Hierarchy Construction for News Summarizations

Jeroen B. P. Vuurens
The Hague University of Applied Science
Delft University of Technology, The Netherlands
j.b.p.vuurens@tudelft.nl

Arjen P. de Vries
CWI
Delft University of Technology, The Netherlands
arjen@acm.org

Roi Blanco
Yahoo Labs, London, England UK
roi@yahoo-inc.com

Peter Mika
Yahoo Labs, London, England UK
pmika@yahoo-inc.com

ABSTRACT

Following online news about a specific event can be a difficult task as new information is often scattered across web pages. In such cases, an up-to-date summary of the event would help to inform users and allow them to navigate to articles that are likely to contain relevant and novel details. Several approaches exist to compose a summary of salient sentences that are extracted from an online news stream for a given topic. Summaries often consist of multiple news stories, that when entwined may make it harder to read. We propose a general approach to convert non-hierarchical temporal summarizations into a hierarchical structure, that can be used to further compress the summary to provide more overview, that allows the user to navigate to specific subtopics of interest, and can be used to provide feedback to improve results. This approach reorganizes the sentences in a summary using a divisive clustering approach to capture the sentences per news story in a hierarchy.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Filtering

General Terms

Hierarchical clustering, Multi-document summarization

1. INTRODUCTION

Internet users are turning more frequently to online news as a replacement for traditional media sources such as newspapers or television shows. Still, discovering news events online and following them as they develop can be a difficult task. Although the Web offers a seemingly large and diverse set of information sources ranging from highly curated professional content to social media, in practice most sources base their stories on previously published works and add a much more limited set of new information. Thus users often end up spending significant amount of effort re-reading the same parts of a story before finding relevant and novel information. Online summarization is a crucial aspect of real-world products such as online live streams for natural disasters, product launches, financial or political events, breaking news notifications on mobile devices and topical daily news summaries like Yahoo News Digest (`https://mobile.yahoo.com/newsdigest`).

Allen et al. formalize the temporal summarization problem as follows. A news topic is made up of a set of events and is discussed in a sequence of news stories. Most sentences of the news stories discuss one or more of the events in the topic [1]. To construct a summary, significant updates are identified for the topics being tracked, relieving the users from having to sift through long lists of similar articles arriving from different news sources, and minimize the time and disruptions to users who wish to follow evolving news stories, similar to the proposal by [3]. For this purpose, several methods have been proposed, e.g. [3, 7, 9, 10]. This research focuses on news topics that are made up of multiple news stories, which are not necessarily discussed in sequence but can also overlap, e.g. for the topic Apple a news story about the iPhone patent case against Samsung can overlap with a news story about the new Apple Watch. This aspect is overlooked in the evaluation of other work for temporal summarizations, which is often focused on recall and precision-like metrics that do not consider the context in which a sentence is reported. Consider the following example:

> Mexico City Mayor Miguel Angel Mancera said many evacuations were reported in the capital but officials received no reports of damage or injuries.
>
> MEXICO CITY - A moderate 5.3-magnitude earthquake shook central Mexico on Friday, causing buildings to sway in the capital and sending hundreds of people into the streets.
>
> According to Pemex's official Twitter account, the platform, called Abkatun Permanente, caught fire early Wednesday in Campeche Sound in the Gulf of Mexico.

The first sentence refers to an event, and therefore is easier to understand when accompanied by a context that informs what this event is. If the first sentence is accompanied (or even better preceded) by the second sentence the information of the first sentence becomes clear. However, if news stories appear entwined and there are several possibilities, this can be more confusing and harder or even impossible to understand. In this example this occurs when Sentence 3 appears close to sentence 1, since evacuations, damage and injuries could also be related to a fire on an oil platform. Therefore, a flat temporal ordering of extracted news sentences may be less efficient to read when a news stream contains entwined news stories, than when there is only a single news story.

SIGIR TAIA workshop, August 13, 2015, Santiago, Chile.

In this work, we aim to improve the comprehensiveness of a summary that contains multiple news stories by capturing each of the underlying news stories in a cluster of the constructed hierarchy. We experimented on query based timelines of sentences that were extracted from news articles. This approach may apply to a broader domain of temporal summarizations, which we leave to study in future work. The hierarchy is constructed by first separating the news articles into pools that are unlikely to discuss the same news story, and then per pool use an divisive clustering approach to further separate sentences that are not likely to refer to the same news. The produced hierarchy provides the user with a more concise overview over a more diverse set of news stories, allowing to drill down the hierarchy to view specific news stories. To evaluate our approach we report the F-Score obtained on a set of news summaries for which we annotated the news stories.

The remainder of this paper is structured as follows: Section 2 discusses related work, in Section 3 we describe our approach, Section 4 discusses the experiment setup, in Section 5 we report the results obtained and finally in Section 6 we present the conclusion.

2. RELATED WORK

A common method for the temporal summarization over multiple texts is to extract their salient sentences, i.e., the sentences that are most *useful* and *novel*. In previous work, we proposed to select sentences from a stream of news articles based on their salience estimated by three factors: relatedness in the 3 nearest neighbor graph, the presence of information the user has not seen, and, a top rank among sentences in the summary when scored using the information seen over the last hour [10]. Similar to other news summarizers (e.g. [3, 7, 9]), the result is a non-hierarchical news summary.

The construction of a hierarchy for such a summary can be seen as a clustering approach over its contents. For instance, Quinlan summarizes an approach to synthesizing decision trees from information that is noisy and/or incomplete [6]. For each attribute he proposes to choose the attribute with the highest information gain, i.e. that results in the lowest entropy when used to divide the data into subsets. Cheng et al. propose to use an entropy-based method for clustering, motivated by the fact that a subspace containing clusters typically has lower entropy than a subspace without clusters [2]. Liu et al. compare the cluster quality of several feature selection methods for K-means clustering and found that in an unsupervised experiment that information gain is one of the best feature selection methods, especially when only a small amount of features is selected. After analyzing the words that were selected by each method they concluded that information gain is more likely to select discriminative terms than other methods [5].

Hierarchic-clustering methods result in tree-like classifications in which small clusters of objects (i.e. documents) that are found to be strongly similar to each other are nested within larger clusters that contain less similar objects. Hierarchical-clustering methods are divided into two broad categories, agglomerative and divisive. Divisive methods normally result in monothetic classifications, where documents in a given cluster must contain certain terms in order to gain membership. On the other hand, in polythetic clustering methods no specific terms are required for membership in a cluster, and such structures are usually the result of agglomerative methods. For information retrieval, polythetic clusterings are preferred [8]. In this research, we propose a polythetic divisive clustering method based on normalized information gain. To the best of our knowledge information gain has not been used in IR research before to cluster high-dimensional data in a noisy collection.

3. DESIGN

We describe how to build a hierarchy for a summary of sentences that were selected from a collection or stream of news articles. We propose the construction as a two step divisive clustering process: (1) separate the news articles that were used for the summary into *document pools* that are unlikely to discuss the same news, and (2) per pool, form clusters based on the most different sentences.

The objective for step (1) is to separate the news articles into pools so that articles that are likely to discuss the same news story are pooled together, while creating separate pools for articles that do not. For example, news articles that contain the word "apple", may partly be related to the computer company and partly to fruit which should ideally be assigned to different pools. At the sentence level there is often insufficient information to make a reliable decision, therefore we use the articles from which the sentences in the summary were selected. The pooling is based on dissimilarity (or impurity) estimated by a normalized version of the *information gain*. Information gain has been successfully used when considering a fixed number of non-sparse dimensions of a data set [6, 2], however, in text collections the information gain is not comparable between subsets that use a different number of features. We introduce *normalized information gain* to address this problem, a measure that returns 0 for identical subsets and 1 for disjoint subsets. Formally, in Equation 1 H is the entropy over the words w in a bag of words s, given the size of the content c and $f_{w,s}$ is the frequency of word w in s. In Equation 2, the information gain IG is defined for separating a group of content $s+t$ into two separate bags of words s and t, with $|s|$, $|t|$ and $|s+t|$ as the number of words contained. In Equation 3, IG_{max} is the maximum information gain that would be obtained given the sizes of data subsets t and s if these are completely disjoint, and in Equation 4 IG is divided by IG_{max} to normalize IG_{norm} to a value in $[0,1]$. The normalized information gain can be computed between sentences and articles and also clusters of sentences and articles by considering these to be the concatenation of the contained elements.

$$H(s,c) = -\sum_{w \in s} \frac{f_{w,s}}{c} log_2 \frac{f_{w,s}}{c} \tag{1}$$

$$\begin{aligned} IG(s,t) =& H(s+t, |s|+|t|) \\ & - \frac{|s|}{|s|+|t|} \cdot H(s,|s|) - \frac{|t|}{|s|+|t|} \cdot H(t,|t|) \end{aligned} \tag{2}$$

$$\begin{aligned} IG_{max}(s,t) =& H(s,|s|+|t|) + H(t,|s|+|t|) \\ & - \frac{|s|}{|s|+|t|} \cdot H(s,|s|) - \frac{|t|}{|s|+|t|} \cdot H(t,|t|) \end{aligned} \tag{3}$$

$$IG_{norm}(s,t) = \frac{IG(s,t)}{IG_{max}(s,t)} \tag{4}$$

We build the hierarchy as follows. In step (1), the news articles from which sentences were used in the summary are processed in order of publication time, comparing the IG_{norm} of a new article to the existing pools of articles, adding it to the pool with which it has the lowest IG_{norm} if this is below a threshold ω_d, or creating a new pool otherwise. Pools are merged when the IG_{norm} between them becomes lower than this threshold.

In step (2), we consider the news articles' sentences that were selected for the summary per pool. We apply a divisive strategy by identifying a set of most different *sentences*, i.e. that have an IG_{norm} that exceeds a threshold ω_s with all other sentences in that set. These sentences will form the initial clusters, and we iteratively add an unassigned sentence from the pool that has the lowest IG_{norm} with any of the clusters to that cluster. When all sentences

have been assigned, clusters are merged when the IG_{norm} between them is below the threshold ω_s. If a pool contains several clusters, a parent node is added to group the pool.

4. EXPERIMENT

To evaluate whether the constructed hierarchy separates the news stories that are contained in the temporal summarization for a query, we use F-Score as proposed by Larsen & Aone [4]. For this metric, the cluster hierarchy is treated as an output from an automatic multi-level routing system, in which for each news story a corresponding cluster will form automatically somewhere in the hierarchy. A parent cluster contains the information of its sub-clusters, therefore different hierarchy levels are tried to find the level that is the best match for each news story. Formally, in Equation 5, $f(t)$ is the total number of sentences that match news story t over all clusters C in the hierarchy, in Equation 6 the Precision for t in C is defined as the number of sentences in C that match news story t divided by the number of sentences in the cluster $|C|$, in Equation 7 the Recall is defined as the number of sentences in C that match the news story t divided by the total number of sentences that match that news story in the hierarchy. Then in Equation 8, for the computation of F, the clusters are considered to include the information of its subclusters, and the F for a news story t is the maximum F-measure over all clusters. In Equation 9, the $F-Score$ is the weighted average of F-measures over all annotated news stories.

$$f(t) = \sum_{C} |\{s \in C | s.topic = t\}| \tag{5}$$

$$Precision(C,t) = \frac{|\{s \in C | s.topic = t\}|}{|C|} \tag{6}$$

$$Recall(C,t) = \frac{|\{s \in C | s.topic = t\}|}{f(t)} \tag{7}$$

$$F(t) = \max_{C} \frac{2 \cdot Precision(C,t) \cdot Recall(C,t)}{Precision(C,t) + Recall(C,t)} \tag{8}$$

$$F-Score() = \frac{\sum_t f(t) \cdot F(t)}{\sum_t f(t)} \tag{9}$$

$$\tag{10}$$

The data set for this evaluation consists of temporal summaries for the queries "apple", "mexico", "cyclone" and "volcano", that were tracked by NewsTracker [10] over the period between March 1st 2015 and June 1st 2015 using the articles published by 70 online news providers (e.g. CNN, NY Times, BBC News). For the original summaries, we identified the news stories contained and assigned every sentence to one news story. We identified entity related news stories such as a named cyclone, a volcano, an apple product, but also event related news stories such as "patent lawsuit apple samsung", "Jalisco Cartel boss arrested". Some news can both be considered a separate news story or part of a bigger news story, as a rule of thumb we annotated the level at which the other sentences provide a useful context, e.g. a violent response to the arrest of the Jalisco Cartel would not be a separate new story reasoning that sentences about the arrest provide a context that makes to subsequent reaction easier to read. In Table 1 we report the number of sentences and news story per query. All news stories were annotated, including for instance the story "an apple a day does not keep the doctor away" for the query "apple". Although we were not interested in this particular story, we focus this evaluation on clustering quality and therefore annotated these news stories as well, and in practice, correctly clustered irrelevant information may at least hurt the quality as perceived by the user less than scattered irrelevant information. Sentences that are not related to a news story are counted in the cluster size and therefore discounts the precision for the cluster they appear in. The news summarizations used for input and the annotated ground truth can be found at `http://hns-dataset.github.io/`. Note that we do not evaluate the effectiveness of the summary used as input.

Table 1: Queries used for evaluation with the number of sentences in the summary and the number of news stories identified.

Query	Sentences	News stories
Cyclone	87	8
Volcano	90	8
Apple	373	60
Mexico	552	145

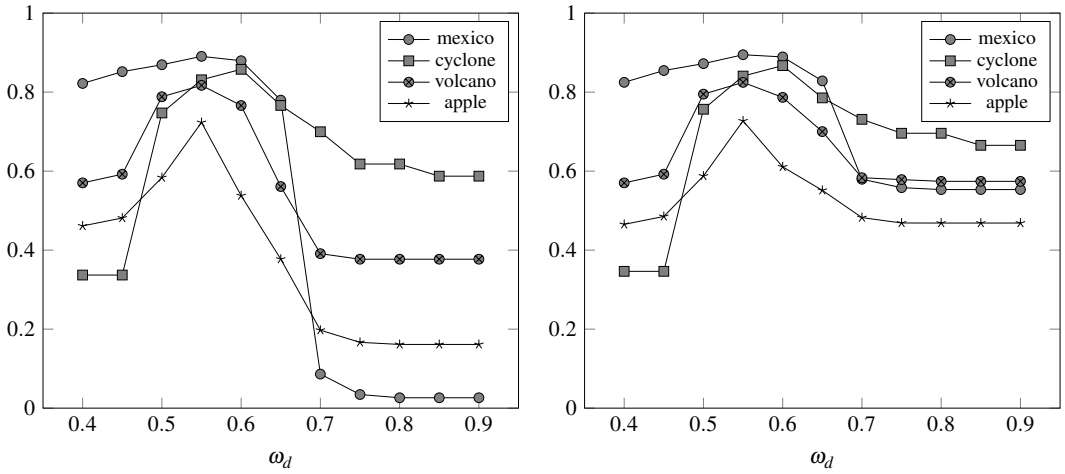
Although for the task we defined there is no state-of-the-art baseline available to compare with, we add a comparison to two simple baselines as reference of the obtained results. The "Single Linkage" method clusters sentences based on nearest neighbor clustering where each sentence is assigned the sentence with the highest cosine similarity between a TF-IDF representation of their contents. The "Multiple Linkage" links all sentences that have a cosine similarity between them that exceeds a threshold. For the latter, we chose the threshold that obtained the highest F-Score per query. For both methods, the connected subgraphs represent the clusters that were formed.

5. RESULTS

We first inspect the effectiveness of document pooling (step 1) in those runs in which we did not use step 2. In Figure 1a, we compare the F-Score per query of each resulting hierarchy when varying the threshold that is to pool documents. Naturally, extreme values for ω_d do not score well, since setting it too high creates a single pool only, without hierarchy, while low values of ω_d establish a separate pool per document. We observe a query dependent "sweet zone", which is possibly related to the coherence between separate news stories of a query; which in our observation was lower for "mexico" than for "apple". In this experiments, a setting for ω_d of 0.55 is (close to) optimal for every query.

Next, we include step (2) and plot average F-Score over all four queries for a sweep of ω_d and ω_s (Figure 2). The results indicate that when the pooling parameter ω_d has an optimal setting, the effect of the grouping parameter ω_s does not further improve clustering quality. However, when ω_s is set too high and unrelated news articles are pooled together, the creation of sub-clusters within its pool comes to rescue. In this experiment, setting ω_d to a value of 0.75 - 0.80 maximizes results for cases with too coarse pooling granularity. By comparing Figure 1a with Figure 1b we see that step (2) is only effective when ω_s is overestimated, and when used with a fixed threshold $\omega_s = 0.75$ this improves the clustering quality.

In Table 2, we list the F-Score that is obtained when the optimal parameter settings are used for each query. For reference, we include the F-Score obtained when using Single Linkage and Multi Linkage. The results indicate a potential of normalized information gain to cluster textual summaries. However, these results are only realistic if in practice near optimal settings for the parameters can be found. An interesting direction for future work is to test the stability of these parameters over a larger collection, or when they appear not stable to predict ideal settings per query, for instance based on observed differences between news articles.



(a) when not using step (2).

(b) when using step 2 with a fixed setting $\omega_s = 0.75$.

Figure 1: Comparison of the F-Score per query when varying ω_d.

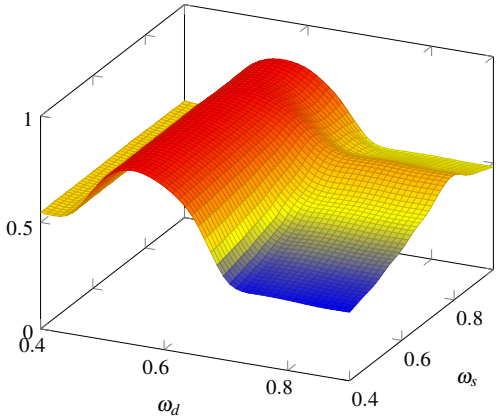


Figure 2: Comparison of the average F-Score over all queries when varying ω_d and ω_s.

6. CONCLUSION

In this study we propose a novel approach to hierarchically cluster news stories that are part of a news summary based on separation of unrelated news as measured by their normalized information gain. We evaluate this approach by assessing the cluster quality produced by our method which was measured by the F-Score over the constructed hierarchy. We compared our approach using a ground truth of news stories that were labeled to the sentences in the summary. We conclude that this approach has the potential to accurately capture news stories in a hierarchy. An interesting direction for future work is to personalize a news summary using the clusters obtained.

Table 2: The F-Score obtained when optimal parameter settings are found for the proposed approach and the Single Linkage and Multiple Linkage baselines.

Query	ω_d	ω_s	F-Score	Single Linkage	Multiple Linkage
Cyclone	0.60	0.75	0.867	0.600	0.644
Volcano	0.55	0.60	0.835	0.425	0.410
Apple	0.55	0.90	0.729	0.458	0.385
Mexico	0.55	0.75	0.895	0.562	0.563

References

[1] J. Allan, R. Gupta, and V. Khandelwal. Temporal summaries of new topics. In *SIGIR*, 2001.

[2] C.-H. Cheng, A. W. Fu, and Y. Zhang. Entropy-based subspace clustering for mining numerical data. In *SIGKDD*, 1999.

[3] E. Gabrilovich, S. Dumais, and E. Horvitz. Newsjunkie: providing personalized newsfeeds via analysis of information novelty. In *WWW*, 2004.

[4] B. Larsen and C. Aone. Fast and effective text mining using linear-time document clustering. In *SIGKDD*, 1999.

[5] T. Liu, S. Liu, Z. Chen, and W.-Y. Ma. An evaluation on feature selection for text clustering. In *ICML*, 2003.

[6] J. R. Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[7] D. Radev, J. Otterbacher, A. Winkel, and S. Blair-Goldensohn. NewsInEssence: summarizing online news topics. *Communications of the ACM*, 48(10):95–98, 2005.

[8] A. Tombros. *The effectiveness of query-based hierarchic clustering of documents for information retrieval*. PhD thesis, University of Glasgow, 2002.

[9] G. B. Tran, M. Alrifai, and E. Herder. Timeline summarization from relevant headlines. In *ECIR*, 2015.

[10] J. B. P. Vuurens, A. P. de Vries, R. Blanco, and P. Mika. Online temporal summarization of news. In *SIGIR Demo*, 2015.