Temporal Analysis of Supply and Demand of Topics on The Web*

Masahiro Inoue
Kyoto University
inoue@dl.soc.i.kyoto-u.ac.jp

Keishi Tajima
Kyoto University
tajima@i.kyoto-u.ac.jp

ABSTRACT

Timing of supply and demand for information on a topic does not always coincide. Sometimes one of them rises first, then the other follows. We show a classification of hot topics on the Web in the past based on the temporal relationship between their supply and demand, and also show that our classification is useful for predicting the timing of supply peaks in some cases.

1 INTRODUCTION

Supply and demand for some information on the Web have correlation, but their timing does not always coincide. Sometimes one of them appears first, and the other follows. The timing they reach their peaks may also be different. In this paper, we show a classification of hot topics on the Web based on the temporal relationship between their supply and demand to investigate their properties.

There have been some research on supply and demand for information. McNie [3] discussed discrepancy between supply and demand of scientific information. McVicar et al. [4] analyzed the geographic discrepancy between supply and demand for musics by independent artists. The keynote talk by Fedyk [1] discussed the behavior of the consumers of the news from major news media. However, this is the first research on temporal relationship between supply and demand for information on the Web.

2 DETECTING SUPPLY AND DEMAND

For our purpose, we need to know when the information on each topic was demanded and when it was supplied.

A good indicator of the demand is the frequency of queries related to the topic submitted to Web search engines. Although query logs of major search engines are not available to the public, we can indirectly know when the demand for a topic rose through Google Trends[1]. We use the date when the topic appeared on Google Trends, denoted by T, to approximate when the demand rose.

On the other hand, there is no easy way to know when the information on a topic was supplied. In this research, we approximate it by retrieving web pages related to the topic through a search engine, producing a timeline showing how many pages newly appeared on each day, and detecting the peak and upsurge on the timeline.

In order to produce such a timeline, we need to know the creation date of each web page. Although creation dates of pages are not always explicit under the present Web, many web pages, e.g., most news articles and blogs, are created with some content management systems (CMSs), and CMSs usually embed creation dates on pages. We developed a system that detects such timestamps on pages [2]. The program is available as an open source library[2].

Table 1: Six Categories and their Ratio

Condition on Relative Position of T	Abbreviation	Ratio
$U \le T = P$	$T = P$	19%
$U \le T < P$, $T = P - 1$	$T = P - 1$	11%
$U \le T < P$, $T = P - 2$ or $P - 3$	$T = P - 2, 3$	1.6%
$P - 3 \le T < U \le P$	$T < U$	1.3%
$U \le P < T \le P + 3$	$P < T$	1.3%
$T < P - 3$ or $P + 3 < T$	$\lvert T - P \rvert > 3$	66%

Our system retrieves top 200 web pages through a search engine, and detects the peak and upsurge on the timeline produced from them. If the ranking by a search engines is influenced by the creation dates of pages, estimating the timelines only by using top 200 web pages given by a search engine would be biased. In order to avoid that, we select topics that are old enough, for which we expect that the temporal factor in the top ranking has become ignorable.

We define the peak day, P, on a timeline as the day when the most web pages appeared. Starting from P, we trace back successive previous days with new page appearances until we reach a day without a page appearance. We define the upsurge date, U, as the last day with page appearances we could trace back. Notice that $U \le P$. If the peak day has less than three pages, we determine the topic has no supply peak, and exclude the topic from our experiments.

Because we only use the top 200 pages, we rarely have pages irrelevant to the given topic, but we miss many relevant pages. Our purpose is, however, to find P and U. Our experiment shows that 200 pages are enough to find the correct P. For about half of queries in our experiment, even top 20 pages are enough to detect the same peak as with 200 pages, and for three fourth of our queries, top 100 pages are enough. On the other hand, if we include more pages, we will obtain earlier days for U. However, it can affect our classification only when the upsurge of supply is later than the date from Google Trends, which is very rare as we explain later.

Our system does not always detect correct timestamps on pages, and it can also affect our classification, but our system had the precision higher than 90% in our experiment, and errors in timestamp detection rarely affected the estimation of P and U.

3 CLASSIFICATION

We collected 4,000 queries from Google Trends, top 20 queries for each of 200 days starting from January 1st, 2011. Eliminating queries for which the peak has less than three pages as mentioned before,

*This work was supported by JSPS KAKENHI JP18H03245.

[1] http://www.google.com/trends/

[2] https://github.com/kkjk21/Timestamp-Extractor

WebSci '19, June 30-July 3, 2019, Boston, MA, USA
© 2019 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-6202-3/19/06.
https://doi.org/10.1145/3292522.3326054

Table 2: Query Examples

#	Category	Queries	U	P	T
1	$T = P$	inside job documentary	02/27	02/27	02/27
2	$T = P - 1$	critics choice awards 2011	01/12	01/15	01/14
3	$T = P - 2$	nfl draft	04/25	04/30	04/28
4	$T = P - 2$	arizona shooting	01/10	01/12	01/10
5	$T < U$	billy walters	01/17	01/17	01/16
6	$T < U$	prader willi syndrome	01/27	01/27	01/26
7	$P < T$	qwiki	01/20	01/20	01/22
8	$P < T$	super bowl food	01/31	02/04	02/05

we classified the remaining 2,695 queries. We classify them into six categories based on the temporal position of T relative to U and P. Table 1 lists the conditions defining the six categories, their abbreviated names, and the ratio of the queries classified into each category. Table 2 shows some examples of the queries. The dates are in 2011 and their format is “mm/dd”. In the following, we explain the details of these six categories.

(1) $T = P$

This is the second largest group of queries in our experiment. Typical topics in this group are news or events occurred on the day P. The query #1 in Table 2 is a typical example. For most queries in this group, $U = P$ also holds. Therefore, we can expect that T and P for this type of topic will be the day of the news or event in most cases. The supply will not increase and have a higher peak later.

(2) $T = P - 1$

This is the third largest group. $T = P - 1$ means the peak of the supply appears on the next day of the rise of the demand. Typically, several web pages appeared on T as prompt reports, after which most web pages with full contents, typically articles on news sites, appeared on $T + 1$. In the case of the query #2 in Table 2, T is the day of the announcement of the award winners, P was the next day, and U, the upsurge of the supply, started several days before T.

In this group, $U = T$ holds only for 7% of queries, while it holds for most queries in $T = P$ group as explained before. It means most queries that satisfy $U = T$ also satisfy $U = T = P$. Therefore, we can predict whether a new topic is in the group $T = P$ or not on the day T in the following way. Suppose a new topic appeared in Google Trends on the day T. We then detect U by tracing back the appearance of related pages starting from T, and if $U = T$, we can expect that $T = P$, i.e., the peak of the supply is also on T, and the supply will decrease next day. On the contrary, if $U < T$, we can expect that $T < P$, i.e., the supply will increase next day.

(3) $T = P - 2$, $T = P - 3$

Although this group is the forth largest, it includes far fewer queries than the previous two groups. For about 70% of queries in this group, $U \leq T$ holds, i.e., the supply started before the query became a trend. For these queries, the supply started early but it continued for long time and reached its peak after the demand rose.

Typical queries in this group are those about long term events or discussions which ended up with some interesting results. The query #3 in Table 2 is an example of such a type. In this group, $P - T$ is larger than in the group $T = P - 1$ by definition, and our experiment shows that $T - U$ is also larger than in the group $T = P - 1$ on average. Therefore, if we find $T - U$ is large on the day T, we can expect that $P - T$ will be also large.

This group also includes topics for which more reports appeared as time went by after the event on T. The query #4 in Table 2 is an example. This happens only for a small number of very big news.

(4) $T < U$

Among queries satisfying $T = P - 1$, $T = P - 2$, or $T = P - 3$, there are a small number of queries where $T < U$ holds. The condition $T < U$ means even the upsurge, usually the earliest reports on the topic, appeared later than the demand for the information.

Most of such queries are related to some real-world events not of type that are reported by news media. Most of them are related to TV programs. The query #5 and #6 in Table 2 are examples of this type. These queries were not triggered by news events but by the information appeared on TV. Because of that, the information was not reported much on the day T, but more web pages appeared later due to the appearance of the query in Google Trends. Such queries are typical examples where demand triggered supply.

(5) $P < T$

There also exist a few queries where $P < T$ holds. We can imagine why $P \geq T$ holds for most queries: it usually takes longer to make a report on an event and publish it than just to submit a query. Therefore, $P < T$ queries are rare, and this group is worth focusing on. We found there are mainly two types of queries in this group.

In one type, there was quick reports on it on the Web, but the search by people rose later because it took time for the people to get to know about it. The query #7 is an example of this type. A new web service “qwiki” was announced and it was reported by many blogs, but it took some time for people to know about the service. This is an example where supply triggered demand.

The query #8 is an example of the second type. In this case, much information on “super bowl food” was published in advance of the date in the near future when the information is expected to be demanded. For this type of queries, supply “foresaw” demand.

(6) $|T - P| > 3$

Many queries in this group are common nouns or popular proper nouns, e.g., names of celebrities or popular places. Such topics are continually demanded and supplied. For many of them, our system failed to find the peak corresponding to the demand, and as a result, $|T - P|$ was large. In our experiment, many queries were classified in this group, but if we can detect the peak more accurately, we can classify more queries in this group into the other groups.

4 CONCLUSION

We classified topics on the web based on the temporal relationship between their supply and demand. Our classification is useful for predicting supply peak in some cases, as explained in Section 3.

REFERENCES

[1] Anastassia Fedyk. 2015. Supply and Demand: Propagation and Absorption of News. In *NewsWWW (WWW 2015 Companion)*. 883–883.

[2] Masahiro Inoue and Keishi Tajima. 2012. Noise robust detection of the emergence and spread of topics on the Web. In *TempWeb (WWW 2012 Companion)*. 9–16.

[3] Elizabeth C. McNie. 2007. Reconciling the supply of scientific information with user demands: an analysis of the problem and review of the literature. *Environmental Science & Policy* 10, 1 (2007), 17 – 38.

[4] Matt McVicar, Cédric Mesnage, Jefrey Lijffijt, Eirini Spyropoulou, and Tijl De Bie. 2015. Supply and Demand of Independent UK Music Artists on the Web. In *WebSci*. 48:1–48:2.