Semantic Connection Based Topic Evolution

Jiamiao Wang,(~) **Xindong Wu,**(+) **Lei Li**(~)
(~)School of Computer Science and Information Engineering, Hefei University of Technology, 230009, Anhui, China
(+)School of Computing and Informatics, University of Louisiana at Lafayette, 70503, Louisiana, USA
wjmzjx@163.com, xwu@louisiana.edu, lilei@hfut.edu.cn

Abstract

Contrary to previous studies on topic evolution that directly extract topics by topic modeling and preset the number of topics, we propose a method of topic evolution based on semantic connection for an adaptive number of topics and rapid responses to the changes of contents. Semantic connection not only indicates the content similarity between documents but also shows the time decay, so semantic connection features can be used to visualize topic evolution, which makes the analyses of changes much easier. Preliminary experimental results demonstrate that our method performs well compared to a state-of-the-art baseline on both qualities of topics and the sensitivity of changes.

Introduction

Using topic evolution technology to handle the vast amounts of data can obtain useful information quickly and avoid the disadvantage of manual processing. However, there are some defects in existing methods and we contribute the following as potential reasons.

First, conventional methods have a characteristic that the number of topics is fixed, which carries two problems: too many redundant topics or a situation that useful topics are not visible. To solve these problems, we introduce semantic connection and use a density-based clustering algorithm (Rodriguez and Laio 2014) to acquire topics according to semantic connection features (SCF). Such selection can bring two benefits: an adaptive number of topics and automatic removals of outliers. Moreover, as the clustering is based on SCF, the scale of a cluster indicates the topic strength. This means that the number of topics is determined by topic strength. Furthermore, semantic connection features make visualization possible, because topic evolution can be depicted in a two-dimensional figure when setting the number of features as 2.

Second, there are two strategies of extracting topics: a simple topic model like LDA (Hindle, Godfrey, and Holt 2009) and an incremental model (Hu, Sun, and Li 2015). However, the former results in independence between topics of different time slices because topic-word distributions are different in different time slices, which raises an open problem on how to connect topics across successive time slices. Directly linking topics which are similar is not convincing when topic-word distributions are different. The latter maintains consistent topic-word distributions when adding data of the next time slice, but it is insensitive to changes. Given these observations, we choose the LDA model to obtain topics instead of an incremental model in order to guarantee the rapid response capability of both the vanishment of old topics and occurrences of new topics. As for the problem of independence between topics of different time slices, we use the sliding window technique and model LDA for every time window for consistent topic-word distributions.

Copyright © 2017, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Semantic Connection

The semantic connection (SC) of documents indicates both content similarity and time decay. More specifically, semantics here represent what topics the documents involve and what proportions of those topics, and the connection not only shows whether there are links between documents but can also spread semantic similarity across successive time slices. The main idea of SC is extracting topics as semantics by LDA and linking the documents that are similar enough by a sliding window to express time decay.

Algorithm 1 Semantic Connection

Require:
data set D, window size wz, threshold T.
Output:
semantic connection graph (SCG) and semantic connection features (SCF).
1: Divide D into D=$\{D_1, D_2, \cdots, D_n\}$ according to the granularity of time slice.
2: **for** w=1 to $n - wz$ **do**
3: $W = \{D_w, D_{w+1}, \cdots D_{w+wz-1}\}$;
4: $\theta_W = LDA(W)$;
5: **for all** $d_i \in W$ **do**
6: $kl = KL_{avg}(\theta_{wd_i}, \theta_{wd_{i+1}})$;
7: **if** $kl < T$ **then**
8: Add an edge between d_i and d_{i+1} into SCG.
9: **end if**
10: **end for**
11: $SCF = DeepWalk(SCG)$
12: **end for**
13: **return** SCG, SCF

Steps of the semantic connection algorithm are represented in Algorithm 1. This algorithm consists of three main components: firstly, acquire semantics via LDA (Lines 3-4),

secondly calculate the similarity of documents on the basis of topic-document distributions θ (Lines 5-10), and we use the symmetrical KL divergence $KL_{avg}(P, Q)$ to calculate the similarity because it is a common measure. Thirdly, transform a semantic connection graph (SCG) into semantic connection features (SCF) by DeepWalk (Perozzi, Al-Rfou, and Skiena 2014) which is unsupervised deep learning for acquiring the representations of topology (Line 11).

Experimental Results

To validate the effectiveness of our method, we have compared our method against a state-of-the-art baseline which acquires content evolution via Online Latent Dirichlet Allocation (On-Line LDA) referred in (Hu, Sun, and Li 2015) with data sets in real applications[1].

Due to space limitations, we select several topics which are marked by big white dots (others are marked by red dots) to demonstrate the effectiveness of our method, and the key words of those topics are shown in Figure 1. Compared with results of the baseline which are shown in Table 1, our method could handle situations where the baseline fails to distinguish similar topics (in Table 1, the five topics detected on December 4 indicate the same event). Meanwhile, our results demonstrate good adaptability to the number of topics, which manifests itself as the different number of different time slices. Furthermore, regarding the sensitivity of changes, the baseline detects protests on December 9 but fails to reveal that Apple is hiring a social media guru in L.A. for warming up social media. In contrast, our method performs well in detecting the emergence of new topics, and Topic 8 and Topic 6 in Figure 1 indicate those two events respectively. Moreover, another advantage of our method is that semantic disparities between those two topics are directly represented as the distances in Figure 1.

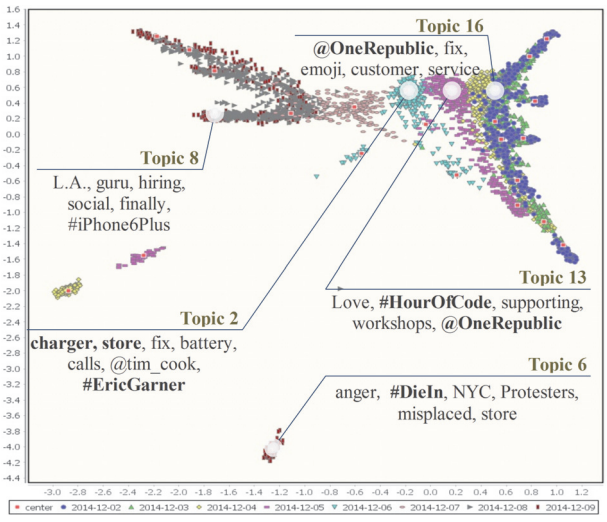

Figure 1: Evolution in Twitter Data

Table 1: Partial Results of Baseline in Twitter Data

2014-12-03	2014-12-04
business, android, release	⇒ 4, 45000, outlet, android
ipad, startup, macbook	⇒ batteries, 45000, outlet
future, stevejobs, ipod	⇒ future, outlet, 45000
note, close, switch	⇒ 4, 45000, outlet, photo
love,amazon, post	⇒ computers, outlet, 45000
2014-12-04	2014-12-05
video, jobs, unlock	⇒ love, workshops, pay
2014-12-08	2014-12-09
store, itunes, watch	⇒ store, protests, anger
ipod, lawsuit, big	⇒ stage, misplaced, iPod

Conclusions

We proposed a visual method based on semantic connection to solve defects caused by presetting the number of topics and extracting topics via topic modeling.

Preliminary experimental studies reveal the effectiveness of our method with Twitter data, including visualizing topic evolution, determining the number of topics adaptively, avoiding redundant topics by semantics, and reflecting changes quickly.

Acknowledgments

This research has been supported by the National Key Research and Development Program of China under grant 2016YFB1000901, and the National Natural Science Foundation of China (NSFC) under award 61503114.

References

Hindle, A.; Godfrey, M. W.; and Holt, R. C. 2009. What's hot and what's not: Windowed developer topic analysis. In *25th IEEE International Conference on Software Maintenance (ICSM 2009), September 20-26, 2009, Edmonton, Alberta, Canada*, 339–348.

Hu, J.; Sun, X.; and Li, B. 2015. Explore the evolution of development topics via on-line LDA. In *22nd IEEE International Conference on Software Analysis, Evolution, and Reengineering, SANER 2015, Montreal, QC, Canada, March 2-6, 2015*, 555–559.

Perozzi, B.; Al-Rfou, R.; and Skiena, S. 2014. Deepwalk: online learning of social representations. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, 701–710.

Rodriguez, A., and Laio, A. 2014. Machine learning. clustering by fast search and find of density peaks. *Science* 344(6191):1492–6.

[1]https://www.crowdflower.com/data-for-everyone/