News Article Position Recommendation Based on The Analysis of Article's Content - Time Matters

Parisa Lak
Data Science Laboratory
Ryerson University
Toronto, Canada
parisa.lak@ryerson.ca

Ceni Babaoglu
Data Science Laboratory
Ryerson University
Toronto, Canada
cenibabaoglu@ryerson.ca

Ayse Basar Bener
Data Science Laboratory
Ryerson University
Toronto, Canada
ayse.bener@ryerson.ca

Pawel Pralat
Data Science Laboratory
Ryerson University
Toronto, Canada
pralat@ryerson.ca

ABSTRACT

As more people prefer to read news on-line, the newspapers are focusing on personalized news presentation. This study is a step towards recommendation of personalized articles to the subscribers of a news agency based on the article's positions in the news website. In this study, we investigate the prediction of article's position based on the analysis of article's content using different text analytics methods. The evaluation is performed in 4 main scenarios using articles from different time frames. The result of the analysis shows that the article's freshness plays an important role in the prediction of a new article's position. Also, the results from this work provides insight on how to find an optimised solution to automate the process of assigning new article the right position. We believe that these insights may further be used in developing content based recommender system algorithms.

Keywords

Content-based Recommender System; Text Analytics; Ranking models; Time-based Analysis

1. INTRODUCTION

Since 1990s the Internet has transformed our personal and business lives and one example of such a transformation is the creation of virtual communities [3]. However, there are challenges in the production, distribution and consumption of this media content [8]. Nicholas Negroponte has contented that moving towards being digital will affect the economic model for news selection and the users' interest play a bigger role for news selection [7]. Therefore, users actively participate in online personalized communities and they expect the online news agency to provide as much personalized services as possible. Such demand, on the other hand, puts pressure on the news agency to employ the most recent technology to satisfy their users.

Our research partner, the news agency, is moving towards providing a more personalized service to their subscribed users. Currently editors make the decision on which article will be placed in which section and who to offer (i.e. the subscription type) on the website based on their experience. Editors also decide on the position of the news within the first page of the section and it is similar for all users. The company would like to first automate this process and in the second step to provide personalized recommendations to their users. They would like to position the news on each page based on the historical behavior of each user. The users explicitly provide their demographic information when they register based on a different level of subscription offered by the agency. The historical news reading behavior of users is also available through the analysis of user interaction logs.

In this work, we investigate different solutions to optimize and automate the process of positioning the new articles. The results of this study may further be used in building a recommender system algorithm to provide personalized news position recommendations to the subscribed users at different tiers. The high level research question that we address in this study is:

RQ- How to predict an article's position in a news website?

To address this question, we evaluate three key factors. First, we compare three text analytics techniques to find the best strategy to analyze the content of the available news articles. Second, we evaluate different classification techniques to find the best performing algorithm for article position prediction. Third, we investigate the impact of the time variable on the prediction accuracy. The main contribution of this work is to provide insights to researchers and practitioners on how to tackle a similar problem by providing the results from a large scale real life data analysis.

RECSYS '16

ACM ISBN 978-1-4503-2138-9…$15.00
DOI: 10.1145/1235

The rest of this manuscript is organized as follows: Section 2 provides a summary of prior work in this area. Section 3 describes the data and specifies the details of the analysis performed in this work. The results of the analysis are provided in Section 4 that is followed by the discussion and future direction in Section 5.

2. BACKGROUND

To automate the process of assigning the right position to a news article, researchers provide different solutions. In most of the previous studies, a new article's content is analyzed using text analytics solutions. The result of the analysis is then compared with the analysis of previously published articles. The popularity of the current article is predicted based on the similarity of this article with the previously published articles. Popularity is considered with different measures throughout literature. For example, Tatar et al. predicted the popularity of the articles based on the analysis of the comments provided by the users [10]. Another study, evaluated the article's popularity based on the amount of attention received by counting the number of visits [1]. Another article popularity measure used in a recent work by Bansal et al. is based on the analysis of comment-worthy articles. Comment-worthyness is measured by the number of comments on a similar article [2].

In the current work we considered the popularity measure to be a combination of some of the aforementioned measures. Specifically, we used measures such as article's number of visits, duration of visit, the number of comments and influence of article's author measure to evaluate the previous article's popularity. The popularity measure is then used towards the prediction of article's position on the news website.

To evaluate the content of the article and find the relevant article topics several text analytics techniques has been used by different scholars [5]. Among all, we selected three commonly used approaches in this study. The three approaches are Keyword popularity, TF-IDF and Word2Vec that will be explained in section 3.

3. METHODOLOGY

In this section we specify the details of our data and we outline the details of the methodology used to perform our analysis. The general methodology framework that was used in this study is illustrated in Figure 1.

3.1 Data

One year historical data was collected from the news agency's archive from May 2014 to May 2015. One dataset with the information regarding the content of the articles as well as its author and its publication date and time was extracted trough this process. This dtaset is then used to generate the keyword vector shown in Figure1

As illustrated in Figure 1, another dataset was also extracted from the news agency's data warehouse. The information regarding the popularity of the article such as,Author's reputation, Article's freshness and Article type were included in this dataset. The dataset also contained the news URL as article related information. This piece of information provides the details regarding the article's section and subscription type. The current position of the article is also available from the second dataset. The popularity of the article is

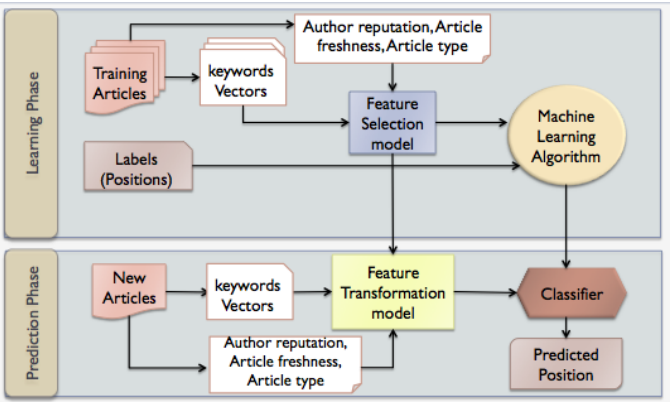


Figure 1: Step by step Prediction Methodology

then calculated based on available features and the position of the article in the website. This information along with the information from keyword vectors are then used as an input to the machine learning algorithms.

3.2 Analysis

We first analysed the content of each article available in the first dataset, using three text analytics techniques. Keyword Popularity, TF-IDF and word2vec was used to perform these set of analysis.

For the Keyword Popularity technique, we extracted the embedded keywords in the article's content and generated keyword weights based on the combination of two factors: the number of visits for a particular keyword and the duration of the keyword on the web site. For instance, if the article had a keyword such as "Canada", we evaluated the popularity of "Canada" based on the number of times it occurred in the selected section and the number of times an article with the keyword "Canada" was visited previously.

In TF-IDF technique, TF measures the frequency of a keyword's occurrence in a document and IDF refers to computing the importance of that keyword. The output from this technique is a document-term matrix with the list of the most important words along with their respective weight that describe the content of a document [9]. We used nltk package in python to perform this analysis over the content of each article.

The last text analytics technique used in this study is word2vec. This technique was published by Google in 2013. It is a two-layered neural networks that processes text[6].This tool takes a text document as the input and produces a keyword vector representation as an output. The system constructs vocabulary from the training text as well as numerical representation of words. It then measures the cosine similarity of words and group similar words together. In another words, this model provides a simple way to find the words with similar contextual meanings [9].

A set of exploratory data analysis was performed on the second dataset to find the most relevant features to define article's popularity. Based on the result from this set of analysis we removed the highly correlated features. The popularity measure along with the position and the keyword vector of each article is then used in 4 main classification algorithms: support vector machine (SVM), Random forest, k-nearest neighbors (KNN) and Logistic regression [4]. The

result of the analysis are only reported for the first two algorithms (i.e. SVM and Random Forest) as they were the best performing algorithms among the four for our dataset.

The steps to perform the prediction analysis also illustrated in Figure1. As shown, the analysis is mainly performed in two phases denoted as "Learning phase" and "Prediction phase". In the learning phase the training dataset is cleaned and preprocessed and the features to be used for the evaluation of popularity are selected based on the exploratory analysis. All observations (i.e. articles) in this dataset are also labeled with their current positions. In the prediction phase, the article content is analyzed and the keyword vectors are created based on the three text analytics techniques. Then, the popularity of the article is calculated based on available features. The test dataset is then passed through the classifier, which predicts the position of the article. The accuracy of prediction is evaluated based on the number of correctly classified instances to the total number of observations and can be computed with Equation 1.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100\% \qquad (1)$$

The result of the analysis is reported in the following section.

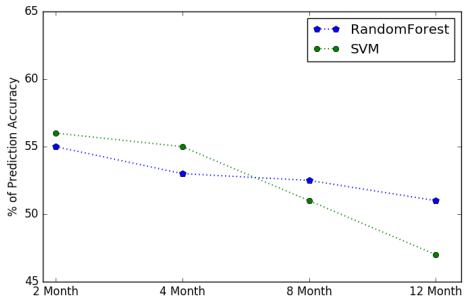
4. RESULTS

The set of graphs in Figure 2 illustrates the percentage of prediction accuracy trend for articles' positions in 4 different scenarios using the two classification algorithms. The blue graph shows the accuracy trend for RandomForest classification algorithm, while the green graph reports the accuracy for the SVM. The 4 scenarios are based on the training data used in the machine learning algorithms. The first points from the left show the accuracy for the scenario, when the training set contains the articles from 2 months prior to the publication of the test article. Similarly the second point from the left shows the scenario in which the training set contains articles from 4 month prior to the publication of the test article and so on for the 8 month and 12 month scenario.

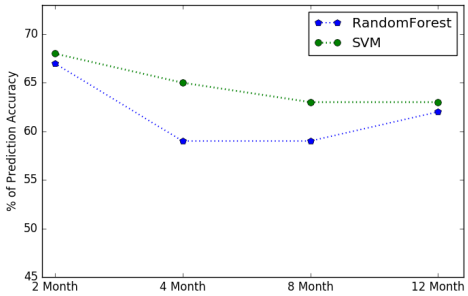
Figure 2(a) shows the accuracy results for the articles when their content (for both training and test dataset) is analyzed based on Keyword Popularity technique. In this graph we observe that the accuracy of the prediction algorithm is related to the time frame factor used to build the training set. More specifically, both algorithms perform best while the most recent articles are used in the training set. The performance of both SVM and Random Forest is dependant on the time frame that is used to define the training set.

Figure 2(b) provides the accuracy for the analysis of the prediction in the case when the articles are analyzed by TF-IDF technique. The result of the analysis for this content analysis technique further confirms that the accuracy of prediction is dependant on the time frame selected to define the training set. For this type of article content analysis, SVM always works superior to RandomForest in terms of accuracy.

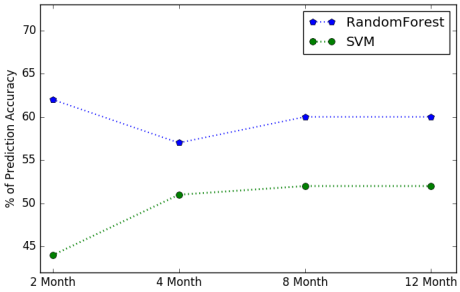
Figure 2(c) shows the result of the prediction for the articles that are evaluated by Word2Vec technique. The result from this graph is slightly different from the previous two graphs. The accuracy for the most recent articles shows to



(a) Keyword popularity



(b) TF-IDF



(c) Word2Vec

Figure 2: Prediction accuracy for SVM and Random forest in 4 time frame scenarios (2, 4, 8 and 12 months) using different article content analysis techniques

be completely different from other scenarios, however the difference between the accuracy of the other scenarios are not shown large enough.The accuracy results for the Random Forest algorithm is consistent with the previous analysis in which the best perfomance is gained through the use of the most recent articles in the training set. However, this is not the case for the SVM algorithm in this set of analysis. The result for this text analytics technique and the use of SVM algorithm works best, while using the older articles. Nevertheless, SVM is not considered as the best performing algorithm for this text analytics technique.

To better visualization the argument that the best performing algorithms for all text nalytics techniques works at best while using the most recent documents we provide the graphs that is illustrated Figure 3.

Figure 3 shows the result from the best performing algorithm for the three content analytics techniques based

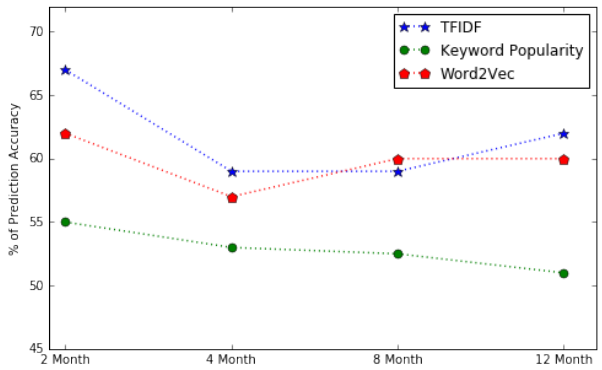


Figure 3: Prediction accuracy based on the three content analysis techniques for the 4 time frame scenarios

on the 4 time based scenarios. Specifically, the blue graph shows the performance of SVM for TF-IDF technique and the green graph and the red graph show the accuracy result for Random Forest for Keyword popularity and Word2Vec, respectively. This figure shows that for all the three content analysis techniques, the best prediction performance is achieved while the fresh articles are used for training purposes. The accuracy is always dropped as more old articles are added to the training set in the 4 month scenario. In Word2vec technique, the accuracy increases when the 8 month prior articles are used for training. This is not the case for the other two text analytics techniques.

Another observation from this analysis is that TF-IDF technique provides the best keyword evaluation that further generates higher prediction accuracy for article's position.

5. DISCUSSION & FUTURE DIRECTION

Personalized news recommendation is a recently emerged topic of study based on the introduction of the interactive online news media. The decision on the news presentation is made based on the assigned position of the article within the news website. The position of the article can be assigned based on the popularity of the article. The popularity of the article can be predicted based on the analysis of its content and the similarity of the article's content to the previously published articles. Previous article's popularity is measured based on different popularity measures. This study is the preliminary investigation on the importance of attributes to be used in the content based personalized news recommendation system. In this study, we used a combination of article's popularity measure attributes as well as the attributes from the analysis of the articles' content to provide prediction of the position of a new article.

Essentially, we evaluated the impact of the three key factors on the prediction of new article's position. The results from the analyses provide evidence that all three factors under investigation in this study plays a role in the accuracy of prediction. One of the important findings from this work is that the result of the analysis of a new articles content should only be evaluated with the recent historical articles. The analysis shows that as the older articles are used as an input to the prediction algorithm the accuracy of the system drops in almost all cases. Also, the best performing prediction algorithm shows to be dependent on the text analytics techniques used in the analysis of the article's content. Regardless of the prediction algorithm the best text analytics technique for the current dataset is shown to be TF-IDF.

The results from this study can cautiously be extended to other datasets. To avoid the impact of sampling biases we used 10 fold cross validation technique in our prediction model. Also, the analysis of the large scale real life data minimizes this threat to the validity of this study. In our future work, we will use the the results from this study as well as the features detected through the exploratory analysis to design a personalized news recommendation system.

6. ACKNOWLEDGMENTS

The authors would like to thank Bora Caglayan, Zeinab Noorian, Fatemeh Firouzi and Sami Rodrigue who worked at different stages of this project. This research is supported in part by Ontario Centres of Excellence (OCE) TalentEdge Fellowship Project (TFP)-22085.

7. REFERENCES

[1] M. Alshangiti. Mining news content for popularity prediction. 2015.

[2] T. Bansal, M. Das, and C. Bhattacharyya. Content driven user profiling for comment-worthy recommendations of news and blog articles. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 195–202. ACM, 2015.

[3] P. J. Boczkowski. *Digitizing the news: Innovation in online newspapers*. mit Press, 2005.

[4] 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.

[5] S. Lee and H.-j. Kim. News keyword extraction for topic tracking. In *Networked Computing and Advanced Information Management, 2008. NCM'08. Fourth International Conference on*, volume 2, pages 554–559. IEEE, 2008.

[6] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[7] N. Negroponte. *Being digital*. Vintage, 1996.

[8] J. V. Pavlik. *Journalism and new media*. Columbia University Press, 2001.

[9] N. Pentreath. *Machine Learning with Spark*. Packt Publishing Ltd, 2015.

[10] A. Tatar, J. Leguay, P. Antoniadis, A. Limbourg, M. D. de Amorim, and S. Fdida. Predicting the popularity of online articles based on user comments. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, page 67. ACM, 2011.