Semi-Supervised Events Clustering in News Retrieval

Jack G. Conrad
Thomson Reuters
Corporate Research & Development
Saint Paul, Minnesota 55123 USA
jack.g.conrad@thomsonreuters.com

Michael Bender
Thomson Reuters
Thomson Reuters Global Resources
Baar, Zug 6340 Switzerland
michael.bender@thomsonreuters.com

Abstract

The presentation of news articles to meet research needs has traditionally been a document-centric process. Yet users often want to monitor developing news stories based on an event, rather than by examining an exhaustive list of retrieved documents. In this work, we illustrate a news retrieval system, *eventNews*, and an underlying algorithm which is event-centric. Through this system, news articles are clustered around a single news event or an event and its sub-events. The algorithm presented can leverage the creation of new Reuters stories and their compact labels as seed documents for the clustering process. The system is configured to generate top-level clusters for news events based on an editorially supplied topical label, known as a 'slugline,' and to generate sub-topic-focused clusters based on the algorithm. The system uses an agglomerative clustering algorithm to gather and structure documents into distinct result sets. Decisions on whether to merge related documents or clusters are made according to the similarity of evidence derived from two distinct sources, one, relying on a digital signature based on the unstructured text in the document, the other based on the presence of named entity tags that have been assigned to the document by a named entity tagger, in this case Thomson Reuters' *Calais* engine.

In: M. Martinez, U. Kruschwitz, G. Kazai, D. Corney, F. Hopfgartner, R. Campos and D. Albakour (eds.): Proceedings of the NewsIR'16 Workshop at ECIR, Padua, Italy, 20-March-2016, published at http://ceur-ws.org

1 Introduction

1.1 Motivations

Thomson Reuters has been exploring alternative models for organizing and rendering articles found in its news repository. Whether the users are editors, financial analysts, lawyers or other professional researchers, a more effective means of examining a set of event-related news articles beyond that of a ranked list of documents was expressly sought. The presentation of news articles based on events aligns well with contemporary research use cases, such as those arising in the finance and risk sectors, where there is a salient need for more effectively organized news content through the lens of events. Other news organizations such as Google have experimented with news clustering, but in the absence of the concrete use cases of Thomson Reuters' professional users.

This project uses semi-supervised clustering capabilities in order to group news documents based upon shared news events. Germinal Reuters stories with editorially assigned labels (a.k.a. 'sluglines') are used as seed documents for event identification and organization. This task addresses the fundamental aim of the project.

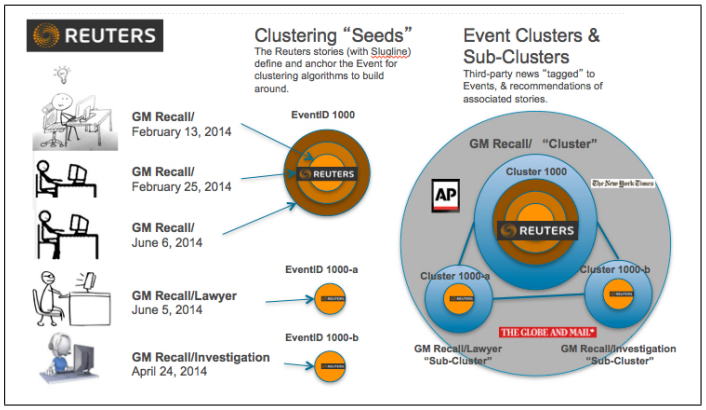

Figure 1: News Events Clustering Process

1.2 Objectives

The main objective of this project is to develop an event-centric news paradigm that solves the challenge of event validation and event story clustering at scale. This goal is in response to feedback received from consumers on news in their products. In addition to organizing news results around events rather than documents, another goal of this study is to provide a mechanism for clustering third-party (non-Reuters) news documents together with corresponding Reuters articles around common news events. This is aided by leveraging metadata tags that exist in Reuters news articles about the same topical event. Since these tags distinguish Reuters news documents from third-party content, it is possible to consider using them as the basis for grouping news articles together. The initial plan for this project was developed in conjunction with R&D's partner, the news asset owner and subject matter expert (SME), to use the initial or top-level story labels known as primary sluglines (e.g., VOLKSWAGEN-EMISSION-FRAUD/) as an organizing principle for top-level clusters, and an algorithmic means for creating lower-level clusters which can incorporate second tier story labels known as secondary sluglines (e.g., VOLKSWAGEN-EMISSION-FRAUD/COMPENSATION).

1.3 Workflow Illustration

In Figure 1, we see an example involving the "General Motors Recall" for faulty ignition switches. Through regular editorial practices, journalists write and tag event-related stories. The first story with the first "GM Recall" tag serves as the seed story for initiating the cluster. As Reuters writes and tags more stories about the GM Recall, the set of tags and text defining the GM Recall event expands. As it expands, so too does the algorithm's grasp of the event, helping it to better identify cluster candidates, in particular, within third-party news. Both the editorially generated slugline responsible for the birth of the cluster and the algorithmic identification and population of subsequent sub-clusters are depicted in the figure.

2 Previous Work

Previous work published on the topic of news events structuring has been largely academic in nature, for example, as in Borglund [6]. This thesis includes three contributions: a survey of known clustering methods, an evaluation of human versus human results when grouping news articles in an event-centric manner, and lastly an evaluation of an incremental clustering algorithm to see if it is possible to consider a reduced input size and still get a sufficient result.

In addition, there have been journal articles that have explored the computational complexity of the algorithms necessary to cluster real-time news articles [5]. But they have focused largely on the math behind the clustering rather than the use case and practitioners benefitting from it.

Some of the earliest work in this area was pursued under DARPA and NIST funding and resulted in reports written by various forums created to advance the state of the art in event detection [3, 1].

There have also been research group work and dissertations on the subject of topic detection and tracking resulting from the above research [12, 11]. Subsequent work has attempted to capture some of the structure of events and their dependencies in a news topic by creating a model of events, a.k.a. 'event threading' [10]. Yet more recently there have been actual forums under large umbrella organizations like ACL focusing on automatically computing news stories (and their titles) [2, 14].

There is also another field of research that addresses event extraction in the ACE tradition[1] that is relevant to the context of our current work, e.g., [9]. What is distinct about our present project, however, is the use of SME-defined seed stories and labels in a semi-supervised manner and the subsequent clustering stages at scale for real world news streams.

Worth noting is that one of the building blocks of the current work is represented by an initial form of 'local' clustering that involves the identification and grouping of exact and fuzzy duplicate documents [8]. This takes place in the stage immediately preceding the final, aggregated clustering step.

3 Data Resources

The news repository under examination in this effort is known as NewsRoom. It is a Thomson Reuters news aggregation platform. It consists of approximately 15-30 million documents per year from 12,000 independent news sources which consist of national and local newspapers, periodic journals, radio program transcriptions, etc. From 2012 to 2015, NewsRoom consisted of approximately 80 million news articles. These were the target of our investigation for this project (Table 1).[2]

In order to test our news workflow and the clustering algorithms that support it, we focus on chunks of data representing approximately three months of documents at a time.

Having investigated baseline news clusters in earlier research efforts (i.e., baseline algorithm, its granularity, speed and complexity) we have subsequently pursued improvements and efficiencies to help us approach

[1] `http://www.itl.nist.gov/iad/mig/tests/ace/`

[2] Thomson Reuters has long made comparably large news collections available for external research: `http://trec.nist.gov/data/reuters/reuters.html`

Table 1: NewsRoom Integrated Data Sources

Year	Sources	Document Count
2012	Reuters / Diverse	14.6M
2013	”	20.3M
2014	”	27.8M
2015	”	20.0M
Total	”	82.7M

our objectives more effectively.

4 Methods

Given our substantial data resources and our goal to build a flexible experimental retrieval environment, we have established three stages for processing and clustering a large set of news documents around news events (Figure 2). These stages include: (1) document extraction (Reuters and non-Reuters articles) from our news repository; (2) local clustering based on duplicate document detection of identical and fuzzy duplicates [7]; and (3) aggregate clustering performed over the result set from stage 2. We have determined empirically that the local clustering stage works highly effectively [8]. It is the aggregate clustering stage that has spawned ongoing research, evaluation and refinement. This stage consists of the application of hierarchical agglomerative clustering, where different types of cluster centroid representations were examined. Although we provide descriptions of each of the three processing stages below, it is the third of these stages that is the principal focus of our latest efforts and this research report.

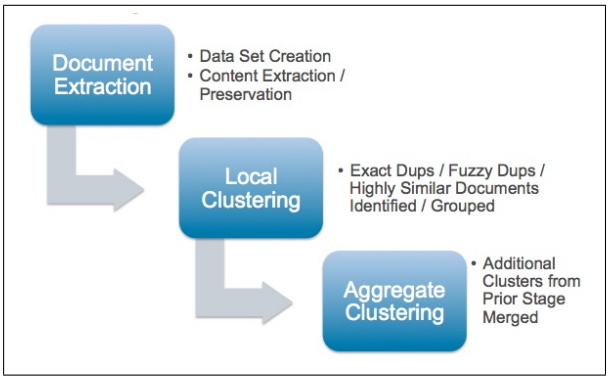

Figure 2: News Events Clustering Functional Stages

4.1 Document Extraction Stage

The document extraction process can be customized to facilitate experimentation such as that undertaken for this study. NewsRoom represents a news repository of both Reuters and non-Reuters sources covering roughly 12,000 news sources. Given a date range, e.g., [20141001T0000000Z 20141231T235959Z], one can extract all of the ‘recommendable’ news documents in the repository, or some user-defined subset of them. Since the repository contains substantial numbers of Reuters and non-Reuters financial documents, for example, some stories are largely non-textual, e.g., containing tabular information only; very short, e.g., stubs for stories in progress stories; or meta-data snippets for topics that were not substantiated. These types of documents would be considered non-recommendable and thus are not retrieved for subsequent processing. In general, over half of the documents in the repository would be classified as recommendable for this use case. The NewsRoom environment comes with a recommendation classifier. Additional details beyond those provided above would be beyond the scope of our current focus.

The extraction process results in all recommendable documents being loaded from the repository to an Apache Derby JDBC relational database. The tabular data structures that store the documents and subsequent clusters contain basic information such as doc_id, dataset_name, doc_date, title, article_source, source_url (if applicable), body, body_length, together with tens of additional features that can be used to discriminate and used by various classifiers, e.g., primary news code, short sentence count, ticker count, quantity of numbers, quantity all-caps, quantity of press releases, etc. These additional features are available for subsequent downstream processing such as classification, routing or clustering.

4.2 Local Clustering Stage

The next process, local clustering, is designed to rapidly and efficiently identify initial clusters based on documents that satisfy criteria for identical or fuzzy duplicates. Documents are compared using two types of digital signatures that harness the most discriminating terms, one, smaller and more compact leveraging O(10) terms, is used to identify identical duplicates; another, more expansive, leveraging O(100) terms, is used to identify fuzzy duplicates. The process being executed uses techniques reported on in [8]. For this application, a rolling window of n days is used, where ($n < 10$). Documents falling within this window are compared. Heuristics relying on features such as doc_length, are also invoked to reduce the number of comparisons required. For example, when a document exceeds the length of another by 20% or more, though they may satisfy a containment relationship, according to our definition, they would not be considered ‘duplicates.’

4.3 Aggregate Clustering Stage

During the third, aggregate clustering stage, the clusters are initiated via seminal Reuters articles contain-

ing slugline tags. These tags are distinct from headlines, as shown in Figure 3. The articles with suglines may be singletons or they may exist in one of the local clusters formed in preceding stage. Both of these 'objects' qualify to serve as a cluster 'seed.'

```
<slugline separator="-">VOLKSWAGEN-EMISSIONS-SCANDAL/</slugline>
<headline>Volkswagen could face $18 billion penalties from EPA</headline>
<dateline>WASHINGTON/DETROIT, September 18 (Reuters)</dateline>
<by>Timothy Gardner and Bernie Woodall</by>
<creditline>Reuters</creditline>
```

Figure 3: Reuters Article - Slugline Illustration

Two main challenges confronted when implementing this hierarchical, agglomerative clustering stage were, first, finding the best set of features and metrics to decide whether a pair of singletons or local clusters justify merging into larger clusters while still remaining sufficiently cohesive, and, second, identifying the optimal sequence for comparing these clusters when considering merging (Figure 4).

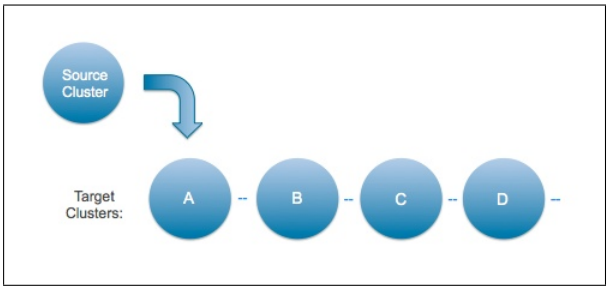

Figure 4: Approaches to Source to Target Merging

Based upon observations made by subject matter experts who created exemplar news clusters to support the project, we determined that there were two, often independent, means by which documents could be identified as belonging to the same news event. One involves the unstructured text of an article; the other involves the structured text, in our case, documents that have been tagged by the Calais named-entity tagging engine [13, 4]. Given that articles involving news events can be found to be similar based on either of these two feature spaces, our approach to aggregate (stage 3) clustering is robust: a decision to merge two of these documents or local clusters can be based on the similarity between the unstructured text of two objects, the tagged named entities that have been identified by Calais (listed below), or both.

- People – person name entities
- Reuters Instrument Codes (RICs) – for companies
- Reuters Classification System (RCS) – for topics & industries
- Topics – domain independent topical phrases
- Smart Terms – topical taxonomy terms

Operationally, the hybrid feature set described above is used to decide whether or not to merge two clusters. It consists of two data structures, both repesented in vector form. The first is a term-based vector. It is used to determine the degree of overlap between two cluster centroids, constituted by two central 'documents' (e.g., longest, most recent, true centroid, etc.). The second is a tag-based vector, representing a set of Calais tags present in the cluster's documents. The similarity measures used in each of these cases is thresholded, with the threshold determined empirically. In the case of the term vectors for the unstructured text, the thresholds are set high, although not as high as those for duplication detection used in stage 2. In the case of the set of Calais tags for the structured text, a weighted sum is used, whereby various combinations of named entities can be assembled to satisfy the threshold for merging.

Table 2: Experimental Processing (4 QTR 2014)

Stage	Name	Type	Count
1.	Document Extract	`documents`	3.63M
2.	Local Clustering	`clusters`	2.10M
3.	Agglom. Clustering	`clusters`	1.67M

5 Evaluation

Given the objectives of this study with respect to retrieval performance and organizational structure, evaluation is an essential piece of the validation process. After having conducted a number of trials to establish various thresholds (document or cluster similarity, named entity similarity, etc.), we conducted a trial which focused on a number of news events chosen by subject matter experts (SMEs) from the final quarter of 2014. We focused on the set of high-level news events shown below.

1. Halliburton Buying Baker Hughes (Nov. 13, 15)
2. Defense Secretary Hagel Resigns (Nov. 24, 25)
3. Air Asia Crash (Dec. 28, 31)
4. Pope Urges Tolerance in Turkey (Nov. 28)
5. Lufthansa Braces for Next Strike (Dec. 3)
6. Iran Rouhani Says Will Try to Clinch Nuclear Deal in Talks (Dec. 15)
7. Alstom Nearing $700M Bribery Settlement (Dec. 16)

For each of the events identified, result sets were created and stored in worksheets (Table 2 presents dataset details). The result sets consisted of numerous clusters on the subject of the event (often involving named entities such as Halliburton, Hagel, the Pope, Rouhani, Alstom, etc.), some of which are on the topic of the news event, some of which address the entity in other contexts. For those that *were* on the subject of the event, the clusters represent sub-topical (second-level) clusters (see VW example in Section 1.2). Re-

garding the result worksheets, in addition to doc ids, they included local cluster and batch cluster ids, date and time stamp, document title, document length and URL link to the complete news article (if available). The worksheets were presented to two evaluators, both subject matter experts from the news domain.[3]

Two metrics were used to evaluate these experiments. First, the assessors scored each cluster for coherence and accuracy, making sure that all of the documents that belong to a specific cluster were present, and that all of the documents that didn't belong were not present. The cluster database was queried broadly, e.g., 'Defense Secretary Hagel', in order to permit the assessors to have access to clusters both about and not about the event in question, again, in order to inspect those documents that belong in the relevant clusters and those that do not. For this task, they used a five-point Likert scale, A (very good) thru F (very weak), codified as 5-to-1.[4] Secondly, the assessors determined a 'cluster edit distance' for each cluster solution, indicating which sub-clusters they would merge and which they would split, if any, to achieve an optimal solution. Each merge or split step would be the cluster equivalent of an 'edit' in the standard character-based edit distance measure. The results of this assessment task are presented in the Table 3.

In general, we see that with few exceptions, the majority of clusters returned for our queries were about the underlying event(s) (Table 3, column 4). In addition, the coherence/accuracy scores for the clusters reviewed were in the 4.0 or 'B' range, some higher, some lower. When the same entities, but out-of-event clusters are included (column 3), their scores are slighty higher, still in the 4.0 or 'B' range.[5] In terms of the cluster edit distances measured, for the seven news events represented in the table, the mean number of 'splits' required for each cluster set was λ=1.15 (σ=1.2) while the mean number of merges was λ=4.7 (σ=4.3).

Clearly the larger numbers appearing in the context of merges have been influenced significantly by a couple of the outliers found in the list of events, i.e., nos. 2 and 7. In the case of the latter, there was greater variety in the news sources and articles reporting on the statements coming from the Iranian leader, and as a result, the algorithm may not have captured the overarching similarity among the documents. In addition, there was a greater variety of persons mentioned in these articles who were responding to President Rouhani.

Regarding the queuing strategy and its impact on agglomerative clustering and merging (Figure 4), we conducted a series of experiments that involved different strategies, including least-recently-used and most-recently-used. Other strategies tended to have a significant impact on computational complexity insofar as it was necessary to perform real-time tracking of dynamic cluster characteristics. Although the spectrum of considerations involved in those experiments may be beyond the scope of the current reporting space, we found that the most-recently-used was as effective a queuing strategy as the majority of others investigated.

There is clearly room for improved performance and additional evaluation. One way of addressing some of the disparities revealed above is by tuning the joint thresholds for document signature and named entities tagged. Alternatively, one could have the thresholds learned and optimized depending on features associated with the documents (e.g., range of idfs in the signatures, number and type of entities in the document). Moreover, one could use a variable weighted sum of the similarity scores, depending on the contribution of the named entities and distinguishing terms present in the articles being compared.

6 Conclusions

The news events clustering efforts summarized in this report and depicted in Figure 1 represent a combination of semi-supervised clustering techniques and human-generated, labeled data. They aim to deliver an effective solution by leveraging Reuters' labels and validating the scope of events at scale. The ultimate goal of the study is to determine to what extent combined human-computer resources can produce event-based clusters that are considerably more useful – i.e., more effective – than exhaustive lists of unstructured documents. In addition, third-party content can be gathered and organized around existing clustered content based upon Reuters' own editorially labeled and classified news events. The variety of challenges confronted – using Reuters' metadata, getting the granularity right, and scaling the solution – all depend on the right mix within this integration. By tracking the steps outlined above, we anticipate having a more robust working model available for evaluation in the near

[3]The first SME assessed the quality of both types of clusters, those about the event and those not; the second SME assessed the quality of the event clusters only.

[4]The five grades used in the American educational system are A-B-C-D-F, which range from exceptional (A) to failure (F). E is not used.

[5]Although in aggregate, the mean of the grades assigned the clusters by the two SMEs were comparable, when we calculated the weighted Kappa score for inter-reviewer agreement, we found that they were not as uniform, as the scores generally fell into the bottom quartile. The reviewers assigned identical grades in only about a third of the cases. In the majority of the other cases, they were one and sometimes two grades apart.

Table 3: Graded Assessments of News Events Clusters

No.	Event Title	No. Clusters	No. Clusters on Event	Mean Avg Score for All Clusters	Mean Avg Score for Event Clusters	
					SME #1	SME #2
1.	Halliburton Buying Baker Hughes	6	5	4.33	4.20	4.00
2.	Defense Secretary Hagel Resigns	24	17	4.02	3.94	2.94
3.	Air Asia Crash	14	7	3.93	3.64	3.50
4.	Pope Urges Tolerance in Turkey	7	6	4.29	4.17	4.33
5.	Lufthansa Braces for Next Strike	5	2	4.00	3.00	4.50
6.	Iran Rouhani Tries to Secure Nuclear Deal	59	47	3.99	3.86	3.93
7.	Alstom Nearing $700M Bribery Settlement	5	3	3.80	3.50	4.50
T.	Total			Avg = 4.05	Avg = 3.73	Avg = 3.95

future. Anticipated amendments or extensions of the model are addressed below.

7 Future Work

In future work, we will extend our evaluations by comparing our results with exemplar clusters identified by our SMEs, both in terms of granularity and in terms of completeness, at the top, topical cluster level and lower, sub-topical level of resulting clusters. This form of assessment addresses overall cluster precision. We will also need to conduct tests that approach evaluating recall, i.e., of all the possible news events in the data set or sample, how many do we capture and represent at top and lower levels of the shallow hierarchy?

8 Acknowledgments

The authors thank Sarah Edmonds at TRGR for her diligent work assessing result sets. We are also grateful to Brian Romer with Reuters Data Innovation Lab for his innovative work on the UI and demo (to be shown at the workshop).

References

[1] *Topic Detection and Tracking Workshops*, Washington, D.C., 2004. NIST.

[2] *First Workshop on Computing News Storylines (CNewS 2015)*, Beijing, PRC, July 2015. ACL.

[3] James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. Topic detection and tracking pilot study final report. In *DARPA Broadcast News Transcription & Understanding Workshop*, Feb. 1998.

[4] Samet Atdag and Vincent Labatut. A comparison of named entity recognition tools applied to biographical texts. In *2nd International Conference on Systems and Computer Science (ICSCS13)*, pages 228–233. IEEE, Aug. 2013.

[5] Joel Azzopardi and Christopher Staff. Incremental clustering of news reports. *Algorithms*, 5:364–378, 2012.

[6] Jon Borglund. Event-centric clustering of news articles. Masters thesis, University of Uppsala, Sweden, Oct. 2013.

[7] Jack G. Conrad, Joanne C. Claussen, and Jie Lin. Information retrieval systems with duplicate document detection and presentation functions. U.S. Patent #7,809,695, Oct. 2010.

[8] Jack G. Conrad, Xi S. Guo, and Cindy P. Schriber. Online duplicate document detection: Signature reliability in a dynamic retrieval environment. In *Proceedings of the 12th Conference on Information and Knowledge Management (CIKM03)*, pages 243–252. ACM Press, Nov. 2003.

[9] Qi Li, Heng Ji, and Liang Huang. Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the ACL*, pages 73–82. Association for Computational Linguistics, Aug. 2013.

[10] Ramesh Nallipati, Ao Feng, Fuchun Peng, and James Allan. Event threading within news topics. In *Proceedings of the 13th Conference on Information and Knowledge Management (CIKM04)*, pages 446–453. ACM Press, Nov. 2004.

[11] Ron Papka. *On-Line New Event Detection, Clustering, and Tracking*. Ph.d. thesis, University of Massachusetts - Amherst, Sept. 1999.

[12] Jakub Piskorski, Hristo Tanev, Martin Atkinson, and Erik van der Gout. Cluster-centric approach to news event extraction. In *2008 Conference on New Trends in Multimedia and Network Information Systems*, pages 276–290, 2008.

[13] Thomson Reuters. Open Calais Named^{TM} Entity Tagging Engine. `http://www.opencalais.com`, 2016.

[14] Piek Vossen, Tommaso Caselli, and Yiota Kontzopoulou. Storylines for structuring massive streams of news. In *Proceedings of the First Workshop on Comparing News Storylines*, pages 40–49. ACL and Asian Federation of NLP, July 2015.