MediaMeter: A Global Monitor for Online News Coverage

Tadashi Nomoto
National Institute of Japanese Literature
10-3 Modori Tachikawa, Japan
nomoto@acm.org

Abstract

This paper introduces MediaMeter, an application that works to detect and track emergent topics in the US online news media. What makes MediaMeter unique is its reliance on a labeling algorithm which we call WikiLabel, whose primary goal is to identify what news stories are about by looking up Wikipedia. We discuss some of the major news events that were successfully detected and how it compares to prior work.

1 Introduction

A long term goal of this project is to build a sociologically credible computational platform that enables the user to observe how social agenda evolve and spread across the globe and across the media, as they happen. To this end, we have built a prototype system we call MediaMeter, which is designed to detect and track trending topics in the online US news media. One important feature of the system lies in its making use of and building upon a particular approach called WikiLabel (Nomoto, 2011). The idea was to identify topics of a document by mapping it into a conceptual space derived from Wikipedia, which consists of finding a Wikipedia page similar to the document and taking its page title as a possible topic label. Further, to deal with events not known to Wikipedia, it is equipped with the capability of re-creating a page title so as to make it better fit the content of the document. In the following, we look at what WikiLabel does and how it works before we discuss MediaMeter.

2 WikiLabel

WikiLabel takes as input a document which one likes to have labeled, and outputs a ranked list of label candidates along with the confidence scores.

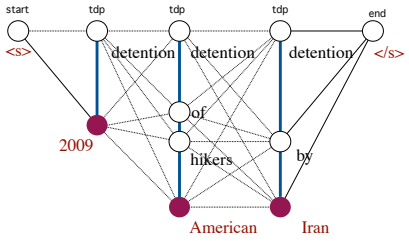


Figure 1: Trellis for enumerating compressions for *"2009 detention of American hikers by Iran"*.

The document it takes as input needs to be in the form of a vector space model (VSM). Now assume that $\vec{\theta}$ represents a VSM of document d. Let us define $l^*_{\vec{\theta}}$, a likely topic label for d, as follows.

$$l^*_{\vec{\theta}} = \arg\max_{l:p[l]\in\mathcal{U}} \text{Prox}(p[l], \vec{\theta}|_N), \qquad (1)$$

where $p[l]$ denotes a Wikipedia page with a title l and $\vec{\theta}|_N$ a VSM with its elements limited to top N terms in d (as measured by TFIDF). $\text{Prox}(p[l], \vec{\theta}|_N)$ is given by:

$$\text{Prox}(p[l], \vec{\theta}|_N) = \lambda Sr(p[l], \vec{\theta}|_N) + (1-\lambda) Lo(l, \vec{\theta}).$$

We let:

$$Sr(\mathbf{r}, \mathbf{q}) = \left(1 + \sum_t^N (\mathbf{q}(t) - \mathbf{r}(t))^2\right)^{-1}$$

and

$$Lo(l, \vec{v}) = \frac{\sum_i^{|l|} I(l[i], \mathbf{v})}{|\ l\ |} - 1$$

where $I(w, v) = 1$ if $w \in v$ and 0 otherwise.

$Sr(\vec{x}, \vec{y})$ represents the distance between $\vec{x}$ and $\vec{y}$, normalized to vary between 0 and 1. $Lo(l, \vec{v})$ measures how many terms l and $\vec{v}$ have in common, intended to quantify the relevance of l to $\vec{v}$. $l[i]$ indicates i-th term in l. Note that Lo works as a penalizing term: if one finds all the terms l has in $\vec{v}$, there will be no penalty: if not, there will

Proceedings of the First Workshop on Computing News Storylines, pages 30–34,
Beijing, China, July 31, 2015.

Table 1: Compressing a Wikipedia title

2009 detention of American hikers by Iran
2009 detention
2009 detention by Iran
2009 detention of hikers
2009 detention of hikers by Iran
2009 detention of American hikers by Iran
. . .

Table 2: Summary of the quality review by humans. '#instances' refers to the number of labels sent to judges for evaluation.

LANGUAGE	RATING	#instances
ENGLISH	4.63	97
JAPANESE	4.41	92

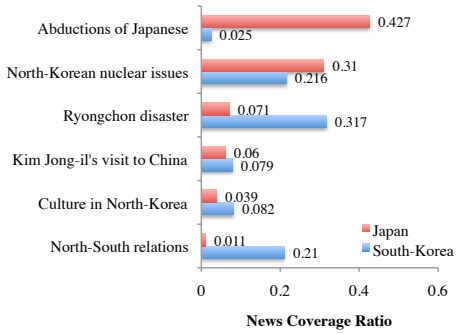


Figure 2: Conflicting media perceptions of North-Korea (E Gwangho, 2006). 'News coverage ratio' indicates the proportion of news articles focusing on a particular topic.

be a penalty, the degree of which depends on the number of terms in l that are missing in $\vec{v}$. $\mathcal{U}$ represents the entire set of pages in Wikipedia whose namespace is 0. We refer to an approach based on the model in Eqn. 1 as 'WikiLabel.' We note that the prior work by Nomoto (2011) which the current approach builds on, is equivalent to the model in Eqn. 1 with λ set to 1.

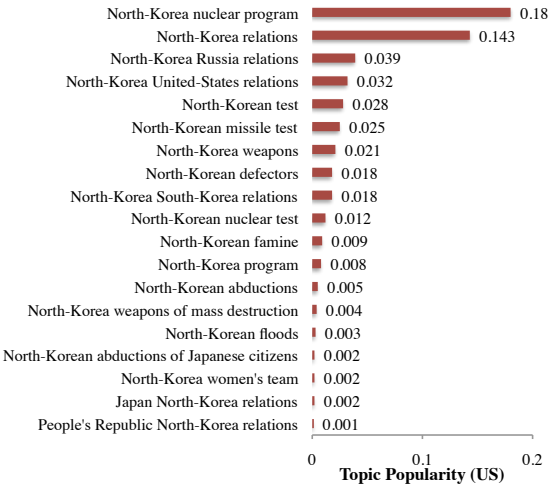
One important feature of the present version, which is not shared by the previous one, is its ability to go beyond Wikipedia page titles: if it comes across a news story with a topic unknown to Wikipedia, WikiLabel will generalize a relevant page title by removing parts of it that are not warranted by the story, while making sure that its grammar stays intact. A principal driver of this process is sentence compression, which works to shorten a sentence or phrase, using a trellis created from a corresponding dependency structure (e.g. Figure 1). Upon receiving possible candidates from sentence compression, WikiLabel turns to the formula in Eqn. 1 and in particular, Lo[1] to determine a compression that best fits the document in question.

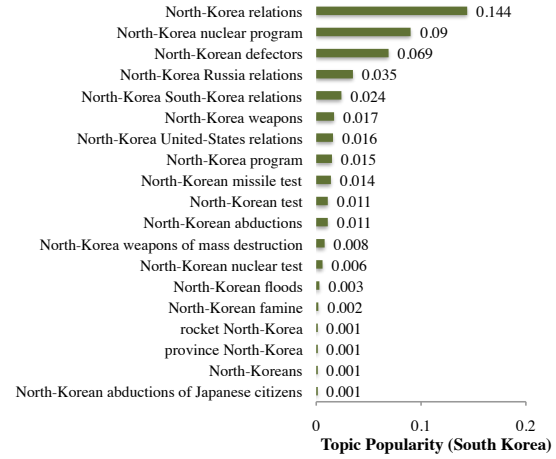
3 North-Korean Agenda

Shown in Figure 3 are most popular topics WikiLabel found among news stories discussing North Korea (DPRK), published online in 2011, in a number of different countries, including US, South-Korea and Japan (the number of stories we covered was 2,230 (US), 2,271 (South-Korea), and 2,815 (Japan)). Labels in the panels are given as they are generated by WikiLabel, except those for the Japanese media, which are translated from Japanese. (The horizontal axis in each panel represents the proportion of stories on a given topic.) Notice that there are interesting discrepancies among the countries in the way they talk about North Korea: the US tends to see DPRK as a nuclear menace while South Korea focuses on diplomatic and humanitarian issues surrounding North Korea; the Japanese media, on the other hand, depict the country as if it had nothing worth talking about except nuclear issues and its abduction of the Japanese. Table 2 shows how two human assessors, university graduates, rated on average, the quality of labels generated by WikiLabel for articles discussing North-Korea, on a scale of 1 (poor) to 5 (good), for English and Japanese.

Curiously, a study on news broadcasts in South Korean and Japan (Gwangho, 2006) found that the South Korean media paid more attention to foreign relations and open-door policies of North Korea, while the Japanese media were mostly engrossed with North Korean abductions of Japanese and nuclear issues. In Figure 2, which reproduces some of his findings, we recognize a familiar tendency of the Japanese media to play up nuclear issues and dismiss North Korea's external relations, which resonate with things we have found here with WikiLabel.

[1]Because the candidates here are all linked to the same Wikipedia page, Sr can be safely ignored as it remains invariant across them.





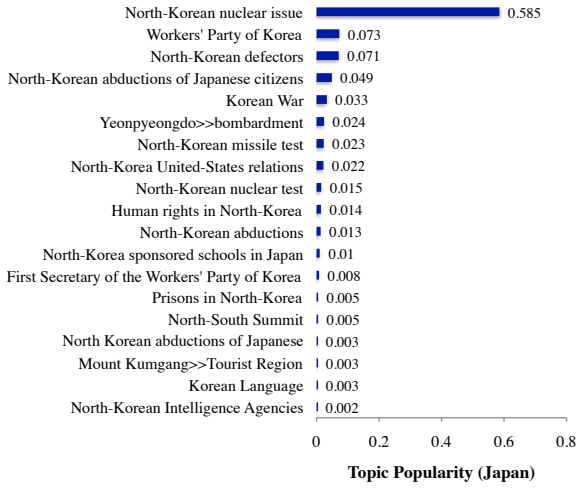


Figure 3: North-Korean agenda across countries

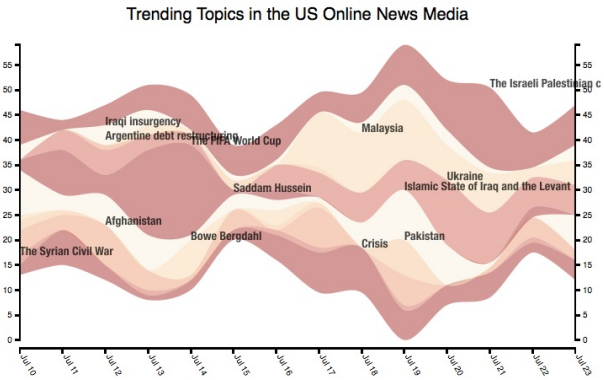


Figure 4: MediaMeter: Overview

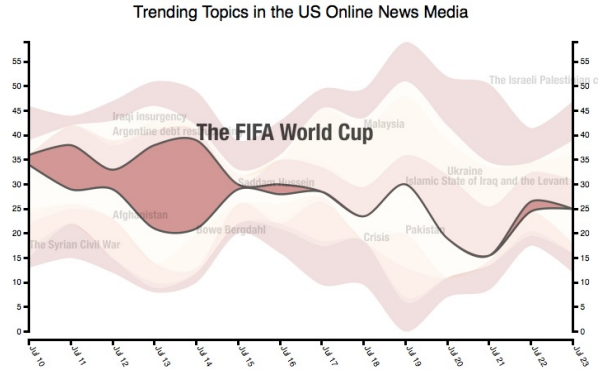


Figure 5: MediaMeter: Focused View 1

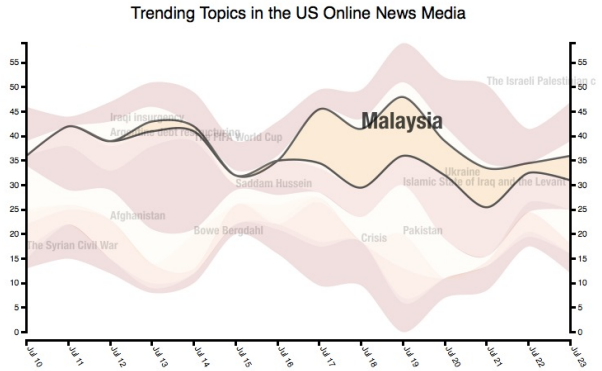


Figure 6: MediaMeter: Focused View 2

4 MediaMeter

MediaMeter[2] is a web application that draws on WikiLabel to detect trending topics in the US online news media (which includes CNN, ABC, MSNBC, BBC, Fox, Reuters, Yahoo! News, etc). It is equipped with a visualization capability based on ThemeRiver (Havre et al., 2002; Byron and Wattenberg, 2008), enabling a simultaneous tracking of multiple topics over time. It performs the following routines on a daily basis: (1) collect news stories that appeared during the day; (2) generate topic labels for 600 of them chosen at random; (3) select labels whose score is 1 or above on the burstiness scale (Kleinberg, 2002); (4) find for each of the top ranking labels how many stories carry that label; and (5) plot the numbers using the ThemeRiver, together with the associated labels. Topic labels are placed automatically through integer linear programming (Christensen et al., 1995).

Figure 4 gives a ThemeRiver visualization of trending topics for the period from July 10 to 23, 2014. Figures 5 and 6 show views focusing on particular topics, with the former looking at the World Cup and the latter at Malaysia. The media's attention to the World Cup mushroomed on July 14th, the day when the final match took place, and fizzled out on the following day. Meanwhile, in Figure 6, there is a sudden burst of stories related to Malaysia on July 17th, which coincides with the day when a Malaysian jetliner was shot down over the Ukrainian air space. While it is hard to tell how accurately MediaMeter reflects the reality, our feeling is that it is doing reasonably well in picking up major trends in the US news media.

5 Evaluation

To find where we stand in comparison to prior work, we have done some experiments, using TDT-PILOT, NYT2013, and Fox News corpora. TDT-PILOT refers to a corpus containing 15,863 news stories from CNN and Reuters, published between July 1, 1994 and June 30, 1995. The Fox News corpus has the total of 11,014 articles, coming from the online Fox news site, which were published between January, 2015 and April, 2015. NYT2013 consists of articles we collected from the New York Times online between June and December, 2013, totaling 19,952. We measured performance in terms of how well machine generated labels match those by humans, based on the metric known as ROUGE-W (Lin, 2004).[3] ROUGE-W gives a score indicating the degree of similarity between two strings in terms of the length of a subsequence shared by both strings. The score ranges from 0 to 1, with 0 indicating no match and 1 a perfect match. In the experiment, we ran TextRank (TRANK) (Mihalcea and Tarau, 2004) – the current state of the art in topic extraction – and different renditions of WikiLabel: RM1 refers to a model in Eqn 1 with λ set to 0.5 and sentence compression turned off; RM1/X is like RM1 except that it makes use of sentence compression; RM0 is a RM1 with λ set to 1, disengaging Lo altogether.

Table 3: Per-document performance@1

	TRANK	RM_0	RM_1	RM_1/X
NYT	0.000	0.056	0.056	**0.069**
TDT	0.030	0.042	0.048	**0.051**
FOX*	0.231	0.264	0.264	**0.298**

Table 3 gives a summary of what we found. Numbers in the table denote ROUGE-W scores of relevant systems, averaged over the entire articles in each dataset. Per-document performance@1 means that we consider labels that ranked the first when measuring performance. One note about FOX. FOX has each story labeled with multiple topic descriptors, in contrast to NYT and TDT where we have only one topic label associated with each article. Since there was no intrinsically correct way of choosing among descriptors that FOX provides, we paired up a label candidate with each descriptor and ran ROUGE-W on each of the pairs, taking the highest score we got as a representative of the overall performance. Results in Table 3 clearly corroborate the superiority of RM0 through RM1/X over TextRank.

6 Conclusions

In this paper, we looked at a particular approach we call WikiLabel to detecting topics in online news articles, explaining some technical details of how it works, and presented MediaMeter, which showcases WikiLabel in action. We also demonstrated the empirical effectiveness of the approach through experiments with NYT2013, FOX News and TDT-PILOT.

[2] http://www.quantmedia.org/meter/demo.html

[3] Each article in all the three datasets comes with human supplied topic labels.

References

Lee Byron and Martin Wattenberg. 2008. Stacked graphs - Geometry & Aesthetics. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1245–1252.

Jon Christensen, Joe Marks, and Stuart Shieber. 1995. An empirical study of algorithms for point-feature label placement. *ACM Trans. Graph.*, 14(3):203–232, July.

E. Gwangho. 2006. *Hutatsu no Kita-Chosen* (Two North Koreas). *Media Communication*, 56:59–71. Keio University.

S. Havre, E. Hetzler, P. Whitney, and L. Nowell. 2002. Themeriver: visualizing thematic changes in large document collections. *Visualization and Computer Graphics, IEEE Transactions on*, 8(1):9–20, Jan.

Jon Kleinberg. 2002. Bursty and hierarchical structure in streams. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 91–101, New York, NY, USA. ACM.

Chin-Yew Lin. 2004. Rouge: a package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into texts. Association for Computational Linguistics.

Tadashi Nomoto. 2011. Wikilabel: an encyclopedic approach to labeling documents en masse. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 2341–2344, New York, NY, USA. ACM.