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 i 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 Wiipedia 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.

(Linguistic Data Consortium, 2004) describes the definition of topics and events used in the TDT dataset.

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 an X 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 by XXX% or more 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 eents.

Mention Wikipedia current events, and why we

did not use them as our gold standard. They are boring events that do not generate much traffic, and that 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 **??** 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". [take the first sentence from Chesley's article] [Superbowl XLIII] will feature the... (the second setence, 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? This could be an aesthetic question, but, in Wikipedia, gaining popularity usually means a big inflow of users looking up an article of a specific topic. More often than not, a popular on-going event accompanies many topics, or Wikipedia articles, collecting sudden increases in page view counts (Figure **??**). The TrendingTopics[1] algorithm (§5) finds a big increase in page view counts and make a list of the articles with the highest increase.

Barack_Obama
Inaugural...
...
The names of past presidents...
...
Hudson_River...

Table 1: Topics for January 27th, 2009.

The wikipedia has a section called current events, in which the recently occurred events are listed manually by Wikipedians[2]. Some of the most charateristic instances from the topics and the events are shown in Table **??**. We evaluated the precision and recall score of the TrendingTopics topics against the Wikepedian Current Events (Table **??**. Their precision and recall scores are extremely low. Probably, wikipedians are more interested in political and scientific topics, and the page view statistics tend to be more dramatic for the societal and cultural events– recent deaths of famous people, recent release of movies and music albums.

Due to the discrepancy between the Wikipedian Current Events and the topics based on page views, it was not realistic to predict from the page view counts the topics that would overlap the Current Events. We set the topics extracted by the TrendingTopics algorithm as the baseline data.

Our system is described as follows. First, the most popular pages, or topics, are collected per each day. Second, clusters are identified by clustering related topics. Third, the main cluster of each cluster is identified. Lastly, the sentence that best describes the cluster are extracted for each cluster.

3 Clustering

Once extracted, the topics include individual articles that can be divided into some categories (see Table 1). The categories, however, are somewhat random and highly related to on-going events. In many cases, there are "central" topics that are directly relevant to an on-going event, and there are "peripheral" topics that are weakly related to the same events. In Figure 1, `Barack Obama` and `Presidential Inauguration` are cen-

[1] Should we put a space between Trending and Topics?

[2] Should Wikipedians be capitalized or not?

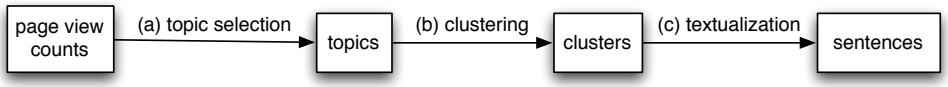

Figure 1: Process diagram. (a) Topic selection. (b) Clustering. (c) Textualization.

tral topics, and the articles of past U.S. Presidents are peripheral topics.

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.

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

[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 2: 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.

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 Cluster-

ing Evaluation Metrics based on Formal Constraints Where can I find cite information for this paper?

Sasa Petrovic, Miles Osborne, and Victor Lavrenko. 2010. Streaming First Story Dectection with application to Twitter NAACL, Los Angeles, USA. June 2010.

Linguistic Data Consortium. 2004. TDT 2004: Annotation Manual Linguistic Data Consortium.