WikiTopics: What is popular on Wikipedia and why.

Abstract

We establish a novel task in the spirit of news summarization and topic detection and tracking (TDT): daily determination of the topics newly popular with Wikipedia readers. Central to this effort is a new public dataset consisting of the hourly page view statistics of all Wikipedia articles over the last three years. We give baseline results for the tasks of: discovering individual pages of interest, clustering these pages into coherent topics, and extracting the most relevant summarizing sentence for the reader. When compared to human judgements, our system shows the viability of this task, and opens the door to a range of exciting future work.

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 –can be used to identify and explain popular trends, including political elections, natural disasters, sports championships, popular films and music, etc.

Our approach is to select a set of articles whose daily page views increase above their average from the previous fifteen days. 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 events in the real world. On any given day, there are many articles whose popularity is spiking: while some of the articles are related to each other, many of them are a coincidence.

In this paper we examine 100 such articles from each of 5 randomly selected days in 2009 and attempt to group the articles into clusters such that the clusters coherently correspond to current events. Quantitative and qualitative analyses are provided along with the evaluation data set.

We compared our automatically collected articles to those in the current events portal of Wikipedia where Wikipedia editors compile current events every day, which mainly consist of armed conflicts, protests, attacks, international relations, law and crime, natural disasters, social, political, and sports events. 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.

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

Figure 1: Automatically selected articles for Jan 27, 2009.

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.

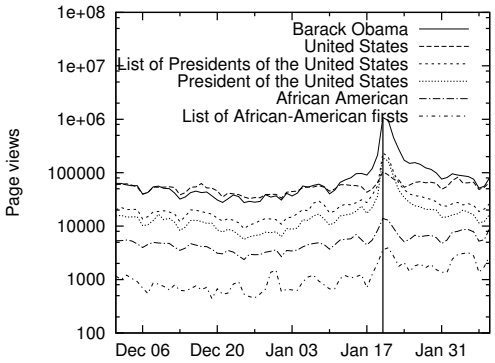

Figure 2: Page views for the articles related to the inauguration of Barack Obama. The articles are linked from an item in the current events portal in Wikipedia. Interestingly, the list does not include the article Inauguration of Barack Obama, the very page about the event that has a spiking page views.

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 information from Wikipedia.

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

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 a different level of popularity: some have high page views, and others low page views on average. This tendency maintains throughout the year, but sometimes an external newsworthy event such as an election, sports event, a natural disaster or a pandemic, occurs and incur that many articles related to that specific event has a significant increase in page views.

Wikipedia has a section called "current events", in which the recent significant events are listed manually by Wikipedia editors. Figure 2 shows the page views of the articles related to the inauguration of Barack Obama, that were manually listed in the Wikipedia current events section. Each event may have a hierarchical structure–there may be multiple minor events related to a given major event. 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 set up a website[1] that you can see the sparkline graphs of pageviews for each day, each link, or each event in the form as Figure 2. In some of the events, you can see clear correlation between the spikes of the page views of the articles and the date on which the articles appear as the current events.

Following Trending Topics[2], we automatically select 100 articles for each day in 2009. The articles are selected based on the changes in page views for the previous 30 days, to detect a spike in page views. We refer to these articles are refered to as the WikiTopics articles.

We compared the automatically selected articles to the articles linked from the Wikpedia current events. When evaluated against the articles linked from the Wikipedia current events, the WikiTopics articles perform badly with precision of 0.13 and the recall of 0.28. There are two main reasons for this. First, 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[3] WikiTopics articles. Second, many of the hand-selected articles turned out having very low page views: 6,294 articles (36.5%) have maximum daily page views less than 1000 in 2009. Naturally, they are not chosen by automatic selection based on page views[4].

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 events descriptions that Wikipedia editors hand-curate? We say no for the following reasons. They are not interesting topics: many of them have too low page views, which does not draw people's attention. Also, the hand-curated articles are too generic and biased to geopoltical names such as the names of countries. Therefore we recommend against this methodology for other researchers.

we establish a more concrete goal of our novel task: to detect recent events from popular Wikipedia articles, summarize the events, and provide the links to the relevant Wikipedia articles just as the hand-curated Wikipedia current events do, except that the events are popular topics that have a significant increase in page views. This work can be used to replace the hand-curated Wikipedia current events, listing the events that in reality many people are interested in.

Our system pipeline and this paper are organized as follows. First, the most popular articles are collected per each day. Second, correlated articles are clustered into the clusters that correspond to intersting events or topics. Lastly, the sentence that best describes the interesting events are extracted. See the process diagram in Figure 4.

[1] http://ANONYMIZED

[2] http://www.trendingtopics.org

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

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

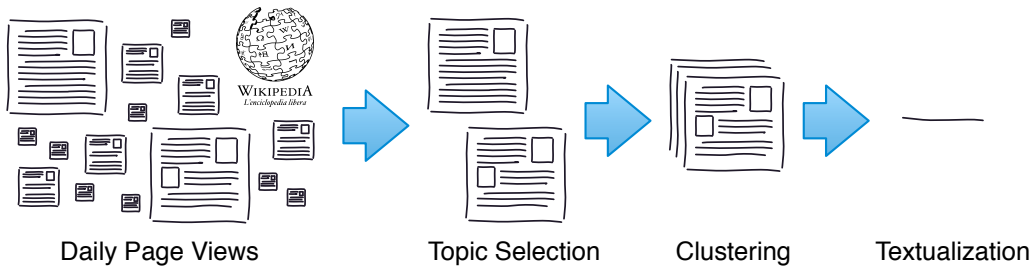

Figure 4: Process diagram. (a) Topic selection: select interesting articles based on increase in page views. (b) Clustering: cluster the articles according to relevent events using topic models or the Wikipedia hyperlink structure. (c) Textualization: select the sentences that best summarizes the relevant events in text.

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: Example articles for January 27, 2009. These are the articles that do not have a counterpart in the other side in a window size of 15 days. The hand-selected articles are linked from a single event "Florida hedge fund manager Arthur Nadel is arrested by the United States Federal Bureau of Investigation and charged with fraud."

Article selection

Design For each day, the 100 articles with the most increase in page views are selected. The difference between the total sum of page views for the past 15 days and the total sum of page views for the previous 15-days period are calculated, and the articles are sorted in the order of decreasing difference. To facilitate the process, the articles with too small page views are ignored.

Evaluation We do not attempt to evaluate the selected articles against the Wikipedia current events for the reasons described in the previous section.

Clustering

More often than not, more than one popular articles are related to an external current event. For example, all the hand-selected articles shown in Figure 3 are related to a Wikipedia 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.

Design As a baseline, two different clustering scheme makes use of the link structure: CONNCOMP and ONEHOP. CONNCOMP is to cluster articles in the same connected components, connecting articles that have a direct link from one to another. ONEHOP is to cluster articles within only one hop. The number of resulting clusters depend on the order in which you choose the next article to cluster. To find the minimum number of such clusters is NP-complete. Instead of attempting to find the optimal clusters, we just tried to cluster in the descreasing order of the number of links: With most links, first clustered. The link structure is downloaded from the website of Henry Haselgrove[5].

We also performed K-means clustering on the set of articles, treating article text as a bag of words. For 100 automatically selected articles on each of the five selected dates, the number of clusters K was set to 50. We used the Mallet software (McCallum 2002) to run K-means clustering. 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.

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 measure the annotator agreement, using

[5]http://users.on.net/ henry/home/wikipedia.htm

Test set	# Clusters	B^3 F-score
Human-1	48.6	0.704 ± 0.083
Human-2	50.0	0.710 ± 0.108
Human-3	53.8	$\mathbf{0.741 \pm 0.103}$
ConnComp	31.8	0.424 ± 0.183
OneHop	45.2	0.580 ± 0.172
K-means tf	50	0.521 ± 0.042
K-means tf-idf	50	$\mathbf{0.584 \pm 0.089}$
LDA	44.8	0.426 ± 0.080

Table 1: Clustering evaluation: CONNCOMP and ONEHOP are clustering using the link strcuture. K-MEANS clustering uses the text of the articles as bag of words. For the B-Cubed metric, exchanging the gold standard and the data set results in the exchange of the precision and the recall score, thus leaving the F-score same.

the multiplicity B^3 metric (Amigó et al. 2009) that can handle overlapping clusters. The results are shown in Table 1.

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:

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

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

The inter-annotator agreement in the B-cubed scores are in the range of 67%–74%. Clustering with the link structure performed the worse, having half of the precision for manual clusters. K-means clustering performs best, achieving 72% precision compared to manual clustering.

Analysis There are two main reason that the link structure performed worse. First, the link structure was too old. The link structure was generated on January 28, 2009 and 13.2% of the WikiTopics articles were created after the date, thus missing from the link structure. Second, there are a few "octopus" articles that have links to many articles, and cause grouping them into one large cluster. The United States on January 27, 2009 was particularly harmful, grouping 79 articles into a single cluster.

K-MEANS has its own defects: it does not distinguish different meanings of words. For example, the automatically selected articles for April 19 include both Piracy in Somalia and The Pirate Bay as well as Piracy. Their resemblance in word spelling might result in confusion in clustering, depending on the clustering method. In fact, K-MEANS correctly clustered The Pirate Bay with The Pirate Bay Trial, but clustered Piracy 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. In contrast, clustering method using the link strcuture, CONNCOMP and ONEHOP correctly clustered Somalia, Piracy in Somalia, and Piracy all in the same cluster.

Clustering the articles according to the relevance to recent popularity is not a trivial work even for humans. 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 merged the articles relevant to Grammy Awards or Super Bowl into a cluster.

Textualization

We attempt to generate textual descriptions for the clustered articles to explain why they are popular and what event is relevant. We consider the date expressions, the reference to the article as features. Currently, our work is limited to select the best sentence that describes the relevant events, but it could be future work to descibe the relationships of the articles to the relevant event, and summarize the description using sentence fusion or paraphrasing. Often, some articles are directly connected to an external event while others are subsidiary topics about the event and may not be relevant to the reason why the event is popular.

Preprocess We preprocess the Wikipedia articles using the Serif system (Boschee, Weischedel, and Zamanian 2005) for date tagging and 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 relative time ("now", "later that year", "The following year"). Some examples are shown in Figure . The entities mentioned in a given article are compiled into a list and the mentions of each entity are linked to the entity as a coreference chain.

Design As a baseline, we picked the first sentence for each article because the first sentence usually is an ovewview of the topic of the article and often relevant to the external event. We refer to this as FIRST.

We also picked the sentence with the most closest date to the date on which the article was selected. Closeness refers to the difference in days between the dates. The dates in sentences may vary in their formats, so we put precedence over the formats so that more specific date i.e. "Feburary 20, 2009" has precedence over more vague date formats such as "February 2009" or "2009". We refer to this scheme as RECENT.

For the third data set, we picked the sentence both with the most recent date and with the reference to the article's

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

Figure 5: Examples of temporal expressions identified by the SERIF system in the preprocess step, selected from 247 such date and time expressions extracted from the article Abraham Lincoln.

2009-01-27: Barack Obama

Before: He was inaugurated as President on January 20, 2009.
After: **Obama** was inaugurated as President on January 20, 2009.

2009-05-12: Eminem

Before: He is planning on releasing his first album since 2004, Relapse, on May 15, 2009.
After: **Eminem** is planning on releasing his first album since 2004, Relapse, on May 15, 2009.

Figure 6: Selected examples of sentence selection and coreference resolution. From each article the best SENTence is selected based on the most recent date expression and the reference of the topic, and the personal pronouns are substituted with their proper names, which are in **bold**.

topic. We refer to this sceme as SELF.

After selecting a sentence for each cluster, we used coreference resolution to substitute personal pronouns in the sentence with their proper names. This step enhances readability of the selected sentence, which often refers to its subject by a pronoun such as "he", "his", "she", or "her'. The examples of substituted proper names appear in Figure 6 in bold face.

The SERIF system tags the type of each entity mention as proper name, nominal position, or pronoun. There may be more than one proper name for each entity and to choose the best one is not a trivial task: proper names vary from *John* to *John Kennedy* to *John Fitzgerald "Jack" Kennedy*. Our algorithm chose the most frequent proper name to substitute with.

Evaluation For ten articles on each of five selected dates, an annotator selected sentences that describes why each article gains in popularity, among 289 sentences per each article on average. The annotator picked the one best sentence, and the possible multiple second best sentences. In the case there is no best sentence among them, he marked none as the best sentence, and listed all the partially explaning sentence as second best sentences.

To see inter-annotator agreement, another annotator selected the best sentence for the ten articles of the first date.

	Best		Second best	
Scheme	Precision	Recall	Precision	Recall
Human	0.50	0.55	0.85	0.75
First	0.14	0.21	0.34	0.39
Recent	**0.33**	**0.48**	**0.55**	**0.61**
Self	0.29	0.36	0.49	0.45

Table 2: Precision of different sentence selection scheme.

The evaluation results for all the selection schemes are shown in Table 2.

Analysis 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 during the preprocess. In other words, the gold standard sentence cannot be selected no matter how well the selection performed. We ran the splitta (Gillick 2009) over the article text and found that it does not split the first sentence at the wrong position.

Selection of the best sentence with a RECENT date seems to work well, with its own 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 year, which is overshadowed by the other dates with year describing prior events in our scheme where more specific dates have precedence over less specific ones.

The selection of the FIRST sentence performs worst, but in a third of the articles the first sentences are either best or second best. It is because the first sentence is an overall introduction about the topic, often including a person's recent achievement or death if the topic is about a person.

The SELF feature, which is about if a sentence has a self reference to the topic, must be used another feature such as RECENT because it is a binary feature and cannot be used to select a sentence among the sentences all having a self reference. When it was forced to select among sentences that have self reference, a different sentence is selected for about a third of data (18 out of 50). Only half of the difference selections (8 out of 18) contributed to make the selection better of worse. Exactly half of them made it better and the other half made it worse. For three out of four cases, selecting a sentence with self reference failed because SERIF failed to find self reference that is in fact in the sentence.

Related work

News summarization systems such as Google News and Columbia Newsblaster (McKeown et al. 2002) are probably the most famous efforts that applies techniques in topic detection and tracking along with various techniques of natural language processing into a big pipeline. Google News

"group news articles into clusters of articles about related events and categorize each event into predetermined top-level categories, finally selecting a single representative article for each cluster." (copied from Lydia paper. Need paraphrasing.) Newsblaster "goes further in providing computer-generated summaries of the day's news from the articles in a given cluster." (copied from Lydia paper. Need paraphrasing.) NewsInEssence from University of Michigan follows the same line of news summarization but allows users to provide an example news and keywords to make a customized cluster of news articles. Lydia project analyzes entities such as people, places, and things that appear in news articles as well as news sources to find relationships between entities and between entities and news sources.

Petrović, Osborne, and Lavrenko (2010) is the first attempt to do first story detection (FSD) on Twitter in a streaming setting. It does not cluster the articles, and just finds the closest nearest neighbor of a new post in approximate fashion using locality sensitive hash that guarantees the probabilty it misses the nearest neighbor to be under a given limit and says that if the distance is under a given threshold, they are about the same event. They use cosine metric as a similarity measure and tf-idf as Allan's paper suggest them to be the best settings for FSD task.

Generating story highlights is a task akin to summarization in that it needs best phrases to cover the original text, but it is different in that it is not entirely grammatical but in telegraphic style omitting some function words. Woodsend and Lapata (2010) addresses the problem using integer linear programming to meet global constraints at the same time, such as summary length, story length, coverage, and grammaticality. To extract the most important words, they extracted the words with the most tf-idf scores.

Conclusion

We aim to automatize locating recent popular events and describing the events. We presented a novel data set and provides the baseline results to improve upon. It is shown that the topics that are automatically selected are very different from Wikipedia's current events that are hand-curated. We could go even further by predicting some novel events' happening using social media such as Twitter before the news hit the traditional newspaper or magazine. We used famous clustering such as K-means and LDA and compared the performance to manual clustering. We provides the gold standard for clustering for evaluation of different clustering methods. We want to discover the best way to take advantage of the page text, link structure, and edit history of the selected Wikipedia articles to find clustering. We did flat clustering but hierarchical clustering is always an option. We showed the value of date taggers and coreference resolution to extract best sentences and revise them to improve readability. We can use hierarchical clustering jointly with to summarize the clusters with fewer sentences and to clearly show the associations between the selected Wikipedia articles.

References

Amigó, E.; Gonzalo, J.; Artiles, J.; and Verdejo, F. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inf. Retr.* 12(4):461–486.

Boschee, E.; Weischedel, R.; and Zamanian, A. 2005. Automatic information extraction. In *Proceedings of First International Conference on Intelligence Analysis*.

Gillick, D. 2009. Sentence boundary detection and the problem with the U.S. In *Proceedings of HLT/NAACL*.

McCallum, A. K. 2002. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

McKeown, K. R.; Barzilay, R.; Evans, D.; Hatzivassiloglou, V.; Klavans, J. L.; Nenkova, A.; Sable, C.; Schiffman, B.; and Sigelman, S. 2002. Tracking and summarizing news on a daily basis with Columbia's Newsblaster. In *Proceedings of HLT*.

Petrović, S.; Osborne, M.; and Lavrenko, V. 2010. Streaming first story dectection with application to Twitter. In *Proceedings of NAACL*.

Woodsend, K., and Lapata, M. 2010. Automatic generation of story highlights. In *Proceedings of ACL*.