Automatic Event Detection on Reuters News

A.Y. Mohamad, *S.M.F.D Syed Mustapha, M.S. Razali

Faculty of Information Technology
Universiti Tun Abdul Razak
Malaysia
**syedmalek@pintar.unirazak.edu.my*

ABSTRACT

Automatic Event Detection is a somewhat recent area of information retrieval research. The event detection is about spotting same news, articles that talking about similar incident and then arranges them under the same group or event. We used event clustering technique to perform Automatic Event Detection. In our experiments on Reuters corpus, we find that the system works out significantly and capable to determine same keywords and cluster the news according to their events. We also experiment our system with several different news providers due to the vagueness of the concept event.

Keywords
Text Extraction, Event clustering, Event Detection

1.0 INTRODUCTION

The constantly growing amount of various data in the world has achieved a point where standard methods of knowledge management and information retrieval are no longer sufficient and need to be combined with other methods. The challenge is to build automated, unsupervised methods which can then deal with huge sets of data without costly and much human efforts.

Automatic Event Detection is the task of spotting same news, articles that talking about similar incident and then arranges them under the same group or event. An event is a dynamic topic, a subject that is discussed deeply in the news at some period of time. Here the definition of an event is in terms of the documents as an attribute or a characteristic that a certain groups of documents share while another group does not. An event is not a single incident one can point to, but rather a topic that springs up and needs to be detected. Thus, we determined to explore the applicability and effectiveness of Natural Language Processing methods for automated event clustering.

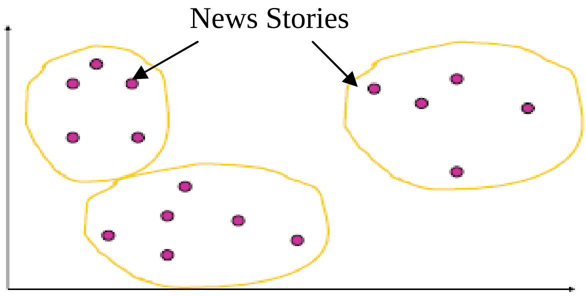

Figure 1: Example of event clustering

Cluster analysis is a well-known technique for creating order in huge sets of data (not only textual data). Event clustering as a method for arranging documents into groups is usually used in combination with the following Information Retrieval (IR) methods (Forster 2006):

- Provide support for presentation of Information Retrieval results - Typical IR methods, even plain text search (for example finding documents containing a specific word or words), are influential and proven methods. But, they are frequently overwhelmed by a large number of irrelevant results, particularly in large data sets. By event clustering, it is feasible to expand this simple query by picking a relevant document from the list of results and apply it as the query. The results of this technique will be more relevant and significantly fewer in number.
- Provide support for document retrieval - Document retrieval using event clustering is centered on the same theories as classic IR but searches organized document collections instead of unorganized clusters. For that reason, the query results are not covered by ambiguous terms.
- Provide direct access to documents – These techniques are basically based on following or mediating user’s actions and queries in a

inadequate environment and do not depend exactly on text retrieval.

Event clustering has two major components: 1) document representation and 2) cluster analysis. Document representation handles with translating documents (web pages, articles, etc.) into structures appropriate for clustering (Forster 2006). This is generally completed by representing documents numerically as vectors and matrices. Cluster analysis contains methods for designing meaningful data clusters from the data structure formed by document representation methods. Clusters are designed after relating (estimating the distance) the numerical representations of the documents.

The motivation of our work is in the similar spirit with the work presented for solving Topic Detection and Tracking tasks (Allan 2002). However, instead of considering to clusters events from the news topics like in a classic topic modeling approach, our goal is to create structured event representations and to clarify the event interactions that present in these representations.

2.0 PROBLEM

The research aims are to automatically manage news articles into clusters where each cluster corresponds to a particular event. This research in literature is also called event detection (Yang et al. 1998, Papka et al. 1999) but without unsupervised methods. One example of the well recognized online resources that arrange information in such method is Google News (Google 2008). Google News offers a service called Google Alerts. Google Alerts are emails automatically sent to user when there are new Google results for user search terms. User can also choose to have user alerts delivered via feed to the feed reader of user choice (e.g., Google Reader or add the feed to user iGoogle page). They currently offer alerts with results from News, Web, Blogs, Video and Groups. For news alert, it offered an email aggregate of the latest news articles that contain the search terms of user choice and appear in the top ten results of user Google News search.

There are some difficulties regarding data clustering, the main problem is combinatorial explosion, a frequent issue when dealing with large data sets. A piece of information used to reduce combinatorial explosion in data clustering is the time of article published. Practically, every event clustering system applies this important piece of information, i.e. event detection evaluates clusters on documents published in a particular time frame. To exactly cluster data, the similarity function between every two data points has to be calculated.

Another piece of information regularly used in news event detection is the source of the article, i.e. the publisher. The assumption is that one online publisher will not create more than one article on a singular event. But, there is certainly an erroneous mapping of reports from different publishers regarding consecutive events (article ^{n1} covering events another publisher covered in more than one article). However this problem is not of interest in this paper.

The evaluation of the efforts is most often carried out through a data sample organized by hand. In this research we used a sample of online news published by Reuters. The results of this process are intended at the representation of the data gathered and all articles covering one event would be presented as one entity.

3.0 EXPERIMENT

The data used in this experiment consist of 925 news stories obtained from Reuter corpus. Reuter's articles contain many unwanted tags and characters that need to be eliminated. This requires a process called HTML (or other) tags removal. Then, each article requires a new ID for the ease of tracking in the future work. After removing HTML tags, the articles will go through a process called stop words removal. A stop word has no great significance and a type of word that come out very frequently in a text collection. In the Information Retrieval (IR) community, stop words are defined as grammatical words. Example of stop words is prepositions, coordinators and determinants (a, an, the, in, of, on, are, be, if, into, which). Those words are worthless for search and retrieval objectives. Thus, to clean them out, a list of predefined stop words must be developed first. The program will then detect and finally remove all the stop words in the corpus based on the predefined list.

Finally, a program to perform word stemming is executed. The mean of word stemming is the process of suffix removal to produce word stems. This is done to group words that have the same conceptual meaning, such as walk, walker, walked and walking. The Porter stemmer (Porter 1980) is a well known algorithm for this duty. Its main purpose is as part of a term normalization process in Information Retrieval systems. Today, this stemmer became the standard algorithm used for English stemming. The Porter stemming algorithm has been used to normalize the terms in the Reuters corpus collection.

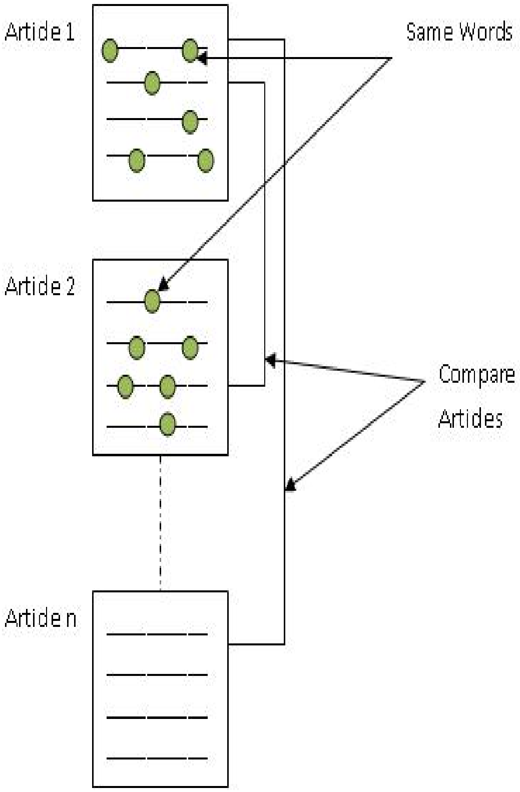

Figure 2: Finding Same Keywords inside Reuters Articles

Therefore a corpus of 925 news articles and 79357 tokens is constructed and the distribution of interest is calculated in order to use tf-idf measure (Jones 1973). Then the words in one article will be compared with other articles to find the same keywords.

Then the formula to find same articles are $(sw^{n1} + sw^{n2}) / (tw^{n1} + tw^{n2})$ where sw is the same word between article ^{n1} and article ^{n2}. Also tw is the total word for article ^{n1} and article ^{n2}. We are playing with the threshold and below are the result we managed to get:

Table 1: Threshold Setting

0.1 – 0.2	Find articles with same keywords
0.3 – 0.7	Find articles with same events
0.8 – 1.0	Find exactly same articles, correction articles.

4.0 RESULTS / DISCUSSION

The goal in Automatic Event Detection is to find a set of keywords that can link to several news articles. A set of keywords are the same words used inside news articles.

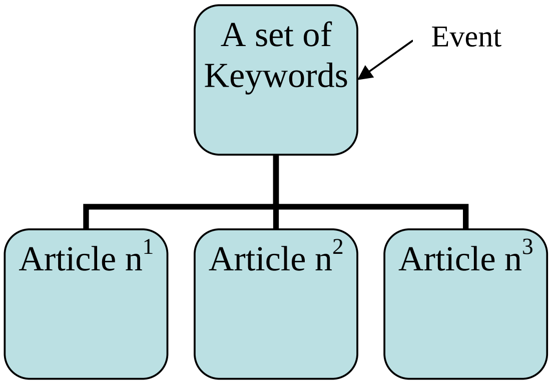

Figure 3: Event and Articles Relation

The estimation of our clustering is completed using the B-Cubed metric (Bagga and Baldwin 1998). Where C_i be the symbol for the cluster that document d_i gets clustered in, and M_i be its manual cluster (i.e. from the ground truth). The B-Cubed metric then evaluates for each document its precision (how many of the other documents in its automatic cluster should be in it?) as $\frac{|C_i \cap M_i|}{|C_i|}$, and its recall (how many of the documents in its manual cluster are in its automatic cluster?) as $\frac{|C_i \cap M_i|}{|M_i|}$. The total clustering precision and recall are gathered as the average for all documents.

Our main observe on the B-Cubed metric is the fact that it gives a singleton clustering (every document in its individual cluster) with a precision of 100%, as no document is clustered together with an unrelated one. Surely, recall will be very low in that matter. So we apply the F1 values, as these offer clear picture on both precision and recall.

Table 2: Results from Reuters

Threshold	# of Events	Articles Captured
0.2	670	911
0.5	90	149
0.8	45	83

To make our experiment more convincing, we also test our system to others online newspapers provider. We tested on 4 events consist of 30 articles taken from The Star and The Malay Mail:

1) Tiger Woods scandal – 10 stories.
2) Bus accident – 5 Sani Express, 5 AB Express.
3) Cristiano Ronaldo transfer to Real Madrid – 5 stories
4) Michael Jackson death – 5 stories.

Table 3: Results from The Star

Threshold	# of Events	Articles Captured
0.2	18	30
0.5	2	4
0.8	-	-

Table 4: Results from The Malay Mail

Threshold	# of Events	Articles Captured
0.2	15	21
0.4	2	4
0.8	-	-

For 0.2 thresholds, the number of events is higher (18 and 15) even we only had 4 events inside the database. This is because our system will give new definition for the event i.e. stories about Tiger Woods scandal also discuss about he involved in car accident and will be connected with the bus accident stories although the two stories are in different event. Our system gets lower number of events (2 instead of 4 events) in 0.4 and 0.5 thresholds because we wanted to extract only the correct stories according to the correct events. Therefore, our system precision will be higher and can be trusted to give users only the articles they wish to seek. There is no result for 0.8 thresholds because the articles taken had no identical or correction stories.

5.0 CONCLUSION

In this research, we automatically detect event for several news providers using event clustering. This is an unsupervised method in order to detect similarities and differences between text documents. Our main hypothesis was to relate one text document with another text document by their word frequency. In order to test the hypothesis, we investigated by manually determined the keywords used for certain events. The result confirmed our hypothesis and then we tried on unsupervised method. We managed to organize the news based on the words that frequently appear inside the news stories.

In the future, we would like to apply these techniques on online news stream and able to alert users when new stories they wanted to follow the progresses are spotted.

REFERENCES

Allan, J. (2002). *Introduction to Topic Detection and Tracking.* In Allan, J., ed., Topic Detection and Tracking, Event-based Information Organization, pp. 1–16. Kluwer Academic Publishers.

Bagga, A., & Baldwin, B. (1998). Algorithms for Scoring Coreference Chains. *In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Conference*, pp. 563–566.

Forster, R. (2006). *Document Clustering in Large German Corpora Using Natural Language Processing.* Thesis presented to the Faculty of Arts of the University of Zürich for the degree of Doctor of Philosophy.

Google News. (2008). *Google.* http://news.google.com/.

Papka, R., Croft, B. W., Barto, A. G., Danai, K., Kurose, J. F. (1999). *On-line New Event Detection, Clustering and Tracking*

Porter, M.F. (1980). *An Algorithm for Suffix Stripping.* Program 14, 3, 130-137.

Sparck, K. (1973). A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *Journal of Documentation.*

Yang, Y., Pierce, T., Carbonell, J. (1998). A Study on Retrospective and On-line Event Detection. *In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM press*