International Journal of Innovations & Advancement in Computer Science
IJIACS
ISSN 2347 – 8616
Volume 7, Issue 4
April 2018

Extracting News from Online Database – News Clustering based on Content Ranking

Prof.(Dr) S.R.Dhore
Mandeep Singh, Pawan Kumar, Sachin Choudhary
Army Institute of Technology, Pune

Abstract— We present NewsMaster, an approach to collect, cluster and categorize and select news articles from Internet.Due to the perception of cheap publishing, organizations have been producing enormous amount of content online since the hidden cost of maintenance and usability has always been neglected. This presents the opportunity for automatically maintaining crisp and usable content, especially in news articles. In this paper, we use machine learning algorithm to extract features from different classes of content and cluster them under an umbrella topic. For each cluster, we then go on to predict popularity of documents using additional features based on the content only. We conduct our experiments on different news corpuses. Our study also serves to remove information redundancy in multiple articles.

Keywords—News, redundancy, content popularity, machine learning

I. INTRODUCTION

News articles are very dynamic in nature due to continuously developing nature of the event and parallel reporting of the same, thus they have a very short span of life. The ease and low cost of online content creation and sharing have changed the traditional rules of competition for public attention. News sources now concentrate a large portion of attention on online mediums where they can disseminate their news effectively and to a large population. Due to the time-sensitive post aspect and intense competition for attention in the socially connected digital platform, accurately estimating the extent to which a news article will spread on the web is extremely valuable to journalists, content providers, advertisers and news recommendation systems. However, predicting the online popularity of online news articles is a challenging task. First, context outside the web is often not readily accessible and elements such as local and geographical conditions and various circumstances that affect the population to make this prediction extremely difficult. Furthermore, network properties such as the structure of social networks that are propagating the news, influence variations among members and interplay between different sections of the web add other layers of complexity to this problem. Most significantly, intuition suggests that content of an article must play a significant role in its popularity. Content that resonates with most of readers such as a major worldwide event can be expected to garner wide attention while specific content relevant only to a few may not be as successful. Content that is up-to-date and highlights all aspect of that article.

The news data for our study has been collected from News Aggregator Dataset from Kaggle. To generate features and classify the articles, we have used Multinomial Naive Bayes. To remove redundant information, we perform specific topic-wise clustering in a certain timeframe. For each cluster, we analyze the contents of new articles and use those for prediction of the popularity prior to publishing. Our work shall also help content writers to remove irrelevant, outdated, trivial and redundant content.

II. BUSINESS

♦ **Content Caching and Traffic Management**There is a hidden cost to publishing content, the cost to review and maintain the content. The millions of articles also affect the usability and maintainability of the site. In the long run, it is necessary to tackle redundant, outdated and trivial content which has been cursing the site.

♦ **Advertising** This work can find its application in content-based advertisement along-side news pieces. It will optimize ad-placement logistics and revenues.

♦ **News Aggregation** With our current event driven clusters knowledge base, we predict the popularity of written articles to be published in that

International Journal of Innovations & Advancement in Computer Science
IJIACS
ISSN 2347 – 8616
Volume 7, Issue 4
April 2018

domain. It will allow content writer to write more relevant and less redundant pieces that can make it different from current flowing articles. We have been aggregating up-to-date content rich articles ignoring social backlinks.

♦ **Trends Forecasting** Since the cache contains most popular pieces from different news events, we can show current trends with no redundancy. This data helps in forecasting future trends.

III. LITERATURE SURVEY

The work by Martin Weber[1] generates a comparison between different news sources and approaches followed by them as in Table 1.

SCOPE	APPROACH	EXAMPLES
Global	Limited amount of news according to user's region & timeframe, Newsblaster, Google News, Yahoo News, Bing News	Newspapers, magazines, news web sites, news broadcast on TV/radio
Resort	Politics / Business / Local / Sports / Feuilleton / Technology, Entertainment / Music / Leisure.	Weblogs, Rivva, Blogrunner, Sections of Newsblaster / Google News
Sub-Resort	Specialized areas / topics within resorts e.g. Sports ~ Soccer, Technology ~ Apple	Specialized Sites (editorial) e.g. football365.com, 9to5mac.com
Social	Most liked / linked / favorited stories in timeframe	Twitter, favstar.fm, Google+, Rivva, zite, trap.it, Fever, Wavii
Social (Personal)	Linked / liked / favorited stories from peers (and evt. outsiders) in social network	Google News (personalized version) Instapaper, zite, trap.it, Facebook, Twitter
Keyword-based	Filter streams of news in timeframe according to fixed keyword(s) / interest(s) or rules	Yahoo Pipes, Google Alerts, Google API, advanced / custom Google Search
Mashups	mix of two or more of the above-mentioned techniques to filter, aggregate, cluster news	Yahoo Pipes, trap.it, Zite, Rivva, Techmeme Newsblaster (content summaries)

Table 1. Classification of news through various media

IV. SOLUTIONFRAMEWORK

Fig 1. Shows high level overview of solution framework

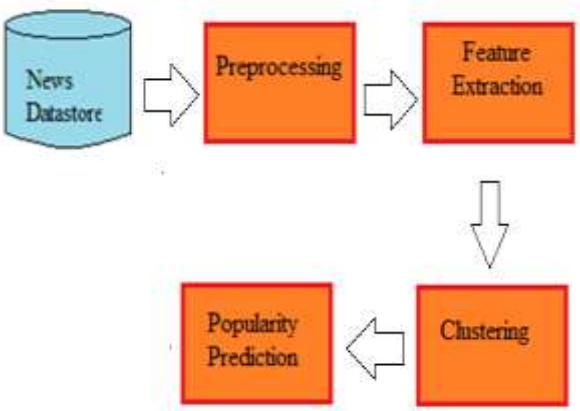

Fig. 1 Solution Framework

A. Preprocessing

We preprocess the data to make processing more meaningful.

- **Filtering** Removal of markup, punctuation and special characters from sentences.
- **Tokenization** Splitting of text into individual units.
- **Stemming** Reduction of words to their base forms
- **Stop words Removal** Deletion of words that do not convey any special meaning.
- **Pruning** Removal of words that do appear with a low frequency throughout the text.

The result of these preprocessing steps is a set of feature words.

B. Text Understanding

Text understanding consists in reading texts formed in natural languages, determining the explicit or implicit meaning of each element such as words, phrases, sentences and paragraphs, and making

International Journal of Innovations & Advancement in Computer Science
IJIACS
ISSN 2347 – 8616
Volume 7, Issue 4
April 2018

inferences about the implicit or explicit properties of these texts.

TF-IDF

To get importance of word in news corpus, we used tf-idf algorithm. The tf-idf value
increases proportionally to the number of times a word appears in the document and is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general. *tf-idf = (1 + log(tf)) * log(N/dfw)*

Defining,

tf: term frequency (the count of words in headlines)

idf: inverse document frequency

N: number of documents (number of news headlines)

dfw: document frequency of term (number of headlines in which word appears)

MODELS

For classification, we created models using deep learning and machine learning classifiers. We made a 3-layer deep learning model (Fig 2) and for machine learning we created classifier using SVM and Naïve Bayes. We achieved maximum accuracy of model (0.89) using Naïve Bayes and with SVM classifier, accuracy achieved is 0.84, the accuracy of deep learning model was much lower(0.36). It is due to nature of dataset, deep learning performs much better on very huge dataset and the data present for out training is already classified into categories(business, medical, entertainment and technology). So, we are using Naïve Bayes Classifier for this paper.

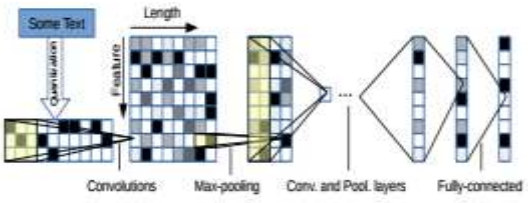

Fig. 2 Deep ConvNets model illustration for feature extraction

C. Clustering

In recent years, internet has become a mainstream medium and offers opportunity for large-scale production and distribution. With more news than ever, it has become increasingly difficult to find relevant news. Regardless which approach is taken and which services are used, one may be confronted with multiple news about the same event within the field of interest. The importance of a news event creates the need for a regular detailed coverage and hence, duplicates and redundant pieces. During high-peak of interest to a topic, there is no limit to number of duplicates produced. We need to manually filter and review the relevant news pieces. Existing approaches like Weber [1]cluster news pieces based on similarity of textual content. We intend to use deep learning feature vectors for clustering news items into highly specific cluster from a news event. Clustering algorithm k-means does not work because it requires number of clusters beforehand. As the number of clusters will never be fixed, we use Average-link agglomerative clustering. We believe that the cluster should be densely connected to an event and thus, average-link distance.

D. Popularity Prediction

We are motivated to predict popularity of article beforehand only from content based features and store only a plausible set of articles from each cluster. We intend to generate a score for each of the articles unlike categorizing them into classes.

Features: The choice of features is motivated by multiple questions. Does the source agent reach many readers? Does the language connect with the reader? Has the article became outdated? Do we have some information in the news piece or not? Is the news worthy of a read? These questions helped us in designing following five features.

- **Age** The date of publication of news given by the dataset. We remove few recordswith missing dates.
- **Text Quality** The ratio of size of document before and after preprocessing.
- **Source Quality** The popularity of source of the content given by initial number of hits provided by the source. If missing, we use the popularity of news agent. This is log-normalized to account for high range of hits.
- **Subjectivity** This examines whether an article is written in more emotional, touchy tone, where it connects with the reader. We make use of subjectivity classifier from Ling pipe, a natural language toolkit.
- **Named Entities** We hypothesize that well-known named entities will cause a further spread of the article. For instance, articles on

International Journal of Innovations & Advancement in Computer Science
IJIACS
ISSN 2347 – 8616
Volume 7, Issue 4
April 2018

Narendra Modi are more likely to be popular among Indian Readers as compared to others. We make use of Stanford CoreNLP to process named entities. We rate entities based on their prominence or past popularity in the media.

E. Experimental Evaluation

A) Dataset:

1) **News Headlines of India**[7] This dataset consists of 16 years of categorized headlines focused on India.
2) **News Aggregator Dataset**[8] This dataset consists of headlines and categories of 400k news stories from 2014.

B) Baseline:

1) **News Aggregators** We conduct an internal survey to verify initial results of the pipeline when compared with different news agents and aggregators like Google News.

V. OUTPUT

NewsMaster is a web page with the latest news(last two days), the news of the day is the lead. Below, cluster of stories are presented. The news is divided into four categories, i.e. Business, Entertainment, Medical and Technology. Further the articles are ranked based on table 2.

Feature	Description
Sentiment Score	Positive articles on top and negative ones below the list
Age	Difference between published date and today's date
Source	The top source is decided based on number of visitors on website, e.g. For technology news, Fig 3
Text Quality	The ratio of article data before and after preprocessing
Named Entities	Country a person or location belongs to

Table 2. Article features

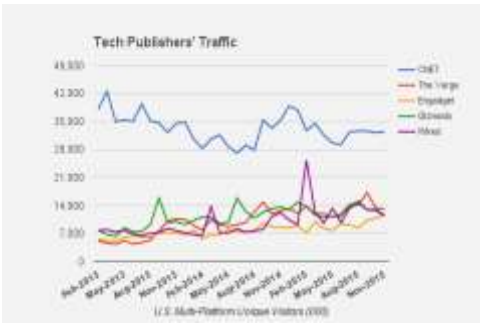

Fig 3. Number of visitors from Feb 2013 to Nov 2015

VI. CONCLUSION

In this paper, we improve the quality of news cache and recommendations by predicting popularity of articles prior to publishing. The need for the same arises from the stiff competition among different news agencies and aggregators. To remove redundant information, we make highly specific clusters of news items. Finally, we predict the most popular pieces in different clusters to provide the set of most popular articles, which is then used for multiple use-cases in content caching, advertising, forecasting and recommendation. With an initial survey, we ensure inceptive results of the pipeline versus different competitors. Lastly, we compare with different baselines to ascertain quality of our work.

REFERENCES

[1] Martin Weber and Maarten H. Lammers "Finding news in Haystack - Event based news clustering with social media based ranking"
[2] Roja Bandari, SitaRam Asur, Bernando A. Huberman"The Pulse of News in Social Media: Forecasting Popularity"
[3] Xianshu Zhu and Tim Oates"Finding story chains in newswire articles using random walks"
[4] Xiang Xang, Junbo Zhao, Yann LeCun"Character-level Convolutional Networks for Text Classification"
[5] Lewis, D. D. Naive (Bayes) at forty: The independence assumption in information retrieval. MachineLearning: ECML-98, Tenth European Conference on Machine Learning 1998.
[6] Kecman V. "Learning and Soft Computing, Support Vector machines, Neural Networks and Fuzzy Logic Models," The MIT Press, Cambridge, MA, 2001.
[7] News aggregator Dataset https://www.kaggle.com/uciml/news-aggregator-dataset
[8] News Headlines of India https://www.kaggle.com/therohk/india-headlines-news-dataset