Leveraging Social Big Data and Machine Learning for Advanced Predictive Analytics

Mahadevi Somnath Namose[1], Dr. Pravin H. Ghosekar[2]
[1]*Research Scholar, Department of Computer Science, Sardar Patel University, Bhopal, MP, India*
[2]*Professor, Department of Computer Science Sardar Patel University, Bhopal, MP, India*

***Abstract*—Instability in key socioeconomic indicators can significantly impact global development. This thesis introduces a suite of novel big data analytics algorithms designed to process unstructured web data streams, enabling the automatic inference of events, knowledge graphs, and predictive models. These methods facilitate improved understanding, characterization, and anticipation of socioeconomic indicator volatility. Four principal contributions are presented. First, novel models are proposed for extracting events and learning their spatio-temporal features from large volumes of unstructured news streams. Two types of event models are explored: one based on event triggers and another probabilistic model that identifies meta-events by extracting named entities from text streams. The second contribution addresses the extraction of knowledge graphs from time-sensitive data such as news and real-time events. Event graphs provide a concise representation of event chronologies relevant to specific news queries by characterizing their interconnections through event-phenomenon graphs, while spatio-temporal article graphs capture inherent relationships among news stories. The results of predictive analysis using these approaches are also presented.**

***Index Terms*—Big data, Machine Learning**

I. INTRODUCTION

Many countries, particularly those in the developing world, have historically exhibited significant volatility across a range of socioeconomic indicators. Such volatility in key economic indicators—including commodity prices, unemployment rates, and currency exchange rates—can have severe consequences for national economies. For example, fluctuations in commodity prices negatively affect GDP growth, national budgets, and wealth distribution, and may exacerbate poverty levels. In several countries, including India, commodity sales constitute more than 90% of total revenue. Currency exchange rate fluctuations also contribute to commodity price volatility. Therefore, maintaining economic stability requires continuous monitoring and analysis of these socioeconomic indicators and their volatility. Despite decades of recorded fluctuations, a fundamental understanding of the causes and interconnections among various socioeconomic indices remains elusive. Economists typically rely on established economic models to interpret and predict indicator volatility, while computer scientists have employed computational modeling techniques to analyze structured time series data. The recent proliferation of unstructured data streams on the internet, coupled with advances in computational linguistics, has enabled the exploration of previously inaccessible data reflecting societal and economic changes. This study contributes to the field of big data analytics by introducing unique algorithms capable of automatically inferring events, knowledge graphs, and predictive models from unstructured news streams.

1.1 Objective

Develop an autonomous inference engine capable of identifying relevant trends from online data for specific regions, time periods, and topics.

To accurately categorize and represent events, it is necessary to process a large corpus of news articles. The goal is to construct knowledge graphs that (a) uncover hidden relationships for predictive analysis, (b) clarify the broader socioeconomic implications of news events, such as food prices, (c) visually depict terms within articles to illustrate their interrelationships, and (d) generate structured, domain-specific content from unstructured text. These

connections enable predictions about economic and social variables.

II. LITERATURE SURVEY

2.1 Big Text Figurative Schemes

Several sparse, low-dimensional representational systems for text documents have been developed as a result of recent developments in text mining techniques. In topic models like pLSI [1] and LDA [2], a huge document corpus with a vocabulary of 100,000 words may be represented by only 100 topics. These models have greatly improved the efficiency with which information may be gleaned from texts written in natural languages. Such document representation has proven useful in internet news items.

2.2 News Analytics and Submissions

New information is used by several current initiatives to improve existing applications. Radinsky and Horvitz [5] offer a paradigm that can be used to predict events like the spread of disease in the wake of natural disasters. Rudin et al. [6] focus on applying association rule mining and Bayesian analysis to predict the next event in a sequentially arranged data set.

Combining a traditional statistical model with machine learning techniques, the New York Times database is used to create forecasts. In order to analyse the Twitter data following the Haiti disaster, FBLG modified their algorithm for discovering temporal dependence from time series data. Luo et al. [7] also provided evidence for the connection between events and time series data.

By examining the effect of uncommon news events on stock prices, Hogenboom et al. [8] improved the efficacy of Value-at-Risk (VaR), a common technique for gauging portfolio risk.

Some efforts have been made to foretell some factors, such stock price, using news data. Presenting a revolutionary approach that incorporates financial news and market reaction into stock price predictions.

2.3 Methodology

In this piece, we'll examine the process that leads to a shortened answer to a user's search query on the internet. Foremost, our summarizing engine pulls the top 64 (customizable) search result pages from Google using the Google Search API. The summarizing engine's primary function is to analyses the text of each search result page in order to compile a streamlined summary page. The final search response is an abbreviated version of the original text that includes only the most important information from the longer text, relevant graphics, and related references to the individual search result pages, should the user feel compelled to click on any result page. Each page's relevant summaries are extracted by the Text Summarization Engine, relevant images are extracted by the Image Extraction Engine, and the aggregated summaries and photos are presented by the Aggregation and Presentation layer. Next, we go into the specifics of each subsystem.

2.4 Extraction Using Domain-Specific Large Datasets:

We will look at how to extract structured information from unstructured data in a specific domain. Each key corresponds to an attribute of the underlying domain, and the value is supplied as a pair consisting of a key and its associated value. Unstructured short-message ad-postings on online portals like Craigslist contain a large quantity of user-defined information in current Web content, making this a pressing concern. Writing style of these ad-postings lacks a well-defined lexicon or syntax, and the quality of the textual material is often fairly unpredictable and noisy due to factors like users using different acronyms and forms for expressing the same information.

Such classified advertising would likely be incomprehensible to a computer programmer, but not to a human. Improving machine analysis of such adverts requires a solution to the pressing problem of converting such advertisements into structured data with defined keys and values.

Our aim is to create a system that can automatically generate a structured document from a collection of adverts on a given topic, such as those found in online car classifieds.

Two distinct algorithms are provided by us for squeezing information out of raw, unlabeled data. Sporadic commercials that don't amount to anything. The advertised item is usually defined by a handful of broad qualities. Apartment features commonly advertised in rental ads include square footage, monthly rent, neighborhood, and number of bathrooms. Private spaces include bedrooms and bathrooms. Each classified ad filed under a particular subject has the same fundamental pattern, and under

any given classified ad, you'll find a detailed analysis of a feature or features presented in a haphazard method. To this end, we will describe a data structure in which "keys" are qualities (such as monthly rent) and "values" are values of $2,000 or more. The elements of the converted, structured form of an advertising are displayed.

Combinations of keys and values that are distinct an unsupervised technique is used to generate a word-affinity network, where each edge represents a different word. These weights make it so that synonyms, antonyms, word pairings, etc., that have a higher amount of mutual information, are more likely to be connected with one another. Binary keys, numeric key phrases, and descriptive keys are the three types of keys defined in the unsupervised algorithm. Binary keys are used to represent features, and their values are expressed by binary outputs on whether or not the feature is present. Keys for numeric attributes have numerical values. Output Descriptive keys are those that characterize a broad category and its associated set of values.

2.5 Datasets

Our algorithms were tested on a collection of Craigslist advertisements for both automobiles and housing. A total of 12,984 ads for automobiles and 10,784 for housing were gathered from Craigslist. These classified ads were culled from the Boston, New York, Chicago, and Los Angeles sections of Craigslist, website 1. About 80% of the ads were used to train the models, and the rest were saved for evaluation. Performance was evaluated using a manually labelled test set.

III. UNSUPERVISED TECHNIQUE

In this section, a graph-based unsupervised algorithm is proposed. The words in the advertising serve as vertices in a graph that captures the associations between those phrases; an edge between the vertices represents an affinity between those terms. This method is based on the premise that, generally speaking, keys may be broken down into three distinct types: descriptive, binary, and numeric. The structure of the affinity graph makes it simple to identify the various categories of keys. To this end, we offer a rule-based inference mechanism that can reliably identify the group of topic- specific keys in the affinity network and the values they represent.

Table 3.1: Different Keys and Their Occurrence in the ads

Label	Type	Example
Color	Numeric	Grey with black interior shiney red paint the color is black Red with black racing stripe
Price	Desc	Best Offer $5000 Cost $2952 Asking $22(K) firm value is 3(XM)S price for quick sale 3500 $
Miles	Numeric	Under 62, 000 miles approx 170k miles has 144,000 miles

Power	Binary	new power steering
steering		Power Steering. Cruise Control Power Steering - Power. Brakes

Table 3.2 Corresponding < key, value> pairs

Label (Type)	Value
Make (Desc)	BMW
Color (Desc)	black
Miles (Num)	114,000
Transmission (Desc)	automatic
Alloy wheels (Bin)	yes
Air conditioning (Bin)	yes
Leather Seals (Bin)	yes
Poor window (Bin)	yes
AM/FM radio (Bin)	yes
CD player (Bin)	yes
Price (Num)	11.488

IV. FINDING CONSISTENT LABELS AND VALUES IN AN AFFINITY GRAPH

The terms within the affinity graph determine the type of vertex it contains: descriptive, binary, or numeric. According to the hypotheses supporting this grouping,

- In a graph, a descriptive key takes the shape of a star with the central word serving as the key and the

surrounding terms as its values.

- In a graph, a descriptive key takes the shape of a star with the central word serving as the key and the surrounding terms as its values.
- All the terms in a binary key component will have comparable edge weights since they all belong to

the same group. All the bits of a binary key are combined into a single value. Binary values can be inferred from their appearance as keys.

- Under these premises, we offer an algorithm that, given an affinity graph as input, can label each vertex as

either descriptive, binary, or numeric. The proposed approach, referred to here as Algorithm 2, is a deterministic one.

The vertex degrees are used to categories the three distinct groups. The hypotheses predict that various key types will exhibit distinct graph orientations. The presence of many edges at a vertex, for instance, suggests that it may serve as a descriptive key. One of the most important clues in finding the (key, value) pairings is the connection between the words (or an edge between the words). Therefore, the graph must be refined by removing the edges that do not provide adequate evidence of semantic similarity.

The vertex degrees are used to categories the three distinct groups. The hypotheses predict that various key types will exhibit distinct graph orientations. The presence of many edges at a vertex, for instance, suggests that it may serve as a descriptive key. One of the most important clues in finding the (key,value) pairings is the connection between the words (or an edge between the words). Therefore, the graph must be refined by removing the edges that do not provide adequate evidence of semantic similarity.

An edge e(wi; wj) was pruned if MI(wi; wj) > λthreshold . By measuring the average edge weights of randomly selected word pairs, an empirical cutoff is established. The cases of strong affinity were sampled separately from those with no known relationships between the couples.

Algorithm

procedure GETKEYVALUE
Input: Affinity graph
Output: Set of keys and their values
    $Labels_{Desc} \leftarrow \{\}$; $Labels_{Bin} \leftarrow \{\}$; $Labels_{Num} \leftarrow \{\}$
    for each w_i in W **do**
        $score(w_i) = 0$
    end for
    for each w_i in W **do**
        for each x in $neighbor(w)$ **do**
            $P(x|w) = \frac{1}{deg(w)}$
            $score(x) = score(x) + P(x|w)$
        end for
        for each $e(w_i, w_j)$ in E **do**
            if $score(w_i) \gg score(w_j)$ **then**
                $Labels_{Desc}.add(w_i)$; $Value[w_i] \leftarrow w_j$
            else $score(w_i) \ll score(w_j)$
                $Labels_{Desc}.add(w_j)$;$Value[w_i] \leftarrow w_i$
            end if
            if $score(w_i) \approx score(w_j)$ **then**
                $Labels_{Bin}.add(w_i)$
                $Labels_{Bin}.add(w_i, w_j)$
            end if
            if $score(w_i) \approx score(w_{num})$ **then**
                $Labels_{Num}.add(w_1)$
            end if
        end for
    end for
    return $Label_{Desc}, Label_{Bin}, Label_{Num}$
end procedure

The difference in distributional dispersion between the two measures of dispersion. If we assume that both distributions are normally distributed, then the threshold is the mean of the values of the data points at the 95th percentile of the weak association distribution and the 5th percentile of the strong association distribution. For a normal distribution with mean and standard deviation, we have the quantile function for a given probability, If q norm(p; μ; σ)

$$\lambda_{threshold} = \frac{qnorm(0.05,\mu_s,\sigma_s)+qnorm(0.95,\mu_w,\sigma_w)}{2}$$

V. PERFORMANCE

In order to learn a set of keys (descriptive, binary, and numeric) for the two separate themes, the graph-based technique was applied to a corpus of 10,000 Craigslist advertising on vehicles and 8,000 ads on apartment rentals. In order to gauge the effectiveness of the training keys, they were applied to a database consisting of 2,984 automobile advertising and 2,784 apartment rental postings. We created a "golden set" by manually extracting key-value pairs from test data. A human annotator reviewed the training and testing data and assigned each label descriptive, binary, and numeric to generate the golden set. Extracting (key, value) pairs from the testing set using the affinity network built during training allowed us to assess the unsupervised method. Precision-recall values were used to evaluate the effectiveness of the extracted key-value pairs in comparison to the manually produced golden set.

The F-value computed from the precision and recall values for the car and apartment ad sets are shown in Figure 5.1 and 5.2 respectively.

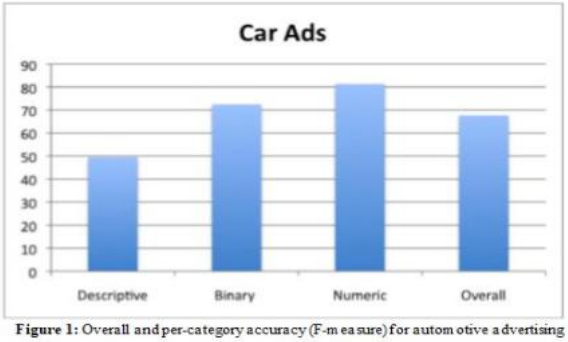


Figure 1: Overall and per-category accuracy (F-measure) for automotive advertising

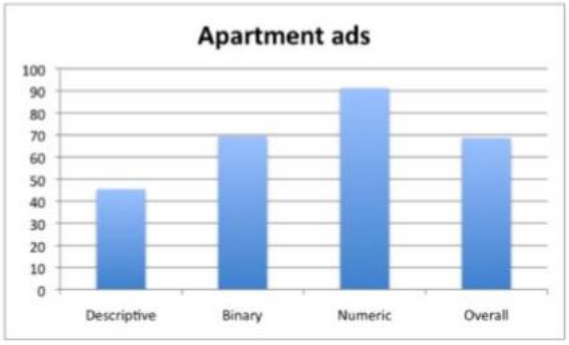


Figure 2 Total accuracy (F-measure) for apartments ads and under each category

VI. RESEARCH AND FINDINGS

600 adverts from a training set were used to train the model. To assess the model, 200 more ads were annotated in the same manner. Due to the significant cost of creating such sets, the training and testing set was tiny in size. Based on the accuracy calculated on the test set of 200 ads, the final findings are presented. The X axis in Figure 5.4 displays the performance for various tests. In ten separate studies, we tested different feature combinations, including varied word window sizes. Based on several word surface traits, we examined the mistake cases and added a few more features. Most of these features were binary. Examples of such attributes include whether a word contains a numeral, a symbol, etc. The performance was enhanced by nearly 0.6 by adding the digit functionality. The symbol feature, however, did not improve the accuracy. On the basis of this finding, we developed features involving symbols like "$" and "-." The addition of these features worked well, increasing accuracy to 73.66%.

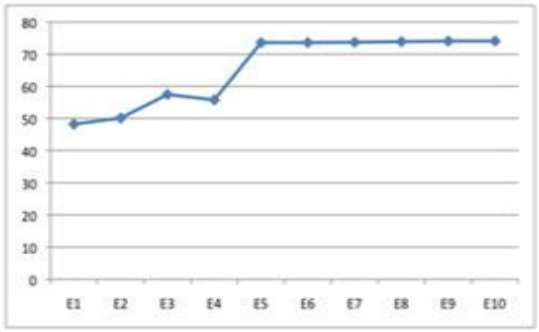


Figure 3 Accuracy of each experiment. There were 10 experiments performed on car ads with combinations of different word-window size and feature sets. Y-axis shows the accuracy in percentage.

6.1 Newscast Trainings for Investigative and Forecasting Future Proceedings

Our objective is to create a system that can automatically extract data about real-world occurrences from news sources and then use that data to foresee changes in broad economic indices. In order to infer and perhaps anticipate changes in selected macroeconomic indices, we aim to create a system that can automatically form event-driven predictive models. Assuming we have access to a large corpus of real-world events that can be extracted from news sources and a structured time series about a specific macro-economic index of interest, we ask if it is possible to learn a predictive model for the macro-economic index by connecting it with the appropriate events that relate closely to it.

Estimating various macroeconomic indexes calls for information from a wide range of sources. However, in research seeking to estimate such variables [62], analysis and forecasting are often performed using structured data sources, with only a select few of such aspects included (and chosen manually).

The indices' extreme swings could be due to unknown reasons or a convergence of known and unknown ones that increase volatility. The sensitivity of these indicators can be better understood by keeping tabs on global happenings. We owe a great deal of our understanding of these events to unstructured text streams, such as those seen in the news, online, and on social media. There have been a plethora of studies that have used information gleaned from the media to predict future events.

6.2 Models of Events

The news provides information about events from all over the world. Therefore, news stories can be used as a database to extract information about the real world. Numerous works have drawn inspiration from events reported in the news. These works provide a variety of interpretations and depictions of the same event.

Case in point: Radinsky et al. [73] based their models on news stories.

- Cyclone in Rwanda, floods in Africa, etc. A different approach involved representing events as a whole, which included objects (or actors), actions, and time. Automatic Content Extraction (ACE) defines 8 main types of events and 33 sub-types of events for usage in event extraction by considering entities, event triggers, timings, event mention argument, etc. In order to isolate the relevant triggers, it is often required to zero in on a certain type of domain-specific event. A few examples of the most common occurrences in the financial world are hiring, firing, resignation, acquisition, loss, growth or decrease in profits, etc.

In this section, we present a novel probabilistic model to characterize events that can be extracted from a corpus of news articles. News is typically assumed to be a brief report of a singular occurrence because most news stories follow a consistent format. On the basis of this assumption, we make an attempt to simulate the incident in question using data from the news account. We also assume that in any given time period, only a small subset of possible event types will actually occur. When it comes to events, news stories fall into a specific category. "Event classes" are a useful tool for organizing a wide variety of events. For example, an accident can be seen of as both a category of events and a single instance of that category, depending on the exact location, nature, and other factors involved (such as actors, objects, etc.). The primary action, or trigger, of an event is the most crucial aspect of the event class. In the preceding instance, the word "accident" functions as the trigger; however, other words or phrases, such as "crash," "collision," "rammed," and so on, may also act as triggers for the same class.

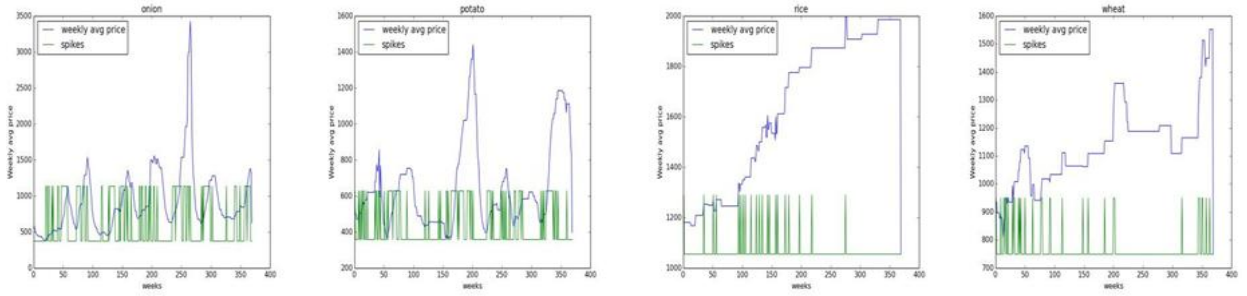
6.3 Prediction Correctness

Our dataset includes 7 years (2557 days). Using the event model, we retrieved daily events over 7 years to anticipate food price spikes and stock price fluctuations.

The model's predictive accuracy is evaluated. Accuracy is the ratio of correct predictions to data points. We compare our event model's performance across baselines and text mining approaches (Section 6.4).

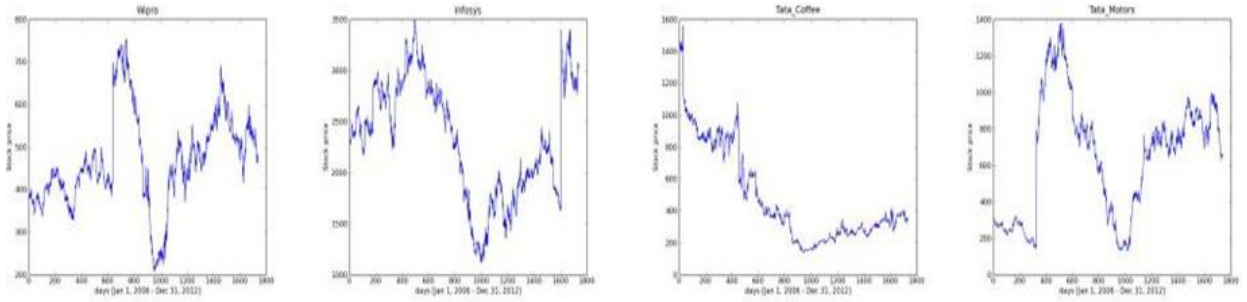
Our event model uses a pipeline method, with event triggers extracted first and subsidiary events added subsequently. First, the benefit of adding subsidiary events to the event model was tested. So, performance was compared to a naive event trigger-based model (TRIGG). We compared performance with LDA-

based predictor (section 6.4.4). We examined adding subsidiary events vs. subjects in our model. We combined TRIGG and LDA to observe the secondary events' benefits. Finally, we assess all these models with another model that uses past and current events to predict indicator changes. Model HIST (6.4.3). Tables 6.2, 6.3, and 6.4 summaries the food price and stock price tests.



(a) Onion price (b) Potato price (c) Rice price (d) Wheat price

Figure 4: 2012-2018 weekly average agricultural prices. The spikes represent a 10% price increase. The horizontal axis shows the weeks between January 2012 and December 2016, from 0 to 364.



(a) Wipro (b) Infosys (c) Tata Coffee (d) Tata Motors

Figure 5: Stock price variation of 4 Indian concerns

wheat, potato (events: attack on railways, railway strike etc.). Wheat and potatoes aren't grown equally across the country and rely largely on transportation for distribution. Any occurrence that affects agricultural movement affects supply, raising prices. Festivals only impact onion and wheat. Festivals enhance demand for these two commodities, driving up prices.

Table 1 Performance for Food Prices

	Potato	Wheat	Rice	Onion	All
TRIGG	0.509	0.512	0.514	0.520	**0.535**
LDA	0.431	0.432	0.471	0.488	**0.492**
TRIGG+LDA	0.539	0.544	0.548	0.553	**0.592**
TRIGG+SUBS-TRIGG	0.567	0.569	0.579	0.592	**0.599**
TRIGG+SUBS-TRIGG+HIST	**0.603**	**0.613**	**0.615**	**0.618**	**0.621**

Table 2 Performance for Stock Prices

	Wipro	Infosys	Tata Motors	Tata Coffee
TRIGG	0.636	0.641	0.541	0.604
LDA	0.433	0.439	0.473	0.413
TRIGG+LDA	0.632	0.573	0.543	0.597
TRIGG+SUBS-TRIGG	**0.679**	**0.662**	**0.593**	**0.611**
TRIGG+SUBS-TRIGG+HIST	0.622	0.553	0.509	0.551

VII. CONCLUSION

To this purpose, we looked into the viability of using news articles as a source for event extraction and the creation of knowledge graphs to characterize event dependencies for different kinds of analytics. Two distinct event model types and five knowledge graph building approaches were provided in the thesis.

These knowledge graphs can be used in a number of automated and human-driven methods to examine historical news archives. An important use of both event models and knowledge graphs is the modelling of the behavior of external variables and indicators. We show how knowledge graphs may be utilized to build predictive models that take into account the relationships between news events.

This paper primarily utilizes unstructured online data to extract events, knowledge graphs, and predictive models for socio-economic indicators. We plan to expand this work by exploring additional news data sources and reporting methodologies to strengthen our findings. It is important to consider potential biases in news organizations, as some evidence may have been overlooked during this research.

REFERENCES

[1] Economic review and statistical appendix, department of statistics and programme implementation, government of west bengal, 2000-2012.

[2] E. Acar, S. A. Camtepe, M. S. Krishnamoorthy, and B. Yener. Modeling and multiway analysis of chatroom tensors. In Intelligence and Security Informatics, pages 256–268. Springer, 2005.

[3] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. Slicsuperpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell., 34(11):2274–2282, Nov. 2012.

[4] J. Allan, R. Gupta, and V. Khandelwal. Temporal summaries of new topics. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '01, pages 10–18, New York, NY, USA, 2001. ACM

[5] A. Balasubramanian, N. Balasubramanian, S. J. Huston, D. Metzler, and D. J. Wetherall. Findall: A local search engine for mobile phones. In Proceedings of the 8th international conference on Emerging networking experiments and technologies, pages 277–288. ACM, 2012.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. J. of machine Learning research, 3:993–1022, 2003

[7] J. Chen and L. Subramanian. Interactive web caching for slow or intermittent networks. In Proceedings of the 4th Annual Symposium on Computing for Development, page 5. ACM, 2013

[8] K. B. Cohen, K. Verspoor, H. L. Johnson, C. Roeder, P. V. Ogren, W. A. Baumgartner, Jr., E. White,H. Tipney, and L. Hunter. High-precision biological event extraction with a concept recognizer BioNLP '09, pages 50–58, 2009.

[9] L. Fang, A. D. Sarma, C. Yu, and P. Bohannon. Rex: Explaining relationships between entity pairs. Proc. VLDB Endow., 5(3):241–252, Nov. 2011.

[10] Y. Fang, L. Si, N. Somasundaram, and Z. Yu. Mining contrastive opinions on political texts using cross-perspective topic model. WSDM '12, pages 63–72