Tweet Stream Summarization for Online Reputation Management

Jorge Carrillo-de-Albornoz, Enrique Amigó, Laura Plaza(✉), and Julio Gonzalo

NLP & IR Group, Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain
{jcalbornoz,enrique,lplaza,julio}@lsi.uned.es

Abstract. Producing online reputation reports for an entity (company, brand, etc.) is a focused summarization task with a distinctive feature: issues that may affect the reputation of the entity take priority in the summary. In this paper we (i) propose a novel methodology to evaluate summaries in the context of online reputation which profits from an analogy between reputation reports and the problem of diversity in search; and (ii) provide empirical evidence that incorporating priority signals may benefit this summarization task.

Keywords: Summarization · Diversity · Tweets · Reputation management

1 Introduction

Since the advent of Social Media, an essential part of Public Relations (for organizations and individuals) is Online Reputation Management, which consists of actively listening online media, monitoring what is being said about an entity and deciding how to act upon it in order to preserve or improve the public reputation of the entity. Monitoring the massive stream of online content is the first task of online reputation experts. Given a client (e.g. a company), the expert must provide frequent (e.g. daily) reports summarizing which are the issues that people are discussing and involve the company.

In a typical workflow, the reputation experts start with a set of queries that try to cover all possible ways of referring to the client. Then they take the results set and filter out irrelevant content (e.g., texts about apple pies when looking for the Apple company). Next, they determine which are the different issues people are discussing, evaluate their priority, and produce a report for the client.

Crucially, the report must include any issue that may affect the reputation of the client (reputation alerts) so that actions can be taken upon it. The summary, therefore, is guided by the relative priority of issues. This notion of priority differs from the signals that are usually considered in summarization algorithms, and it depends on many factors, including popularity (How many people are commenting on the issue?), polarity for reputation (Does it have positive or

N. Ferro et al. (Eds.): ECIR 2016, LNCS 9626, pp. 378–389, 2016.
DOI: 10.1007/978-3-319-30671-1_28

negative implications for the client?), novelty (Is it a new issue?), authority (Are opinion makers engaged in the conversation?), centrality (Is the client central to the conversation?), etc. This complex notion of priority makes the task of producing reputation-oriented summaries a challenging and practical scenario.

In this context, we investigate two main research questions:

RQ1. Given the peculiarities of the task, what is the most appropriate evaluation methodology?

Our research is triggered by the availability of the RepLab dataset [1], which contains annotations made by reputation experts on tweet streams for 61 entities, including entity name disambiguation, topic detection and topic priority.

We will discuss two types of evaluation methodologies, and in both cases we will adapt the RepLab dataset accordingly. The first methodology sticks to the traditional summarization scenario, under the hypothesis that RepLab annotations can be used to infer automatically entity-oriented summaries of near-manual quality. The second evaluation methodology models the task as producing a ranking of tweets that maximizes both coverage of topics and priority. This provides an analogy with the problem of search with diversity, where the search system must produce a rank that maximizes both relevance and coverage.

RQ2. What is the relationship between centrality and priority?

The most distinctive feature of reputation reports is that issues related with the entity are classified according to their priority from the perspective of reputation handling (the highest priority being a *reputation alert*). We want to investigate how the notion of priority translates to the task of producing extractive summaries, and how important it is to consider reputational signals of priority when building and appropriate summary.

We will start by discussing how to turn the RepLab setting and datasets into a test collection for entity-oriented tweet stream summarization. Then we will introduce our experimental setting to compare priority signals with text quality signals and assess our evaluation methodology, discuss the results, link our study with related work, and finish with the main conclusions learned.

2 A Methodology to Evaluate Reputation-Oriented Tweet Stream Summarization

A reputation report is a summary – produced by an online reputation expert – of the issues being discussed online which involve a given client (a company, organization, brand, individual... in general, an entity). In reputation reports produced daily, microblogs (and Twitter in particular) are of special relevance, as they anticipate issues that may later hit other media. Typically, the reputation expert follows this procedure (with the assistance of more or less sophisticated software):

- Starts with a set of queries that cover all possible way of referring to the client.
- Takes the results set and filter out irrelevant content.
- Groups tweets according to the different issues (topics) people are discussing.

- Evaluates the priority of each issue, establishing at least three categories: reputation alerts (which demand immediate attention), important topics (that the company must be aware of), and unimportant content (refers to the entity, but do not have consequences from a reputational point of view).
- Produces a reputation report for the client summarizing the result of the analysis.

The reputation report must include any issue that may affect the reputation of the client (reputation alerts) so that action can be taken upon it. This (extractive) summary, therefore, is guided by the relative priority of issues. However, as we pointed out in the introduction, this notion of priority differs from the signals that are usually considered in summarization algorithms, and it depends on many factors, including: popularity, polarity for reputation, novelty, authority, and centrality. Thus, the task is novel and attractive from the perspective of summarization, because the notion of which are the relevant information nuggets is focused and more precisely defined than in other summarization tasks. Also, it explicitly connects the summarization problem with other Natural Language Processing tasks: there is a filtering component (because it is entity-oriented), a social media component (because, in principle, non-textual Twitter signals may help discovering priority issues), a semantic understanding component (to establish, for instance, polarity for reputation), etc.

2.1 The RepLab 2013 Dataset

The RepLab 2013 task is defined as (multilingual) topic detection combined with priority ranking of the topics. Manual annotations are provided for the following subtasks:

- *Filtering*. Systems are asked to determine which tweets are related to the entity and which are not. Manual annotations are provided with two possible values: related/unrelated. For our summarization task, we will use as input only those tweets that are manually annotated as related to the entity.
- *Polarity for Reputation Classification*. The goal is to decide if the tweet content has positive or negative implications for the company's reputation. Manual annotations are: positive/negative/neutral.
- *Topic Detection*: Systems are asked to cluster related tweets about the entity by topic with the objective of grouping together tweets referring to the same subject/event/conversation.
- *Priority Assignment*. It involves detecting the relative priority of topics. Manual annotations have three possible values: Alert, mildly_important, unimportant.

RepLab 2013 uses Twitter data in English and Spanish. The collection comprises tweets about 61 entities from four domains: automotive, banking, universities and music. We will restrict our study to the automotive and banking domains, because they consist of large companies which are the standard subject

of reputation monitoring as it is done by experts: the annotation of *universities* and *music bands and artists* is more exploratory and does not follow widely adopted conventions as in the case of companies. Our subset of Replab 2013 comprises 71,303 tweets distributed as in the following table.

Table 1. Subset of RepLab 2013 dataset used in our experiments

	Automotive	Banking	Total
Entities	20	11	31
# Tweets (training)	15,123	7,774	22,897
# Tweets (test)	31,785	16,621	48,406
# Tweets (total)	46,908	24,395	71,303
# Tweets (EN)	38,614	16,305	54,919
# Tweets (ES)	8,294	8,090	16,384

2.2 Automatic Generation of Reference Summaries

We investigate two alternative ways of evaluating tweet stream summaries using RepLab data: the first one consists in automatically deriving "reference" or "model" summaries from the set of manual annotations provided by RepLab.

The goal of a reputation report is to cover all issues referring to the entity (in our dataset, a bank or a car manufacturer) which are relevant from a reputational perspective. RepLab manual annotations group relevant tweets according to fine-grained issues related to the company, and assign a three-valued priority to them. If we select only alerts and mildly important topics, and we pick randomly one tweet per topic, the result would be equivalent to a manual (extractive) summary under certain simplifying assumptions about the data:

- In a topic, all tweets are equally representative. This is a reasonable assumption in the RepLab dataset, because selected tweets are very focused, every tweet is independently assigned to a topic, and topics are fine-grained and therefore quite cohesive.
- A tweet is enough to summarize the content of an issue appropriately. This is certainly an oversimplification, and reputation experts will at least rewrite the content of a topic for a summary, and provide a logical structure to the different topics in a report. However, we may assume that, for evaluation purposes and as an average observation, most tweets are representative of the content of a topic.

Under this assumptions, variability between model summaries depends on which tweet we choose from each relevant topic. Therefore, we use a simplified user model where an expert may randomly pick any tweet, for every important topic (alerts and mildly relevant issues), to produce a reputation report. In our

experiments, we generate 1,000 model summaries for every entity using this model. Note that the excess of simplification in our assumptions pays off, as we are able to generate a large number of model summaries with the manual annotations provided by the RepLab dataset.

Once we have created the models (1,000 per test case), automatic summaries can be evaluated using standard text similarity measures. In our experiments we use ROUGE [2], a set of evaluation metrics for summarization which measure the content overlap between a peer and one or more reference summaries. The most popular variant is ROUGE-2, due to its high correlation with human judges. ROUGE-2 counts the number of bigrams that are shared by the peer and reference summaries and computes a recall-related measure [2].

2.3 Tweet Summarization as Search with Diversity

Our second approach to evaluate summaries does not require model summaries. It reads the summary as a ranked list of tweets, and evaluates the ranking with respect to relevance and redundancy as measured with respect to the annotated topics in the RepLab dataset. The idea is making an analogy between the task of producing a summary and the task of document retrieval with diversity. In this task, the retrieval system must provide a ranked list of documents that maximizes both relevance (documents are relevant to the query) and diversity (documents reflect the different query intents, when the query is ambiguous, or the different facets in the results when the query is not ambiguous).

Producing an extractive summary is, in fact, a similar task: the set of selected sentences should maximize relevance (they convey essential information from the documents) and diversity (sentences should minimize redundancy and maximize coverage of the different information nuggets in the documents). The case of reputation reports using Twitter as a source is even more clear, as relevance is modeled by the priority of each of the topics. An optimal report should maximize the priority of the information conveyed and the coverage of priority entity-related topics (which, in turn, minimizes redundancy).

Let's think of the following user model for tweet summaries: the user starts reading the summary from the first tweet. At each step, the user goes on to the next tweet or stops reading the summary, either because she is satisfied with the knowledge acquired so far, or because she does not expect the summary to provide further useful information. User satisfaction can be modeled via two variables: (i) the probability of going ahead with the next tweet in the summary; (ii) the amount of information gained with every tweet. The amount of information provided by a tweet depends on the tweets that precede it in the summary: a tweet from a topic that has already appeared in the summary contributes less than a tweet from a topic that has not yet been covered by the preceding tweets. To compute the expected user satisfaction, the evaluation metric must also take into account that tweets deeper in the summary (i.e. in the rank) are less likely

to be read, weighting the information gain of a tweet by the probability of reaching it. We propose to adapt Rank-Biased Precision (RBP) [3], an Information Retrieval evaluation measure which is defined as:

$$RBP = (1-p)\sum_{i=1}^{d} r_i * p^{i-1}$$

where r_i is a known function of the relevance of document at position i, p is the probability of moving to the next document, and RBP is defined as utility/effort (expected utility rate), with utility being $\sum_{i=1}^{d} r_i * p^{i-1}$ and $1/(1-p)$ the expected number of documents seen, i.e. the effort.

We prefer RBP to other diversity-oriented evaluation metrics because it naturally fits our task, the penalty for redundancy can be incorporated without changing the formula (simply defining r_i), and because it has been shown to comply with more desired formal properties than all other IR measures in the literature [4], and can be naturally adapted to our task.

Indeed, the need to remove redundancy and the relevance of priority information can be incorporated via r_i. We will model r_i according to two possible scenarios. In the first scenario, incorporating more than one tweet from a single topic still contributes positively to the summary (but increasingly less than the first tweet from that topic). This is well captured by the reciprocal of the number of tweets already seen from a topic (although many other variants are possible):

$$r_i = \frac{1}{|\{k \in \{1 \ldots i-1\} | \text{topic}(i) = \text{topic}(k)\}|}$$

We will refer to RBP with this relevance formula as **RBP-SUM-R** (RBP applied to SUMmarization with a Reciprocal discount function for redundancy).

In the second scenario, each topic is exhaustively defined by one tweet, and therefore only the first tweet incorporated to the summary, for each topic, contributes to the informative value of the summary. Then the relevance formula is simply:

$$r_i = \begin{cases} 1 \text{ if } \forall k \in \{1..i-1\} \text{topic}(i) \neq \text{topic}(k) \\ 0 \text{ otherwise} \end{cases}$$

We will refer to RBP with this relevance formula as **RBP-SUM-B** (RBP applied to SUMmarization with a Binary discount function for redundancy). With respect to the parameter p (probability of going ahead reading the summary after reading a tweet), we must aim at large values, which better reflect the purpose of the summary. For instance, a value of $p = 0.95$ means that the user has only a 60 % chance of reading beyond the first ten tweets, and a value of $p = 0.5$ decreases that probability to only 0.1 %. Figure 1 shows how the probability of reading through the summary decays for different values of p. We will perform our experiments with the values $p = 0.9$ (which decays fast for a summarization task) and $p = 0.99$ (which has a slower but still representative decay).

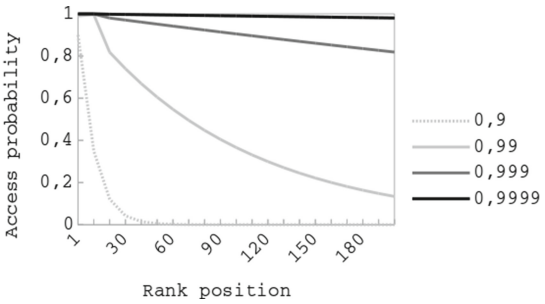


Fig. 1. Probability of reading through the summary for different p values

3 Experimental Design

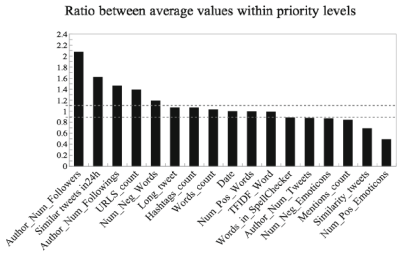
Our first research question (how to evaluate the task) is partially answered in the previous section. We now want to compare how the two alternative evaluation metrics behave, and we want to investigate the second research question: what is the relationship between centrality and priority, and how priority signals can be used to enhance summaries. For this purpose, we will compare three approaches (two baselines and one contrastive system):

LexRank. As a standard summarization baseline, we use LexRank [5], one of the best-known graph-based methods for multi-document summarization based on lexical centrality. LexRank is executed through the MEAD summarizer [6] (http://www.summarization.com/mead/) using these parameters: `-extract -s -p 10 -fcp delete`. We build summaries at 5, 10, 20 and 30 % compression rate, for LexRank and also for the other approaches.

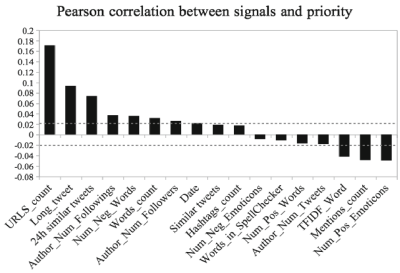
Followers. As a priority baseline, we simply rank the tweets by the number of followers of the tweet author, and then apply a technique to remove redundancy. The number of followers is a basic indication of priority: things being said by people with more followers are more likely to spread over the social networks. Redundancy is avoided using an iterative algorithm: a tweet from the ranking is included in the summary only if it has a vocabulary overlap less than 0.02, in terms of the Jaccard measure, with any of the tweets already included in the summary. Once the process is finished, if the resulting compression rate is higher than desired, discarded tweets are reconsidered and included by recursively increasing the threshold in 0.02 similarity points until the desired compression rate is reached.

Signal Voting. Our contrastive system considers a number of signals of priority and content quality. Each signal (computed using the training set) provides a ranking of all tweets for a given test case (an entity). We follow this procedure:

– Using the training part of the RepLab dataset, we compute two estimations of the quality of each signal: the ratio between average values within priority



(a) Ratio between average values for priority vs unimportant topics



(b) Pearson correlation between signal values and manual priority

Fig. 2. Signal assessment

values (if priority tweets receive higher values than unimportant tweets, the signal is useful), and the Pearson correlation between the signal values and the manual priority values. The signals (which are self-descriptive) and the indicators are displayed in Fig. 2.

- We retain those signals with a Pearson correlation above 0.02 and with a ratio of averages above 10 %. The resulting set of signals is: `URLS count` (number of URLs in the tweet), `24h similar tweets` (number of similar tweets produced in a time span of 24 hours), `Author num followers` (number of followers of the author), `Author num followees` (number of people followed by the author), `neg words` (number of words with negative sentiment), `Num pos emoticons` (number of emoticons associated with a positive sentiment), and `Mentions count` (number of Twitter users mentioned).
- Each of the selected signals produces a ranking of tweets. We combine them to produce a final ranking using Borda count [7], a standard voting scheme to combine rankings.
- We remove redundancy with the same iterative procedure used in the *Followers* baseline.

4 Results and Discussion

We have evaluated all systems with respect to the test subset of RepLab 2013. Figure 3 (left) compares the results of LexRank, the followers baseline and the signal voting algorithm in terms of ROUGE-2. For each entity and for each compression rate, systems are compared with the set of 1,000 reference summaries automatically generated. Figure 3 (right) shows the recall of relevant topics at different compression ratios. Finally, Fig. 4 evaluates the summaries in terms of RBP-SUM-R and RBP-SUM-B directly with respect to the manual assessments in the RepLab 2013 dataset.

In terms of ROUGE, the combination of signals is consistently better than both LexRank and the Followers baseline at all compression levels. All differences are statistically significant according to the t-test, except at 5 % compression

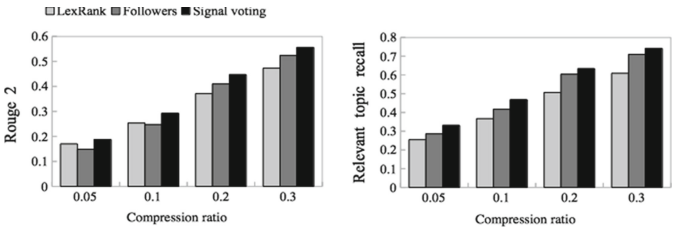


Fig. 3. Results in terms of Rouge and recall of important topics

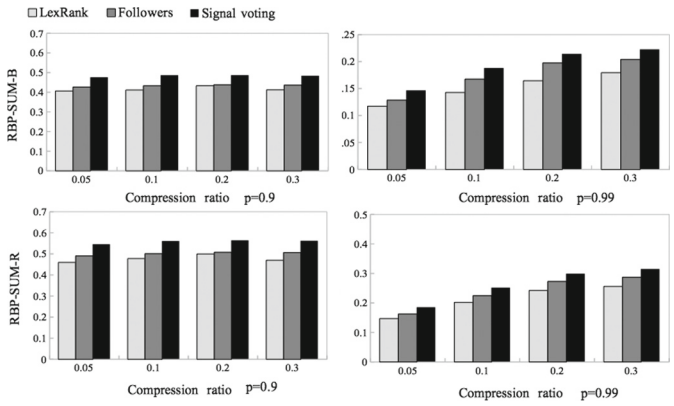


Fig. 4. Value of priority signals according to RBP-SUM

rate where the difference between signal voting and LexRank is not significant ($p = 0.08$). Remarkably, at 20 % and 30 % compression rates even the Followers baseline – which uses very little information and is completely unsupervised – outperforms the LexRank baseline. Altogether, these are clear indicators that priority signals play a major role for the task.

In terms of recall of relevant topics, the figure shows that Signal voting > Followers > LexRank at all compression ratios. In terms of RBP-SUM, results are similar. With both relevance scoring functions, signal voting outperforms the two baselines at all compression rates, and all differences are statistically significant. The only difference is that this evaluation methodology, which penalizes redundancy more heavily (tweets from the same topic receive an explicit penalty), gives the followers baseline a higher score than LexRank at all compression levels (with both relevance scoring functions).

Relative differences are rather stable between both p values and between both relevance scoring functions. Naturally, absolute values are lower for RBP-SUM-B, as the scoring function is stricter. Although experimentation with users would be needed to appropriately set the most adequate p value and relevance scoring schema, the measure differences seem to be rather stable with respect to both choices.

5 Related Work

5.1 Centrality Versus Priority-Based Summarization

Centrality has been one of the most widely used criteria for content selection [8]. Centrality refers to the idea of how much a fragment of text (usually a sentence) covers the main topic of the input text (a document or set of documents). However, the information need of users frequently goes far beyond centrality and should take into account other selection criteria such as diversity, novelty and priority. Although the importance of enhancing diversity and novelty in various NLP tasks has been widely studied [9,10], reputational priority is a domain-dependent concept that has not been considered before. Other priority criteria have been previously considered in some areas: In [11], concepts related to treatments and disorders are given higher importance than other clinical concepts when producing automatic summaries of MEDLINE citations. In opinion summarization, positive and negative statements are given priority over neutral ones. Moreover, different aspects of the product/service (e.g., technical performance, customer service, etc.) are ranked according to their importance to the user [12]. Priority is also tackled in query (or topic)-driven summarization where terms from the user query are given more weight under the assumption that they reflects the user relevance criteria [13].

5.2 Multi-tweet Summarization

There is much recent work focusing on the task of multi-tweet summarization. Most publications rely on general-purpose techniques from traditional text summarization along with redundancy detection methods to avoid the repetition of contents in the summary [14]. Social network specific signals (such as user connectivity and activity) have also been widely exploited [15].

Two different types of approaches may be distinguished: feature-based and graph-based. Feature-based approaches address the task as a classification problem, where the aim is to classify tweets into important/unimportant, so that only important tweets are used to generate the summary. Tweets are represented as sets of features, being the following the most frequently used: term frequency [16], time delay [16], user based features [17] and readability based features [15]. Graph-based approaches usually adapt traditional summarization systems (such as LexRank [5] and TextRank [18]) to take into consideration the particularities of Twitter posts [14,15,19]. These approaches usually include both content-based and network-based information into the text graph.

Concerning the subject of the input tweets, most works have focused on those related to sport and celebrity events [14,19]. These events are massively reported in social networks, so that the number of tweets to summarize is huge. In this context, simple frequency based summarizers perform well and even better than summarizers that incorporate more complex information [14]. The problem of summarizing tweets on a company's reputation has been, to the best of our knowledge, never tackled before and presents additional challenges derived from the less massive availability of data and the greater diversity of issues involved.

6 Conclusions

We have introduced the problem of generating reputation reports as a variant of summarization that is both practical and challenging from a research perspective, as the notion of reputational priority is different from the traditional notion of importance or centrality. We have presented two alternative evaluation methodologies that rely on the manual annotation of topics and their priority. While the first evaluation methodology maps such annotations into summaries (and then evaluates with standard summarization measures), the second methodology establishes an analogy with the problem of search with diversity, and adapts an IR evaluation metric to the task (RBP-SUM).

Given the high correlation between Rouge and RBP-SUM values, we advocate the use of the latter to evaluate reputation reports. There are two main reasons: first, it avoids the need of explicitly creating reference summaries, which is a costly process (or suboptimal if, as in our case, they are generated automatically from topic/priority annotations); the annotation of topics and priorities is sufficient. Second, it allows an explicit modeling of the patience of the user when reading the summary, and of the relative contribution of information nuggets depending on where in the summary they appear and their degree of redundancy with respect to already seen text.

As for our second research question, our experiments indicate that priority signals play a relevant role to create high-quality reputation reports. A straightforward voting combination of the rankings produced by useful signals consistently outperforms a standard summarization baseline (LexRank) at all compression rates and with all the evaluation metrics considered. In fact, the ranking produced by just one signal (number of followers) also may outperform LexRank, indicating that standard summarization methods are not competitive.

In future work we will consider including graded relevance with respect to priority levels in the data. In our setting, we have avoided such graded relevance to avoid bias in favor of priority-based methods, but RBP-SUM directly admits a more sophisticated weighting scheme via r_i.

Acknowledgments. This research was partially supported by the Spanish Ministry of Science and Innovation (VoxPopuli Project, TIN2013-47090-C3-1-P) and UNED (project 2014V/PUNED/0011).

References

1. Amigó, E., Carrillo-de-Albornoz, J., Chugur, I., Corujo, A., Gonzalo, J., Martín, T., Meij, E., de Rijke, M., Spina, D.: Overview of RepLab 2013: Evaluating online reputation monitoring systems. In: Forner, P., Müller, H., Paredes, R., Rosso, P., Stein, B. (eds.) CLEF 2013. LNCS, vol. 8138, pp. 333–352. Springer, Heidelberg (2013)
2. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Proceedings of the ACL Workshop on Text Summarization Branches Out, pp. 74–81 (2004)

3. Moffat, A., Zobel, J.: Rank-biased precision for measurement of retrieval effectiveness. ACM Trans. Inf. Syst. (TOIS) **27**(1), 2 (2008)
4. Amigó, E., Gonzalo, J., Verdejo, F.: A general evaluation measure for document organization tasks. In: Proceedings of ACM SIGIR, pp. 643–652. ACM (2013)
5. Erkan, G., Radev, D.R.: Lexrank: Graph-based lexical centrality as salience in text summarization. J. Artif. Int. Res. **22**(1), 457–479 (2004)
6. Radev, D., Allison, T., Blair-Goldensohn, S., Blitzer, J., Çelebi, A., Dimitrov, S., Drabek, E., Hakim, A., Lam, W., Liu, D., Otterbacher, J., Qi, H., Saggion, H., Teufel, S., Topper, M., Winkel, A., Zhang, Z.: MEAD – A platform for multidocument multilingual text summarization. In: Proceedings of LREC (2004)
7. Van Erp, M., Schomaker, L.: Variants of the borda count method for combining ranked classifier hypotheses. In: Proceedings of Seventh International Workshop on Frontiers in Handwriting recognition. pp. 443–452 (2000)
8. Cheung, J.C.K., Penn, G.: Towards robust abstractive multi-document summarization: A caseframe analysis of centrality and domain. In: Proceedings of ACL, Sofia, Bulgaria. pp. 1233–1242 (2013)
9. Mei, Q., Guo, J., Radev, D.: Divrank: The interplay of prestige and diversity in information networks. In: Proceedings of ACM SIGKDD. pp. 1009–1018 (2010)
10. Clarke, C.L., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I.: Novelty and diversity in information retrieval evaluation. In: Proceedings of ACM SIGIR 2008, pp. 659–666 (2008)
11. Fiszman, M., Demner-Fushman, D., Kilicoglu, H., Rindflesch, T.C.: Automatic summarization of medline citations for evidence-based medical treatment: A topic-oriented evaluation. J. Biomed. Inform. **42**(5), 801–813 (2009)
12. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retr. **2**(1–2), 1–135 (2008)
13. Nastase, V.: Topic-driven multi-document summarization with encyclopedic knowledge and spreading activation. In: Proceedings of EMNLP, pp. 763–772 (2008)
14. Inouye, D., Kalita, J.: Comparing twitter summarization algorithms for multiple post summaries. In: Proceedings of the IEEE Third International Conference on Social Computing, pp. 298–306 (2011)
15. Liu, X., Li, Y., Wei, F., Zhou, M.: Graph-based multi-tweet summarization using social signals. In: Proceedings of COLING 2012, pp. 1699–1714 (2012)
16. Takamura, H., Yokono, H., Okumura, M.: Summarizing a document stream. In: Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudoch, V. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 177–188. Springer, Heidelberg (2011)
17. Duan, Y., Chen, Z., Wei, F., Zhou, M., Shum, H.Y.: Twitter topic summarization by ranking tweets using social influence and content quality. In: Proceedings of COLING 2012, Mumbai, India, pp. 763–780 (2012)
18. Mihalcea, R., Tarau, P.: Textrank: Bringing order into texts. In: Proceedings of EMNLP 2004, Barcelona, Spain pp. 404–411 (2004)
19. Sharifi, B., Hutton, M.A., Kalita, J.: Summarizing microblogs automatically. In: Proceedings of NAACL, pp. 685–688 (2010)