News Headlines: What They Can Tell Us?

Sahisnu Mazumder, Bazir Bishnoi and Dhaval Patel
Indian Institute of Technology - Roorkee, Uttarakhand, India.
{sahisnumazumder, bazirbishnoi}@gmail.com, patelfec@iitr.ac.in

ABSTRACT

News headlines represent the key idea of news articles published in online news media and act as a great resource for discovering news concepts and their relationships. Moreover, the temporal information associated with the news headlines can be utilized to capture the temporal dynamics of the news concepts and their relationships which facilitates the development of many time-aware news analytics applications. Existing works on news data analytics have mostly dealt with news articles, but none of them has talked about the usefulness of news headlines in news data analytics research. In this paper, we analyze the potentiality of news headlines in inferring interesting facts of the news world. We show how news headlines can help us to capture the temporal dynamics of the news concepts and their relationships. We introduce the notion of `Time-aware News Concept Graph` to capture the said temporal dynamics and show how it opens the doorway of developing numerous interesting news analytics applications. The results of our analysis conform to the facts of the reality and advocate for the success of our effort.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Theory, Management, Experimentation.

Keywords

News Headlines, News Concept and Relationship Discovery.

1. INTRODUCTION

Online news media has grown to be a reliable and widely-accessed source of information about the latest happenings around the world. With the escalating popularity of smartph one-enabled internet browsing in recent years, a significant number of internet users currently acquire news from these sources [1]. The information published in these news sources are primarily news articles associated with a news headline for each of them. The news headlines are summarized, audited textual information and represents the key idea of the corresponding news article. Such news headlines are useful for studying relationships between various news concepts and analyzing their temporal dynamics of the news concepts and their relationships over a course of time [5].

The news concepts are basically nouns, collocation of nouns and collocation of nouns and numbers that co-occur in the news headline to convey some fact of the real world. For example, in the news headline- "*Goa Police to file chargesheet against Tarun Tejpal by Feb 5*", "*Goa Police*", "*chargesheet*", "*Tarun Tejpal*", "*Feb 5*" are the news concepts that have co-occurred together. We can utilize the notion of the co-occurrence of various news concepts to discover relationships among themselves. For an instance, the two concepts "*Goa Police*" and "*Tarun Tejpal*" are related to each-other in the given news headline along with "*Goa Police*" and "*chargesheet*", "*chargesheet*" and "*Tarun Tejpal*" etc. Thus, we can discover news concepts their relationships from the news headlines and use them to build a `News Concept Graph` where nodes are the news concepts and two concepts are connected by a link if they co-occur in some news headline.

Such `News Concept Graph` can be empowered further to capture the temporal dynamics of the news concepts and their relationships. With the advent of new news information, new concepts appear, rule the news world for a considerable amount of time and then, gradually disappear from the limelight. We can track the temporal information of the news headlines and augment the `News Concept Graph` with these information to capture the said temporal dynamics. Thus, we introduce the notion of a `Time-aware News Concept Graph` for studying the temporal evolution of news concepts and their relationships. Such graph is self-evolving in nature and can be leveraged to develop myriads of news analytics applications like time-aware query expansion, news trend analysis, time-aware entity relationship mining etc.

In this paper, we analyze the potentiality of news headlines in inferring temporal dynamics of the news concepts and their relationships. To aid this task, we develop a `Time-aware News Concept Graph` to represent the knowledge discovered from the news headlines. Finally, we present some news analytics applications where such graph can be used to infer interesting facts of the news world. We have collected 1.5 million Indian news headlines from 87 different online news sources over a period of 5 months (from 26th January,

I-CARE 2014, October 9-11, 2014, Banglore, India.

http://dx.doi.org/10.1145/2662117.2662121.

2014 to 29th June, 2014) and used this large-scaled dataset for the potentiality analysis purpose. Considering the related works [5] [4] [3] done in the area of news data analytics, the proposed idea is novel to the best of our knowledge and a topic of demand in present research scenario.

2. NEWS HEADLINES DATASET

Dataset Collection. To collect the news headlines dataset, we have built a news crawler. The crawler crawls each of a predefined list of 87 news websites at an interval of 30 minutes and then, collects information about the news headlines published during that time interval. It runs on an Intel Xeon Machine with 32 processors and 64 GB RAM with an Ethernet access of 100 Mbps. Table 1 shows some of the major news sources frequently crawled by the said crawler.

Table 1: Examples of major news sources.

News Source	URL
Indian Express	`http://indianexpress.com`
The Hindu	`http://www.thehindu.com`
India Today	`http://indiatoday.intoday.in/`
CNN	`http://edition.cnn.com/`
BBC	`http://www.bbc.com/news/`

The news crawler visits each website, crawl it and store the data into html format. To extract the headlines from the html pages, we use jsoup parser. The news headlines are stored in the database in the form of records where each record is primarily consisted of a unique *news headline ID*, *news headline text*, *start_timestamp* (the time stamp when the news headline appeared in the news media) and *end_timestamp* (the time stamp when the news headline disappeared from the news media), *source_id* (ID of the news website from which the news headline has been extracted), *source_url* (url of the news website from which the news headline has been extracted) and *category* of the source (e.g., business, sports, technology etc.). Examples of some sample news headlines are shown in Table 2.

Table 2: Sample of news headlines data.

News Headline Text	Start Timestamp	End Timestamp
Cricket websites delay live score updates after Delhi HC order	2014-01-27 17:47:47	2014-01-29 10:47:47
Satya Nadella named Microsoft CEO	2014-02-04 20:47:01	2014-02-05 10:17:01
Modi to address rally in Meerut today amid tight security	2014-02-02 10:17:05	2014-02-02 15:17:05

Statistics. Table 3 presents the summary of news headlines data set collected over a time span of 5 months (from 26th January, 2014 to 29th June, 2014).

3. TIME-AWARE NEWS CONCEPT GRAPH

After considerable amount of data is accumulated, we proceed to form the `Time-aware News Concept Graph`. We process news headlines one by one, extract concepts and their relationships and form the time-aware news concept graph to represent the discovered knowledge. In the concept extraction process, we first tag each word of a news headline with their corresponding parts-of-speech using Stanford

Table 3: Statistics of the news headlines dataset.

Parameters	Value
Total No. of HTML pages	6,05,085
Memory Consumed by HTML pages	69 GB
Total No. of News Headlines	15,30,129
Memory Consumed by News Headlines Dataset	550 MB
News Headlines collection rate per day	10,200

POS tagger. Next, we infer the nouns, collocation of nouns and collocation of nouns and numbers as news concepts. If two news concepts co-occur in a news headline, we pair them to represent the relationships. Then, the extracted news concepts become nodes and the relationships become edges in the time-aware news concept graph. The details of the `Time-aware News Concept Graph` formation is given in our technical report.

The `Time-aware News Concept Graph` is modelled in the form of a property graph which consists of a set of concept nodes, each labelled with a unique news concept denoted by *Concept_Name* and a set of links that connect the related concept nodes. Figure 1 shows the generic structure of the graph where 6 concept nodes $\{C_i \mid 1 \leq i \leq 6\}$ are interconnected with links with label"*Is_Related_To*" which show the relationships among themselves. Some concept nodes are special in the sense that they carry information about a personal news entity. In this case, the *Concept_Type* property of the concept node is set as "*Personal Entity*"; otherwise, it is simply set as "*concept*". The concept nodes having *Concept_Type* as "*Personal Entity*" also accompany with a property namely *Attribute_List* which is basically a list of strings that describe the personal entity.

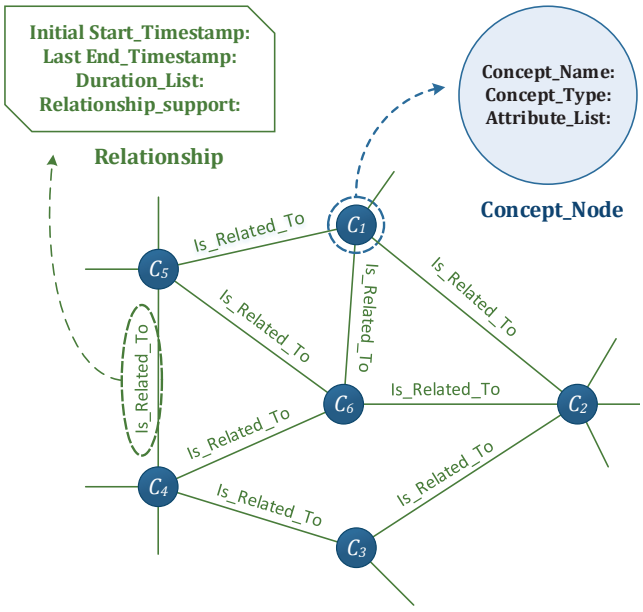


Figure 1: Generic Structure of Time-aware News Concept Graph.

Each link in time-aware news concept graph is associated with the following set of properties, for tracking the temporal dynamics of news concepts and their relations:

- **Initial Start_Timestamp.** It specifies the time-stamp when the concept pair involved in the relationship co-occurred for the first time in any news headline, i.e.,

the time when the news concept took birth in the concept graph.

- **Last End_Timestamp.** It specifies the time-stamp when the concept pair involved in the relationship co-occurred for the last time in any news headline, i.e., the time when the relationship become inactive in the concept graph.

  The quantity (*Last End_Timestamp - Initial Start _Timestamp*) gives the *Life-Time of the concerned relationship* during which the concept pair has co-occurred in different news headlines at different times in the news media.

- **Duration_List.** It holds the detailed information about the co-occurrences of the concept pair in various news headlines during the *Life-Time of the concerned relationship*. It is basically a list of records of the form *<headline_id, start_timestamp, end_time stamp>* where the *headline_id* denotes the id of a news headline where the concept pair has co-occurred and *start_timestamp* and *end_timestamp* are set to the corresponding time-stamp values of that news headline.

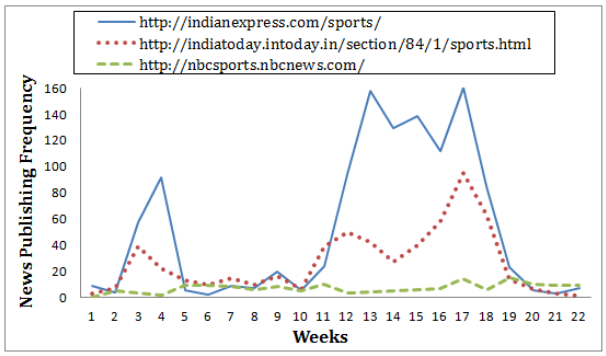
- **Relationship_Support.** It denotes the number of different news headlines where the concept pair has co-occurred during the *Life-Time of the concerned relationship* and is given by the size of the *Duration_List*.

We have used Neo4j graph database [2] to store all the information while constructing the news concept graph. Using the collected news headlines dataset, we have built a Time-aware News Concept Graph having 22,704 concept nodes and 1,54,318 relationships and used this knowledgebase for news analytics purpose.

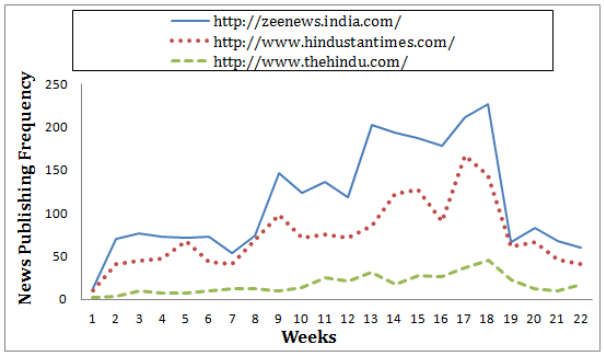
4. NEWS DATA ANALYTICS AND APPLICATIONS

In order to analyze the potentiality of the news headline dataset and the usefulness of the proposed `Time-aware news Concept Graph`, we have performed our analytical study in two major perspectives-

Perspective I: News Source Analytics. In new source analytics, we have analyzed the potentiality of different news sources in publishing news information. In this analysis, we have discovered various statistics about news sources like identifying top-k news sources that publishes information about a given concept, top-k different news concepts that are involved in the news published by a given news source etc. For example, Figure 2(a) and 2(b) show statistics of three news sources in publishing news related to the concepts "*ipl*" and "*narendra_modi*" respectively over past 5 months. Among these three news sources shown in each graph, one is the top-ranked news source that publishes news frequently, one is average-ranked that publishes news in moderate rate and other one is low-ranked news source that rarely publishes news considering the respective concepts. In Figure 2(a), we can observe that the news source `http://indianexpress.com/sports/` has published more news about "*ipl*" compared to `http://indiatoday.intoday.in/section/84/1/sports.html` and `http://nbcsports.nbcnews.com/`. Moreover, during the period of week 11 to 20, the rate of publishing "*ipl*" related news has raised to a significant extent due to the occurrence of IPL tournaments in India. Similarly, in case of "*narendra_modi*", news



(a) "**ipl**" (related to sports)



(b) "**narendra_modi**" (related to politics)

Figure 2: Frequency of publishing news related to two news concepts by two news sources over 22 weeks of past 5 months.

source `http://zeenews.india.com/` has dominated `http://www.hindustantimes.com/` and `http://www.thehindu.com` with respect to the rate of news publishing. Such analysis are very useful in inferring the **biasness** of different news sources in publishing news of certain concepts and can be leveraged to build various **news source recommendation** applications.

Perspective II: News Headline Analytics. In News headline analytics, we have analyzed the potentiality of news headlines and inferred interesting facts of the news world which mainly involves mining and analysis of the news concepts and their relationships. We have extracted information about the temporal variation of popularity of 15 news concepts from the `Time-aware news Concept Graph` over past 22 weeks and prepared the stack graph shown in the Figure 3. Here, only 10 concepts' temporal dynamics are clearly visible in the Figure and the remaining 5 concepts' information is not visible due to their lower popularity distributions. Figure 4(a), 4(b) and 4(c) show the magnified view of the popularity distributions of three concepts "*kkr*", "*rahul_gandhi*" and "*fifa*" respectively inferred from the stack graph in Figure 3.

Careful observation of the mentioned figures shows that, "*narendra_modi*" was the most popular concept in the news world over past 5 months followed by "*rahul_gandhi*", "*aap*", "*ipl_7*" etc. Moreover, considering the temporal dynamics of news concepts, "*rahul_gandhi*" is showing decreasing popularity trends in recent weeks (see Figure 4(b)), whereas newly emerged concept "fifa" is showing growing popularity due to the on-going FIFA Worldcup news updates (see Figure 4(c)). Besides that, concept "*narendra_modi*" has persisted in news media over past 5 months with significant

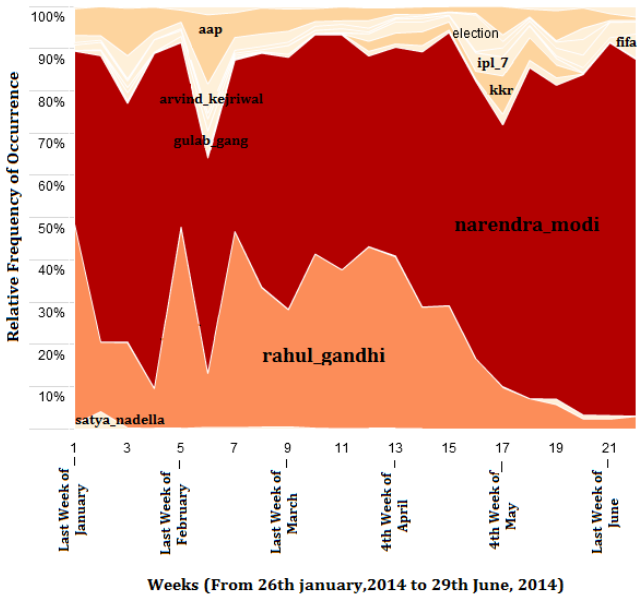


Figure 3: Temporal dynamics of different news concepts over past 5 months.(Best viewed in colour.)

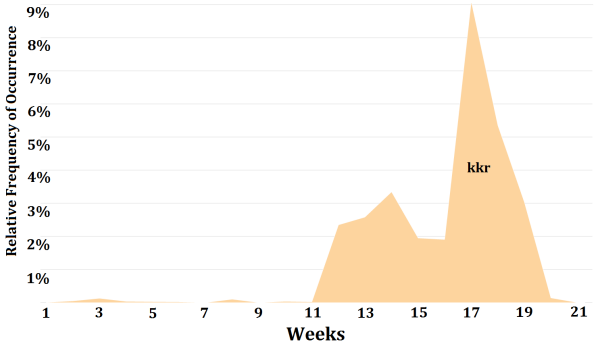
popularity distributions (see Figure 3) and concept "*kkr*" has shown significant popularity only during week 11 to 20 (see Figure 4(a)) when IPL-7 matches were going on. Similarly, we have analyzed the temporal popularity variations of relationships between news concepts and discovered interesting facts like "*narendra_modi*" was mostly cited with "*rahul_gandhi*" in last 5 months compared to "*arvind_kejriwal*" which implies that "*rahul_gandhi*" was the stronger opponent of "*narendra_modi*" compared to "*arvind_kejriwal*" in Indian Lokshabha Election 2014.

Using all these analytical results and `Time-aware news Concept Graph`, we can build following news analytics applications:

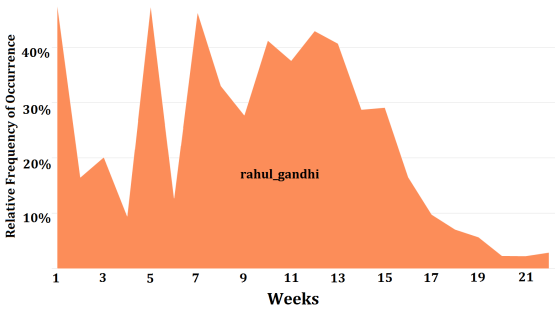
- **Time-aware Entity Relationship Mining.** for finding top-k names that have co-occurred frequently with a person over a specified time-span, how their relationships have evolved over a course of time etc.
- **Temporal Rank-aware News Concept Recommendation.** for recommending other high-ranked concepts of a week based on a given concept (searched by the user) that has got top rank in that week.
- **Time-aware Query Expansion.** Such query expansion returns a set of related concepts co-occurred with an input news concept in any news headline within a specified input time-span and is used for recommending time-specific related news concepts.
- **Concept-based Community Discovery.** Such application concerns about finding concept clusters consisting of strongly related concepts in `Time-aware news Concept Graph` and is used for context-aware related concept recommendation purpose.

5. CONCLUSIONS

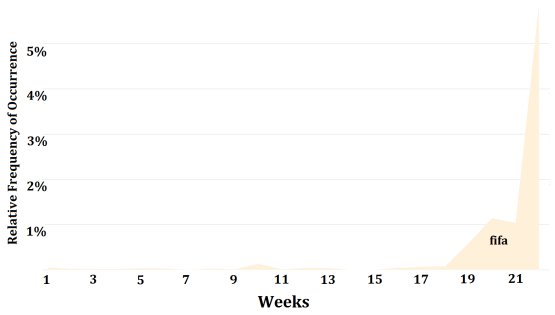
In this paper, we have investigated the potentiality of the news headlines in discovering interesting facts about the real-world events. We have collected a large-scale news headlines data set and discovered and represented the knowledge about the news concepts and their relationships using



(a) "**kkr**" (related to sports)



(b) "**rahul_gandhi**" (related to politics)



(c) "**fifa**" (related to sports)

Figure 4: Temporal dynamics of the news concepts over past 5 months.

a `Time-aware News Concept Graph` for news data analytics and real-world applications development purpose. In summary, the paper shows the usefulness of news headlines in news analytics research and indicates several promising research directions in this area.

6. REFERENCES

[1] Americans spending more time following the news. source:`http://www.people-press.org/2010/09/12/americans-spending-more-time-following-the-news/`.

[2] Neo4j graph database. source:`http://www.neo4j.org/`. Accessed on June, 2014.

[3] F. Goossen, W. IJntema, F. Frasincar, F. Hogenboom, and U. Kaymak. News personalization using the cf-idf semantic recommender. In *In WIMS*. ACM, 2011.

[4] R. Goyal, R. Malla, A. Bagchi, S. Mehta, and M. Ramanath. Esthete: a news browsing system to visualize the context and evolution of news stories. In *In CIKM*, pages 2529–2532. ACM, 2013.

[5] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *In SIGKDD*, pages 497–506. ACM, 2009.