WikiTopics: 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

We establish a new task and pipeline to find interesting topics from Wikipedia and summarize them and introduce a novel data set: with the hourly page view statistics of all Wikipedia articles for three years and evaluation data for the results. Our pipeline consists of three steps: to find the best articles, cluster them, and extract the best sentences. Our K-means clustering and clustering using the link structures performs as well as 75% human annotators do. Our sentence selection make use of the link structure, named entity and time expression recognition and works as well as 75% humans do. The result shows promise for explaining what's currently popular on Wikipedia and for automatically creating a timeline of past newsworthy events.

1 Introduction

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, traffic 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

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.

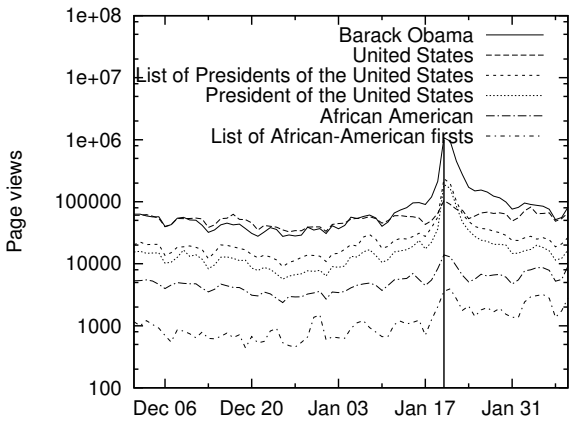

Figure 2: Page views for the articles related to the inauguration of Barack Obama. The articles are linked from an item in the Wikipedia current events. Interestingly, the list does not include the article Inauguration of Barack Obama, the very page about the event that has a spiking page views.

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 hierarchical 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.

Should we try to predict the current event descriptions that Wikipedia editors hand-curate? We say no for the following reasons. Therefore we recommend against this methodology for other researchers.

We set up a website[3] that you can see the sparkline graphs of pageviews for each day, each link, or each event in the form as Figure 2. You can see the clear correlation between the spikes of the page views of the articles and the date on which the articles appear as the current events.

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.

WikiTopics
Inauguration of Barack Obama
Joe Biden
Notorious (2009 film)
The Notorious B.I.G.
Lost (TV series)
Wikipedia current events
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."

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 Article selection

Dataset The Wikipedia Traffic Statistics dataset is originally publicized and made available at http://dammit.lt/wikistats/ by a Wikipedian Domas Mituzas. This data is only kept up to a several months that the space allows. For the previous statistics, two sets of the statistics are published at Amazon Public Datasets (http://aws.amazon.com/datasets/, http://aws.amazon.com/datasets/2596, http://aws.amazon.com/datasets/4182). This dataset consists of the files that each has hourly page view statistics for every article in every language. Each line of the files contains the language or the project name, the title, the hourly page views, and the numbers of bytes of the text of an Wikipedia article.

We limited the work only to the English

[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.

[3]See http://ANONYMIZED

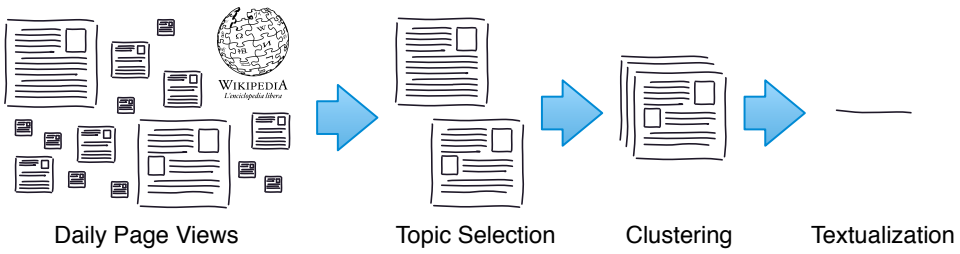

Figure 4: Process diagram. (a) Topic selection. (b) Clustering. (c) Textualization.

Wikipedia. These statistics are collected from the Wikipedia cache server as requested by users, and it includes many wrongful or malicious requests. Many requested pages are also redirect pages that automatically refer the requester into another page. The redirect pages are usually the ones that are different names of an entity. To process these difficulties, we downloaded the English Wikipedia dump on June 22nd, 2010 from Wikimedia dump (http://download.wikipedia.org/) and from the database dump extracted the list of the titles of all articles and the redirect articles. Using these data, we filtered out the request for non-existing articles and merged the page views for the redirect pages into the main articles. Also the title of the Wikipedia articles has to be normalized according to a specific format, that is, the first letter of each title are capitalized and a space in it has to be replaced with an underscore, and so on.

4 Clustering

More often than not, some of the popular articles are about a current event. For example, the hand-selected articles shown in Figure 3 are about a current event that "Florida hedge fund manager Arthur Nadel is arrested . . . and charged with Fraud." Among the automatically selected articles, main events such as Inauguration of Barack Obama and release of the file Notorious (2009 film) involves making popular the incidental articles about the players of the events such as Joe Biden and The Notorious B.I.G. along with the articles about the main events themselves.

We attempt to cluster the automatically selected articles into mutually related articles and find the article that describes the main event for each cluster. For clustering, we make use of the unigram bag-of-words model and the link structures of articles. To find the centroid articles that describes the main event, we used K-means model and the link structure of articles.

Dataset For each day of the five selected dates in 2009, we downloaded the text of the 100 automatically selected articles from Wikipedia. The downloaded texts are the latest texts as of the date on which the article is selected. We use the Wikipydia module, which is a python module to make use of the Wikipedia API. As preprocessing, we stripped out all HTML tags from the article text, and replaced the Wikipedia-specific tags as the corresponding text using the mwlib library, and finally split sentences using the NLTK splitter.

Design As a baseline, K-means clustering was performed on the set of articles, treating the article texts as bag-of-words. There are 100 automatically selected articles on each of the five selected dates and we set 50 as the number of resulting clusters. We used the Mallet[4] software to run K-means software. Normalization and tokenization are not performed before running K-means. The algorithm calculates the mean of each cluster in word-vector space, and we chose the centroid article that is closest to the center in the vector space.

We also used the link structures of the Wikipedia articles. The link structures are downloaded from the website of Henry Haselgrove (http://users.on.net/ henry/home/wikipedia.htm).

Evaluation Three annotators performed manual clustering on the topics for the five specified dates to get the gold standard clusters. The three manual clusters were evaluated against each other to mea-

[4]Reference?

sure the annotator agreement, using the multiplicity B-cubed metric(Amigó et al., 2009) that can handle overlapping clusters.

The B-cubed metric is one of the extrinsic clustering evaluation metrics, which need a gold standard set of clusters to evaluate the set of clusters of interest against. Each item e has potentially multiple gold standard categories, and also potentially multiple clusters. Let $C(e)$ is the set of the clusters that e belongs to, and $L(e)$ is the set of e's categories. The multiplicity B-cubed scores for a pair of e and e' are evaluated as follows:

$$\text{Prec}(e, e') = \frac{\min\left(|C(e) \cap C(e')|, |L(e) \cap L(e')|\right)}{|C(e) \cap C(e')|}$$

$$\text{Recall}(e, e') = \frac{\min\left(|C(e) \cap C(e')|, |L(e) \cap L(e')|\right)}{|L(e) \cap L(e')|}$$

The overall B-cubed scores are evaluated as follows:

$$\text{Prec} = \text{Average}_{e \neq e'}\text{Prec}(e, e')$$

$$\text{Recall} = \text{Average}_{e \neq e'}\text{Recall}(e, e')$$

Analysis The bag of words model we used does not distinguish different meanings of words. This could result in wrong clustering. For example, the automatically selected articles include both Piracy in Somalia and The Pirate Bay as well as Piracy. Their resemblance in word spelling might result in confusion in clustering. In the real results, The Pirate Bay was correctly clustered with The Pirate Bay Trial, but Piracy was wrongly clustered with USS Bainbridge (DDG-96) and MV Maersk Alabama, both of which are the names of vessels. Instead of Piracy, Moldova wrongfully ended up in the same cluster as Somalia and Piracy in Somalia.

Trivial as it may sound, but it is not. In automatically selected articles for February 10, 2009, Journey (band) and Bruce Springsteen may seem to be relevant to Grammy Awards, but in fact they are relevant on this day because of the Super Bowl. The K-means clusters wrongfully clustered the articles relevant to Grammy Awards and Super Bowl altogether into one large cluster.

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 the 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.

5 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 (§4) 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[5] for the temporal expression identification and the coreference resolution. The identified temporal expressions are in various formats such as exact date ("February 12, 1809"), a season ("spring"), a month ("December 1808"), a date without a specific year ("November 19"), and even ralative time ("now", "later that year", "The following year"). Some examples are showed in Figure 5. The coreferences are analyzed into a list of the entities in the article and all the mentions of each entity in the article are compiled as co-ref chains.

Design As a baseline, we picked the first sentence for each article because the first sentence generally explains the article. As a second test set, we picked the sentence with a temporal expression that is most closest on which the article was selected. Among various date and time forms, the following policy was adapted to define the closeness: The most spe-

[5]References?

February 12, 1809	Later that year
1860	about 18 months of schooling
now	November 19
the 17th century	that same month
some time	The following winter
December 1808	The following year
34 years old	April 1865
spring	late 1863
September	later that year

Figure 5: The temporal expressions identified by the SERIF system in the preprocess step. They are examples selected from 247 such date and time expressions extracted from the article about Abraham Lincoln

cific date with year, month, and day ("February 12, 1809") has the most precedence over the other ones, and the next is a month with a year ("December 1808"), then a season with a year, and the date without a year.

After selecting a sentence for each cluster, we used coreference resolution to replace the important pronoun in the sentence with its proper name. This step is important and necessary because the selected sentence often refers to its subject by a pronoun such as "He", "She", or "It".

- Discussion of which proper name – most frequent one? most recent one?

Evaluation For ten articles on a specific day, an annotator selected sentences that best describes why the article is popular from all the sentences in each article. Each article has about 289 sentences on average. The annotator picked the best single sentence, and the second best sentences that could potentially be multiple. In the case there are no best sentence among them, he marked none as the best sentence, and listed all the partially explaning sentence as second best sentences.

The results will be added soon.

Analysis We quantitatively analyzed the output of the system, and it suggested a couple of current problems and future work.

Serena Williams is an example that the error in sentence splitting propagates to the sentential selection. The best sentence manually selected was the first sentence in the article "Serena Jameka Williams ..., as of February 2, 2009, is ranked World No. 1 by the Women's Tennis Association ...." The sentence was disastrously divided into two sentences right after "No." by the NLTK splitter through our preprocess. It means no matter how well the sentential selection is done, it cannot choose the gold standard sentence. We ran the splitta (Gillick, 2009) over the article text and found that it does not split the first sentence at the wrong position. The better the splitting is, The better the sentential selection works.

Selection of the best sentence with dates seems to work well, with some problems. Farrah Fawcett is a nice example of multiple sentences with dates, in a single section, that could potentially be spliced together into a timeline (the final event, that she was released from the hospital, makes more sense if we included why she was there). Furthermore, the sentence describing the most recent event contains a date without the year, which has less precedence over the other dates with the year even when it is closer to the date of interest than the others are. So having precedence over the date forms might not always work well.

An improved baseline for sentence selection would include the opening sentence of the page from which the date-stamped sentence comes from. For example, in "pick0419" :

Grey Gardens # "It is scheduled to air on HBO on April 18, 2009."

as compared to:

Grey Gardens # Grey Gardens is a 1975 documentary film by the direction/cinematography/editing team of Albert and David Maysles, Susan Froemke, Ellen Hovde, and Muffie Meyer. ... It is scheduled to air on HBO on April 18, 2009.

From there we'd want to compress the opening sentence, as in this case:

Grey Gardens # Grey Gardens is a 1975 documentary film. ... It is scheduled to air on HBO on April 18, 2009.

Which then brings up the potential flaw: it is an *adaptation* of this film that is being released on April 18th, not the original.

Looking deeper, we see the earlier sentence: "Grey Gardens, a film for HBO, starring Jessica Lange and Drew Barrymore as the Edies, with Jeanne Tripplehorn as Jacqueline Kennedy, and Daniel Baldwin as Julius Krug.". This is hard: even

if we run a co-ref system over the page to resolve the "It" in "It is scheduled to air", then we'll know that "Grey Gardens" is scheduled to to air. But it isn't the same Grey Gardens, it is the new adaptation, which we know only from the section header. Or we would have to have background knowledge that could enable reasoning about movies not having two different lists of starring actors, etc. (even harder!)

6 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.

7 Future work

Each step of the WikiTopics system can be developed further.

First, article selection. Should we predict what the Wikipedia current events will be? No. But if we can predict which will come next before any newspaper or traditional media, it would be awesome. But should we be? The point is in automaticity anyway. We currently use the absolute difference in page views with the window of size 15-days. Using relative difference might be useful.

Second, clustering. Hierarchical clustering... There a lot of clusters that have hierarchical structures. For example, on the day of the inauguration of Barack Obama, the automatically selected articles include the former presidents of the United States, family of Barack Obama, and the appointed staff in the new goverment. Using hierarchical clustering, the summary of each cluster will be better.

Third, sentence selection. Sentence fusion could be useful. Most of the Wikipedia current events contain more than two links to articles. We currently select one sentence per each article, but it might be true sometimes that fewer sentences can well explain the events. We can use hierarchical clustering to summarize the clusters with fewer sentences.

Acknowledgments

We appreciate Wikipedians Domas Mituzas and Frédéric Schuütz for making the page views statistics avaialble and Peter Skomoroch for prividing Trending Topics and answering questions.

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.

D. Gillick. 2009. Sentence Boundary Detection and the Problem with the U.S. In *Proceedings of NAACL: Short Papers*.

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