Over the course of our meetings with the client at Skyscanner, she pointed out that there are two main metrics in our data that are actionable- Unique Pageviews (UPVs) and Average Time on Page (ATOP). These two metrics are now the focus of our analysis in the sense that they would be used to define 'performance'. Thus our analytical problem has shifted to understanding the factors that trend to drive these two main metrics. The implication of this is that UPVs and ATOP will now be the target or dependant variable and the goal of our further analysis would be to understand the relationship between these target variables and the other attributes in our dataset. To build on this change, we first need to define our data cube to better explain the relevance of each attribute.

Data

Extraction from Data Sources

Our final data cube was derived from 3 main raw data sources, each requiring a different method to pull and transform the raw data.

1. **Skyscanner News Site (Crawling):** This data is pulled from the Skyscanner news site, for each article, including attributes such as article text, number of links, images, published date etc. This data constitutes public information that is visible to the user and hence did not need to be requested for from the client. These attributes tell us about the nature of the actual content that is being seen by the user.

2. **Google Analytics:** This data was pulled for each article being tracked via the Skyscanner Google analytics account. This contained metrics regarding the performance of each URL on the Skyscanner news site. These attributes tell us about the performance metrics of each article mainly through Unique Page Views and Average Time on Page. It also provides the different sources of traffic such as Facebook paid media or Google organic and the contribution of each of the different online channels.

3. **Social Media Shares**: This data was pulled via continuous scraping through a java program of a website called http://linktally.com that provides an API of social media

shares for a given URL. These attributes tell us about how widely each article is shared across different social media platforms.

Transformation and Loading - Merging of 3 Raw Data Sources

Once these three raw datasets had been extracted, there was need to merge them so that the relationship between all of these attributes can be analysed in the next stage of our project. This required many of the 'dirty' URLs to be cleaned out as part of the merging process. The article URL was used as a primary key in order to join all the attributes from the three datasets. The diagram below shows the entire ETL process showing how the three sources of data were extracted, transformed and then stored in a MySQL database.

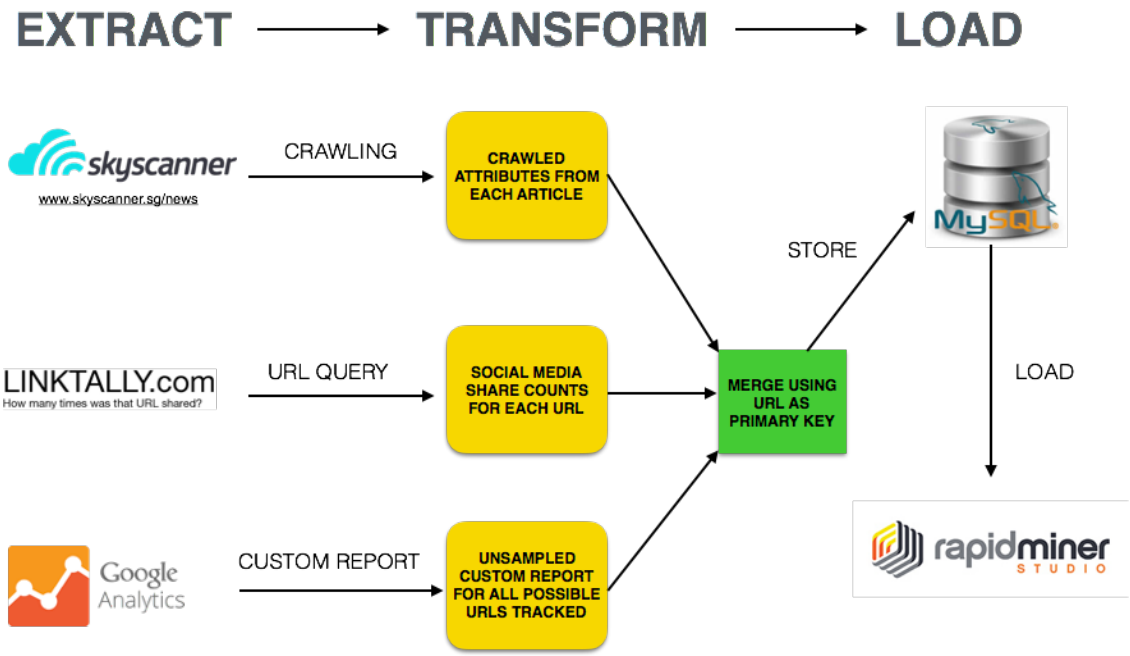

Figure 1 Data Transformation and Merging Process

Data Dictionary

Attribute Name	Description	Type
URL	The original URL of the article on the Skyscanner news site	Categorical
Source/ Medium	The source of the traffic to the website	Categorical
Unique Page Views	The total number of unique views for the given article	Numerical
Average Time on Page	The average time in seconds that a user spends on the given article	Numerical

No. of links	The number of out-links embedded in a given article	Numerical
No. of images	The number of images in a given article	Numerical
No. of shares	The number of social media shares for a given article	Numerical
Published Date	The date at which the article was published	Categorical
Article Text	The body of text of the entire news article	Categorical

Aggregation of data

The final dataset contains about 9621 rows at the most disaggregated level. The identifier for each row is the article URL as well as the 'source/medium' for each URL. In order to analyse different aspects of the business problems, this dataset has been divided into different aggregated levels. The most important levels would be based on the traffic source- mainly 'organic' (non-paid) VS 'inorganic' or 'paid' media. The main reason behind this division is that 'paid' traffic numbers tend to usually be higher than non-paid ones and hence skew the data in favour of articles that have been distributed through paid channels such as Facebook, Taboola and StumbleUpon. The diagram below shows the different levels of aggregation building up from the most disaggregated (10k rows) to the most aggregated level.

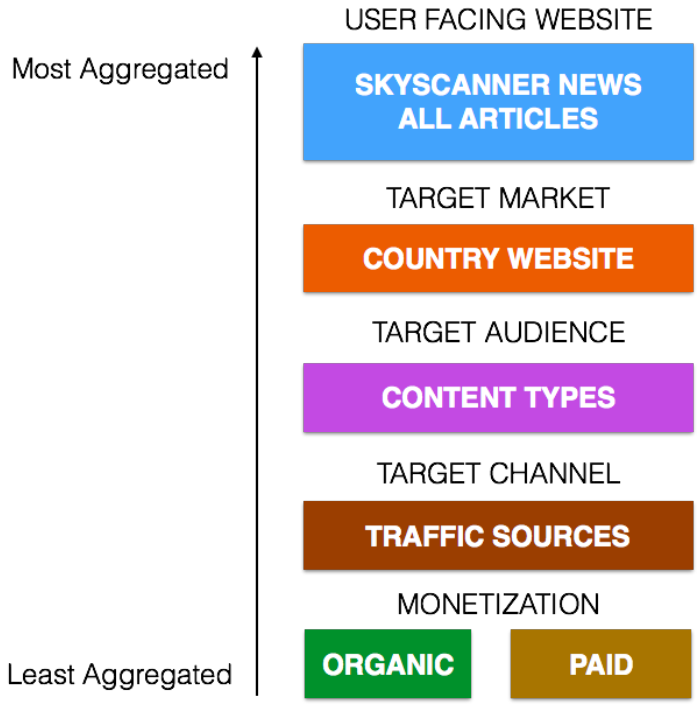

Figure 2 Levels of aggregation of Data

Methodology

Segmenting the Data

In the course of our meetings with the client, we learned that there are two broad categories for sources of traffic- Organic and Paid. Articles that appear on paid channels tend to have much higher traffic numbers than ones that don't. This leads to the data getting skewed in favour of paid sources. Hence there is a need to divide the data based on traffic source and analyse the two types separately. There are three main divisions or versions of the data that we will explore differently as each one answers a different question:

1. **Complete News Article Set**
   This consists of both Organic and Paid sources. This can be used for analysis where the skew in traffic numbers does not impact the question being answered by the analysis. One example of this would be text mining of the body or title of the article.

2. **Organic Traffic Only**
   This consists of the all articles but only with organic numbers. This dataset can be used where performance all articles need to be compared directly without any bias. Since organic numbers are considered ‘natural’ traffic, they better reflect the performance of an article independent of the money spent on promoting it.

3. **Paid Traffic Only**
   This consists of the all articles but only with paid traffic numbers. This is useful when comparing different paid sources in order to gain a better understanding of the return on investment of different online marketing channels. This helps get a better understanding of how to efficiently use resources to maximize performance.

Complete News Article Set

Content Theme Analysis

We had previously highlighted the 7 Content Themes (CT) Skyscanner believes its articles belong to. The aim of this analysis is 3 fold. To validate if these 7 CT are representative of the article content being written. To identify the top 3 CT with the greatest yield. Lastly, to understand the performance across each CT. As mentioned earlier, we will be measuring yield and performance by the metrics UVP and ATOP.

It was not going to be possible to read each and every single article in order to identify the various CT, hence verifying our client list of CT. Hence, we would employ the use of the K-means clustering algorithm to identify the latent groups of CT within our dataset.

Preparing the Dataset

Our database contains the html for each of the 399 articles hosted on Skyscanner Singapore’s travel news site. RapidMiner was used to clean this data. HTML tags were removed from the html content, leaving only the article content. The content was then

tokenized, transformed to lowercase, filtered for stop words from the English dictionary, then filtered for tokens with character length between 3 and 41. Following which, a tf-idf matrix was generated for each every token in each article. Tf-idf was used because it accentuates the value of rare word in distinguishing an article from another, thereby augmenting our goal of discovering the latent CT.

Applying K-Means Clustering and Discovering more CT

This matrix was then used to do K-Means Clustering. Based on Can & Ozkarahan's paper, we derived the number of clusters to be 71. Each of the 71 clusters were evaluated based on its representative words before assigning them a descriptive label of what its CT might be. Once done, these descriptive labels were then binned into the relevant CT provided.

Validating the Representativeness of the 7 Identified CT

In the course of doing so, we found all 7 CT to be represented in the article. However, we felt it appropriate to generate 2 new CT, 'Activity/topic discussion' and 'Food' since they represented 17% and 8% of articles respectively. The following pie chart shows the proportion of each CT.

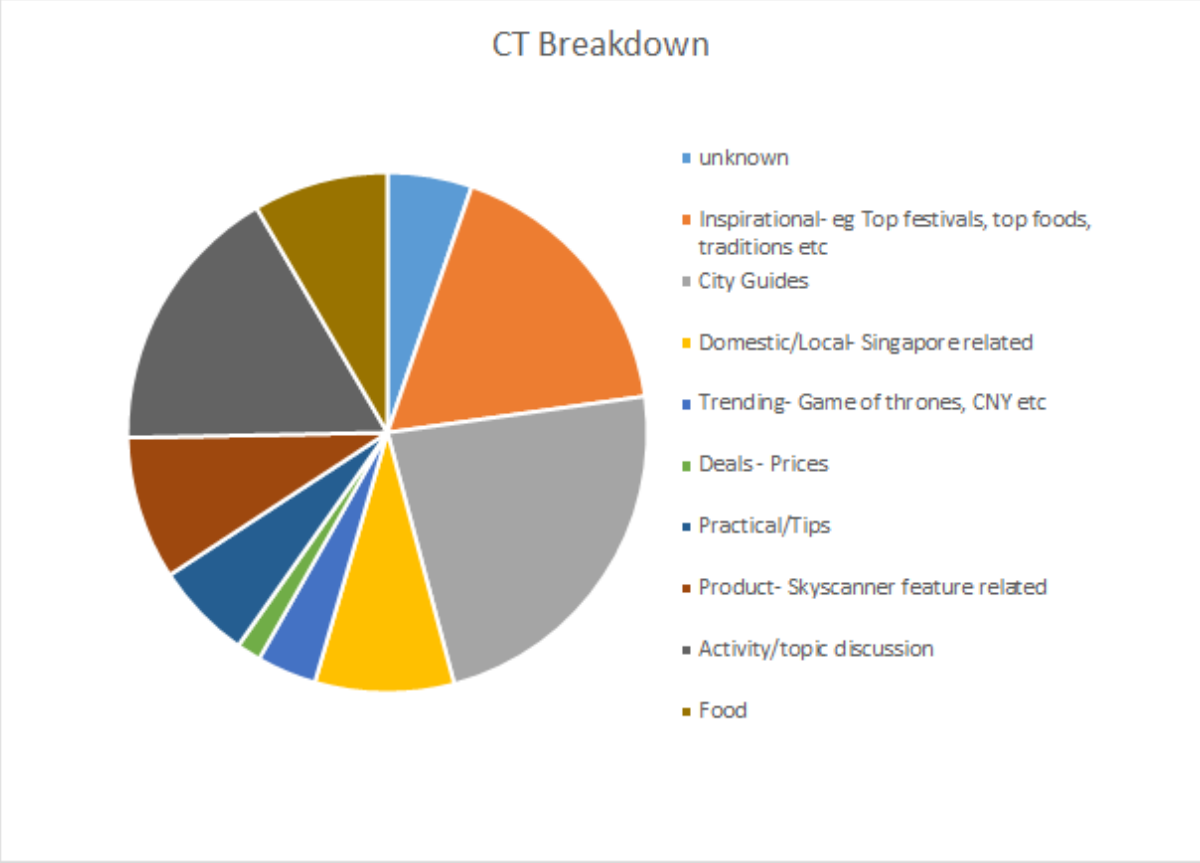

Figure 3 Proportion of Articles in each CT

City Guides, Inspirational and Activity/topic discussion are the top 3 most represented CT at the moment.

Identifying the Top Performing 3 CT

Under UPV, City Guides, Trending and Food are ranked highest in descending order.
Under ATOP, Inspirational, Food and Trending are ranked highest in descending order.
It is interesting to note that only Food made it to the top 3 place across both metrics. The client has expressed her opinion that strong performance in both these fields would be a good indicator of high quality traffic since it is both able to get high viewership rates as well as sustain viewer interest in reading the article. Thus, Skyscanner might choose to focus its efforts on writing food articles.

Understanding each CT Performance (Z score)

We used the z-score to evaluate the relative performance of each CT against one another, represented in Figure 4 below. Taking the mean and the baseline for comparison, it is interesting to note that a CT that fares well in one metric tends to do badly in the other. 'Food', 'Product', 'Trending', 'Domestic/Local', 'City Guides' and 'Inspirational' are such CT. This negative correlation between UPV and ATOP is also represented in the correlation analysis done in Figure 5 below.
It is interesting to note that the 'Practical/Tips' and 'Products' CT constantly fare relatively poorly under both the UPV and the ATOP metric. Skyscanner might consider diverting their resources away from these CT.

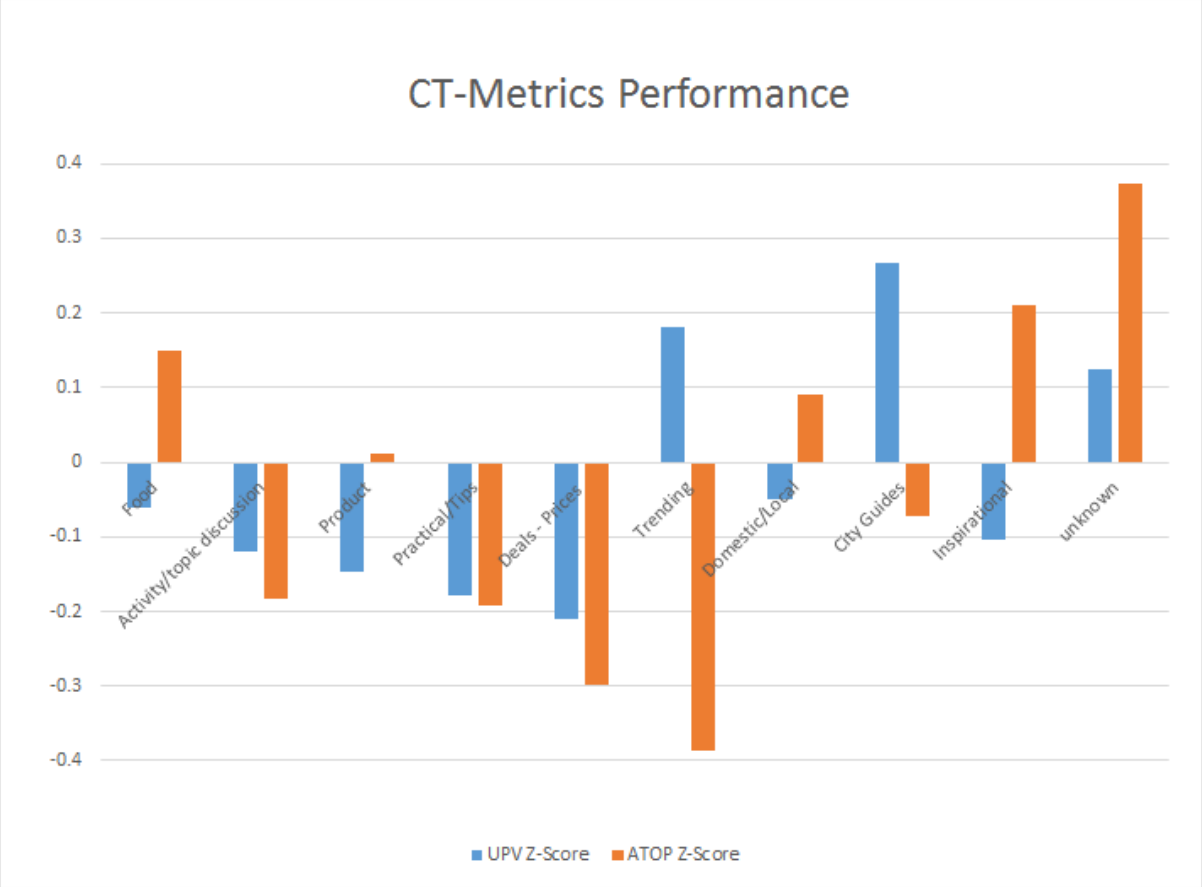

Figure 4 CT UPV and ATOP Z-Score Performance

Organic Articles Dataset

In order to understand the relationship between the target variables and other attributes of the dataset, we have considered running a regression analysis with UPVs and ATOM as the dependant variables. This method would be appropriate as the goal here is to understand the factors that can help predict the performance of a news article.Since most of the data is numerical in nature regression is an appropriate modelling method that will help determine the incremental impact of a unit increase in one variable on the target variables that define performance. Before conducting this, we ran a correlation analysis to check the relationship of the variables with each other, in order to prevent multicollinearity.

Correlation Analysis

	no_of_words	no_of_links	no_of_imgs	no_of_videos	facebook_shares	unique_pageviews	pageviews	sessions	organic_searchs	bounce_rate	exit_rate	avg_time_in_sec
no_of_words	~~1.00~~	0.54	0.51	0.06	0.06	0.13	0.13	0.13	0.13	-0.07	-0.13	-0.11
no_of_links	0.54	~~1.00~~	0.58	-0.03	0.04	0.15	0.15	0.14	0.17	-0.01	-0.08	-0.13
no_of_imgs	0.51	0.58	~~1.00~~	-0.08	0.08	0.20	0.20	0.19	0.22	-0.05	-0.11	-0.15
no_of_videos	0.06	-0.03	-0.08	~~1.00~~	-0.01	-0.01	-0.01	-0.01	-0.02	0.02	-0.02	0.02
facebook_shares	0.06	0.04	0.08	-0.01	~~1.00~~	0.16	0.16	0.17	0.14	-0.05	-0.12	-0.09
unique_pageviews	0.13	0.15	0.20	-0.01	0.16	~~1.00~~	~~1.00~~	1.00	0.83	-0.09	-0.68	-0.68
pageviews	0.13	0.15	0.20	-0.01	0.16	~~1.00~~	~~1.00~~	1.00	0.83	-0.09	-0.68	-0.68
sessions	0.13	0.14	0.19	-0.01	0.17	1.00	1.00	~~1.00~~	0.81	-0.11	-0.69	-0.69
organic_searchs	0.13	0.17	0.22	-0.02	0.14	0.83	0.83	0.81	~~1.00~~	0.20	-0.41	-0.42
bounce_rate	-0.07	-0.01	-0.05	0.02	-0.05	-0.09	-0.09	-0.11	0.20	~~1.00~~	0.26	0.26
exit_rate	-0.13	-0.08	-0.11	-0.02	-0.12	-0.68	-0.68	-0.69	-0.41	0.26	~~1.00~~	0.83
avg_time_in_sec	-0.11	-0.13	-0.15	0.02	-0.09	-0.68	-0.68	-0.69	-0.42	0.26	0.83	~~1.00~~

Figure 5 Article Attributes and Metrics Correlation Analysis

Given this correlation matrix, some of the variables that are highly correlated can be removed, in order to prevent multicollinearity. Some examples of these are unique page views, page views and sessions. Since these are highly correlated, we can drop sessions and page views from the dataset for the regression model and use only unique page views. One interesting insight here is that the two target variables ‘Unique page views’ and ‘Average time on page’ are negatively correlated with a correlation coefficient of -0.68. This shows an inverse relationship and hence indicates that there might be a need to give up focus on one of these metrics in order to drive the other one up.

Regression Model Analysis

Since we have two dependant variables UPVs and ATOP, we will create two regression models. All other attributes in the dataset will be the independent variables. The goal here is to better understand the relationship between these attributes and the content performance attributes - UPVs and ATOP

Unique Page Views as Target Variable

Attribute	Coefficient	Std. Error	Std. Coeffici...	Tolerance	t-Stat	p-Value
no_of_words	-0.017	0.008	-0.017	0.990	-2.085	0.038
no_of_links	0.005	0.008	0.005	1.000	0.628	0.531
no_of_imgs	0.006	0.008	0.006	0.992	0.700	0.485
organic_searchs	0.995	0.007	0.995	0.948	151.690	0
bounce_rate	0.008	0.010	0.008	0.999	0.769	0.442
exit_rate	-0.041	0.011	-0.041	0.988	-3.828	0.000
avg_time_in_sec	0.001	0.006	0.001	0.987	0.114	0.909
facebook_sha...	0.063	0.006	0.063	1.000	9.734	0
(Intercept)	0	0.006	?	?	0	1

Figure 6 Regression Model with UPV as Dependent Variable

The first regression model, using UPVs as the dependent variable, gives us three attributes with p-value greater than 0.05, indicating that the coefficients can be used to predict the relationship with the dependant variable. At this stage we can drop the attributes that give us a high p-value(0.05). The three attributes that seem to have a relationship with UPV are:

1. **Organic Searches**: An increase in organic searches leads to a positive increase in UPVs
2. **Exit Rate**: An increase in exit rate leads to a negative change in UPVs
3. **Facebook Shares**: An increase in Facebook shares leads to a positive increase in UPVs.

One issue with using these coefficients from these 3 attributes is the fact that they may not be in control of the client and hence cannot be considered ‘actionable’. Given that many of the other attributes do not have a predictive linear relationship with UPVs, it is possible that regression may not be the best approach to understand the performance of articles.

Average Time on page as target

Attribute	Coefficient	Std. Error	Std. Coeffici...	Tolerance	t-Stat	p-Value
no_of_words	-0.050	0.065	-0.050	1.000	-0.772	0.441
no_of_links	0.035	0.067	0.035	0.984	0.533	0.594
no_of_imgs	0.040	0.065	0.040	0.995	0.611	0.542
unique_pagev...	0.045	0.397	0.045	0.537	0.114	0.909
organic_searchs	0.053	0.399	0.053	0.528	0.134	0.893
bounce_rate	-0.051	0.081	-0.051	0.724	-0.633	0.527
exit_rate	0.151	0.085	0.151	0.838	1.781	0.076
facebook_sha...	0.043	0.056	0.043	1.000	0.765	0.445
(Intercept)	0	0.050	?	?	0	1

The regression results show a high p-value (>0.05) indicating that the attributes in the model are not good predictors of average time on page and hence should be dropped. This again indicates that regression may not be the best approach to measuring the performance of these variables.

Conclusion

Our regression analysis has shown that from a business point of view that there is no particular attribute that tends to have a strong linear relationship with content performance. This indicates that there is no one ‘winning formula’ when it comes to designing an article, especially in terms of ‘changeable’ attributes such a number of images, links, total words and videos. While organic searches, Facebook shares and exit rate did show a significant relationship with Unique Page views, these metric are not very ‘actionable’ meaning they will be difficult for the client to change in the short term, making the regression results not very impactful from a business point of view. This indicates that there might be move value in the qualitative nature i.e. Content themes and text analysis results in determining what makes a ‘high performing’ news article.
One interesting insight from the correlation matrix was that the two target variables ‘Unique page views’ and ‘Average time on page’ are negatively correlated with each other with a correlation coefficient of -0.68. This shows an inverse relationship and hence indicates that there might be a need to give up focus on one of these metrics in order to drive the other one up. This is an important question to consider from a business point of view.

Paid Source Dataset

Data visualization and exploration

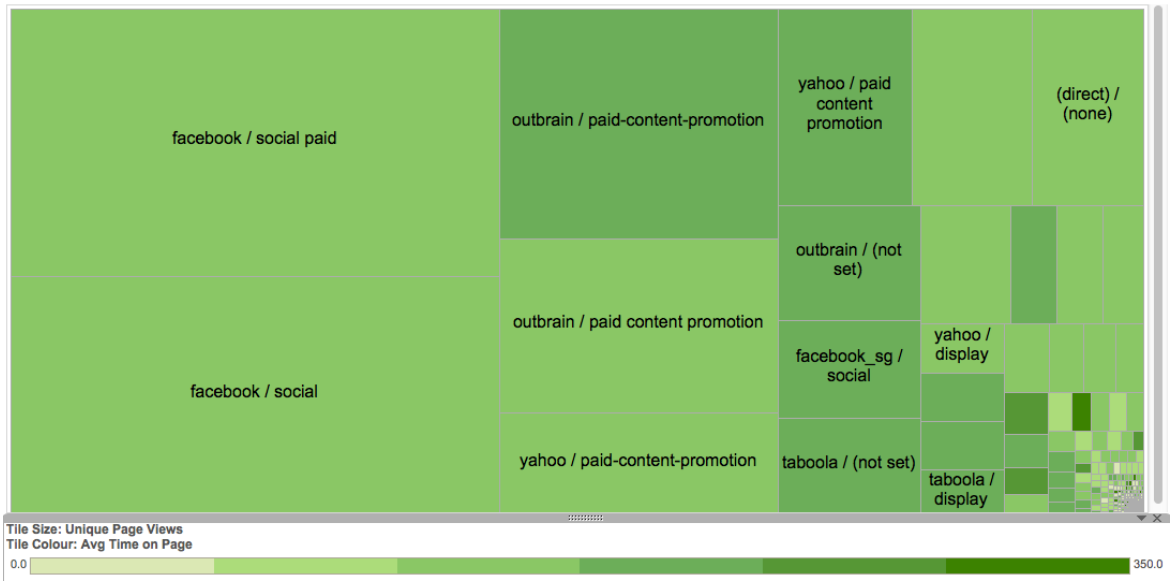

Figure 7 Tree Map of Sources against UPV against ATOP

In order to look at the different paid sources and compare them, we conducted exploratory analysis using tableau charts. One such chart can be found above where each tile represents a traffic source. The tile size indicates the Unique Page views for the source while the colour represents the Average Time on page. This can help the client decide which paid source to use depending on whether she wants to focus on high traffic through page views or readership engagement through greater average time on page.

Analysis Plans Moving Forward

At this stage, we have completed a preliminary level of analysis and would now proceed to our next iteration where we would like to expand our methodology to include both a deeper and broader approach to our questions.

1. **Text Analysis for titles:** The text analysis method used to group the articles based on content theme will be replicated on the titles for each article. Since the titles have a different role and impact compared to the news article body text, we expect these groups to be different from the content themes and hence analysing them will provide additional value in the content planning process.
2. **Create derived attributes:** As a result of our clustering and grouping process, there will be two additional derived attributes in our dataset namely content theme and title category. These will later be included in our analysis to better understand their relationship with content performance.
3. **Conversion of numerical to categorical**: Some of our attributes such as bounce rate, exit rate and UPVs will be converted to categorical attributes based on threshold values provided by the client. These new categorical attributes will be then used to create a new predictive model (possibly logistic regression) in order to understand the improvement in performance between switching from one category to another.
4. **Replicating analysis to Malaysia and Thailand market:** Once our analysis is complete, we will proceed to extend this same analysis to two new similar dataset namely for the Malaysia and Thailand markets. There is a need to analyse each country market independently due to the wide difference in market dynamics and customer behaviour.

Software Tools Assessment

We considered the pros and cons of 3 data analysis tools, namely RapidMiner, SAS Enterprise Miner (EM) and R. While EM offers greater visualization capabilities it is not necessary for our analysis. Considering that we only had 399 articles to deal with, each K-means clustering run (most expensive analysis) averages at a mere 7 minutes. Hence, it would not be reasonable to expect the client to invest in a license.
On top of the benefits of RapidMiner being a free open-source software, it has a wide selection of operators available for immediate and relevant use in our project. More specifically, they can be used for the ETL process in preparation for K-means clustering. RapidMiner is identified to be capable of accomplishing our project objectives at a lower learning curve and at no monetary cost, hence selected as our tool of choice.

Software	Advantages	Disadvantages
RapidMiner	● Fully sufficient ecosystem for complete ETL process. ● The drawing board is editable during the runtime. Hence you can prepare a new experiment while RapidMiner is still performing calculations on the last experiment. ● Open source software with many plugin options for extension of feature base.	● Too easy to setup the flows/operators that they would take eternity to calculate hence leading to loops. ● Have to rerun whole flow even for a small refinement. ● You can't stop running a node. You can only tell RapidMiner to prevent execution of subsequent nodes. Hence forced application restarts are common.
SAS Enterprise Miner	● Nice interactive visualizations. Whenever you click on a label, the corresponding data in the chart gets highlighted. ● Selective execution in EM is a great feature, which accelerates development, because if you make a mistake at the end of the flow, you don't have to recalculate everything from the scratch	● Uninformative error messages. You often have to blindly test many different things until you find the root of the problem. ● Commercial software hence, high cost of setup and service
R	● Rich functionalities, packages and development tools especially for statistical analysis. ● Flexible and easy to extend with unmatched charting capabilities ● Open source software	● Issues related to security, speed, efficiency and memory management since it emanates from languages built in the 1960s ● Steep learning curve and disorganized documentation

Scope of Work

As part of the Travel-News Traffic Optimization Project the team will be responsible for performing tasks throughout various stages of this project. The following is a list of these tasks, which will result in the successful completion of this project:

Project Sourcing

- Meeting with prospective clients to understand project requirements
- Team assessment of project feasibility
- Project confirmation with client
- Obtain dataset from client

Problem Definition

- Understanding business problem
- Understanding analytical problems
- Understanding analytical methodologies
- Document all meeting minutes

Data Preparation

- Data cleaning
- External data source crawling
- Data merging
- Data loading
- Research on data analysis tools
- Research data analysis methodologies

Data Exploration

- EDA analysis

Design phase

- Finalize business problem
- Finalize analytical problems
- Finalize analytical tools
- Finalize analytical methodology

Implementation Phase

- Analysis
    - Complete news article set analysis
        - K means clustering
        - Cluster labelling
        - Cluster label – Client content theme (CT) analysis
            - Validation of present set of CT
            - Mapping of clusters to CT
            - CT metric analysis
                - Identification of top 3 CT
                - Z score comparison across CT
    - Organic growth analysis

        - Attribute Metric Correlation modelling
        - Regression modelling
    - Paid source analysis
        - Data visualization and exploration
            - Identifying most valuable (UVP and ATOP) paid source
        - Odds ratio source switching sensitivity analysis
    - Titles Text Analysis
        - K means clustering
        - Cluster labelling
        - Cluster metric analysis
            - Identification of top 3 clusters
            - Z score comparison across clusters
- Proposal report
- Mid-term report
- Final report

Control & Monitoring

- Fortnight basis
    - Review with client
    - Review with Prof Kam
- Weekly basis
    - Team meeting update
    - Wiki update

Work Plan

Task	W1	W2 Milestone 1	W3	W4	W5	W6	W7	W8	W9	W10 Milestone 2	W11	W12	W13	W14	W15	W16 Milestone 3
	28 Dec-3 Jan	*4-10 Jan*	*11-17 Jan*	*18-24 Jan*	*25-31 Jan*	*1-7 Feb*	*8-14 Feb*	*15-21 Feb*	*22-28 Feb*	*29 Feb-6 Mar*	*7-13 Mar*	*14-20 Mar*	*21-27 Mar*	*28 Mar-3 Apr*	*4-10 Apr*	*11-17 Apr*
Discussion with prospective project sponsors	A															
Prospect Analysis	TEAM															
Prospect Confirmation	TEAM															
Detailed discussion on project motivation and objectives		TEAM														
Obtain Project Data		A														
Exploration of possible analysis methodology		TEAM														
Research on Data Retrieval Techniques: 1. Google calendar public holiday pull 2. Google trends data pull		H														
Complete Proposal		TEAM														
Complete Wiki Page		TEAM														
Data Cleaning			H													
Research on Analysis Techniques: RapidMiner			J													
Research on Analysis Techniques: SAS Text Mining			A													
Perform Exploratory Data Analysis				TEAM												
Project review before consultation with Prof Kam				TEAM												
Meeting with Prof Kam					29-Jan											
Meeting with Client to clarify project objectives, understand client team operations briefing on analysis methodology						02-Feb										
Team review of project objectives and analysis methodology						ALL										
Data pull from GA (paid and unpaid sources)							A									
Web crawling of article attributes based on urls pulled from GA								H								

Task	W1	W2 Milestone 1	W3	W4	W5	W6	W7	W8	W9	W10 Milestone 2	W11	W12	W13	W14	W15	W16 Milestone 3
	28 Dec-3 Jan	*4-10 Jan*	*11-17 Jan*	*18-24 Jan*	*25-31 Jan*	*1-7 Feb*	*8-14 Feb*	*15-21 Feb*	*22-28 Feb*	*29 Feb-6 Mar*	*7-13 Mar*	*14-20 Mar*	*21-27 Mar*	*28 Mar-3 Apr*	*4-10 Apr*	*11-17 Apr*
Web crawling of share counts for articles based on urls pulled from GA and merging with huy's dataset								J								
Update Wiki Page								H								
Perform Exploratory Data Analysis								H								
Attribute Metric Correlation modelling									A							
Regression modelling on organic sources									A							
Odds ratio senstivity analysis across various sources									A							
Identifying most valuable (UVP and ATOP) paid source									J							
K-means clustering (article content)									J							
Cluster Labelling (article content)									J							
Validation of present set of CT									J							
Mapping of clusters to CT									J							
Identification of top 3 CT									J							
z score comparison across CT									J							
Complete Mid Term Report									A & J							
Update Wiki Page									H							
Model Findings and Revision with Client										01-Mar						
Integration of newly formed article content CT label into data cube										H						
Conversion of numerical metrics to categorical variables										A						

Task	W1	W2 Milestone 1	W3	W4	W5	W6	W7	W8	W9	W10 Milestone 2	W11	W12	W13	W14	W15	W16 Milestone 3
	28 Dec-3 Jan	*4-10 Jan*	*11-17 Jan*	*18-24 Jan*	*25-31 Jan*	*1-7 Feb*	*8-14 Feb*	*15-21 Feb*	*22-28 Feb*	*29 Feb-6 Mar*	*7-13 Mar*	*14-20 Mar*	*21-27 Mar*	*28 Mar-3 Apr*	*4-10 Apr*	*11-17 Apr*
K-means clustering (title)										J						
Cluster Labelling (title)											H					
Identification of top 3 clusters											H					
z score comparison across clusters											H					
Update Wiki Page											H					
Model Revision works											TEAM					
Update Wiki Page											A					
Model Revision with Client												15-Mar				
Update Wiki Page												H				
Model Revision works													TEAM			
Update Wiki Page													J			
Model Revision works														TEAM		
Update Wiki Page														A		
Team review before client meeting															TEAM	
Update Wiki Page															H	
Final Client Meeting															05-Apr	
Paper review for Final Report																TEAM
Presentation review for Final Report																TEAM
Final Wiki Updates																H
Final Report Presentation																17-Apr

Research Material

1. Can, F & Ozkarahan, E. (1987). Concepts and Effectiveness of the Cover Coefficient Based Clustering Methodology for Text Databases. Retrieved from https://sc.lib.miamioh.edu/bitstream/handle/2374.MIA/246/fulltext.pdf?sequence=2

2. Jensen. Public Engagement with Research Online. Retrieved from http://www2.warwick.ac.uk/fac/soc/sociology/staff/jensen/ericjensen/pero/pero-google-analytics-guide-v4.pdf

3. Krill, P. (2015, 30 Jun). Why R? The pros and cons of the R language. Retrieved from http://www.infoworld.com/article/2940864/application-development/r-programming-language-statistical-data-analysis.html

4. Ledna. (2014) Comparison of popular analytics tools. Retrieved from http://datafilos.blogspot.sg/2014/01/comparsion-of-sas-enterprise-miner-and.html

5. Omidvar, M.A , Vahid, R.M & Shokry N. (2011, 1 January) Analyzing the Impact of Visiors on Page Views with Google Analytics. Retrieved from http://arxiv.org/pdf/1102.0735.pdf