* اهمیت مهم بودن یک خبر برای کاربر و موتورهای جستجو

The amount of news content on the Web is increasing, enabling users to access news articles coming from a variety of sources: from newswires, news agencies, blogs, and at various places, e.g. even within Web search engines result pages. Anyhow, it still is a challenge for current search engines to decide which news events are worth being shown to the user (either for a newsworthy query or in a news portal). Being able to predict event impact will, for example, enable a newspaper to decide whether to follow a specific event or not, or a news search engine which stories to display [1].

* معیارهای انتخاب خبر(new worthy)

According to news factors theory news is about people, preferably members of the power elite, elite institutes, and elite countries. Other news factors that are considered important in news selection are: the presence of a negative angle, the amount of drama, surprise, frequency, continuity, etc. [1]

Very simple measures of impact can be imagined as, for example, the duration of an event (i.e., the time between the first and the last news about this event). More interesting and complex impact definitions can be rendered as well. For example, a definition could try to model the visibility of an event to the general public (e.g., number of news article reads), to measure how strong are the effects of an event (e.g., measure the number of deaths), or, by means of blog postings written about the event, which controversial discussions an event triggers in the public opinion (e.g., positive vs negative views).

به‌طورکلی، اهمیت یک خبر هم در روزنامه ها تاثیرگذار است، مثلا، یک روزنامه مطلبی را در تیتر اصلی عنوان میکند، برخی مطالب را در صفحه اول و برای برخی مطالب نیز اهمیت زیادی قائل نیست.

* اینکه چرا اصلا برای یک رخداد، زیر رخداد به وجود میاد

As global connectivity continues to grow, the worldwide relevance of local events has become much more significant. Events that decades ago would be isolated news reports, now have the potential to create a cascading effect resulting in global reaction[2].

* اهمیت فیلتر کردن خبر از منابع مختلف:

Since people are increasingly relying on multiple online sources of information, it is important to support users in filtering news automatically. In this work, we consider the problem of anticipating news story importance, i.e. given a news item, predicting if it will be of interest for a majority of users. Such ranking is currently done manually for newspapers, and we explore automatic approaches and indicative features for the same. Our main conclusion is that importance prediction is a hard problem, and pure textual features are not sufficient for classifiers with 90% accuracy[3].

* موضوع story link detection

Topic Detection and Tracking (TDT) research is sponsored by the DARPA TIDES program. The research has five tasks related to organizing streams of data such as newswire and broadcast news (Wayne, 2000). A link detection (LNK) system detects whether two stories are “linked”, or discuss the same event. A story about a plane crash and another story about the funeral of the crash victims are considered to be linked. In contrast, a story about hurricane Andrew and a story about hurricane Agnes are not linked because they are two different events. A new event detection (NED) system detects when a story dis- cusses a previously unseen event. Link detection is considered to be a core technology for new event detection and the other tasks[4].

* تعریف task های مختلف در حوزه TDT :

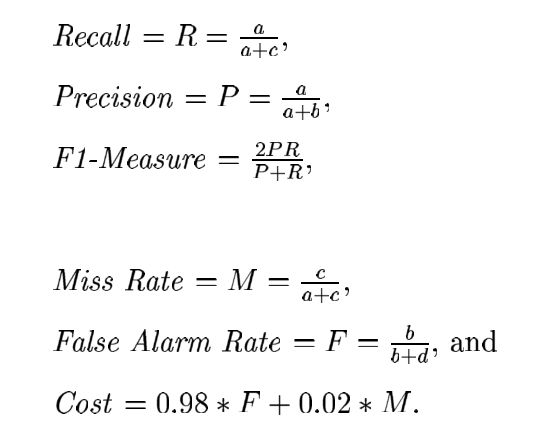
Topic Detection and Tracking (TDT) is an integral part of the DARPA Trans lingual Information Detection, Extraction, and Summarization (TIDES) program [2]. TDT mainly contains two sub-task: Topic Detection and Topic Tracking. Topic detection (Event detection) aims at detecting novel topics/events from text corpus while topic tracking is dedicated to tracking the evolution of existing topics over temporal dimension. Topic detection has attracted much attention in machine learning, information retrieval and social media modeling [1,3,6,9,12–20]. Specifically, topic detection can be classified into two types: New Event Detection (NED) and Retrospective Event Detection (RED)[5].

* معروف تر بودن Traching:

This task has been the most popular task within TDT evaluations, with more sites participating in it than in any other task. Its attractiveness is probably the result of its being similar to an existing task, and possibly also because it is much easier to work with multiple on-topic stories than to try to guess the topic from a single story[6].

* ارزیابی task های Detection :

|  |  |
| --- | --- |
| C:\Users\Mostafa\Desktop\normalized detection cost.JPG  0.02مخرج= | C:\Users\Mostafa\Desktop\detection cost.JPG |



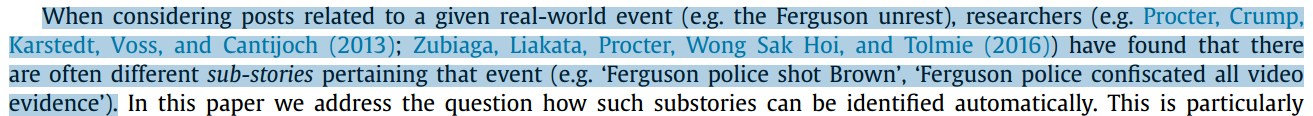
* تعریف tracking و new event detecion در مقاله خوب james allan

The problems discussed in this study are new event detection and event tracking. The goal of those tasks is to monitor a stream of broadcast news stories so as to determine the relationships between the stories based on the real-world events that they describe. New event detection requires identifying those news stories that discuss an event that has not already been reported in earlier stories. Event tracking means starting from a few sample stories and finding all subsequent stories that discuss the same event[7].

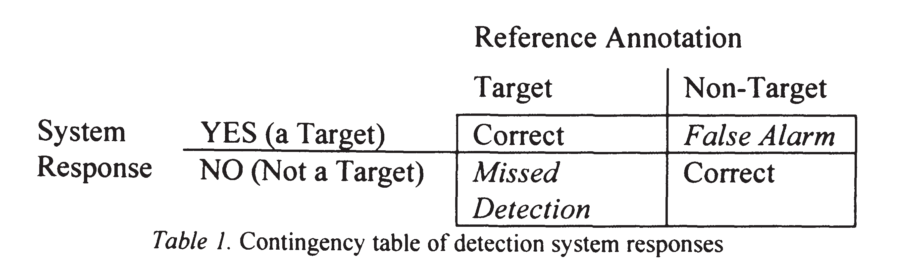
* مقالات خوب

|  |  |
| --- | --- |
| موضوع event detection و tracking  تعریف Event و مسائلی که در تعریف آن وجود دارد | On-line new evetn detection and tracking |
| یه تعریف متفاوت از topic و tracking | Topic detection, tracking, and trend analysis using self-organizing neural networks |
| نگاه فراتر به موضوع. عالی برای مقدمه و انگیزه | Topic detection and tracking: event clustering as a basis for first story undestanding |
| تعریف newEvent و نیز تعریف Event و ویژگی هایی که یک event رو از بقیه مجزا میکند. | On-line new event detection using single pass clustering[8] |

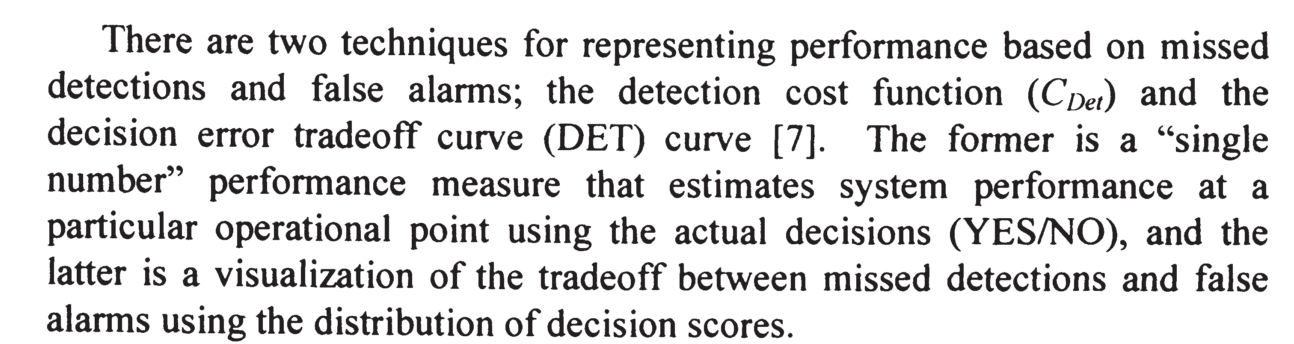
* Sub event با کلی منبع خوب و نیز یه رفرنس خوب از IPM[9]



* ارزیابی

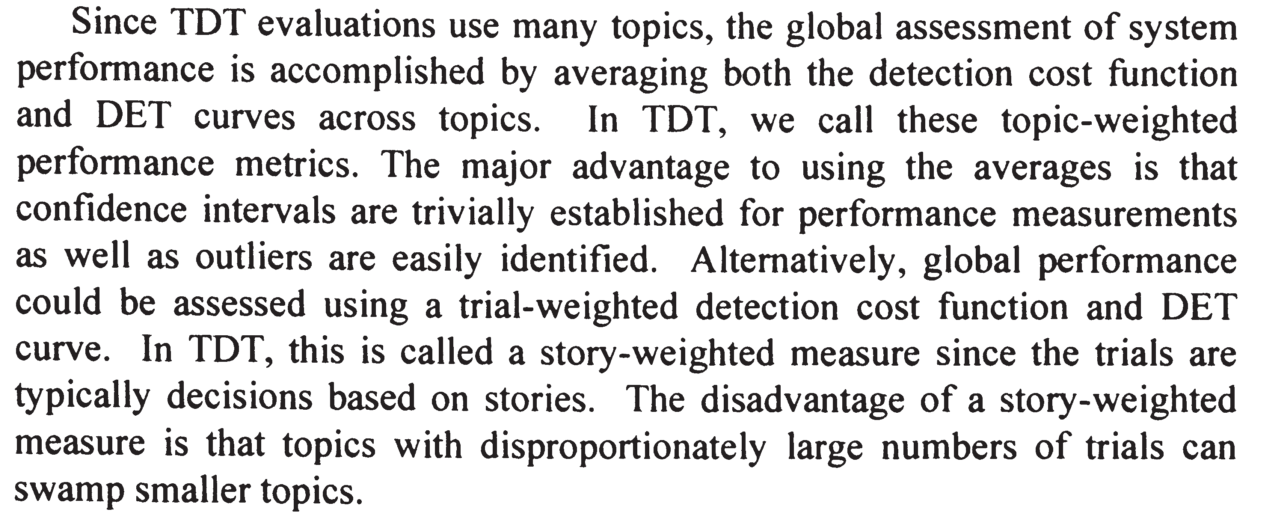


تصویر از کتاب Topic Detection and Tracking - Event-based Information Organization



این تصویر از همان(همانطور که در [10] هم آمده است.)

مزایا و معایت دو نوع ارزیابی هم اینجا مطرح شده (منبع همان قبلی)



* Subevent که تاحالا بیشتر روی twitter کار شده و کم روی general کار شده و اینکه ما میخوایم روشی ارائه بدیم که عمومی تر باشه[11]:

Subevents have been studied for event tracking applications (Shen et al., 2013; Meladianos et al., 2015). However, ک like Twit- ter, in terms of both its definition of subevents and methodologies. For example, in previous research by (Shen et al., 2013), a subevent is defined as a topic that is discussed intensively in the Twitter stream for a short period of time before fading away. Accordingly, the subevent detection method relies on modeling the “burstiness” and “cohesiveness” properties of tweets in the stream. We instead aim to provide a more general definition of subevents as well as present a method for identifying subevent at the article level.

* در [8] از time به عنوان روشی برایthresholding استفاده شده است. من از time diff به عنوان یک feature استفاده کردم.
* روشی برای برچسب زدن و ساخت judgment در [8] آمده است. دو گروه مجزا برجسب میزنند. در نهایت برای ساخت برچسب نهایی، reconciled صورت میگیرد.
* یک motivation خیلی خوب از جای خوب و مقاله خوب برای اینکه یه خبر عموما تکراری از خبرهای گذشته و ارائه یه بخش خیلی کوچیک از بقیه خبره[12]:

Although the Web offers a seemingly large and diverse set of information sources ranging from highly curated professional content to social media, in practice most sources base their stories on previously published works and add a much more limited set of new information. Thus users often end up spending significant amount of effort re-reading the same parts of a story before finding relevant and novel information.

* کار من اینه تقریبا:

Our approach works in an online fashion and provides previously unseen information related to predefined ad-hoc information need, expressed as a user query.[12]

* در عین حال:

The temporal summary of news stories can help a person monitor changes in the coverage of news stories over time, which are typically very redundant and increase the effort required to identify genuinely new information.

* تعریف از جمله novel:

Two general criteria to select the best candidates sentences are the most useful and novel sentences, i.e., related to the topic and non-redundant.

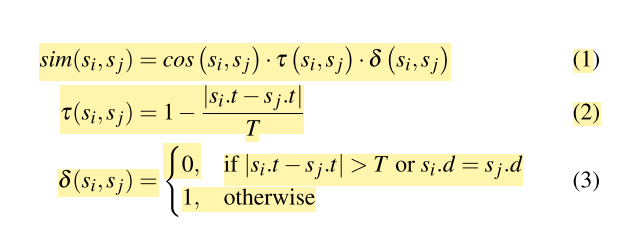
توصیف کلی تکراری بودن خبر در رسانه های مختلف:

the instant they are published; or, alternatively, to present the user a summary of the most important news facts over a timeline. Additionally, presenting the most important news facts on a timeline may also be useful to help keep update knowledge bases up-to-date, such as Wikipedia or the knowledge graphs used by search engine companies. From an end-user perspective, we consider it important that a high percentage of results is on-topic, and therefore this study uses news articles as the sole source. We expect their content to be mostly factually correct, timely, and presented in an accessible form. Events that are of interest to many people are naturally reported in different news articles, from different sources. In our approach, we leverage the redundancy between news articles, clustering sentences that are likely to discuss the same news facts to select salient sentences and to avoid biased information.

اهمیت publish time

The latter is primarily a practical argument when developing an online news summarization approach. The use of proximity in publication times is motivated by the observation that stories about the same event often occur in proximate time, most particularly for unexpected events where the news media exhibit strong interest in a story .

فرمول شباهت از [12]:



شکل 1: منظور از d ، خبرگزاری است.

مقاله ی دیگه ای که کار شبیه به ما انجام داده[13]

We shall present methods for identifying information novelty and show how these methods can be applied to manage content that evolves over time.

برچسب زدن به اخبار[13]:

When we debriefed several users who completed the experiments described in Section 4, several of them reported that it was difficult to judge the novelty of articles because of their varying relevance to the seed story. In some cases this had to do with errors in the tagging of the news stories by the newsfeed we relied upon, while in one or two extreme cases a failure of the Moreover parser caused grossly unrelated articles to be glued to the relevant ones. In other cases the variance in relevance was due to the differences in writing styles and policies among different publications.

These comments led us to believe that the novelty scores we compute should not be relied upon as a sole selection criterion; some articles are identified as novel by virtue of changing the topic. To further refine the analysis of informational novelty, we have formulated a classification of types of novelty, based on different relationships between an article and a seed story. These classes of relationships include:

1. **Recap** articles are those that are clearly relevant, but only offer reviews of what has already been reported and carry little new information.

2. **Elaboration** articles add new, relevant information on the topic set forth by the seed article.

3. **Offshoot** articles are also relevant to the mainstream discussion, but the new information they add is sufficiently different from that reported in the seed story to warrant a new trail.

4. **Irrelevant** articles are those due to clustering and parsing mistakes.

* دلیل انتخاب آنلاین بودن کار:[14]

The usage model that we envision requires that the technology produce a revised summary at regular time intervals—e.g., every hour or at the start of each day. It is neither possible nor meaningful to wait until the topic is done to produce a summary. Nor does it make sense to produce an up-to-date overall summary at every time interval: the summary must indicate only what has changed. After all, the user has already been informed about everything that happened earlier.

از منبع قبل:

تاکید بر sub- event داشتن درون یک سری از اتفاقات:

For this study, we assume that news topics can be broken into a sequence of events, and that it is those events that interest users. In that case, it is sufficient for a system to produce a list of the events within the topic in the order those events are reported. Better summaries will list the events in the order reported, will capture all of the events.

* شباهت TDT و temporal summarization

This work also arises out of Topic Detection and Tracking (TDT), a body of research and an evaluation paradigm that addresses event- based organization of broadcast news. TDT investigation has been carried out over five years by about a dozen academic and industrial research institutions, and explored in the context of four “cooperatively competitive” evaluations sponsored by the U.S. government [1, 7, 28, 29]. **The problems tackled by TDT are all story-based rather than sentence based. In many ways, the temporal summarization problem is an event- and sentence-level analogue of TDT’s “first story detection” problem, where the task is to identify the first story that discusses each topic in the news.**

* اشاره صریح به این موضوع که مفهوم جدید یه چیز سخته برای تعریف کردنش:

Time-based Summarization There has been very little work on time-based summarization to date. Some researchers have focused on how to extract temporal expressions from text, looking for and normalizing references to dates, times, and elapsed times [22]. That work is important for analyzing the content of the text, but does not directly address summarization itself. In the summer of 1999, the Novelty Detection workshop at Johns Hopkins University’s Center for Speech and Language Processing defined and explored new information detection (NID) [2]. **The NID task was to identify the onset of new information within a topic by flagging the first sentence that contained it. The NID task is obviously very similar to the time-based summarization work proposed here. The summer workshop was unable to make significant progress because of problems with the definition of “new”: when the team looked at an evaluation corpus they constructed, they discovered that 80% of the sentences were marked to contain new information. It turns out that almost every sentence in the news contains some new information—even if it is just the age of a person in the news**. In this research, we have chosen a looser definition of “event” that makes this less of a problem.

به همین دلیل مجبوریم از چیزای دیگه برای بررسی جدید بودن استفاده کنیم. مثل کامنت ها

* اینجا یه ارتباط ضعیفی با storyline extraction مطرح شده که باز بد نیست بیان بشه:

This research is also related to work on automatic timeline construction [38]. That work focused on using the χ2 measure to ex- tract unusual words and phrases from a stream of news, and on grouping those features to isolate topics within the news. They suggested the idea of looking within the topics to create an event- level timeline, but have not yet done so. Further, since the timeline work is driven by graphical visualization, it will not take the same form as this text-driven approach.

* خوب برای مقدمه[15]:

News forms a major portion of information disseminated in the world every day. Common people and news analysts alike are very interested in keeping abreast of new things that happen in the news, but it is becoming very difficult to cope with the huge volumes of information that arrives each day. Hence there is an increasing need for automatic techniques to organize news stories in a way that helps users interpret and analyze them quickly. This problem is ad- dressed by a research program called Topic Detection and Tracking (TDT) [3] that runs an open annual competition on standardized tasks of news organization. One of the shortcomings of current TDT evaluation is its view of news topics as flat collection of stories.

* ارتباط تویتر و اخبار[16]

Studies have shown that 85% of the tweets are news affiliated

* اصلا چرا داریم روی خبر کار میکنیم هنوز؟ تویتر که خیلی خفنه![17]

Social media has become a popular platform for publishing, sharing and consuming news. However, it is not a replacement for traditional sources of news — they are complementary. While news sites provide in-depth and comprehensive coverage of events and topics, social media postings include comments, opinions and rumors about facts publicized on the news. Social media can thus serve as a useful sensor for how popular a story (or topic) is, for how long, and people’s sentiments about it.

* شباهت و تفاوت با تسک near duplicate detection:

در تسک NDD فرض بر این است که قستمی از دو متن کاملا مشابه هستند، یعنی به فضای معنایی توجه نمشه و فرض بر این است که یک متن اصلی وجود دارد، و یک متن دیگر قسمتهایی از متن اصلی را کپی کرده و با افزودن بخشهایی به آن که این بخش ها میتواند شامل مطالبی باشد که عملا ارتباطی به متن اصلی ندارد و شامل تبلیغات و غیره باشد، متن جدیدی تولید کند. خیلی فصا به سمت semantic پیش نمیره. علاوه بر این، ممکنه که یه خبر 90 درصدش کپی باشه اما فقط در یک پاراگراف، مطلب جدیدی ذکر کرده باشه، در این صورت در تسک NDD به احتمال زیاد این متن Duplicate تشخیص داده میشه اما در تسک ما بستگی به عکس العمل کاربران داره☺

[1] J. Gaugaz, P. Siehndel, G. Demartini, T. Iofciu, M. Georgescu, and N. Henze, “Predicting the Future Impact of News Events,” in *ECIR*, 2012.

[2] R. M. Patton and T. E. Potok, “Identifying event impacts by monitoring the news media,” in *Proceedings of the International Conference on Information Visualisation*, 2008.

[3] R. Krestel and B. Mehta, “Predicting news story importance using language features,” in *Proceedings - 2008 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2008*, 2008.

[4] F. Chen, T. Brants, and A. Farahat, “Story Link Detection and New Event Detection are Asymmetric,” in *HLT-NAACL*, 2003.

[5] C. Zhang, H. Wang, L. Cao, W. Wang, and F. Xu, “A hybrid term–term relations analysis approach for topic detection,” *Knowledge-Based Syst.*, vol. 93, pp. 109–120, 2016.

[6] A. James, *Introduction to Topic Detection and Tracking*, vol. 12. 2002.

[7] J. Allan, R. Papka, and V. Lavrenko, “On-line New Event Detection and tracking,” in *Proceeding SIGIR*, 1998.

[8] R. Papka and J. Allan, “On-Line New Event Detection Using Single Pass Clustering TITLE2:,” 1998.

[9] P. K. Srijith, M. Hepple, K. Bontcheva, and D. Preotiuc-Pietro, “Sub-story detection in Twitter with hierarchical Dirichlet processes,” *Inf. Process. Manag.*, vol. 53, no. 4, pp. 989–1003, 2017.

[10] T. Brants and F. Chen, “A System for new event detection,” in *SIGIR*, 2003.

[11] A. Badgett and R. Huang, “Extracting Subevents via an Effective Two-phase Approach,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 906–911.

[12] J. B. P. Vuurens, R. Blanco, and P. Mika, “Online news tracking for ad-hoc Information Needs,” in *ICTIR ’15*, 2015, pp. 2–4.

[13] E. Gabrilovich, S. Dumais, and E. Horvitz, “Newsjunkie: providing personalized newsfeeds via analysis of information novelty,” in *WWW*, 2004, pp. 482–490.

[14] J. Allan, R. Gupta, and V. Khandelwal, “Temporal Summaries of News Stories,” *Proc. ACM SIGIR 2001*, pp. 10–18, 2001.

[15] R. Nallapati, A. Feng, F. Peng, and J. Allan, “Event threading within news topics,” in *Proceedings of the Thirteenth ACM conference on Information and knowledge management - CIKM ’04*, 2004, pp. 446–453.

[16] A. Mogadala, D. Jung, and A. Rettinger, “Linking Tweets with Monolingual and Cross-Lingual News using Transformed Word Embeddings,” 2017.

[17] S. Phuvipadawat and T. Murata, “Breaking news detection and tracking in Twitter,” *Proc. - 2010 IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol. - Work. WI-IAT 2010*, pp. 120–123, 2010.