Building a smart news annotation system for further evaluation of news validity and reliability of their sources

Natalia Shakhovska[1], Volodymyr Berbelyuk[2]

Information Systems and Networks Department,
Lviv Polytechnic National University,
UKRAINE, Lviv, S. Bandery street 12,
E-mail: [1]shakhovska@lp.edu.ua, [2]v.berbelyuk@gmail.com

Abstract – The article describes the elements of news annotating information systems to further assess the news validity and reliability of their sources. Considers the process of system operation and algorithms that can be used in the implementation of such a system.

Key words – information system, NLP, news aggregation, multiple documents annotation, the theory of trust.

I. Introduction

Modern information society allows to easily obtain information about current events in the world. Numerous media give them to his audience through the press, radio, television and the Internet. The latter distribution channel is currently the most promising (and for many has become a major), which is not surprising, given his negotiations, such as distribution, flexibility, convenience.

Numerous media provides them to his audience through print, radio, television and the Internet. The last distribution channel is currently the most promising (and for many has become a major), which is not surprising, given his negotiations, such as spread, flexibility, convenience.

However, there are a number of problems facing the audience when receiving media. These include: (1) selection of topical (on the subject of the request and the time) to the recipient scenes from a huge amount of available ones, (2) eliminating repetitions of topics that arise in the provision of information about the same event in different media providers, and (3) filtering of received news of those who in one way or another are not true (include gross inaccuracy, misrepresentation or outright lie).

The latter problem is particularly important in the field of modern information technology, especially given the unsolid of information on the Internet. It is connected with the previous one; since annotation precedes and is interwoven with existing strategies assess the reliability of sources.

Information system being developed is designed to take a step towards addressing these problems. Particular are considered algorithms of annotation and assessment of the reliability of the news.

II. Analysis of recent research, publications and existing solutions

Problem of news annotation currently has a number of practical solutions. Most of them, however, are based on manual labor of specialists and has limitations in terms of objectivity, speed, amount of work and uniformity. There are services that use automated approach, such as Google News and Yandes.Novosty. The last one in particular even offers a Ukrainian version.

However, there is not application services in a broad use that would allow to assess the reliability of sources or individual events, although there are developments that lead us to it.

III. News aggregation algorithm

The procedure for aggregating the feed may include the following steps (based on procedure used by Yandes.Novosty [1]).

1. Download news via RSS feeds or other;
2. Reports segmentation (allocation of title, description, body text, images, videos, etc.);
3. Select of stories — news referring to a particular event (by clustering messages based on analysis of their texts);
4. Annotation of subjects (presentation of subjects core content in abbreviated form);
5. Allocation reports within the stories that confirm or refute it.

Download and segmentation (for structured representation of reports, for example, XML/RSS) is a trivial technical problem.

Select of stories described below.

Annotate is to create a shorter version of a text or set of texts. Creating annotations by a human is common task. There are two kinds of annotation task: annotation of one document and annotation of several documents. Goal of the first case is a brief overview of the main contents of given document. The second case relies to identify different viewpoints on the story and possibly considering time factor, which is that certain documents may lose their relevance. News annotation problem is just a task relate to second case with taking into account the time factor.

IV. Select of stories

Select of stories is kind of Topic Detection and Tracking task (TDT), occurred in 1996-1997. The concept of topics closely related to the concept of event: the subject – an event or activity along with all directly related events and activities. The objective is to extract from the stream event information.

Vector Space Model (VSM) is a common way for presentation of documents. According to it, each word is associated with a weight in accordance with the chosen weight function. With such representation for all documents one may, for example, to find the distance between points in space thereby solving the problem of similarity references – the closer the points, the more similar the corresponding documents.

The classic method of words weighting is TF-IDF:

$$TFIDF(t,d,D) = TF(t,d) \times IDF(t,D) \qquad (1)$$

$TF(t,d)$ (term frequency) is the normalized frequency of term in the text:

$$TF(t,d) = \frac{freq(t,d)}{\max_{w \in d} freq(w,d)} \qquad (2)$$

Here $freq(t,d)$ — the number of occurrences of t in the document d.

$IDF(t, D)$ (inverse document frequency) defined as next:

$$IDF(t, D) = \log \frac{|D|}{|\{d \in D; t \in d\}|}. \quad (3)$$

Here the numerator is the number of documents in the set, and the denominator is the number of documents in which the term t occurs.

Various modifications of TF-IDF are used depending on the problem. For example, one of the solutions uses the following weights:

$$w(t, D) = (1 + \log_2 TF(t, D)) \times \frac{IDF(t)}{\| \vec{D} \|}. \quad (4)$$

Here $\| \vec{D} \|$ is the norm of the vector representing the document D.

More recent studies used, for example, the following TF-IDF modifications:

$$TF' = \frac{TF}{TF + 0.5 + 1.5 \frac{len_d}{len_{avg}}}. \quad (5)$$

Here len_d is the length of the document d, and len_{avg} is the average length of the documents.

$$IDF' = \frac{\log(IDF)}{\log(N + 1)}. \quad (6)$$

For comparison of documents' vectors both methods used metrics such as cosine, Kullback-Leibler divergence, and other methods (weighted sum of the components of the document, simple language models). In total, there are more than 70 ways to calculate the similarity of vectors.

Two types of tasks can be considered: detection of events from the data set for a certain period and event detection in real time.

The first type of problem is to partition the original data into groups relevant events, as well as in determining whether the text describes document sets out an event. The basic idea of all solutions was the use of clustering algorithms (incremental clustering, K-means method, etc.). It is assumed that each cluster contains documents that describe an event.

The task of the second type is to determine for a new document whether it describes an event that has occurred in the raw data. To track events used classification algorithms (the method k-nearest neighbors, decision trees, etc.). The classification was made with using two classes: YES – the document describes the event, NO – does not describes [2].

V. Assessment of the reliability of news

The next step is to estimate the reliability of the news. One of its options can be based on the model proposed in [3]. The model assumes the recipient A gives two marks to news provider B:

$t_{AB} \in (0, 1)$ – assessment of confidence in the source and $u_{AB} \in (0, 1]$ – uncertainty in own estimation of confidence.

The initial values are taken $t_{AB} = 0.5$ and $u_{AB} = 1$, i. e. 50% confidence in the source and the maximum possible uncertainty.

To clarify the evaluated values recipient A chooses from news reports, reseved by him (in quantity n_B) those which he can say with some certainty whether they are true (let's denote this number by m_B). Let r_B among them are true and false s_B , $r_B + s_B = m_B$. Then the frequency of false news as a part of $r_B / (r_B + s_B)$. Then in [3] define t_{AB} as the expected value of a beta distribution with parameters $\alpha = r_B^* + 1$, $\beta = s_B^* + 1$, where r_B^* and s_B^* are averaging of pre-assembled values of r_B and s_B, made on the basis of so-called aging factor, and u_{AB} – as a normalized value of the variance of that distribution.

Another possible way to estimate is given, in particular, in [5] and uses trust model, which uses a pair of independent coefficients of confirmation (k_1) and retraction (k_2) of some event according to different providers ($k_1, k_2 \in [0, 1]$). The initial values are taken as $k_1 = k_2 = 0$. Then for provider that has determined rate of confirmation or refutation of the event W_i, the coefficients are recalculated using the Shortliffe's formula:

- if the event is confirmed, then $k_1 = k_1 + W_i (1 – k_1)$,
- otherwise $k_2 = k_2 + W_i (1 – k_2)$.

The resulting confidence factor k is calculated by the difference of k_1 and k_2: $k = k_1 - k_2$, $k \in [-1, 1]$. The advantage of this method is the lack of need for the recipient to directly assess confidence of some of the news reports.

Conclusion

The result of the proposed system will be annotated news reports from different providers, as well as the assessments of the reliability of news and credibility for individual media. They will help users to better and faster navigate the available media topics that interest them.

References

[1] A. Shahraev, "Avtomatycheskoe annotyrovanye novostnoho potoka." [Automatic annotation of news flow], Natalia Ostapuk's page on slideshare.net, 29 Nov 2011. [Online]. Available: SlideShare.net, http://www.slideshare.net/NataliaOstapuk/ss-10380447 [Accessed: 1 Oct 2013]

[2] A. Korshunov, A. Homzyn, "Tematycheskoe modelyrovanye tekstov na estestvennom yazyke" [Topical modeling of natural language texts], Trudy ISP RAN [Proceedings of ISP RAS], vol. 23, 2012. Available: Open Access Library "KyberLenynka", http://cyberleninka.ru/article/n/tematicheskoe-modelirovanie-tekstov-na-estestvennom-yazyke [Accessed: 1 Oct 2013]

[2] E. Staab, V. Fusenig and T. Engel, "Towards Trust-Based Acquisition of Unverifiable Information", in Proc. 12th International Workshop, CIA 2008, Prague, Czech Repb.: Springer, vol. 5180, pp. 41-54

[4] V. V. Litvin, N. B. Shahovska, "Pro zadachu avtomatizovanogo anotuvannya podiï on osnovi prostoru danih" [On the problem of automatic event annotation based on data space], Naukovyy visnyk Chernivetskogo universitetu. Seriya: Fizika. Elektronika – Scientific Bulletin of Chernivtsi University – Chernivtsi, no. 426, 2008, pp. 58-62.