Topic Detection and Tracking in Social Media Platforms

Riccardo Cantini, Fabrizio Marozzo

University of Calabria

Email: rcantini@dimes.unical.it

EAI International Conference on Pervasive knowledge and collective intelligence on Web and Social Media
EAI PerSoM 2022, November 17-18, 2022

Outline

- Introduction to Topic Detection and Tracking (TDT)
- State-of-the-art techniques and topic chains
- Main limitations
- Proposed methodology
- Experimental results
- Final remarks

Topic Detection and Tracking (TDT)

Motivation and goals

- The large amount of information available on the Web can be effectively leveraged to keep up with the latest news around the world.

- Traditional keyword-based techniques make it difficult to understand what has been happening over an extended period.
  - They do not organize retrieved information
  - Lack of semantics

- This issue can be overcome by using **Topic Detection and Tracking** (TDT) systems, which provide automated techniques for organizing large amounts of news streams in a way that helps users quickly interpret and analyze relevant information over time.

Topic Detection and Tracking (TDT)

Main concepts

- Main goal of a TDT system: ***detect** the main topics of discussion from stream of news in different formats, **tracking** their evolution over time without human intervention*.

- Key concepts
  - *Event*: something occurring at a precise time and place
  - *Activity*: groups of events with the same purpose
  - *Topic*: seminal event or activity, together with all closely related events and activities
  - *Story*: multimodal source of information (social media post, newspaper article, radio, tv broadcast, …)

Topic Detection and Tracking (TDT)

Main tasks

- **Story segmentation**: divide a multimodal stream of input data into stories.
- **First story detection**: recognize, within a flow of chronologically ordered stories, a story that deals with a new topic for the first time.
- **Topic detection**: identify the topics discovered through first story detection, by clustering stories related to the same topic.
- **Topic tracking**: identify stories related to a specific topic, given a stream of input stories.
- **Story link detection**: determine whether two given stories deal with the same topic or not.

Main approaches to the realization of TDT systems

Clustering-based

- Identify a **topic-based clustering structure** that represents a significant grouping of the processed documents, generally represented within a vector space.

- **Agglomerative hierarchical clustering**
  - Find topics at different levels of granularity, by identifying a topic hierarchy.
  - Assign a story to different topics, located at different level of the clustering structure.

- **Single-pass clustering**
  - Simplicity, high efficiency and low cost make it suitable to process a large amount of data.
  - An online clustering algorithm is used to build a clustering structure incrementally.
  - A sliding window can be introduced to address the topic drifting issue.

Main approaches to the realization of TDT systems

Semantic-based

- Represent documents using **semantic classes**, i.e., classes of terms with similar meaning.
  - *Names*: express the subjects involved in an event.
  - *Terms*: express the occurrence of an event (nouns, verbs, and adjectives).
  - *Time expressions*: represent points mapped on a time axis.
  - *Places*: indicate the places involved in an event.
- **Class-based comparison**
  - *General Term Weight*: a weight for each term is computed based on its occurrences and position in the document.
  - *Temporal Similarity*: temporal information (e.g., the distance between a news and the related first story) is used to scale the similarity values.
  - *Spatial Similarity*: a geographical ontology is used in order to measure the similarity of the spatial references present in the documents.

Main approaches to the realization of TDT systems

Probabilistic topic modeling

- A stream of news is analyzed in an online fashion, by dividing it in windows (mini-batches).

- **Online topic modeling (OLDA)**
  - A topic model is constantly updated by using mini-bathes of data, i.e., subsequent windows of a stream.
  - The current topic model is used as a prior for Latent Dirichlet Allocation (LDA) at the successive time slice, when a new window of the data stream is available for processing.

- **Topic chain**
  - It builds a temporal organization of similar topics that appear within an analyzed period.
  - Topics are extracted from subsequent time slices using Latent Dirichlet Allocation (LDA).
  - A set of chains is derived which link similar topics through time, based on their similarity.

Topic chain

Main elements

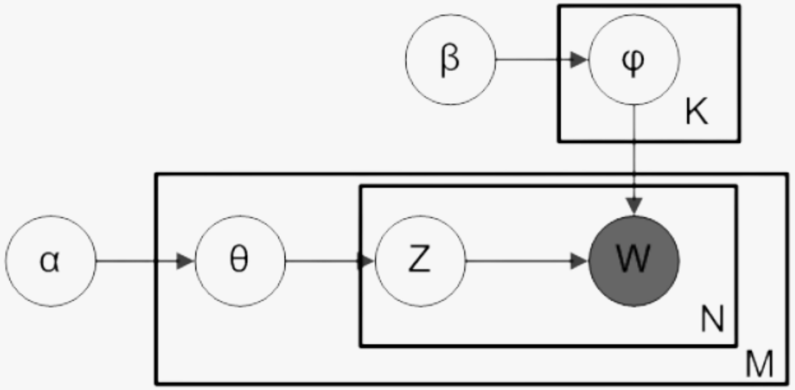
- A **topic chain** is a temporal organization of similar topics, detected from a stream of news, which is useful to understand how detected topics emerge, evolve, and disappear over time.

- Key concepts
  - *Long-term topic*: it is a general topic, present in social media conversation or online news over a long period of time, such as the discussion about Covid19 pandemic.
  - *Temporary issue*: it is a specific topic, that is talked about for a short period of time. It can be related to sporadic events like manifestations, or a part of a broader topic.
  - *Focus shift*: it is the change over time of the particular aspect of a long-term topic on which news about that topic is focused. As an example, contagion-prevention rules and vaccination within the general topic of Covid19.

Topic chain

Topic discovery

- The stream of news is analyzed in a windowed fashion, dividing it into subsequent time slices.
- For each time slice, a set of topics is extracted by using **Latent Dirichlet Allocation** (LDA).
- Main concept regarding LDA
  - A topic is a multinomial distribution over the words of the corpus vocabulary.
  - A document is a mixture of the *K* latent topics underlying the corpus.
  - The number of latent topic *K* must be given in input to LDA.



Topic chain

Choice of the topic similarity measure

- A topic can be seen as a multinomial distribution over the words of the vocabulary, a ranked list of words, or a vector in which each word of the vocabulary is associated to that topic with a certain probability.
- Metrics for **topic similarity**
  - *Cosine similarity*: it measures the cosine of the angle between two n-dimensional vectors.
  - *Jaccard coefficient*: it measures the cardinality of the intersection divided by the cardinality of the union.
  - *Kendall's τ coefficient*: it is a non-parametric measure of the rank correlation between two sets of items.
  - *Discounted cumulative gain*: it penalizes highly relevant documents appearing at a low position in the ranking.
  - *Kullback-Leibler divergence*: it measures the dissimilarity, in probabilistic terms, between two given distributions.
  - *Jensen-Shannon divergence*: it is the symmetric version of the KL divergence between two given distributions, obtained as the average divergence from their mixture distribution.

Topic chain

Chain construction

- Given the topic distribution $\phi^t = \phi_1^t, \dots, \phi_k^t$ at the current time t, the construction process proceeds as follows:

  1. The topic distribution at the previous time $t-1$, i.e. $\phi^{t-1} = \phi_1^{t-1}, \dots, \phi_k^{t-1}$ is found and the similarity between ϕ_i^t and each topic ϕ_j^{t-1} is computed.
  2. For each pair (ϕ_i^t, ϕ_j^{t-1}) such that their similarity is greater than a threshold, a link between them is created and the process moves to the next topic ϕ_{i+1}^t.
  3. If no link was created by comparing ϕ_i^t with the topics of ϕ^{t-1}, a comparison is made with topics of ϕ^{t-2}.
  4. This process is iterated backward, until at least one connection is found, or the size of the time window is exceeded.

- Note that if a divergence measure is directly used (see step 2), such as *JS divergence*, the measured value between the two considered topics must be less than the threshold for a link to be created.

Topic chain

Chain analysis and interpretation

- The obtained chains are analyzed in order to find:
    - Long-term topics such as politics and economics.
    - Temporary issues related to specific events and topics that do not last for a long period.
    - Focus shifts in long-term topics.

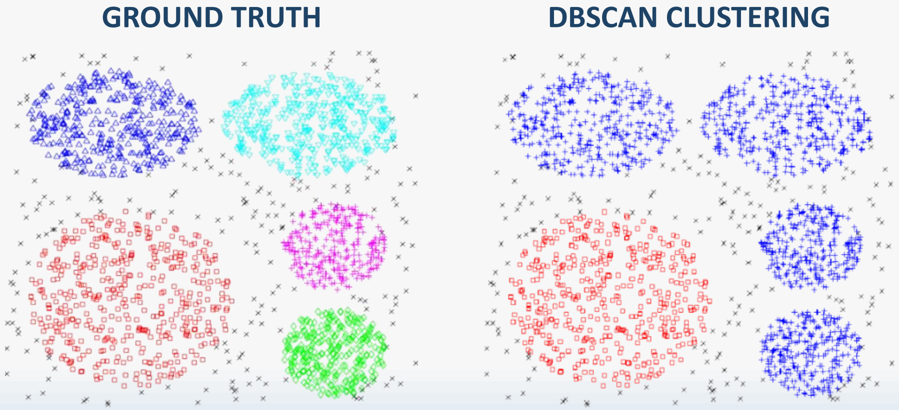
- This step involves an interpretation phase in which the data provided by the topic chains must be understood in relation to real-world events and trending topics of discussion.

Main limitations of topic chain

Resolution-based issue

- It is difficult to detect links at different probability levels and at an arbitrary distance in time, due to the joint action of the threshold and window size.
  - Connections between time slices distant from each other need a wide time window. Consequently, they may not be isolated but included in broader chains along with other noisy connections.
  - Lowering the threshold to remove noisy connections can cause the loss of weaker links, which does not allow to find chains at a lower probability level.

- This is a **resolution-based** issue that also present in other application domains, such as density-based clustering.
  - The *DBSCAN* algorithm struggles to identify a global clustering structure composed of clusters at different density levels.



Proposed methodology

Length-weighted topic chain

- The time window is removed, potentially allowing connections between topics distant from each other.
- Connection probability does not go to zero instantaneously, when the fixed size of the window is exceeded, but decreases smoothly as the length of the chain increases.
- This effect is obtained by dynamically adjusting the threshold using an exponential decay mechanism.

❖ Let's consider the following elements:

- *Topic pair:* $({\phi_i}^t, \phi_j^{t'})$*, with* $t > t'$
- *Initial value of the threshold:* th_0
- *Jensen-Shannon divergence:* JS
- *Constant decay factor:* λ

❖ The threshold undergoes an exponential decay based on the current length L of the chain:

$$th_L = th_0 \cdot e^{-\lambda L}$$

❖ A link between ${\phi_i}^t$ and $\phi_j^{t'}$ is added to the topic chain if:

$$JS({\phi_i}^t, \phi_j^{t'}) \le th_L$$

Proposed methodology

Length-weighted topic chain

- Our approach enables the identification of high-quality links.
  - Topics belonging to time slices distant from each other can be connected.
  - Connections at different probability levels can be extracted.
  - An overall reduction of noise in the discovered chains is achieved.

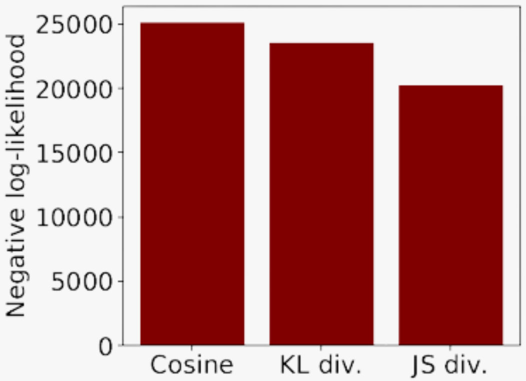
- Connection probabilities are dynamically adjusted.
  - As the length of the chain increases, a current topic encounter greater resistance in forming a connection.
  - The chain is extended only if the candidate link is significant enough to overcome this resistance.
  - Otherwise, the process iterates backward trying to link that topic to another in an earlier time slice.
  - In that case, the topic will be connected to a shorter chain which is thus forked into a new, separate one.

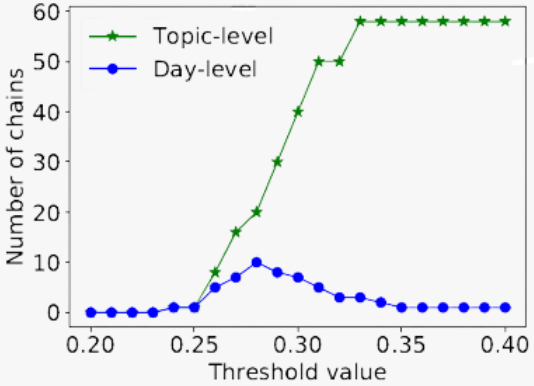
The 2016 US Presidential Election

Experimental setting

- This corpus comprises about 2.5 million tweets, posted by 521, 291 users regarding the **2016 US presidential elections**, published from October 10, 2016, to November 7, 2016.
  - We only considered tweets published in the main US swing states, characterized by high political uncertainty.
  - In this way we obtained a politically-balanced corpus, containing pro-Clinton and pro-Trump posts.

- **Hyper-parameter tuning**
  - The best measure was the JS divergence, which minimizes negative log-likelihood.
  - We selected a threshold value $\mathrm{th}_0 = 0.28$
  - This value is a good trade-off between the number of chains at day and topic level.





The 2016 US Presidential Election

Discovered chains

- *Macro-topic*: **Sexism**
- *Connected days*: 11, 12, 13, 15, and 16 October
- *Micro-topics*:
  1. Trump was criticized for his sexually aggressive comments, which he justified by defining them locker room talk.
     - *words*: locker, room, talk　　*hashtags*: #nevertrump
  2. Trump's supporters posted content in favor of repealing Amendment 19, which grants women the right to vote.
     - *words*: sexism, women　　*hashtags*: #women, #repealThe19, #nevertrump
  3. Trump was compared to Mary Shelley's Frankenstein in a derogatory way.
     - *words*: women, inappropriate, predator　　*hashtags*: #frankentrump

The 2016 US Presidential Election

Discovered chains

- *Macro-topic*: **Disputes over the Clintons**
- *Connected days*: 17, 18, 22, 23, and 24 October
- *Micro-topics*:
    1. Disputes related to Hillary Clinton’s six-years tenure as a director of Walmart, following the leak of Podesta's emails.
        - *words*: walmart, board *hashtags*: #corrupt, #podestamails
    2. Hillary Clinton was accused of being supported by the American elite, pursuing the interests of a few influential people.
        - *words*: elite *hashtags*: #neverhillary
    3. The connection between Bill Clinton and Jeffrey Epstein, a millionaire accused of sexual abuse and child trafficking.
        - *words*: pedophile, island *hashtags*: #lockherup, #draintheswamp

The 2016 US Presidential Election

Discovered chains

- *Macro-topic*: **Support from prominent public figures for Hillary Clinton**
- *Connected days*: 28, 29 October, and 1, 2 November
- *Micro-topics*:
  1. Michelle Obama supported Hillary Clinton's candidacy by speaking at the rally held by Clinton on October 27 in Winston-Salem, North Carolina.
     - *words*: women, rally *hashtags*: #imwithher, #strongertogether
  2. Support from senator Jeanne Shaheen.
     - *words*: hillary *hashtags*: #senatorshaheen, #imwithher
  3. The billionaire Richard Brandson sided with Hillary Clinton, releasing an interview in which he criticized Trump's violent temper, calling him irrational and aggressive.
     - *words*: richard, branson, quote, trump *hashtags*: #branson, #imwithher

The 2016 US Presidential Election

Discovered chains

- *Macro-topic*: **Trump's rhetoric**
- *Connected days*: 26 and 27 October
- *Micro-topic*: the republican candidate was criticized for his rhetoric, often considered violent, homophobic, and racist, following also Richard Branson interview.
  - *words*: rhetoric, violent, trump *hashtags*: #voteblue

- *Macro-topic*: **US elections and propaganda**
- *Connected days*: 3 and 4 November
- *Micro-topics*: pro-Clinton and pro-Trump supporters published content in favor of the two main candidates.
  - *words*: election, clinton, trump *hashtags*: #maga, #votehillary, #vote, #vote2016, #election2016

Coronavirus Pandemic (Covid19)

Experimental setting and discover chains

- This case study analyzes the tweets published in December 2020 related to the Covid19 pandemic.

- Discovered chains span the entire month under consideration, covering almost every day, with no strict boundaries.

- Discovered topic chains (Macro-topics):
  - **General conversation about Covid19**, *words/hashtags*: global, covid, #covid19, #coronaviruspandemic.
  - **Anti-contagion protocols**, *words/hashtags*: prevention, #washyourhands, #socialdistancing, #wearamask.
  - **Remote job**, *words/hashtags*: work, job, #workfromhome, #wfh, #remotejob.
  - **Vaccination and medical personnel**, *words/hashtags*: vaccine, #vaccine, #covidvaccine, #healthcare, #frontlineheroes.
  - **Christmas holidays during the pandemic**, *words/hashtags*: christmas, #covidchristmas, #christmas2020.

Final remarks

- The proposed methodology leverages an exponential decay mechanism to dynamically adjust topic connection probability.
- Main benefits:
  - It overcomes the resolution-based issues of the original topic chain model, allowing the identification of links at different probability levels, between topics located at any point within the global time axis.
  - Discovered chains are less noisy and quite coherent, with no conflicting topics within the same chain, which is desirable in the case of politically-oriented news.
- These good properties do not hold when applying the original approach:
  - As an example, the first chain about sexism is merged with the first part of the second chain, about Walmart.
  - Due to this, the resulting chain, characterized by anti-Trump themes, is polluted by a topic against Clinton.
  - To avoid the noisy links, the cut value for JS divergence should be lowered, but this would result in the loss of other significant links and small chains, such as the one referring to Trump's violent rhetoric.

thank you
gracias
merci
danke
obrigado
bedankt
dziękuję
go raibh maith agat
teşekkür ederim
ngiyabonga
謝謝
спасибо
рахмат
Баярлалаа
sukriya
kop khun krap
terima kasih
감사합니다
ありがとう
arigatô
grazie
dakujem
mochchakkeram
tapadh leat
takk
мерси
ευχαριστώ
dank je
mersi
barka
welalin
tack
spas
kia ora
blagodaram
misaotra
matondo
paldies
grazzi
mahalo
хвала
asante
manana
obrigada
tenki
murakoze
chokrane
mamnun
trugarez
shukriya
merce
dhanyavadagalu
diolch
xiexie
rahmet
tanemirt
taiku
sulpáy
grácies
gratias ago
chnorakaloutioun
akun dankon
aciú
дякую
djiere dieuf
tau
sagolun
najis tuke
kam sah hamnida
didi madloba
rahmat
mési
sobodi
dekuji
hvala
maururu
köszönöm
gracie
bayarlalaa
enkosi
nanni
nandri
kiitos
dhanyavad
dankie
faafetai lava
vinaka
спасибі