Finding news in a haystack

Event based news clustering with social media based ranking

Martin Weber · martin.weber@gmail.com · Graduation thesis · June 2012 · Supervised by Maarten H. Lamers
Media Technology MSc program · Leiden University, The Netherlands

Abstract

We present NewsClu, an approach to collect, cluster, categorize and select news articles from the Internet. The framework is not limited to a certain field, region or division of news. It is capable to work on any topic and runs on real-time live data on the Internet. NewsClu keeps its users informed about that particular topic by filtering, organizing, ranking and clustering of news according to static, dynamic and social factors. Through event based clustering the framework filters duplicate and less relevant news, making it easier to read more relevant news in less time. NewsClu is innovative in terms of clustering and ranking compared to other approaches.

Keywords: NewsClu, News, Aggregation, Clustering. Text Similarity

NewsClu in detail

The software consists of modular components for gathering, collecting, filtering, clustering, ranking, and selecting news items that match a desired topic. It runs live on the Internet and is capable of processing textual data in English.

Crawling

We crawl news from Google News, Bing News and Yahoo News. Besides those sites we use Google's Advanced Search, Google News Search and barrier to crawl content from parts of the web that might not be covered by the mainstream search engines or news aggregators.

Similarity

To cluster news, a similarity-measurement for all documents D of D is to be generated. To make the measurement more meaningful, we pre-process the text:

- text pre-processing
- filtering
- tokenization
- stemming
- stop word removal
- pruning

The result of these four pre-processing steps is a set of tuples containing golden words and their frequency. For each document D we compute a term frequency vector out of these tuples using cosine similarity:

\[ \text{similarity}(A, B) = \frac{A \cdot B}{||A|| \cdot ||B||} \]

Clustering

We cluster together all \( d \in D \) that have a high similarity, since we presume them to deal with the same event. Since NewsClu does not focus on a specific approach, domain or keyword, the similarity measure and clustering must be independent from those. Therefore, we use Agglomerative Single-Link Clustering. The main structure of the clusters is set with stage one; stage two acts as a refinement of the clusters.

Clustering, Stage One

1. For each pair \((d_1, d_2)\) with similarity \( s(d_1, d_2) \) do the clustering as follows:
   - Find a cluster \( \text{cid} \):
     - if either \( d_1 \) or \( d_2 \) is used this \( \text{cid} \)
     - if \( d_1 \neq d_2 \) and \( d_1 \neq \text{cid} \) and \( d_2 \neq \text{cid} \), use \( \text{cid} \)
     - if \( d_1 = d_2 \), use smaller \( \text{cid} \) if \( d_1 \neq \text{cid} \) and assign the \( \text{cid} \)
       - if \( d_1 \) and \( d_2 \) are in different clusters, get a new \( \text{cid} \) by incrementing the max \( \text{cid} \) by 1.
     - Assign \( \text{cid} \) to \( d_1 \) and \( d_2 \), repeat.

Clustering, Stage Two

1. For each pair \((d_1, d_2)\) with similarity \( s(d_1, d_2) \), do the clustering as follows:
   - Check the criterion for appending: If either \( d_1 \) or \( d_2 \), use this \( \text{cid} \)
   - Assign \( \text{cid} \) to \( d_1 \) and \( d_2 \), repeat.

In addition, we apply link-based clustering by connecting entities based on the existence of hyperlinks.

Ranking

The ranking for each cluster is based on a number of fixed factors and the social inclination of the df's mentions on social networks over time. Since the clusters are computed earlier, the ranking sorts the members of a cluster and all clusters themselves by comparing their highest ranking members (cluster heads).

- Fixed factors – twitter & Facebook
  - Age (publication date)
  - Text-Tag-Ratio
  - Alexa
  - Length
  - Links

We also measure the inclination of mentions of \( \text{df} \) on Facebook and Twitter over time:

\[ \text{Social inclination} = \frac{M(\text{df}) - M(\text{other})}{\text{sst}} \]

The fixed factors and the social inclination are combined to a rank of each document \( D \).

The Internet as Environment

It is of utter importance to realize that NewsClu runs in the wild on the Internet. This is fundamentally different from letting the framework run with a predefined (and somehow predictable) set of textual test data. All sources are unknown a priori.

Things. Will. Go. Wrong!

Experiments and Evaluation

We let different people test NewsClu on their favorite topics. The evaluation was informal due to the low number of participants (\( n=15 \)) and the qualitative approach. Overall, the users were satisfied with the results. They felt sufficiently informed about their topics and had not the feeling that important news were missing. Nearly all testers said they found relevant news through our framework that they would not have found with their current method of news consumption. Furthermore, the majority of users said that NewsClu kept them informed faster and more complete than their current method. The users also liked the lead news on top, saying that it covered the most important event within the topic.

In Fig. 1, histograms of NewsClu running on different topics show the clusters with the highest Tex-Tag-Ratio and the number of comparison pairs on the Y-axis. One can see the difference between well clusterable topics (Fig. 1a and 1b) and not so well clusterable topics (Fig. 1c and 1d) by comparing the spread of values. The fuzziness on not well clusterable topics was represented through one more peaks with medium similarity, making clustering using this similarity values less meaningful and harder.

Further Information

If you want to know more about the topic, technical details or the software, please feel free to contact me:

email: martin.weber@gmail.com
twitter: @eew_mar

You can also download this poster as a PDF:


Fig. 1. Histograms of similarity pairs for different topics.

(a) topic A
(b) topic: North Korea
(c) topic Flu
(d) topic: Clinic Garnering

Fig. 2. Screenshot of NewsClu showing news on the topic “north korea”