Detecting Trending Topics Using Page Visitation Statistics

Sayandev Mukherjee Ronald Sujithan Pero Subasic

DOCOMO Innovations Inc.
3240 Hillview Ave
Palo Alto, CA 94304, USA
{smukherjee,rsujithan,psubasic}@docomoinnovations.com

ABSTRACT

Many applications including realtime recommenders and ad-targeting systems have a need to identify trending concepts to prioritize the information presented to end-users. In this paper, we describe a novel approach that identifies trending concepts using the hourly Wikipedia page visitation statistics freely available for download. We describe a MapReduce framework that analyzes the raw hourly visitation logs and generates a ranked list of trending concepts on a daily basis. We validate this approach by extracting hourly lists of trending news articles, mapping these articles to Wikipedia concepts, and computing the similarity of the two lists according to several commonly used measures.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Information networks*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Abstracting methods*

General Terms

Human factors, Economics

Keywords

Trending concepts; Wikipedia; page visitation

1. INTRODUCTION

It is known that "trending topics," meaning topics of aggregate interest and attention, can provide insight into the interests, attitudes, and behavior of global and local communities. Mobile commerce applications such as DOCOMO dmarket [3] can greatly benefit from the ability to be first to identify such trending topics. Up to this time, nearly all investigation of "trending topics" has concentrated on detecting such topics on social media, especially micro-blogging services such as Twitter [4] and Weibo [2].

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577303 .

However, it has been found that the vast majority of trending topics on Twitter live and die over a timescale of at most a few hours [5]. In the present work, we are interested in trends arising from news-related events. A news-worthy event is likely to spawn one or more trending topics on Twitter, but identifying such trends from hashtags in tweets is problematic because of spam and non news-related topics with high levels of activity (as measured by numbers of tweets and retweets). Thus we need other measures to identify such longer-lived news-related topics.

A measure of the "importance" of a trending topic (and therefore correlated with the "life" of this topic) is the number of search queries based on this topic that are entered into search engines such as Google and Bing. For concreteness, we shall use *trending concept(s)* to denote the trending topic(s) contained, in a semantic sense, in a search query.

2. IDENTIFYING TRENDING TOPICS

Direct access to the query logs is unavailable to anyone other than the search engine companies themselves. However, hourly lists of the most popular links (to news articles) returned by search queries, grouped under some reasonably descriptive categories to which these links belong, are published by the search engines, including Google News and Bing News (e.g. the left-hand panel, titled "Top Stories," on `http://news.google.com`).

2.1 Using an inverted index and a dictionary

"Concepts" can be extracted from the contents of these linked news articles using machine learning techniques. A common implementation, available in commercial tools such as Lucene [1], is to scan all documents in a *corpus* in order to create: (i) a *dictionary* mapping *keywords* (concepts) to *values* (frequent phrases or words in the articles), and (ii) an *inverted index* mapping certain phrases and words to keywords in this dictionary. Then, when the appropriate phrases and words are extracted from any document of interest, the inverted index will yield some keywords (i.e., concepts) that pertain to this document.

The Wikipedia document corpus is a popular choice for generating such dictionaries and inverted indices, as it is stable, accurate, and comprehensive. The keywords (concepts) in the resulting dictionary are (titles of) articles in Wikipedia, called "concept pages" in Wikipedia terminology.

Given a list of news articles provided by Google News or Bing News, we extract the text from each of the news articles in the list, pass that text through an inverted index,

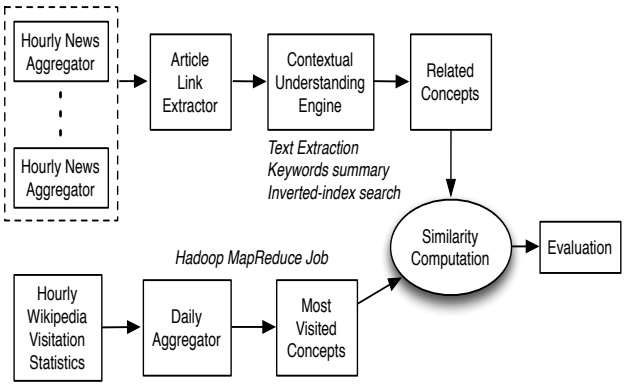

Figure 1: A block diagram showing the various modules and processing steps in obtaining two lists of candidate "trending topics," one from several news aggregator sites and the other from Wikipedia page visitation counts.

and obtain a list of Wikipedia concepts that should define a list of trending concepts.

2.2 Using Wikipedia page visitation statistics

A related approach is to identify trending concepts with *the most-visited* Wikipedia concept pages. Wikipedia links rank high in the results from search engine queries, so a substantial fraction of users making these queries may click on the Wikipedia link in the offered results.

In short, we expect Wikipedia page visitation statistics to provide another way to obtain lists of trending concepts. Since a visit to a Wikipedia page usually represents more than just a passing interest in the topic concerned, we conjecture that trending concepts obtained from such Wikipedia page visitation statistics are likely to be of the kind whose popularity rises and falls over longer timescales than the majority of Twitter trending topics. We also conjecture that the list of trending concepts obtained from Wikipedia page visitation statistics will have a good extent of overlap with the list of trending concepts obtained by processing lists of news articles through an inverted index as described above. Fig. 1 shows a block diagram of an architecture that enables us to test both the above conjectures.

3. RESULTS AND DISCUSSION

In Fig. 2, we plot the mean (over the period December 06-26, 2013) Jaccard similarity and Kendall τ-rank correlation coefficients of the two lists of trending concepts obtained by the two processing pipelines shown in Fig. 1 (normalized with respect to the ratio of the lengths of the two lists). The solid curves are for various lags relative to the visitation derived list, whereas the dashed curves are for various lags relative to the news-aggregation derived list. We observe that for both Jaccard and Kendall coefficients, with reference to the visitation list, the highest mean similarity is with the news-aggregation derived list of the same day (zero lag), with the next highest similarity being with the news-aggregation list of the day before (lag $= -1$). This is plausible as the news cycle is expected to be slightly ahead of the visitation activity. Similarly, with reference to the news-aggregation list, the highest mean similarity is with

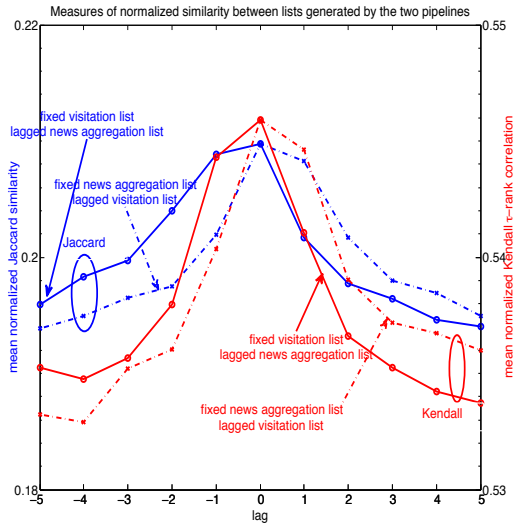

Figure 2: Plots of mean normalized Jaccard similarity and Kendall τ-rank correlation coefficients between the lists of trending concepts obtained from the two processing pipelines of Fig. 1.

the visitation list of the same day, with the next highest similarity for the next day (lag $= 1$).

4. CONCLUSIONS

We implemented a hardware and software architecture to get lists of trending concepts: (i) by processing the text from web pages listed on news aggregation sites through an inverted index, and (ii) by accessing Wikipedia concept visitation statistics. Our results provide empirical support for our conjecture that trending news topics, as defined by concepts embodied in popular news articles linked to by news aggregators, can also be identified directly from Wikipedia page visitation statistics. Thus, either method may be used to identify trending topics. Early identification of such trending topics is essential in designing marketing, advertising, and awareness campaigns. It is also useful in enhancing the effectiveness of recommenders and ad-targeting systems.

5. REFERENCES

[1] Apache Lucene. `lucene.apache.org`.

[2] L. Chen, C. Zhang, and C. Wilson. Tweeting Under Pressure: Analyzing Trending Topics and Evolving Word Choice on Sina Weibo. In *Proceedings of the 1st Annual Conference on Online Social Networks (COSN 2013)*, pages 89–100. ACM, October 2013.

[3] DOCOMO dmarket. `nttdocomo.co.jp/service/entertainment/dmarket`.

[4] S. Kairam, M. Morris, J. Teevan, D. Liebling, and S. Dumais. Towards Supporting Search over Trending Events with Social Media. In *Proceedings of the 7th International Conference on Weblogs and Social Media (ICWSM 2013)*, pages 283–292. AAAI, July 2013.

[5] S. Nikolov and D. Shah. A Nonparametric Method for Early Detection of Trending Topics. In *Proceedings of the Interdisciplinary Workshop on Information and Decision in Social Networks (WIDS 2012)*. MIT, November 2012.