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

T. Sobh (ed.), Innovations and Advanced Techniques in Computer and Information Sciences and Engineering, 219–224.

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).

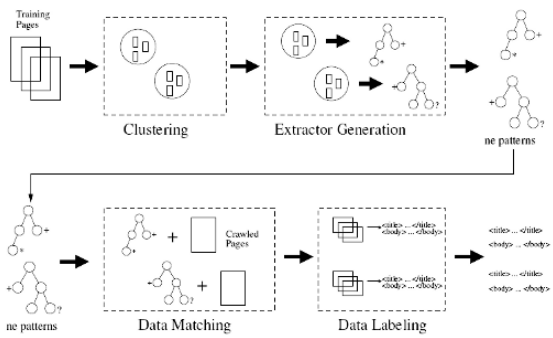

Fig. 1. Overall exteracting steps

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.