What's up, Wikipedia? What is popular on Wikipedia and why.

Byung Gyu Ahn, Chris Callison-Burch, Benjamin Van Durme
Center for Language and Speech Processing
Johns Hopkins University
Baltimore, Maryland
{bahn, ccb, vandurme}@cs.jhu.edu

Abstract

The popularity that social networks such as Twitter and Facebook have recently gained calls more research into the use of metadata that they generate. In this paper, we make use of the page view counts of Wikipedia articles to predict topics gaining in popularity and explain why these topics are gaining users' attention. Our proposed approach clusters related topics to distinguish the main topics of on-going events from secondary related topics and explains the relationship between them. We explain why this is an interesting problem and compare approaches using the link structure of Wikipedia articles and the bag of words model of them to process each step of the process. By combining each step that performs reasonably as well as human annotators, the methods put forth here make a nice summary of what people in the real world are really insterested in.

For sets of articles that show a sharp increase in page view on the same date, we do the following: (1) cluster articles into subsets that correspond to the same current event, (2) select sentences from the articles that describe why the event is popular. We evaluate against manually clustered articles and hand selected sentences. We show the value of bag-of-words topic models, link structure, named entity and time expression recognition. The result shows promise for explaining what's currently popular on Wikipedia and for creating a timeline of past newsworthy events.

1 Introduction

Recently, social media such as Twitter and Facebook gained attention from researchers. By looking at the topics recently prevailing in a lot of Twitter data at the same time, one may know the recent popular topics (Petrović et al., 2010).

In this paper we analyze a novel data set: we have collected the hourly page view statistics for every Wikipedia page in every language for a three year period. We show how these page view statistics–along with a whole host of other features like inter-page links, edit histories, mentions in contemporaneous news stories–can be used to identify and explain popular trends, including political elections, natural disasters, sports championships, popular films and music, and other current events.

Our approach is to select a set of articles whose daily page views increase above their average from the previous two week period. Rather than simply selecting the most popular articles for a given day, this selects articles whose popularity is rapidly increasing. These popularity spikes are presumably due to some external current event in the real world. On any given day, there are many articles whose popularity is spiking.

In this paper we attempt to cluster 100 such articles from each of 5 randomly selected days in 2009, such that the clusters coherently correspond to current events.

We compared our automatically collected clusters to the Wikipedia current events. Wikipedia editors compiles current events every day, which mainly consist of social and political events, traf-

Barack Obama
Joe Biden
White House
Inauguration
...
US Airways Flight 1549
Chesley Sullenberger
Hudson River
...
Super Bowl
Arizona Cardinals

Figure 1: The automatically selected articles for January 27th, 2009. The underscores are used in place of spaces by Wikipedia.

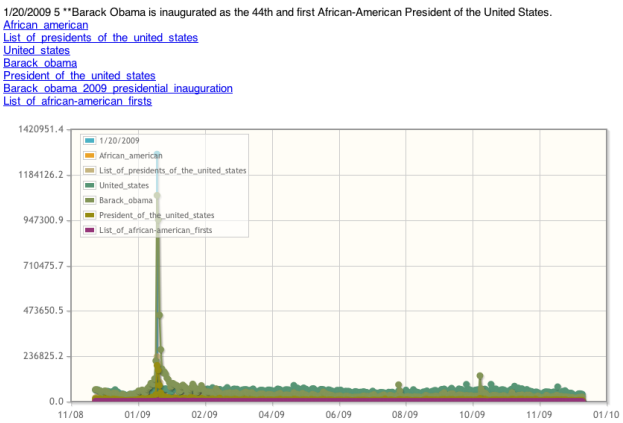

Figure 2: Page views for the articles related to the inauguration of Barack Obama. The articles are linked from an item in the Wikipedia current events. Interestingly, the list does not include the article `Inauguration of Barack Obama`, the very page about the event that has a spiking page views.

fic accidents and disasters. More often than not, they do not generate much traffic, and link to pages that are too general like "United States" or "Israel". We view this work as an automatic mechanism that could potentially supplant the hand-curated method of selecting current events that is currently done by Wikipedia editors.

For instance, we would attempt to cluster the articles in Figure 1 into 3 clusters, { `Barack Obama`, `Joe Biden`, `White House`, `Inauguration` } which corresponds to the inauguration of Barack Obama, { `US Airways Flight 1549`, `Chesley Sullenburger`, `Hudson River` } which corresponds to the successful ditching of an airplane into the Hudson river without loss of life, and { `Superbowl`, `Arizona Cardinals` } which describes the then upcoming Superbowl XLIII.

We further try to explain the clusters by selecting sentences from the revision of the Wikipedia articles on that date. For the first cluster, a good selection might be "the inauguration of Barack Obama as the 44th president of the United States took place on Jan 20, 2009". For the second cluster, "Chesley Burnett "Sully" Sullenberger III (born January 23, 1951) is an American commercial airline pilot, . . . , who successfully carried out the emergency water landing of US Airways Flight 1549 on the Hudson River, offshore from Manhattan, New York City, on January 15, 2009, ...". For the third cluster, "[Superbowl XLIII] will feature the American Football Conference champion Pittsburgh Steelers (14-4) and the National Football Conference champion Arizona Cardinals (12-7) .", which makes clear the relationship with `Arizona Cardinals`.

To generate the clusters we can make use of the text of the articles on that date, versions of the articles from previous dates, the link structure and category info from Wikipedia, and potentially external info like newspaper articles published before the date.

To select sentences we may want to make use of NLP technologies such as coref resolution, named entity and date taggers, and dependency parsers to identify subjects of sentences.

2 Motivation

What are interesting topics? In an online encyclopedia such as Wikipedia, the page view counts for each article reflects the popularity of the article. Each article has its own level of normal popularity: some has high page views, and others low. Sometimes, a newsworthy event such as a major political or sports events, or natural disasters, or pandemic, occurs and incur that many articles related to the specific event has a major increase in page views.

The wikipedia has a section called current events, in which the recently occurred events are listed manually by Wikipedia editors. Figure 2 shows the page views of the articles related to the inauguration of Barack Obama. Each event may have a hierarchi-

cal structure–there may be a major event that related to minor events. Each event is described in a line of text with a possibly multiple links to the related Wikipedia articles. The figure shows the spikes in page views of the related articles around the date on which the event took place–January 20th, 2009.

We have selected the articles that has a major increase in page views. The increase in page views are measured as follows. For each day, the total sum of the page views for the past 15 days and the total sum of the page views for the previous 15 days are summed, and the difference Δ between the two is calculated for each article. The articles are sorted in order of decreasing Δ, and the top 100 articles are selected. The difference Δ is formularized as below:

$$\Delta = S_{15} - S_{30}, \text{ where } S_{15} = \sum_{i=1}^{15} v_i \text{ and}$$
$$S_{30} = \sum_{i=16}^{30} v_i$$

where v_i is the page views of the article on the past ith day from the date on which the articles are selected.

We compared the automatically selected articles to the articles linked from the Wikpedia current events in some aspects. First of all, note that the hand-curated articles are less than half of the automatically selected articles: There are 17,253 hand-selected articles and 36,400[1] auto-selected articles. Only 28% of those hand-selected articles are automatically selected. When analyzed, it was found that many of the hand-selected articles have very low page views: 6,294 (36.5%) have maximum daily page views less than 1000 in 2009. Naturally, they are not chosen by automatic selection based on page views[2].

Figure 3 shows the comparison of the selected articles. Automatically selected articles include an newly created article about a politcal event (`Inauguration of Barack Obama`), a recently released film, a popular TV series and related articles and tend to be specific than hand-selected articles. The hand-selected articles include more generic articles related a specific event, most of which are personal, organizational or geopolitical names. The hand-generated event describes the relationships between related articles.

Automatically selected articles
`Inauguration of Barack Obama`
`Joe Biden`
`Notorious (2009 film)`
`The Notorious B.I.G.`
`Lost (TV series)`
Hand-selected articles
`Fraud`
`Florida`
`Hedge fund`
`Arthur Nadel`
`Federal Bureau of Investigation`

Figure 3: The example articles for January 27th. These are the articles that do not have a counterpart with a window size of 15 days. The hand-selected articles are linked from an event "Florida hedge fund manager Arthur Nadel is arrested by the United States Federal Bureau of Investigation and charged with fraud."

At this point, we set the goal of our novel task: to summarize recent events related to popular articles just as the hand-curated Wikipedia current events, but with more popular articles with high page views. This work can be used to replace the hand-generated current events listing the events that many people are interested in.

Our system pipeline is explained as follows. First, the most popular pages, or topics, are collected per each day(§2). Second, clusters are identified by clustering related topics and the main article of each cluster is identified(§3). Lastly, the sentence that best describes the cluster are extracted for each cluster(§4). See Figure 4.

3 Clustering

We cluster the automatically selected articles. In Figure 1, the articles may be divided into three articles. Among the articles, some seem to be more centric articles; For example, among the articles in Figure 1, `Inauguration`, `US Airways Flight 1549`, and `Super Bowl` seem to be centric articles

[1]The year 2009 has 365 days and one day is missing from our daily statistics.

[2]The automatically selected articles has an increase in page views of at least 10,000.

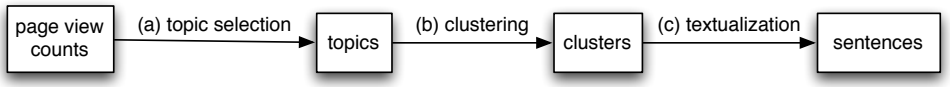

Figure 4: Process diagram. (a) Topic selection. (b) Clustering. (c) Textualization.

that have close relationship with the rest of the articles. The other articles seem to be more local.

We make use of various information about the articles.

These topics are naturally subject to clustering. As a baseline, we ran K-means clustering on the articles extracted by the TrendingTopics algorithm, using the articles as bags of words, which were fetched on the date on which they emerged as the TT topics.

Three annotators performed manual clustering on the topics for the five specified dates to get the gold standard clusters. The three clusters were evaluated against each other. The metric used for evaluation was B-cubed metric. This will be the oracle score for any clustering algorithm.

The modified B-cubed metric (Amigó et al., 2009) is used for evaluation of overlapped clusters. The B-cubed metric is one of the extrinsic clustering evaluation metrics, which need gold standard dataset to evaluate the clustering of interest against. The original B-cubed metric is for clustered

4 Textualization

Textual description of the topics are intended to explain why the topics are popular. Specifically, the textual description for each cluster may consist of the sentences (1) that describe why the cluster is popular at the time and (2) that describe why each topic in the cluster is popular. The concepts of central and peripheral topics (§3) are important in that different types of topics contribute to the description in different ways. Often, the central topics are directly related to the event that caused the recent popularity and often includes a direct explanation of the recent event.

Preprocess We preprocess the Wikipedia articles through the Serif system[3] for the temporal expression identification and the coreference resolution.

[3]References?

Gold standard	Evaluated Set	B-Cubed F-score
Manual-1	Manual-2	0.67 ± 0.076
Manual-1	Manual-3	0.74 ± 0.085
Manual-2	Manual-3	0.75 ± 0.129
Manual-1	K-means	0.53 ± 0.034
Manual-2	K-means	0.53 ± 0.050
Manual-3	K-means	0.49 ± 0.035

Table 1: Clustering evaluation. For B-Cubed metric, exchanging the gold standard and the evaluated dataset incurs the exchange of the precision and the recall score, thus leaving the F-score same.

Design We simplified this task as selecting one sentence from the corresponding wikipedia article. Our baseline

Evaluation Three annotators picked the best sentence and next best sentence that describes why the topic recently got popular for ten chosen topics for the five dates. Qualitative examination was done.

5 Related work

There should be lots of citation to related work here. Examples are NewsBlaster by McKeown and Barzilay; Snippet Selection by Lapata, and a list of ACL/EMNLP papers that study wikipedia.

6 Future work

There are still a lot of challanges that remain. For textualization, sentence fusion is an option for summarizing the relevant topics as one event.

Acknowledgments

The first author was supported by Samsung Scholarship.

Thank you to Wikipedians X and Y for making the page view statistics available.

References

Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inf. Retr.*, 12(4):461–486.

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2010. Streaming first story dectection with application to twitter. In *NAACL-2010*.