# Survey on News Mining tasks

Hassan Sayyadi
Web Intelligence Laboratory
Sharif University of Technology
sayyadi@ce.sharif.edu

Sara Salehi
Web Intelligence Laboratory
Sharif University of Technology
sarasalehi@ace.tju.ir

Hassan AbolHassani
Web Intelligence Laboratory
Sharif University of Technology
abolhassani@sharif.edu

***Abstract*—nowadays, there are plenty of online websites related to news. Hence, new technologies, tools and special search engines are created for having access to the news on these websites. Online news is a special type of public information which has exclusive characteristics. These characteristics contribute news engines tasks such as discovering, collecting and searching to be different with similar tasks in traditional web search engines. Clustering plays conspicuous role in news engines tasks. In this paper we study various tasks in news engines and also focusing on clustering applications in them.**

***Index Terms*— News, Clustering, News Retrieval, News Mining**

## I. Introduction

NOWADAYS, there are plenty of online websites related to news. Traditional news agencies give their information to their clients via corresponding websites. Hence, new technologies, tools and special search engines are created for having access to the news on these websites. As an instance, News Feeder softwares, RSS standard and Google news website (using 4500 source news) can be mentioned.

Furthermore, online news is a special type of public information which has exclusive characteristics. These characteristics contribute news engines tasks such as discovering, collecting and searching to be different with similar tasks in traditional web search engines. The existence of numerous reliable news sources (high trust) and fast news update are the two most important differences.

News engines provide many services and contain various tasks, the quality of each task can affect the other tasks quality. The most important tasks are:

- Collecting News
- News Retrieval
- Categorizing Search Result
- Summarization
- Automatic Event Detection

Moreover, Clustering is a practical and useful solution for all of news mining tasks and lead to efficient and better results.

In this paper we study various tasks in news engines and also focusing on clustering applications in them. The remainder of this paper is organized as follows: In sections 1 and 2 we will explain collecting news and news retrieval. In the next section we focus on categorizing search results. Next, in section 4 and 5 summarization and automatic event detection will be explained.

## II. Collecting News

The first necessity of each news service is to collect news to perform other tasks. Like other web search engines, news engines categorize to three groups, each uses one strategy to collect news corpuses:

1) The engines in which news are submitted to the system by humans manually.
2) Meta-search engines
3) The engines which crawl and discover news sources in the internet and extract news articles automatically.

The engines such as Vivisimo[1] and NewsInEssence[2] are the meta-search engines which don't have collecting process. In these engines, after receiving a user query, query will pass to the other search engines and their output will treated and showed to the user. On the other words, these engines receive the ranked news related to the user's query from other engines via libraries, web services or by processing other engines output pages.

The third group uses different methods for collecting news from available resources in internet. For this type of engines, one of the first and simplest practical ways is to generate news pages URL automatically. For example, a news website contains some fixed groups. Each group includes some news web pages which have a URL with a fixed format. As an instance, news in sport group has an address in the form of http://example.org/sport/n123.html. Consequently, by knowing different groups in each news website, it is possible to create all addresses just by changing news number from 1 to the number of the last news web page. This can help us in collecting the news. Because the news has distinct parts as date, title, and body which are remarkable in other tasks such

Manuscript received October 13, 2006. This work was supported in part by the Web Intelligence Research Laboratory, Sharif University of Technology.

[1] http://www.vivisimo.com/
[2] http://www.newsinessence.com/ili

as retrieval, one of the main weaknesses of this method is its disability for extracting these parts. Hence, the format of collected news pages of each source should be detected for extracting each part. Therefore, for collecting news with this method, the human's helps and manual operations are needed. By virtue of this weakness, the way of automatic news extraction for the whole process including news corpuses and their identifications such as date, title and body is much more concerned and different methods are proposed in this way.

Authors of [12] proposed novel automatic news extraction from news sites using Tree Edit Distance measure. Since the structure of a web page can be nicely described by a tree (e.g., a DOM tree), they have resorted to the concept of tree edit distance to evaluate the structural similarities between pages. Intuitively, the edit distance between two trees TA and TB is the cost associated with the minimal set of operations needed to transform TA into TB. To extract the desired news, their approach recognizes and explores common characteristics that are usually present in news portals. Their approach relies on the basic assumption that the news site content can be divided in groups that share common format and layout characteristics. This set of common layout and format features is called a template. According to this approach, the extraction task is performed in four distinct steps: (1) page clustering, (2) extraction pattern generation, (3) data matching and (4) data labeling (see figure 1).

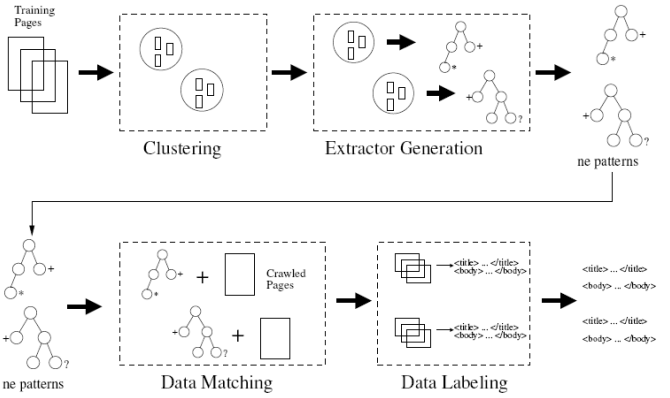


Fig. 1. Overall exteracting steps

The first step takes as input a previously crawled set of pages (a training set) and generates clusters of pages that share common formatting/layout features, i.e., share the same template. The similarity of templates which used for clustering is the Tree Edit Distance measure. Each cluster is later generalized into an extraction structure for a template, in the extraction pattern generation step. After extracting common patterns in templates, the next step is data matching which uses extracted patterns for classification of the newly added news pages to find their templates for extracting news information from those pages. Then, in the data labeling step, for each pattern they will find various parts of each template. So they try to find body, title and date for each pattern. In other words, the passage elected to be the body of the news is the longest one with more than 100 words. Further, the passage selected to be the title is one that has ranges from 1 to 20 words, has a maximum intersection with a body passage, and is the closest one to the body. The intuition behind the title selection is that most of the times the title is placed near the body and its terms usually appear in the news body.

By the advent of RSS standard and related technologies, automatic news extraction methods are no longer useful.

## III. Retrieval

After collecting news corpuses, the next step is retrieval task. In the field of news retrieval, most of the engines use traditional ways of information retrieval such as TF-IDF and PageRank. However, the special characteristics of news such as time, topic and importance have strong influence in news ranking of retrieval step. According to the special properties of news, some criteria are proposed for ranking which are appropriate in this domain. One of the important criteria is the time of news. The more the news is new, the more it is significant and the hot news is more attractive to news readers. On the other hand, the news rank can be affected by its cluster because the importance of a news is arises from the number of news related to it. Hence, the more the number of news about one subject is, the more the news is hot. On the other words, the more one cluster is large, the more its news are hot. For this reason, most of the time, news is clustered and the size of each cluster shows the importance of its news.

The affect of clustering in information retrieval was first studied by van Rijsbergen clustering theory [21]. This theory says that documents which are similar to each other will have similar results for similar queries. On the other hand, related documents are much more similar to each other than unrelated documents. Relevant to this theory, clustering can be used before retrieval which is preprocessed like [16] which creates a list for documents set. So by retrieving pages from each cluster it seem rational to retrieve other pages from respect cluster and list in result related to query.

In commercial area, there are many works done for ranking and retrieving news, but there are a few in researches. The few collegiate researches in this field are done in [1,3] and in [11] for finding news articles on the web that are relevant to news currently being broadcast. Gulli et. al [1] proposed the model for ranking news and source news. They assume 5 specifications for their model:

- Ranking for News posting and News sources: the algorithms should assign a separate rank for news articles and news sources.
- Important News articles are clustered: more important news is announced by the large number of news sources and the more the size of cluster is big, the more its news is significant.
- Mutual reinforcement between news articles and news source: hot news is announced by important source and important source announce hot news.
- Time awareness: The importance of a piece of news changes over the time. They are dealing with a stream of information where a fresh news story should be considered more important than an old one.

- Online processing: The time and space complexity of the ranking algorithm allows online processing, i.e. at some time the complexity can depend on the mean amount of news articles arriving but not on the time since the observation started.

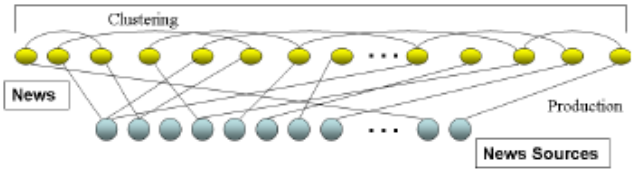


Fig. 2. Model specifications

Based on these criteria, and which is showed in figure 2, they proposed a repeated model derived from the mutual reinforcement between news and news sources. In this model the rank of a news source and news article are computed as follow:

$$R(s_k,t) = \sum_{s(n_i)=s_k} e^{-\alpha(t-t_i)} R(n_i,t) + \sum_{s(n_i)=s_k} e^{-\alpha(t-t_i)} \sum_{t_j>t_i \text{ and } S(n_i)\neq s_k} \sigma_{ij} R(n_j,t_i)^{\beta}$$

$$R(n_i,t_i) = [\lim_{\tau\to 0^+} R(S(n_i),t_i-\tau)]^{\beta} + \sum_{t_j<t_i} e^{-\alpha(t_i-t_j)} \sigma_{ij} R(n_j,t_j)^{\beta}$$

Where *R(s,t)* is the rank of news source *s* , and similarly, *R(n,t)* is the rank of news article *n* which has been released at time *t*. The value α is obtained from the half-life decay time $e^{-\alpha p}=1/2$, that is the time required by the rank to halve its value, with the relation *ρ* which is a model parameter. *β* is also a model parameter. *S(n)* indicates the news source for news article n and *σij* shows the correlation between two news article *i* and *j* which is obtained from the similarity of snippets.

## IV. Categorizing Search Result

One of the prominent problems of current search engines is showing output result in the form of ranked list without categorizing search results. If the user's query is exact enough, the results are exact and few. But for general and vague queries, the results range is very wide. On the other hand, the queries generally contain three words and users usually check the first three pages of the output result [5]. For example, assume that you want Google search engine to find related results for "Jaguar" query. "Jaguar" is a name of kind of cat and is also the drink brand, if you want the results related to the first meaning you should see all the results and extract your desire results, but due to the fact that users commonly check the first three pages, they never reach to their desired results. Consequently, it is necessary to put the output results of every query in different groups and give the categorized results to the user. By doing this method, user will have a comprehensive vision about the results. On the other hand, due to the fact that the first pages always contain the results related to one group, algorithms based on link analysis deprive users from discovering about different groups of results. Consequently, categorization results in better display and easier exploration of search results.

Classification and clustering are the main methods of search result categorization. Clustering on the results of search engines has obvious differences with traditional text clustering. One of these differences is the existence of links between web pages. Also, Due to the fact that search result clustering is an online process, fast computation is needed. The final difference is that clustering is based on small snippets instead of whole document. According to [18] good characteristics for clustering on the search results are defined as bellows:

- No necessity for all pages to be clustered
- Cluster Overlapping
- Incremental clustering

One of the reasons that commercial search engines don't do clustering is the high execution time cost. Some search engines such as Altavista[3] do aggregation in a very simple way. In "Northern Light"[4], results are classified to some custom folders. This classification is out of any intelligence and it is done on the basis of specifications such as page type, language, domain, and site and… Better commercial examples are included "Kartoo"[5], "Grokker"[6], "Mooter"[7], and specially "Vivisimo" which uses clustering algorithms. These engines have suitable qualification in respect to user interface. "Clusty"[8] is also use Vivisimo, but its clustering method is not obvious enough.

In spite of many usages of clustering in dynamic engines [8,9,14,15,16,17,19], Classification may be used in higher levels or just in one level. The reason for that idea is the need of classification to manual operations, and different groups related to the different queries. Owing to these constraints, in this section we are focusing on clustering.

Another important task which should be done after clustering is cluster labeling. Labels are essential for each cluster, so the better the labels describe clusters, the better the quality of clustering algorithm is. Referring to [8] there are two specifications for selecting cluster labels:

1. Label Readability
2. Label describing respective cluster accurately

Regarding the aforementioned statements, clustering approaches fall into two categories:

- Document-based techniques
- Label-based techniques

[3] Http://www.altavista.com/
[4] Http://www.northernlight.com/
[5] Http://www.kartoo.com/
[6] Http://www.grokker.com/
[7] Http://www.mooter.com/moot/
[8] Http://www.clusty.com/

Document-based techniques are such traditional methods that do clustering by defining similarity measures between document specifications like keyword vector. After that some words and sentences of the documents in each cluster are extracted and shown to the users as a cluster label like [15,16,19]. Clusters created by this technique do not have overlap and the quality of clusters label is under the influence of clustering accuracy. On the other hand, controlling the number of clusters and defining the similarity thresholds which determine the quality of cluster is difficult. Consequently, labels quality is not satisfactory for users. For this reason, these methods are no longer used for clustering search engines results. Works done in Document-based techniques are different from each other in two aspects:

1) Clustering algorithm.
2) Clusters and documents distance measure

Since the output shown to the user is hierarchical, a tendency for using hierarchic algorithms is high. This causes less runtime overhead for creating hierarchy (because the output is in the form of tree itself) but its quality is less than other algorithms.

In Label-based techniques informative words and expressions like high frequent words are extracted from news corpuses via statistical analyses, and among these candidate labels, those which result in better clusters are selected as final clusters labels. In this approach each label creates one cluster and each page which contains that label fall into respective cluster. This approach results in overlapped clusters. Due to the fact that one news can point to several events, overlapping seems more rational.

One of the Document-based methods is the clustering used in [9]. The first step was to remove the stopwords from each article. A stopword occurs so often that it creates no significance to a particular document For instance, the removal of the article “a” and the word “however” does not hinder uniqueness regardless of how often they appear. The second step in data cleansing was converting the resulting documents, containing nonstop words, into stemmed words. After that, they used the Porter stemming algorithm which is a process for removing the commoner morphological and in flexional endings from words in English to prevent the program from treating the word "go" and "went" differently.

The final step involved the extracting of the n most frequently occurring words in each article. If *S1* and *S2* are the sets which includes these extracted words of two news, the similarity between two news is defined as bellow:

$$Sim(n_1, n_2) = \frac{|s_1 \cap s_2|}{n}$$

$$Dissim(n_1, n_2) = 1 - Sim(n_1, n_2)$$

Where *Dissim(n1,n2)* is the distance criterion of two news. By virtue of these criteria, K-Nearest Neighbor is done on news. Single-link algorithm is also used in that paper, but the combination of both algorithms leads to the better results.

On the other hand, Authors of [8] proposed efficient Label-based model by using novel criteria for phrase ranking and Named Entities. This method indicates that the TF-IDF criterion is influenced by term frequency and doesn’t sufficiently reject high frequency terms. Hence, some new criteria are proposed which are much more efficient. We consider the significance of the labels and propose new Local and Global Factors. In traditional method of TF-IDF, TF defined as local factor while IDF defined as global factor. In this paper, two new and efficient local factors named LRDF and OLF are computed as follow:

$$LF_i^{LRDF} = \log(1 + DF_{R,i})$$

$$LF_i^{OLF} = DF_{R,i} * \log(\frac{|R|}{DF_{R,i}})$$

$$DF_i^{OGF} = \frac{DF_{r,i} / |R|}{DF_{D,i} / |D|}$$

Where OGF is a global factor. Their experiments showed the combination of OGF with any local factor lead to better results than all combination of local factors with IDF and the OLF-OGF combination gave the best results. Using Named Entities is very efficient in the filed of news, because an event but a regular web page points to the distinct person, location, organization or date. Consequently, extracting persons or organizations and dates are very productive. The difficulty of their work is that they only select named entities as cluster labels but they are not sufficient and named entities can not describe the clusters as well as none phrase.

Vivisimo and Mooter, two of the best engines, are also use Label-based techniques for their clustering algorithm. So we can conclude that label-based approaches are much more applicable in search results clustering. Furthermore, user interface for result presentation in engines which performing result categorization (Vivisimo, Kartoo, Grokker) is important. A good study about this concept is done in [20].

## V. SUMMARIZATION

As we mentioned, one of the usages of clustering is in summarizing news. One of the successful works done in this field is proposed in [7]. In this research an engine named “SimFinder” is implemented which can give summarization of some news document to the users by clustering paragraphs. The reason for choosing paragraph in clustering is that more specialized information can be utilized in working with smaller units of text (sentences or paragraphs). They first identified 43 features for text which could efficiently extract from the text and that could plausibly help determine the semantic similarity of two short text units. Finally they select 11 of the 43 features by using data mining tasks. Afterwards, paragraphs are clustered by using these features. They cast the clustering problem as an optimization task and seek to minimize an objective function $\phi$ measuring the within-cluster dissimilarity in a partition:

$$\phi(\rho) = \sum_{i=1}^{k} [\frac{1}{|C_i|} \sum_{x,y \in C_i \text{ and } x \neq y} d(x,y)]$$

Then MULTIGEN goes beyond sentence extraction into reformulation and analyzes the sentences in each cluster produced by SIMFINDER and regenerates instead a new sentence containing just the information common to almost all sentences in a cluster.

Moreover, another engine named NewsInEssence [4], a fully deployed digital news system. A user selects a current news story of interest which is used as a seed article by NewsInEssence to find in real time other related stories from a large number of news sources. The output is a single document summary presenting the most salient information gleaned from the different sources. NewsInEssence first perform focused crawling which start from given news page. Then, some keywords which are more descriptive will be extracted from crawled pages contents. Afterward, keywords will be passed to some search engines and their result will be fetched. Finally, the summarization of fetched pages will be displayed to the users.

## VI. Automatic Event Detection

Another task in news engines is automatic event detection [6,2,13]. Authors of [6] generate sentence level clusters using hierarchical algorithms such as single-link, complete-link, and GroupWise-average. By keeping in mind that news can point to different events, they proposed that by clustering news in sentence level, each cluster point to an event. In this paper, WORDNET is used to heighten the efficient of comparison between words and sentences. It also use one learning automata for concerning the location of sentences in the document in clustering. Inasmuch as sentences relating to one event should be near each other, clustering and summarization are done over the sentences related to one event.

In [10], there is also one method for dividing news to its event parts and assigning one topic to each. The focus of this model is on automatic speech recognizer (ASR) scripts.
Segmentation system is a two stage process: the first stage hypothesizes boundaries, and the second stage removes boundaries. The first stage of the segmentation system uses a binary decision tree based probabilistic model to compute the probability of a boundary at every point in the ASR transcript that has been labeled a non-speech event. The features proposed for the decision tree are extracted from finite windows to the left and right of the current point. The features used by the tree are selected automatically. After the story boundaries have been hypothesized, a second stage (within the deferral period) removes some of them in order to reduce the false-alarm rate. The second stage uses the document-document similarity score of our detection system to determine if adjacent stories are similar topically, and reject the hypothesized boundary between them. The refinement step is applied iteratively.

Furthermore, a topic detection algorithm for detected stories proposed which is an incremental clustering algorithm that employs a novel dynamic cluster-dependent similarity measure between documents and clusters used for topic detection algorithm and decision tree segmentation which is a classification model. As soon as the document is added, its similarity degree with other clusters is measured. If the similarity degree of that document to one cluster is higher than defined threshold, then it entered that cluster.
The similarity measure used to obtain similarity of document $d^1$ and $d^2$ is computed as follow:

$$Sim(d^1, d^2) = \sum_{w \in d^1 \cap d^2} t_w t_w^2 idf(w, cl)$$

Where $t_w^i$ is the term count of word *w* in document *i* and *idf(w,cl)* is the cluster-dependent inverse document frequency of word *w* in cluster *cl*.

## VII. Conclusion

In this work we studied various tasks and services in news engines and survey on models and approaches applied in each task. Furthermore, we focused on clustering usages in each task in detail. As showed in this paper, clustering has many productive applications in all steps. For example, clustering helps to automatic news collection and cluster size is an important measure for news ranking. Clustering leads to a better display and easier explore of the search results. Consequently, clustering contributes to many efficient methods and results in news summarization and event detection.

## References


[1] G. M. D. Corso, A. Gulli, F. Romani, "Ranking a stream of news", In: WWW '05: Proceedings of the 14th international conference on World Wide Web, New York, NY, USA, ACM Press (2005) 97–106

[2] M. Atallah, R. Gwadera, W. Szpankowski, "*Detection of significant sets of episodes in event sequences*". icdm 00 (2004) 3–10

[3] N. Maria, M. J. Silva, "*Theme-based retrieval of Web news*", Lecture Notes in Computer Science 1997

[4] D. R. Radev, S. Blair-Goldensohn, Z. Zhang, R. S. Raghavan, "*Interactive, domain-independent identification and summarization of topically related news articles*".

[5] C. Silverstein, M. Henzinger, H. Marais, and M. Moricz, "*Analysis of A Very Large Web Search Engine Query Log*", SIGIR Forum, 33(1), 1999.

[6] M. Naughton, N. Kushmerick, J. Carthy, "*Clustering sentences for discovering events in news articles*", In: ECIR. (2006) 535–538

[7] V. Hatzivassiloglou, J. Klavans, M. Holcombe, R. Barzilay, M. Kan, K. McKeown, "*Simfinder: A flexible clustering tool for summarization*", (2001)

[8] H. Toda, R. Kataoka, "*A search result clustering method using informatively named entities*", In: WIDM '05: Proceedings of the 7th annual ACM international workshop on Web information and data management, New York, NY, USA, ACM Press (2005) 81–86

[9] N. A. Shah, E. M. ElBahesh, "*Topic-based clustering of news articles*", In: ACM-SE 42: Proceedings of the 42nd annual Southeast regional conference, New York, NY, USA, ACM Press (2004) 412–413

[10] S. Dharanipragada, M. Franz, J. McCarley, S. Roukos, T. Ward, "*Story segmentation and topic detection in the broadcast news domain*", (1999)

[11] M. Henzinger, B. Chang, B. Milch, S. Brin, "*Queryfree news search*", (2003)

[12] D. Reis, P. Golgher, A. Silva, A. Laender, "*Automatic web news extraction using tree edit distance*", (2004)

[13] Li, Z., Wang, B., Li, M., Ma, W.Y., "*A probabilistic model for retrospective news event detection*", In: SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM Press (2005) 106–113

[14] Z. Jiang, A. Joshi, R. Krishnapuram, and L. Yi., "*Retriever: Improving Web Search Engine Results Using Clustering*", In Managing Business with Electronic Commerce 02.

[15] O. Zamir and O. Etzioni, "*Grouper: A Dynamic Clustering Interface to Web Search Results*". In Proceedings of the Eighth International World Wide Web Conference, Toronto, Canada, May 1999.

[16] M. A. Hearst and J. O. Pedersen, "*Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results*", in Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR96), 1996, pp 76-84.

[17] H. Chen and S. Dumais, "*Bringing Order to The Web: Automatically Categorizing Search Results*", in Proceedings of the CHI 2000 Conference on Human Factors in Computing Systems, pp. 142–152, 2000.

[18] Y. Wang and M. Kitsuregawa, "*Link Based Clustering of Web Search Results*", in Second International Conference on Advances in Web-Age Information Management (WAIM), 2000.

[19] H. Zeng, Q. He, Z. Chen, W. Ma, and J. Ma, "*Learning to Cluster Web Search Results*". In Proceedings of ACM SIGIR '04, 2004.[19] Douglas Cutting, Jan O. Pedersen, David Karger, and John W. Tukey. "Scatter /Gather: A Cluster-Based Approach to Browsing Large Document Collections". In Proceedings of SIGIR'92, pages 318-329, Copenhagen, Denmark, June 21-24 1992.

[20] W. Rivadeneira, and B. B. Bederson**,** "*A Study of Search Result Clustering Interfaces: Comparing. Textual and Zoomable User Interfaces*". University of Maryland HCIL Technical Report, HCIL-2003-36

[21] V. Rijsbergen, C. J., "*Information Retrieval*". London: Butterworths; 1979.