Topic Tracking Using Chronological Term Ranking

Bilge Acun, Alper Başpınar, Ekin Oğuz, M. İlker Saraç
and Fazlı Can

Abstract Topic tracking (TT) is an important component of topic detection and tracking (TDT) applications. TT algorithms aim to determine all subsequent stories of a certain topic based on a small number of initial sample stories. We propose an alternative similarity measure based on chronological term ranking (CTR) concept to quantify the relatedness among news articles for topic tracking. The CTR approach is based on the fact that in general important issues are presented at the beginning of news articles. By following this observation we modify the traditional Okapi BM25 similarity measure using the CTR concept. Using a large standard test collection we show that our method provides a statistically significantly improvement with respect to the Okapi BM25 measure. The highly successful performance indicates that the approach can be used in real applications.

1 Introduction

News portal web sites deal with huge amount of data from different sources. As the number of sources and events increase, news-consumers are overwhelmed with too much information. Different organizational techniques have been employed for more effective, efficient, and enjoyable browsing [1]. Studies on new event detection and topic tracking aim to organize news with respect to events or topics. In this work, we study topic tracking (TT) which aims to find all news articles that follow an initial event/topic.

In topic detection and tracking (TDT) studies, an event is defined as something that happens at a given "place and time, along with all the necessary preconditions and unavoidable consequences," like a car accident. Topic is a connected series of events

B. Acun · A. Başınar · E. Oğuz · M. İ. Saraç (✉) · F. Can
Computer Engineering Department, Bilkent Information Retrieval Group,
Bilkent University, 06800 Ankara, Turkey
e-mail: ilker1486@gmail.com

E. Gelenbe and R. Lent (eds.), *Computer and Information Sciences III*,
DOI: 10.1007/978-1-4471-4594-3_36, © Springer-Verlag London 2013

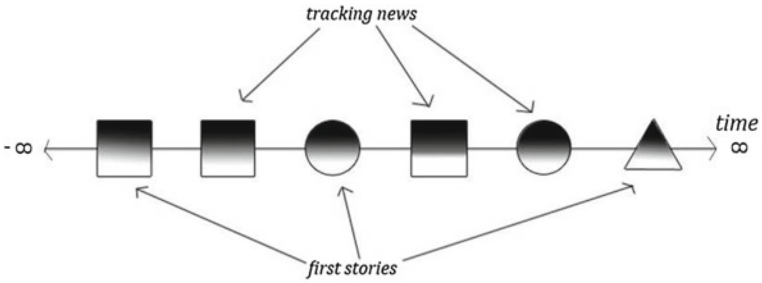

Fig. 1 Illustration of topic tracking (TT) and chronological term ranking (CTR), in the figure different *shapes* indicate stories related to different topics. *Darker gray* initial positions in each story indicate positions of more important words. CTR is based on the inverted pyramid metaphor that implies that most newsworthy information is provided at the beginning and as we go down to lower portions of a story importance of words gradually decreases

that have a common focus or purpose [1]. It is not a broad concept like "accidents," it is limited to a specific accident. In this study, we investigate the use of chronological term ranking (CTR) in TT to identify tracking stories of an event, i.e. we aim to find all tracking news articles of a topic based on a few number of sample initial news articles. To the best of our knowledge, prior to our work CTR approach has not been used for TT. This lack in literature, i.e., not having a CTR-based TT study in literature, is surprising since CTR is a natural fit to the TT problem domain. We use the Okapi BM25 (in short Okapi) similarity function in TT. It is a commonly used similarity function that gives good results in information retrieval and works well when paired with CTR [2]. Figure 1 illustrates the TT process and the CTR concept.

The contributions of this study are the following. (1) Our work extends the previous studies on TT by introducing the CTR approach to similarity calculation for identifying tracking stories in TT, (2) We experimentally show that term weighting component of Okapi can be altered by term position information (as suggested in [2]) such that its performance in TT can be statistically significantly improved, and (3) Successful results we obtain with the CTR approach show that the approach provides a performance that is compatible with those of previous studies [1] and can be used in practical applications with a high user satisfaction.

2 Related Work

Topic Detection and Tracking (TDT) studies were initiated in the second half of the 1990s by researchers from DARPA, Carnegie Mellon University, Dragon Systems and University of Massachusetts Amherst. Later some other groups also joined to the research initiative [3]. In TDT there are five tasks and they are New Event Detection (NED), Topic Tracking, Story Segmentation, Topic Detection (Cluster Detection), and Story Link Detection. Two of them, NED and TT, are more frequently studied in literature and are also studied in Turkish in our previous research [1]. In NED, one common approach is keeping a sliding time-window due to performance concerns [1, 4]. In this approach, each incoming news article is compared with the previous articles stored in the window. For each article pair, if similarity value of the new article is below a similarity threshold value (usually obtained by training) for all other articles then it is identified as new. In TT, initial story (or a few numbers of initial stories) is provided then all incoming subsequent stories are compared with that initial story on the basis of a threshold value. If similarity value is above the threshold this article is flagged as a tracking article (follower). All tasks were understood as detection tasks and evaluated using miss and false alarm error rates [3]. A Detection Error Tradeoff (DET) plot [5] is the primary tool for describing the tracking errors.

This study follows our earlier studies on new event detection and topic tracking in Turkish [1] and information retrieval on Turkish texts [6]. In the experiments we use a large standard test collection BilCol2005 from [1]. In order to have objective initial idf (inverse document frequency, explained later in Sect. 4) values, a retrospective corpus need to be used [4], for that purpose we have another data collection, *Milliyet* Collection from [6].

CTR concept is previously only used to enhance relevance scoring between documents in information retrieval [2]. This idea fits perfectly to news stories since news reporters write articles using inverse pyramid style which consists of writing most important words in the initial sentences. The work reported in [2] shows that the CTR approach works well when paired with Okapi; however, as indicated earlier has not been used in topic tracking applications.

3 Test Collection and Topic Tracking Algorithm

We use two test collections which are BilCol2005 [1] and *Milliyet* Collection [6] in our TT system. BilCol2005 is a large TDT collection that contains 209,305 news stories and 80 topics spanning through 12 months in 2005. In our experiments the topics and stories of the first 8 months (141,910 articles containing several topics 50 of them have been annotated) in the collection are used for training and the remaining 4 months (67,395 articles in 30 topics) are used for testing. Details about the BilCol2005 collection can be found in [1]. *Milliyet* Collection contains 408,305 news articles from *Milliyet Gazetesi* between 2001 and 2004. We use *Milliyet* Collection to calculate idf values. Using these idf values, we start our experiments in an independent unbiased environment for BilCol2005. The details about the *Milliyet* Collection can be found in [6].

The TT algorithm uses a few number of sample stories about a given topic and aims to find the tracking stories for that particular topic in a news stream. In the literature, usually between 1 and 4 documents are used as sample stories [1].

We employ the traditional TT algorithm [1]. In the algorithm a similarity function is used to determine if an article of the news stream that follows the sample story is a tracking story or not. If the similarity value is higher than the threshold value obtained during training it is classified as a tracking story. During training, for each training topic we calculate miss and false alarm rates using a threshold sweep within a range of similarity values (in our case it is between 0 and 121.1). The threshold values for the individual training topics that make the corresponding C_{det} value minimum are determined; where C_{det} signifies TT cost and calculated as a function of miss and false alarm rates [7] (defined in Sect. 5). The average of these threshold values are used during testing to obtain the performance. During the sweep, as in [2] we use 20 threshold values and go with the increments of 6.055 (which is equal to 121.1/20).

4 Experimental Design

News articles should be preprocessed by extraction, tokenizing and stemming before the similarity calculation process. Extraction includes getting the news articles from the collection, elimination of punctuation marks and stopwords. Stopwords are the words which are meaningless on their own but essential within the language (typical examples for English include words such as "the," "is," "at," "which," "on"). We eliminate all punctuation marks and use 217 stopwords which gives the best results in Turkish TDT experiments [8]. We use a Turkish NLP library, Zemberek [9] as a lemmatizer-based stemmer to eliminate the suffixes and prefixes and turn the words into their roots.

We perform TT experiments using the pure Okapi similarity function which we take as a baseline and CTR-based Okapi similarity functions in two different forms (additive and multiplicative).

4.1 Okapi Similarity Function

Okapi is a term frequency-inverse document frequency (tf-idf) based similarity function which calculates the similarity measure between two vectors (d, q) with the following formula [2].

$$sim(d, q) = \sum_{t \in d,q} w_{tf,d} \cdot w_{tf,q} \cdot w_{idf}$$

In Okapi, idf calculation is as follows.

$$w_{idf} = \text{In} \frac{N - df + 0.5}{df + 0.5}$$

Where df = number of documents that includes term t, N = total number of documents in document collection.

In Okapi, tf calculation is as follows.

$$w_{tf} = \frac{(k+1) \cdot tf}{k\left[(1-b) + b.\frac{dl}{avdl}\right] + tf}$$

Where b = 0.75, k = 1.2, dl = document length in terms of number of words (tokens), avdl = average document length.

4.2 Okapi Similarity Function with CTR

We add the term rank component additively and multiplicatively into the Okapi similarity function. They are defined as follows.

Additive CTR

$$sim(d,q) = \sum_{t \in d,q} (w_{tf,d} + R_{t,d}) \cdot (w_{tf,q} + R_{t,q}) \cdot w_{idf}$$

Multiplicative CTR

$$sim(d,q) = \sum_{t \in d,q} (w_{tf,d} \cdot R_{t,d}) \cdot (w_{tf,q} \cdot R_{t,q}) \cdot w_{idf}$$

The rank coefficient R ($R_{t,q}$. $R_{t,d}$) can be calculated as an inverted absolute rank: C/tr or as a percentage rank: C · tr/dl, where C is a constant generally between 0 and 1 giving the best experimental results for term rank *tr*. The C values used in the experiments are adopted from [8] (see Appendix E). In [8] C values are determined for NED; however, since NED and TT are the two sides of the same coin it makes sense to use the C values obtained for NED for TT. All function combinations for the rank coefficient R in both additive and multiplicative CTR are adopted from [2]. In total 21 different formulas are evaluated.

5 Evaluation Methodology

In TT the most common evaluation measures are false alarm (FA) and miss rate (MR), more specifically their probabilities P_{FA} and P_{MR}. They are defined as follows.

- $P_{FA} = \frac{FA}{Number\ of\ non{-}tracking\ stories}$
- $P_{Miss} = \frac{MR}{Number\ of\ tracking\ stories}$

where

- FA = number of non-tracking stories labeled as tracking stories,
- MR = number of tracking stories labeled as non-tracking stories.

From the combination of these FA and MR values a detection cost formula is formed as a single metric for measuring the effectiveness.

$$C_{det} = C_{Miss} \cdot P_{Miss} \cdot P_{Target} + C_{FA} \cdot P_{FA} \cdot (1 - P_{Target})$$

where

- $\mathrm{C_{Miss}} = 1$ and $\mathrm{C_{FA}} = 0.1$ are the prespecified costs of a missed detection and a false alarm
- $\mathrm{P_{Target}} = 0.02$, the $a\ priori$ probability of finding a target as specified by the application [7].

In all calculations we use normalized $\mathrm{C_{det}}$ because in the given formula for $\mathrm{C_{det}}$ has a dynamic range of values which is difficult for relative comparison. In normalized $\mathrm{C_{det}}$, $\mathrm{C_{det}}$ is divided by the minimum expected cost [7].

$$(C_{det})_{Norm} = \frac{C_{det}}{Minimum\{C_{Miss} \cdot P_{Target}, C_{FA} \cdot (1 - P_{Target})\}}$$

- Improvements of the functions with respect to the Okapi baseline are calculated as follows.

$$Improvement(\%) = \frac{(C_{det})_{okapi} - (C_{det})_{compared}}{(C_{det})_{okapi}} \cdot 100$$

6 Experimental Results and a Real Life Application

As illustrated in Table 1 the experimental results show that CTR-based approach is highly effective and in most of the cases does improve the performance of TT with respect to the baseline Okapi performance. Note that due to rounding of the $\mathrm{C_{det}}$ values Improvement (%) and $\mathrm{C_{det}}$ values do not match. In this table, for repeatability, the threshold values obtained by training are also provided. (The details for the other 19 formulas are not shown due to space limitation.) The CTR-based Okapi function that gives the highest improvement, which is 18 %, is the additive function *ai*. The p values are obtained by two-tailed t-tests when the results of a particular CTR-based Okapi measure is compared with those of the baseline. The results (several of them are not listed due to space limitation as we indicated above) show that in most of the cases the difference, in this case improvement, provided by the CTR approaches is either statistically significant or very strongly statistically significant. Furthermore,

Table 1 Test results of the experiment (Among 21 different formulas [2, 8], only the best additive (ai) and only the best multiplicative (mdbl) formulas are provided.)

	Training		Testing				
	$(C_{det})_{min}$	Threshold	P_{miss}	P_{FA}	C_{det}	Improvement (%)	P value
Okapi	0.188	48.440	0.075	0.009	0.120	N/A	N/A
ai	0.170	54.495	**0.064**	**0.007**	**0.098**	**17.934**	0.001**
mdbl	0.163	60.550	0.063	0.008	0.101	**15.870**	0.002*

* significant; ** very strongly significant

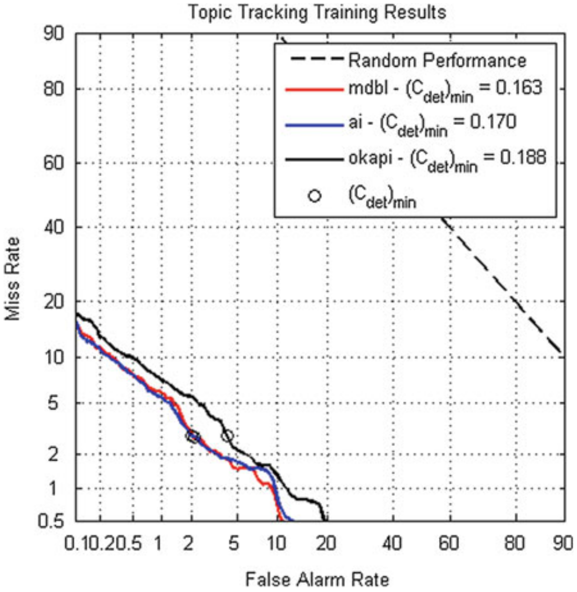

Fig. 2 TT training performance with mdbl-Okapi, ai-Okapi, and Okapi in terms of DET plots

ai gives the most desirable performance as indicated by its C_{det} value. The two-tailed t-test results also show that the improvement of C_{det} with *ai* with respect to the baseline Okapi function is very strongly significant (p = 0.001).

We also use Detection Error Tradeoff curve (DET) for representing the system performance. DET is a curve to see the tradeoff between miss and false alarm rate as shown in Fig. 2. It is obtained during training by using the threshold sweep method [1] and reveals what to expect during testing. DET curves are plotted on a Gaussian (normal) deviate scale which has advantages with respect to linear scales since it expands the high performance region [1, 7, p. 24]. The minimum C_{det} values of the functions are shown in circles. If the circle (or line) is closer to the origin, it means C_{det} value of corresponding point (or line) is smaller. We plotted baseline Okapi similarity function and functions with the greatest improvement in additive CTR and multiplicative CTR, which are *ai* and *mdbl*. These *ai* and *mdbl* functions mostly overlapped in Fig. 2; both their curve and min C_{det} points. As seen from the figure and from C_{det} values, *ai* and *mdbl* functions are significantly better than Okapi similarity function. The curve of baseline Okapi function is above the other curves and its min C_{det} point is more separated from the origin than those of *ai* and *mdbl*.

As shown in Table 1, in the testing phase *ai* function provides miss and false alarm rates of 6.40 and 0.70 %, respectively. This means that out of 100 tracking stories we would miss approximately 6 of them and only one of the stories would be an incorrect choice. These results are compatible with those presented in [1]. In that work refer to Table 1 and look at the best static TT performance using the stand alone cosine similarity measure that provides miss and false alarm rates of 4.94 and 0.71 %, respectively.

Based on experimental results, we also developed an Android application [10] for end users. It aims to show news stories to the users according to their choice for tracking. On the server side our application continuously fetches news stories from news sources and makes the decision of tracking on the fly and saves them in our server.

7 Conclusions and Future Work

Topic tracking (TT) is an important component of TDT systems in various information aggregation applications like news and blog portals. We investigate the TT problem within the framework of the Okapi measure by employing the concept of chronological term ranking (CTR). We propose an alternative method for TT using the CTR concept. For this purpose, we extend the Okapi similarity measure with a CTR component in various ways. The experimental results and statistical tests show that in the majority of the cases CTR significantly improves the performance obtained by the original Okapi similarity measure and is highly successful and can be used in real life applications.

In future work, the cosine similarity coefficient or other similarity functions can be extended with the CTR approach. The results of the CTR-based Okapi approaches and other approaches can be combined to improve the effectiveness to an even higher level [1]. Furthermore, the CTR-based approach can be used in the implementation of various information aggregators for topic tracking and also for story link detection or information filtering.

Acknowledgments This work is partially supported by the Scientific and Technical Research Council of Turkey (TÜBİTAK) under the grant number 111E030. We thank Süleyman Kardaş of Sabancı University for his helps in performance evaluation.

References

1. Can, F., Kocberber, S., Baglioglu, O., Kardas, S., Ocalan, H.C., Uyar, E.: New event detection and topic tracking in Turkish. J. Am. Soc. Inf. Sci. Technol. **61**(4), 802–819 (2010)
2. Troy, A.D., Zhang, G.: Enhancing relevance scoring with chronological term rank. In: Proceedings of the ACM SIGIR'07 Conference, pp. 599–606 (2007)

3. Allan, J.: Introduction to topic detection and tracking. In: Allan, J. (ed.) Topic Detection and Tracking: Event-based Information Organization, pp. 1–16. Kluwer Academic Publishers, Norwell (2002)
4. Yang, Y., Pierce, T., Carbonell, J.: A study on retrospective and on-line event detection. In: Proceedings of the ACM SIGIR'98 Conference, pp. 28–36 (1998)
5. Topic Detection and Tracking Evaluation: NIST Information Access Division. DET-curve plotting software tool. http://www.itl.nist.gov/iad/mig//tests/tdt/ (2007). Accessed 14 April 2012
6. Can, F., Kocberber, S., Balcik, E., Kaynak, C., Ocalan, H.C., Vursavas, O.M.: Information retrieval on Turkish texts. J. Am. Soc. Inf. Sci. Technol. **59**(3), 407–421 (2008)
7. Fiscus, J.G., Doddington, G.R.: Topic detection and tracking evaluation overview. In: Allan, J. (ed.) Topic Detection and Tracking: Event-based Information Organization, pp. 17–31. Kluwer Academic Publisher, Norwell (2002)
8. Baglioglu, O.: New event detection using chronological term ranking. Master thesis, Computer Engineering Department, Bilkent University, Ankara, Turkey (2009). http://www.cs.bilkent.edu.tr/canf/bilir_web/theses/ozgurBagliogluThesis.pdf
9. Zemberek, open source NLP library for Turkic languages. http://code.google.com/p/zemberek/. Accessed 5 Jan 2012
10. BilTracker Android Application Beta Demo Extended. http://youtu.be/MnyTO8bendU. Accessed 5 May 2012