Multi-criterion real time tweet summarization based upon adaptive threshold

Abdelhamid Chellal
IRIT Laboratory
University of Toulouse III
Toulouse, France
Email: abdelhamid.chellal@irit.fr

Mohand Boughanem
IRIT Laboratory
University of Toulouse III
Toulouse, France
Email: mohand.boughanem@irit.fr

Bernard Dousset
IRIT Laboratory
University of Toulouse III
Toulouse, France
Email: bernard.dousset@irit.fr

***Abstract*—Real time summarization in microblog aims at providing new relevant and non redundant information about an event as soon as it occurs. In this paper, we introduce a new tweet summarization approach where the decision of selecting an incoming tweet is made immediately when a tweet is available. Unlike existing approaches where thresholds are predefined, the proposed method estimates thresholds for decision making in real time as soon as the new tweet arrives. Tweet selection is based upon three criterion namely informativeness, novelty and relevance with regards of the user's interest which are combined as conjunctive condition. Only tweets having an informativeness and novelty scores above a parametric-free threshold are added to the summary. The evaluation of our approach was carried out on the TREC MB RTF 2015 data set and it was compared with well known baselines. The results have revealed that our approach produces the most precise summaries in comparison to all baselines and official runs of the TREC MB RTF 2015 task.**

***Index Terms*—Tweet Summarization, entropy, novelty**

I. Introduction

Sharing information on social networks is very common practice and even more a reflex. Users publish valuable information that provide in many cases live coverage of scheduled (sport games) and unscheduled events (natural disaster). In this case, following up the evolution of an event through the generated stream can be very gainful. In reputation management, monitoring what is being said about an entity (organization or individual) in social media may guide decision on how to act upon in order to preserve or improve the public reputation of this entity.

Due to the volume of generated information, monitoring all published posts that describe the development of a given event over time or referring to an entity is time-consuming and it may overload users with irrelevant and redundant posts. In many scenarios where unexpected events occur such as earth-quake or terrorist attack; user seeks for updates to be issued over time. To cope with this issue, we believe that building a summary that highlights, in real time, the most salient (relevant, non redundant) information related to an event or entity as soon as it occurs would be very beneficial to fulfill the user's information needs.

In this paper, we focus on the problem of microblog real-time summarization which has been a popular research topic in the last few years especially on Twitter [1] [2][3] [4] [5] [6]. Recently, a new task dedicated to microblog real time filtering (MB RTF) has been introduced in TREC 2015 [7].

The goal of real time summarization is to present to the user a condensed form of the most important content with a minimum of latency. In addition to relevance, an optimal summary should be short and cover the most important sub-events with no redundancy and where each new piece of information is added to the summary as soon as it becomes available.

In offline summarization, a summary is generated by selecting top weighted tweets iteratively but with discarding those having similarity with regard to a current summary above a certain threshold. However, unlike offline summarization where all documents (tweets) are available, in real time summarization the documents are not known in advance and the decision to select/ignore an incoming tweet needs to be taken immediately as soon as a tweet is available with respect to the previous ones on the stream and a tweet cannot be recalled once rejected. Hence, in existing approaches, the decision depends on whether the relevance and redundancy scores fall above a predefined threshold [8] [2]. Various studies were carried out on how tweet's relevance and novelty score are evaluated. However, threshold estimation at the arrival time of new tweet has never been investigated to the best of our knowledge in tweet summarization. In fact, statistics used to estimate these scores vary while new tweet arrives and the predefined threshold could be inappropriate. In addition, we believe that we cannot learn this threshold because it may depend on the type of event.

For novelty (non redundancy) detection, the main drawback of comparing incoming tweets with all previous tweets in the stream is the computational complexity. For that a pairwise comparison is conducted between an incoming tweet with those already selected in the summary. Nevertheless a pairwise comparison with the current summary is less effective since this summary has only a punctual view of the event history.

Hence, as the decision to select/ignore a tweet depends on a threshold and since a pairwise comparison for novelty detection does not fit real time summarization, the main issue here is how to cope with novelty and threshold. This paper attempts to overcome these issues as follows:

- To estimate the novelty score, the incoming tweet is

978-1-5090-4470-2/16 $31.00 © 2016 IEEE
DOI 10.1109/WI.2016.44

compared with the whole set of tweets of the current summary which are considered as one document and thus the pairwise comparison is avoided leading to reduce the computational complexity;

- The threshold is adaptive and it is estimated at the time of a new tweet arrives;
- The decision of selecting an incoming tweet is taken immediately in real time as soon as it occurs in order to reduce the latency between the publication time and the notification time;
- The decision is made according to three criterion namely relevance, informativeness and the novelty, which are combined as conjunctive constraints.

To achieve this task, we consider the real time summary generation process as multi criteria decision process where the decision of selecting/ignoring the incoming tweet is taken in real time without using any external knowledge. We propose to select incoming tweet if it passes three filters related to the following criteria: relevance, informativeness and novelty. A tweet passes a filter only if its related score is above a certain threshold. The underlying intuition behind this proposition is that only tweets within high informativeness and novelty scores must be selected to yield to a short summary with high coverage.

In order to evaluate the proposed approach, we carried out several experiments on TREC Microblog Real Time Filtering 2015 (MB-RTF) data set [7].

The rest of the paper is organized as follows: section 2 reviews related work. Section 3 describes the proposed approach. The experimental evaluation and results are given in section 4. We end with the conclusion in section 5.

II. Related Work

In this section, we present a brief overview of the related work on microblog summarization followed by novelty detection approaches.

A. Microblog summarization

Multi-document summarization techniques can be categorized into two classes: abstractive and extractive. The former may generate sentences not appeared in the original documents whereas the latter consists of selecting the most salient sentences from the documents [4]. Our approach falls within extractive class because our goal is to select salient tweets not to paraphrase them.

In Micro-blog summarization, most of abstractive techniques are graph based approach. The first one was Phrase Reinforcement (PR) algorithm proposed by Sharifi *et al.*, 2010 [3]. The algorithm builds up a word graph using the topic keywords as root node and words of an incoming tweets as nodes. Each word node is weighted proportionally to its distance to the root and to its frequency. The summary sentence is selected as one of the highest weighted path.

In [9] authors introduced the Multi Sentence Compression which builds a directed word graph from the input sentences. The summary is built by selecting the sentences that are given by the path having the smallest edge weight. In [10] the same authors extended their original approach by considering each node as tri-gram in which the summary is generated in real time. The summary is built from the path that contains the highest weight node and maximizes a score function. The main disadvantage of this approach is that the use of tri-grams leads to increase significantly the number of nodes.

Unlike abstractive summarization where only graph-based approaches were proposed, in extractive based methods, two categories can be distinguished, graph-based and feature-based. In graph-based approach, a vertex denotes a tweet and the weight of an edge is computed by combining the content similarity between two tweets and their social similarity based on features such as the number of followers and retweet [11]. The summary is built from vertices that have the greatest salient score.

Feature based approaches are mostly based on statistical features such as term frequency as term frequency [12] TF-IDF [13], hybridTF-IDF [2]. The approach proposed in [5] is one of the first real-time summarization approaches for scheduled events. It is based on term frequency in order to measure the salience of tweets and kullback-leibler divergence [14] to reduce redundancy. Sharifi et al [2] introduced a hybridTF-IDF approach where TF component is calculated over the overall set of tweets (considered as one document). Top weighted tweets are iteratively extracted, but excluding those having cosine similarity above a predefined threshold with the current summary. Sumbasic approach [15] initially proposed for document summarization was reported to be efficient for microblog summarization [6]. In this approach, the sentence that contains more frequent words has higher probability of being selected for summaries than the sentence with words occurring less frequently.

The TREC MB RTF-2015 official results reveal that run PKUICSTRunA2 [8] and UWaterlooATDK [16] are the two best performing ones. In the former, the relevance score of tweets is evaluated by using the normalized KL-divergence distance and the decision to select a tweet is based on predefined threshold set using Human assist selection. They manually scan the ranked list of top-10 selected tweet of previous day from top to bottom, and once not relevant tweet is found, its relevance score is chosen as the relevance threshold of the query in the next day. In UWaterlooATDK run, the vector space model was applied to compute relevance score of tweets and only those having relevance score above a threshold are selected. The threshold is fixed for each day according to the score of top 10 tweets returned in the previous day.

The comparison of several tweet summarization approaches conducted in [1] has revealed that simple term frequency performs well for topic-sensitive microblog summarization because of the unstructured and shortness nature of tweets. Thereby, HybridTF-IDF was reported as the best summarization approach in microbloging. Recently, mackie et al. [6] compared 11 summarization approaches using 4 microblog data sets. The results indicate that SumBasic [15] and HybridTF-IDF [2] have outperformed the other approaches.

Our approach falls within this line of research, it differs from the previous ones by: (i) it focuses on real time summarization of tweets while [11], [2] and [13] are off-line approaches; (ii) it relies on three criteria combined as conjunctive constrains for decision making where related works are based on relevance; (iii) its decision is based on parametric-free threshold where in [2] [16] the threshold is predefined and in [8] it is set manually according to the score of selected tweets in the previous day. In addition, the proposed approach is applicable for any kind of event, while the approach introduced in [5] is dedicated to scheduled events.

B. Novelty detection

Novelty detection is based on similarity/divergence measures such as the Manhattan, cosine similarities and language models. The novelty detection is usually used with redundancy in the task related to Topic Detection and Tracking (TDT) [17]. According to the way that similarity metric is used, two kinds of approaches can be distinguished, the document-to-document approaches and the document-to-summary approaches [18]. We adopt the document-to-summary approach based on language models to measure the novelty of an incoming tweet. The document-to-summary method choice is motivated by the need of overcoming the limit related to the complexity of document-to-document comparison in real time summarization task.

III. Real time summarization

The problem of real time summarization can be considered as a instance of secretary problem which is described as hiring the best secretary out of n rankable applicants for a position. The applicants are interviewed one by one and the employee has to make immediate decision after each interview. An applicant cannot be recalled once rejected. The problem of real time event summarization can be defined as follows:

Given an event described by keywords and a stream S of timestamped tweets T, output a set R of representative tweets such that:

1) $\forall T_i, \forall T_j \in S$ with the publication time t_i and t_j respectively $t_i < t_j$.. It means that the two tweets are provided in chronological order.
2) $\forall T_i \in R,\ \Delta t = \tau_i - t_i$, is very low. where τ_i is a notification time (time of making decision to select the tweet T_i).
3) $\forall T_i, \forall T_j \in R,\ T_i \nsim T_j$; it means that the two tweets T_i and T_j provide different information in order to keep the summary from being redundant and cover all sub-events (coverage);
4) $R \prec R'$, summary R is preferred to R' if R covers at least same sub-events than R' with less number of tweets (shortness properties).

The main challenges in this task, beside that the summary has to contain relevant tweets, are: (i) the summary has to be concise and covers all essential information about the event without any redundancy; (ii) new information nuggets should be added to the summary as soon as they become available.

These requirements (low latency, minimum of redundancy and shortness) are fulfilled by our approach as follows:

- The outlined approach is a fully real-time that makes select/ignore decision immediately as soon as a tweet appears in order to reduce a latency between publication time and selection time;
- The decision is based upon three criterion namely relevance, informativeness and novelty. The two first aims to detect tweets with important words that bring about high amount of information regarding previous seen information in stream. The latter is used to avoid pushing an information already selected which prevents the summary from being redundant and leads to improve coverage.

For the novelty (non redundancy) requirement two kinds of solutions can be considered. The naive one is to evaluate the similarity/divergence of an incoming tweet with all previous tweets of the stream. The second solution is to compare the incoming tweet with those that are likely to be relevant or with only tweets of the current summary. Both solutions are not effective since, in the first case, the computational complexity depends on the number of previous tweets and in the second case there is not enough data (particularly at the starter) to take effective decision.

A. Tweet filter

To reach the aforementioned requirements, our approach acts like a filter with the three levels. Let us consider a new tweet T and a stream collection S^t, the current summary R^t at time t (time of publication of tweet T) for a given query $Q = \{q_1, q_2, ...q_{|Q|}\}$. The incoming tweet T will be added to the summary if and only if:

$$\begin{cases} RSV(T,Q) = |T \cap Q| \geq K \\ IS(T) \geq Info\_Threshold(t) \\ NS(T) \geq Nov\_Threshold(t) \end{cases} \quad (1)$$

Where $RSV(T,Q)$, $IS(T)$ and $NS(T)$ are the relevance score regarding to query Q, the informativeness and the novelty scores of an incoming tweet T respectively. S^t is the stream at t (time of publication of tweet T). K is the minimum number of overlapping words between tweet T and query required to pass the relevance filter. $Info\_Threshold(t)$ and $Nov\_Threshold(t)$ are thresholds estimated at the arriving time of incoming tweet.

The two first filters select candidate tweets and the third level evaluates the novelty score for only tweets that pass the two first filters. To fit a real time scenario, the novelty score is evaluated with respect to the current summary instead of stream which reduces the computational complexity.

Algorithm 1 describes the overview of our incremental tweet summarization approach. First for relevance criteria, we consider a very simplistic approach, we filter out all tweets that do not contain at least two words of the query. The second filter is the informativeness, which aims at detecting tweets that bring a high amount of information regarding previous tweets in stream. The third filter is novelty to avoid pushing an information already selected. We assume that a tweet containing new information compared to all already seen tweets is likely to supply new sub-topics leading to improve summary's coverage.

One of the main issue here is how to leverage all these criteria. A linear combination between these criteria can be considered and only tweets that obtain a score above a certain threshold will be added to the summary. This solution may yield adding to the summary tweets with high informativeness and low novelty scores or inversely. To overcome this issue, we propose to combine these criteria as a conjunctive condition in order to provide complementarity between them. A tweet is added to a summary only if these three criteria are satisfied. The last issue concerns the threshold, we propose to set it adaptively.

Algorithm 1 Incremental tweet summarization

```
Require: $tweet\ Stream\ S,\ Query\ Q$
  $Summary\ R \leftarrow \varnothing$
  while $!S.end()$ do
    $Tweet\ T \leftarrow S.next()$
    if $T\_words \cap Q\_words \geq 2$ then
      $IS(T) \leftarrow Entropy(T)$ ; $\delta_1 \leftarrow threshold(IS, t)$
      if $(IS(T) \geq \delta_1)$ then
        $NS(T) \leftarrow 1 - \frac{|R \cap T|}{|T|}$ ; $\delta_2 \leftarrow threshold(NS, t)$
        if $(NS(T) \geq \delta_2)$ then
          $R \leftarrow R \cup T$
        end if
      end if
    end if
  end while
```

B. Informativeness score

In information theory, the amount of information carried by a message can be evaluated through the entropy of Shannon [19]. To evaluate the informativeness of an incoming tweet, we use the entropy measure. Thus, the informativeness score (IS) of tweet T at the time t is measured as follows:

$$IS(T,t) = \frac{-\sum_{w_i \in T} P^t(w_i) \times log_2(P^t(w_i))}{max[Minimumthreshold, |T|]} \tag{2}$$

Where $|T|$ is the size of tweet T and $P^t(w_i)$ represents the probability of occurrence of term w_i at time t in the stream. This probability is estimated as follows:

$$P^t(w_i) = \frac{\#TweetInWhich\ w_i\ Occurs\ AtTime\ t}{\#Tweet\ in\ stream\ AtTime\ t} \tag{3}$$

we divide the entropy of a tweet by a normalization factor which allows to carefully control the overall target summary length. Entropy of Shannon is very sensitive to document length and often overweighs terms from longer tweets. Without a normalization factor, Shannon's entropy awarded the most weight to the longer tweets since the weight of a tweet is the simple sum of the weights of the composing words. Alternatively, for tweets shorter than the target summary length (minimum threshold), these tweets will also get penalized since they will be divided by a number larger than that of terms in the tweet.

C. Novelty score

The intuitive way to estimate the novelty of an incoming tweet is to measure its similarity to all tweets in the current summary using cosine similarity or KL-divergence. A drawback of this method is its computational complexity which depends on the length of the summary and how to aggregate the different scores of similarity/divergence of the incoming tweet with regards to all tweets of the current summary. In literature, the mean cosine similarity, the max cosine similarity and the minimum kl-divergence are considered as the traditional method for novelty detection.

We believe that cosine similarity and Kl-divergence are not suitable for evaluating the distance between two tweets because in most cases the term frequency is 1, since tweets are typically very short. Hence, word overlap seems to be more suitable for evaluating the distance between two tweets. In order to avoid the pairwise comparison between summary's tweets and incoming tweet, we propose to merge all summary's tweets into one "summary word set" and compute the number of overlapping words between the incoming tweet and summary word set. Assume that RW is the set of words that occur in current summary then the novelty score of the incoming tweet is evaluated as follows:

$$NS(T, RW) = 1 - \frac{|RW \cap T|}{|T|} \tag{4}$$

The main advantage of this measure is the avoidance of the pairwise comparison and the fact that it is not based upon statistics which change when new tweet occurs. The intuition behind this measure is the incoming tweet is considered novel if it contains words that do not occur in the current summary independently of word frequency in the summary. Also, notice that the number of overlapping words is divided by the size of tweet $|T|$ instead of the size of the summary word set $|RW|$ which leads to take into account the size of tweet. Indeed, if the number of overlapping words is divided by $|RW|$ a long tweet and short tweet with the same number of overlapping words $|RW \cap T|$ will have the same novelty score.

D. Threshold setting

The statistics used to estimate the informativeness and the novelty vary while new tweet arrives. In addition, we believe that we cannot learn the threshold because it may depend on the type of event. To handle this issue, we suggest to set the threshold adaptively by considering previous tweets.

1) Relevance threshold: To set the relevance threshold k (equation 1), we carried out an experimental evaluation of the quality of the relevance filter based on TREC MB RTF-2015 data set. Two values were tested (i) at least one word (k=1) and (ii) at least two words (k=2). Table 1 reports the results by precision and recall obtained by each filter. As shown in the last row the filter (at least 2 query words) increases significantly the precision. The number of tweets that pass this filter is 15878 while the number of tweets that pass the filter (at least 1 query word) is 140 times larger. The filter (at least 2 query words) captures about 40% of relevant tweets while the filter (at least 1 query word) return 74% of relevant tweets but it also brings up a lot of noise. These results motivated our choice to use (at least 2 query words) as threshold. Since our goal is to generate a conics summary, we thing that having 40% of relevant tweets is enough to reach this purpose.

TABLE I
QUALITY OF THE RELEVANCE FILTER ON MB RTF-2015 DATA SET.

		K = 1				K = 2			
DAY	**R**	**S**	**RS**	**Precision**	**Recall**	**S**	**RS**	**Precision**	**Recall**
DAY1	931	222964	823	0,0037	0,8810	1821	393	0,2158	0,4221
DAY2	853	228322	716	0,0031	0,8393	1934	415	0,2145	0,4865
DAY3	728	233224	604	0,0026	0,8296	1639	321	0,1958	0,4409
DAY4	603	229566	518	0,0022	0,8590	1762	297	0,1685	0,4925
DAY5	605	221128	525	0,0023	0,8677	1418	282	0,1988	0,4661
DAY6	642	198784	576	0,0029	0,8971	1210	282	0,2330	0,4392
DAY7	939	204134	491	0,0024	0,5228	1243	233	0,1874	0,2481
DAY8	652	221128	525	0,0023	0,8052	1407	245	0,1741	0,3757
DAY9	1229	223078	682	0,0030	0,5549	1640	411	0,2506	0,3344
DAY10	982	243016	641	0,0026	0,6527	1804	465	0,2577	0,4735
ALL	**8164**	**2225344**	**6101**	**0,0027**	**0,7473**	**15878**	**3344**	**0,2106**	**0,4096**

Note. R, S and RS are the number of relevant tweets, selected tweets and relevant selected tweets per day respectively.

2) Informativeness threshold: Considering the entropy of the query as the amount of information that user has already about an event, an incoming tweet is considered informative with respect to the query and worthy to be added to the summary if its entropy is higher than the entropy of the query. Hence, we propose to set the informativeness threshold to the entropy of the query regarding previous seen tweets in the stream at the publication time of the incoming tweet. The intuition behind this proposition is that to be added to the summary, the incoming tweet should increase the amount of information of user with respect to what he already know.

$$Info\_Threshold = \frac{1}{|Q|} - \sum_{q_i \in Q} P^t(q_i) \times log_2(P^t(q_i)) \quad (5)$$

Where $|Q|$ is the size of query Q and $P^t(w_i)$ represents the probability of occurrence of term q_i at time t in the stream. This probability is estimated as follows:

$$P^t(q_i) = \frac{\#TweetInWhich\ q_i\ Occurs\ AtTime\ t}{\#Tweet\ in\ stream\ AtTime\ t} \quad (6)$$

3) Novelty threshold: The novelty score is based on the number of overlapping words between the incoming tweet and the summary word set RW. This number decrease from $|T|$ to 0 while the size of the summary increase. In the beginning when the summary set is empty $|T \cap RW| = |T|$ and each time a new tweet is added to the summary, the number of overlapping words between the next tweet and RW will likely be low than its size. The probability that $|T \cap RW| = 0$ (which means that $T \subset RW = 0$) increase with the size of RW. From this observation, we think that the novelty threshold value should be relaxed according to the number of tweets already selected in the current summary and the size limit of the summary. Also, in order to avoid a comparison with an empty set in the beginning of the selection process, the summary word set is initialized to the query word . Hence, at the time a new tweet arrives, the novelty threshold is computed as follows:

$$Nov\_Threshold = 1 - \frac{Min(|T| - 1, |T \cap Q| \times \exp^{N/K})}{|T|} \quad (7)$$

Where K is the maximum number of tweets in the summary and N is the number of selected tweets in the current summary with $N \leq K$.

As long as the summary set is empty ($N = 0$) the novelty threshold is equal to $1 - \frac{|T \cap Q|}{|T|}$. This threshold decreases with N and when the limit size of the summary is reached ($N = K$) the threshold value reaches its minimum value. Notice here that the number of overlapping words should not exceed the number of words in tweet. Hence, the number of overlapping words is set to the minimum of either $|T| - 1$ or $|T \cap Q| \times \exp^{N/K}$.

IV. EXPERIMENTAL EVALUATION

To evaluate the effectiveness of our approach, we carried out threefold objectives based experiment:

1) Evaluate the impact of the threshold ;
2) Evaluate the impact of each component and compare the performance of different configurations of the proposed approach;
3) Evaluate and compare the performance of the outlined approach with those obtained in TREC MB-RTF 2015 task and some commonly used baselines.

Therefore, we evaluated different configurations of our approach by considering the two components (informativeness and novelty) separately or together with different combinations (linear, product and conjunctive conditions). Besides, we also evaluate three different ways to estimate the threshold.

A. Data set

Experiments were conducted on TREC MB RTF 2015 data set. This collection was generated by each participant independently by crawling tweets using Twitter's streaming API during the evaluation period (10 days : 20 July to 29 July 2015) with considering English tweets only. After the evaluation period, 51 topics were selected. Two scenarios were defined namely "Push notifications on a mobile phone" and "Periodic email digest". In the latter, a system is allowed to return a maximum of 10 tweets per day per interest profile and these tweets are pushed in real time while in the former a system is tasked with identifying a batch of up to 100 ranked tweets per day per interest profile and the these tweets are delivered to the user daily at the day ends. Hence, the second scenario is more like a TOP-100 retrieval task based on a one-day tweet collection. In our experiments, we focus on the first scenario which corresponds to a real time task and where system was requested to record the time at which a tweet was selected; this information is used to compute a temporal penalty between publication time and notification time.

B. Evaluation metrics

Two temporally discounted gain measures were adopted. The primary metric is the expected latency-discounted gain (ELG) in which a latency penalty is applied. The second metric is the normalized cumulative gain (nCG). These two metrics are defined as follows:

$$ELG(T) = \frac{1}{N} \times \sum G(T) \times MAX(0, (100 - delay)/100) \quad (8)$$

$$nCG(T) = \frac{1}{Z} \times \sum G(T) \quad (9)$$

Where N is the number of returned tweets and Z is the maximum possible gain (given the 10 tweet per day limit). The delay is the gap (in minutes) between the tweet creation time and the selection time. G(T) is the gain of each tweet which is set as follows:

- Irrelevant tweets receive a gain of 0.
- Relevant tweets receive a gain of 0.5.
- Highly relevant tweets receive a gain of 1.0.

C. Baselines

We compare our method with the following baselines.

1) Baseline 1: In this baseline, we compare the proposed thresholds (section III.D) for the informativeness and novelty filter with the average, the maximum and the upper bound of the confidence interval (CI) of the previous seen values. The upper bound of CI is defined as follows:

$$Thershold(X,t) = \frac{\sum_{T_j \in S^t} X(T_j)}{|S^t|} + Z_{a/2} \times \frac{\sigma(X)}{\sqrt{|S^t|}} \quad (10)$$

Where t is the publication time of tweet T and X represents (IS or NS) the informativeness and the novelty scores of tweet T respectively. S^t is the stream at t. $Z_{a/2}$ is the confidence coefficient with degree a. $\sigma(X)$ is the standard deviation of X. The confidence coefficient is fixed to $Z_{a/2} = 1.65$ which corresponds to the confidence degree $a = 90\%$.

2) Baseline 2: In this baseline, we compare our estimation approach of novelty score with the tradition min KL-divergence measure and mean cosine similarity. In min KL-divergence, to evaluate the divergence between two tweets T and T', we use the Kullback-Leibler (KL) divergence [14] between their language models as follows:

$$KL(T, T') = \sum_{w_i \in T \cup T'} \theta_T(w_i) \log \frac{\theta_T(w_i)}{\theta_{T'}(w_i)} \quad (11)$$

where θ_T is the unigram language model of tweet T and $\theta_T(w_i)$ is the probability of occurrence of term w_i in T.

The novelty score of incoming tweet with respect to the current summary set can be measured in different ways. We can consider a global score that aggregates the divergence score between incoming tweet T and all tweets of the current summary R^t. This will provide tweet that is divergent from all tweets of R^t. The second approach is to consider only the divergence of T with the most similar tweet of R^t which is defined as the one having the lowest divergence with T. We choose the second approach because it is the most restrictive. Thereby, the novelty score (NS) is defined as follows:

$$NS(T) = \min_{\forall T' \in R^t} KL(T, T') \quad (12)$$

Where R^t is the summary at time t. We used Dirichlet (D) smoothing to estimate the tweet language model as follows:

$$\theta_T(w_i) = \frac{TF_T(w_i) + \mu P_S^t(w_i)}{|T| + \mu} \quad (13)$$

Where $TF_T(w_i)$ is the frequency of term w_i in tweet T and μ is the smoothing parameter. $P_S^t(w_i)$ is the probability of occurrence of the term w_i in the stream S at the time t, it is estimated using the maximum likelihood estimation (ML). For our experiment, the smoothing parameter μ has been set to 1000, after performing several experiments where μ was varied from 10 to 2000 with increments of 50.

For mean cosine similarity, the novelty score of incoming tweet is evaluated as follows:

$$NS(T) = 1 - \frac{\sum_{T' \in R^t} cossim(T, T')}{|R^t|} \quad (14)$$

3) Baseline 3: In this baseline, we adopt state-of-the-art functions to estimate the informativenss score of incoming tweet. We compared our approach to three approaches namely TF-IDF, SumBasic [15] and hybridTF-IDF[2]. These methods were recommended by [6] to be considered as baselines since it turned out to be the best one among 11 different tweet summarization approaches. These baselines are adjusted to real time selection of tweets and are evaluated with the proposed method of novelty detection. The same thresholds described in section III.D are adopted with these baselines.

The equations below describe the formula used in HybridTF-IDF and in Sumbasic for a given tweet T respectively:

$$HybridTF - IDF(T) = \frac{\sum_{w_i \in T} TF(w_i) \times IDF(w_i)}{\max[Minimum\ threshold, |T|]} \quad (15)$$

$$TF(w_i) = \frac{\#(w_i)\ InAllPosts}{\#WordInAllPosts} \quad (16)$$

$$IDF(w_i) = log_2(\frac{\#Tweet}{\#Tweet\ w_i\ Occurs}) \quad (17)$$

$$Sumbasic(T) = \sum_{w_i \in T} \frac{P(w_i)}{|T|} \quad (18)$$

$$P(w_i) = \frac{\#w_i\ InAllTweets}{\#Tweet\ w_i\ Occurs} \quad (19)$$

D. Results and Discussion

Thresholding impact: In this section, we report the comparative effectiveness of the proposed threshold with the average, the maximum and the upper bound of the confidence interval (CI) of previous seen values as threshold estimation methods. Table 3 reports the performance by ELG and nCG obtained by the proposed threshold estimation against the aforementioned thresholds estimation baselines. The best performing one with respect to each measure is highlighted in bold. $|R|$ represents the size of the summary R.

TABLE II
COMPARATIVE EVALUATION OF EFFECTIVENESS WITH CLASSICAL THRESHOLD ESTIMATION.

Threshold	ELG	nCG	$\|R\|$	%ELG	%nCG
AVG	0.3182‡	0.2563†	1214	+7.28	+29.00
MAX	0.3377	0.2433†	456	+1.60	+32.60
$Z_{a/2} = 1.65$	0.3400	0.2510†	660	+0.93	+29.63
Equation 5, 7	**0.3432**	**0.3610**	2328	-	-

Note. % indicates the proposed thresholds improvements in terms of ELG and nCG. The symbols *, †, and ‡ denote the Student test significance: $*0.01 < t \leq 0.05$, $\dagger t \leq 0.01$, $\ddagger 0.05 < t \leq 0.1$.

As shown in Table 2, our threshold setting model outperforms all baselines in both expected latency-discounted gain (ELG) and normalized cumulative gain (nCG). In order to evaluate the significance of our threshold setting model improvement, we conducted a paired two-tailed t test. Significance testing based on the Student t-test statistic is computed on the basis of both metrics (ELG, nCG). The p values are marked with the symbols *, †, and ‡ statistically significant differences. The positive improvements obtained by our approach were found to be statistically significant with p values < 0.01 for nCG and between 0.05 and 0.01 for ELG metric. From Table 3, we also notice that the performances' improvements in terms of nCG are important for the classical threshold estimation. We found performance improvements up to nCG values of about 32.60 % for maximum and of 29.63% for the upper bound of the CI. These results show that the proposed threshold lead to improve coverage (nCG) without decreasing the precision (ELG). This can be explained by the high number of selected tweets in the summary since the use of adaptive thresholds which depend on the number of selected tweets (for novelty threshold) and entropy of the query by the time new tweet arrives (for informativeness threshold) is less restrictive than the maximum and the upper bound of CI of previous seen values. Indeed, the use of such restrictive thresholds reduces the number of pushed tweets, which decrease significantly the cumulative gain (coverage).

Comparative evaluation with state-of-the-art novelty detection methods: We present a comparative evaluation of word overlap method versus conventional state-of-the-art novelty detection approaches namely min KL-divergence and mean cosine similarity as presented in (baseline 2). For all these methods, the informativeness threshold defined in equation 5 was used and for novelty threshold we test three classical threshold estimations (average, maximum and upper bound of CI). As reported in table 3, word overlap function outperforms significantly all baselines in both precision (ELG) and coverage (nCG) over all three type of thresholds. These results can be explained by the shortness of tweets and the fact that novelty estimation based upon word overlap does not use statistics which may change significantly when new tweets arrives particular in the starter.

TABLE III
COMPARISON OF NOVELTY ESTIMATION.

	AVG			MAX			$Z_{a/2} = 1.65$		
Novelty	**ELG**	**nCG**	$\|R\|$	**ELG**	**nCG**	$\|R\|$	**ELG**	**nCG**	$\|R\|$
Min KLD	0,2918$*$	0,3507	3050	0,2963$*$	0,3307$\ddagger$	2734	0,2952$*$	0,3426$\ddagger$	2938
mean CosSim	0,293$*$	0,3128$\dagger$	2408	0,298$*$	0,2759$\dagger$	1573	0,3027$\ddagger$	0,3027$\dagger$	2043
Overlap	**0,3353**	**0,3783**	2749	**0,3357**	**0,3641**	2452	**0,3354**	**0,3713**	2614
% Change	+12,61	+17,31	-	+11,23	+24,22	-	+9,75	+18,47	-

Note. The last row % Change shows the improvement in terms of ELG and nCG with the best baseline in terms of ELG (i.e., mean CosSim). The symbols *, †, and ‡ denote the Student test significance: $*0.01 < t \leq 0.05$, $\dagger t \leq 0.01$, $\ddagger 0.05 < t \leq 0.1$.

Components combination: In table 4, we compare the impact of each criterion taken alone as well as the impact of the different combinations of the three criterion (product, linear and conjunctive condition denoted by the symbol ×, +, & respectively). As shown in table 4, the conjunctive combination outperforms both linear and product combinations. The improvement in terms of ELG is up to 14.18% for linear combination and 13.81% for product combination. This result is expected sine conjunctive combination is more restrictive than the two other combinations which leads to reducing the number of selected tweets in summary. We also notice that informativeness is more significant than novelty. Informativeness increases the precision (ELG) and the coverage (nCG) and the generated summary is longer than the one generated when the novelty is used alone.

TABLE IV
COMPARATIVE EVALUATION OF SUMMARY QUALITY OF DIFFERENT COMBINATION.

Combination	ELG	nCG	$\|R\|$	%ELG	%nCG
$RSV\&IS\&NS$	**0.3432**	**0.3610**	**2328**	-	-
$RSV \times IS \times NS$	0.2958*	0.3508	2975	+13.81	+2.82
$RSV + IS + NS$	0.2945†	0.3505	2959	+14.18	+2.90
$RSV\&IS$	0.3145‡	0.3590	2772	+8.36	+0.55
$RSV\&NS$	0.3025*	0.3364	2620	+14.18	+6.81
$RSVOnly$	0.2926†	0.3528	2939	+14.74	+2.27

Note. RSV, IS, NS represent the relevance, informativeness and novelty scores respectively. % indicates the conjunctive combination improvements in terms of ELG and nCG.The symbols *, †, and ‡ denote the Student test significance: $*0.01 < t \leq 0.05$, $\dagger t \leq 0.01$, $\ddagger 0.05 < t \leq 0.1$.

TABLE V
COMPARATIVE EVALUATION OF SUMMARY QUALITY WITH STATE-OF-THE-ART SUMMARIZATION APPROACHES.

Method	ELG	nCG	$\|R\|$	%ELG	%nCG
Entropy-Overlap	**0.3432**	**0.3610**	2328		
HybridTFIDF-Overlap	0.3037*	0.3200*	2428	+11.50	+11.35
TFIDF-Overlap	0.3038*	0.3197*	2152	+13.86	+4.82
sumbasic-Overlap	0.2956*	0.3436	2828	+12.23	+13.24
TREC MB RTF 2015 official Results					
PKUICSTRunA2	**0.3175**	**0.3127**	-	+7.48	+13.37
UWaterlooATDK	0.3150	0.2679	-	+8.21	+25.78

Note. % indicates the conjunctive combination improvements in terms of ELG and nCG.The symbols *, †, and ‡ denote the Student test significance: $*0.01 < t \leq 0.05$, $\dagger t \leq 0.01$, $\ddagger 0.05 < t \leq 0.1$.

Comparative evaluation with state-of-the-art summarization approaches: In this section, we compare our approach with some traditional state-of-the-art summarization approaches more particularly with (HybridTF-IDF, TF-IDF and Sumbasic) and with the two best performing runs in TREC MB-RTF 2015 task namely PKUICSTRunA2 [8] and UWaterlooATDK [16]. Table 5 reports the results by ELG, nCG and size of the summary $|R|$ obtained by our method against the aforementioned summarization baseline approaches.

Table 5 shows that our summarization model (Entropy-overlap) outperforms all baselines in terms of ELG and nCG.

The improvement is up to 11.5% and 11.35% for the best baseline for ELG and nCG respectively while the size of the summary is smaller (2328 tweets for Entropy-overlap against 2427 tweets for Hybrid-TFIDF-overlap). These results may be explained by several factors. First, in TFIDF, TF component has no effect because most term frequencies will be equal to 1 which leads to reduce TFIDF to IDF component. The HybridTF-IDF function can be seen as adding a little complexity to word frequency in stream by including information regarding the IDF component. However, IDF component can be considered as novelty score since it awards most score to infrequent words in the stream. It seems that IDF is not particularly helpful in real time summarization since in the binning the IDF score is high leading to select the first tweets for summary that pass the relevance filter. In sumbasic function tweets that contain more frequent words has higher probability of being selected for summaries. We notice that the SumBasic method has a higher nCG than ELG whereas the other methods have a closer balance between ELG and nCG. This suggests that the SumBasic algorithm may be biased towards longer tweets. In the proposed approach, the entropy measure is based upon the number of occurrence of words in the stream. It seems that simple word frequency calculations are particularly important for summarizing twitter topics.

We also notice that the proposed summarization method outperforms TREC MB TRF-2015 runs in which the thresholds were predefined. In UWaterlooATDK the threshold was set for each day according to the score of top-50 selected tweets in the previous day and in PKUICSTRunA2 Human assist selection is used to set relevance threshold according to top-10 selected tweets of previous day. These results can be explained by the fact that entropy gives a high score to words that occur frequently in the stream and the novelty filter discards any tweet that contains frequent words. Hence, only tweets that contain a good mix of frequent terms and new term will be selected to the summary. Also the capacity of our method to adapt the threshold values according to statistics and the current summary while new tweet arrives helps enhancing the quality of the generated summary.

V. Conclusion

In this paper, we proposed a new method for microblog real-time summarization which aims to be independent from the event and provides a summarized stream in incremental way instead of categorizing sub-events. The decision to select/ignore an incoming tweet is made in real time according to whether its informativeness and novelty scores are above an adaptive threshold which is estimated at the time tweet arrives. Experiments were carried out on TREC MB RTF-2015 data set show that best performances are observed when using a conjunctive combination of the three selection criteria. The obtained results give evidence that measured based on stream statistics can be used alone to generate, in real time, a concise summary with a good precision as well as coverage. However, to improve the efficiency, further research need to be carried out in order to identify other selection criteria as well as threshold estimation.

References

[1] D. Inouye and J. K. Kalita, "Comparing twitter summarization algorithms for multiple post summaries," in *PASSAT/SocialCom 2011, Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on Social Computing (SocialCom), Boston, MA, USA, 9-11 Oct., 2011*, 2011, pp. 298–306.

[2] B. Sharifi, M.-A. Hutton, and J. K. Kalita, "Experiments in microblog summarization," in *Proceedings of the 2010 IEEE Second International Conference on Social Computing*, ser. SOCIALCOM '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 49–56. [Online]. Available: http://dx.doi.org/10.1109/SocialCom.2010.17

[3] B. Sharifi, M.-A. Hutton, and J. Kalita, "Summarizing microblogs automatically," in *Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics*, ser. HLT '10, 2010, pp. 685–688.

[4] Z. Ren, S. Liang, E. Meij, and M. de Rijke, "Personalized time-aware tweets summarization," in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '13. New York, NY, USA: ACM, 2013, pp. 513–522. [Online]. Available: http://doi.acm.org/10.1145/2484028.2484052

[5] A. Zubiaga, D. Spina, E. Amigó, and J. Gonzalo, "Towards real-time summarization of scheduled events from twitter streams," in *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*, ser. HT '12, 2012, pp. 319–320.

[6] S. Mackie, R. McCreadie, C. Macdonald, and I. Ounis, "Comparing algorithms for microblog summarisation," in *Information Access Evaluation. Multilinguality, Multimodality, and Interaction - 5th International Conference of the CLEF Initiative, CLEF 2014, Sheffield, UK, September 15-18, 2014. Proceedings*, 2014, pp. 153–159.

[7] J. Lin, M. Efron, Y. Wang, G. Sherman, R. McCreadie, and T. Sakai, "Overview of the trec 2015 microblog track," in *Text REtrieval Conference, TREC, Gaithersburg, USA, November 17-20*, 2015.

[8] F. Fan, Y. Fei, C. Lv, L. Yao, J. Yang, and D. Zhao, "Pkuicst at trec 2015 microblog track: Query-biased adaptive filtering in real-time microblog stream," in *Text REtrieval Conference, TREC, Gaithersburg, USA, November 17-20*, 2015.

[9] A. Olariu, "Hierarchical clustering in improving microblog stream summarization," *Computational Linguistics and Intelligent Text Processing*, vol. V.7817, 2013.

[10] ——, "Efficient online summarization of microblogging streams," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, Gothenburg, Sweden, April 2014.

[11] X. Liu, Y. Li, F. Wei, and M. Zhou, "Graph-based multi-tweet summarization using social signals," in *Proceedings of COLING 2012*. Mumbai, India: The COLING 2012 Organizing Committee, December 2012.

[12] F. Liu, Y. Liu, and F. Weng, "Why is "sxsw" trending?: Exploring multiple text sources for twitter topic summarization," in *Proceedings of the Workshop on Languages in Social Media*, ser. LSM '11, 2011, pp. 66–75.

[13] D. Chakrabarti and K. Punera, "Event summarization using tweets," in *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*, 2011.

[14] S. Kullback and R. A. Leibler, *The Annals of Mathematical Statistics*, no. 1, pp. 79–86, 03.

[15] A. Nenkova and L. Vanderwende, "The impact of frequency on summarization," MSR-TR-2005-101, Tech. Rep. MSR-TR-2005-101, January.

[16] L. Tan, A. Roegiest, and C. L. Clarke, "University of waterloo at trec 2015 microblog track," in *Text REtrieval Conference, TREC, Gaithersburg, USA, November 17-20*, 2015.

[17] M. Markou and S. Singh, "Novelty detection: A review-part 1: Statistical approaches," *Signal Process.*, vol. 83, no. 12, pp. 2481–2497, Dec. 2003.

[18] M. Karkali, F. Rousseau, A. Ntoulas, and M. Vazirgiannis, "Using temporal IDF for efficient novelty detection in text streams," *CoRR*, vol. abs/1401.1456, 2014.

[19] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423 , 623–656, Oct. 1948.