Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 122 (2017) 715–719

Procedia Computer Science

www.elsevier.com/locate/procedia

Information Technology and Quantitative Management (ITQM 2017)

News clustering based on similarity analysis

Ilya Blokh [a], Vassil Alexandrov [b,*]

[a] *Perm State University, Bukireva st.15, 614990 Perm, Russia*
[b] *ICREA-Barcelona Supercomputing Centre, Spain*

Abstract

This paper's focus is to continue the research on Internet psychological warfare analysis, where the authors faced a necessity to propose an accurate algorithm for news clustering that could be able to group news into semantically close sets. A two stage approach to reach that goal is proposed. First a similarity estimation between news messages is performed using semantic similarity metric based on WordNet. Next, the most suitable for given data structure clustering algorithms is selected in order to obtain thematic news clusters and observe their size distribution over time. Experiments were made on news volumes from several news mass media official pages in Facebook.

Peer-review under responsibility of the scientific committee of the 5th International Conference on Information Technology and Quantitative Management, ITQM 2017.

Keywords: psychological warfare; ontology; similarity analysis; news clustering, news analysis.

1. Introduction

In previous research [1] we introduced the concept of psychological warfare in electronic medium (Internet, etc.) as the demonstration of multi-directed regular actions forwarded into influencing the social opinion in some area. We are focusing at news mass media analysis trying to investigate patterns in news publishing, to find out how actively different topics could be discussed and if any kind of value judgement takes place. In this paper we suggest an algorithm for clustering news from mass media in social networks. After composing clusters we then analyze their parameters in order to understand the news spreading processes in mass media. Thus the purpose of this research is to define an effective algorithm of mass media news clustering and obtain the corresponding theme distribution.

The algorithm has two main stages. At first stage, we preprocess news data by configuring similarity matrix. We use ontology to estimate similarity between news sentences based on sense and context. At the second stage, we perform clustering and find out how news could be clustered.

2. Related work on clustering methods

Text clustering algorithms could be classified in several groups: vector space models, k-means variations, generative algorithms, spectral algorithms, dimensionality reduction methods and phrase-based methods [19]. Vector space model is a classic approach which shows better results on homogeneous topics and needs to know the number of clusters [20]. K-means algorithm and its extensions are historically most popular approaches for hierarchical and partitioned clustering. However they have a number of drawbacks: effectiveness decreases on large data corpora and relies on random initialization. Also they are susceptible to outliers and noise and needs to know the number of clusters as well [21]. Generative algorithms are also sensitive to outliers and it makes them less effective on heterogeneous data and have cluster count as input [19]. Spectral clustering shows high accuracy when data's vector model could be presented as bipartite graph. The advantage of this group is that it doesn't need the number of clusters and can find this value during the processing [18, 22, 23]. Dimensionality Reduction methods are originally developed for computer vision applications, has been effectively used for document clustering. Their main disadvantages are that they rely on random initialization which may lead to different results across runs on same data. However they have

* *E-mail address*: vassil.alexandrov@bsc.es

1877-0509
Peer-review under responsibility of the scientific committee of the 5th International Conference on Information Technology and Quantitative Management, ITQM 2017.
10.1016/j.procs.2017.11.428

high performance and some of them can estimate the optimal count of clusters [24, 25]. Phrase-base methods are improved by encoding word order information. However, it doesn't guarantee higher accuracy than other clustering methods [26]. Some specific approaches were suggested in regard to short text and news clustering. In [27], Yungquing et al, present discriminative bi-term topic model to perform clustering based on news headlines. Social network analysis used to cluster topics in Twitter network is presented in [28]. In [29] special kernel function provided to measure short text semantic similarity applied for search engine query analysis. Also clustering accuracy for short texts could be improved using feature generation from Wikipedia in [30]. In [31] Conrad and Bender showed that agglomerative clustering technique may be used to implement event centric news clustering algorithm. Also cosine similarity based clustering applied to propose a method for news collecting and clustering [32].

3. News similarity estimation

We purpose to amplify clustering accuracy by estimating similarity between news data based on ontology. Using ontology could give a better understanding of information spreading and impact. We aim to obtain some news clusters were each cluster contains information concerning one theme or even one point of view regarding this theme.

We use WordNet – lexical database of English. Words in WordNet are united in synsets (sets of cognitive synonyms) which are interlinked by means of conceptual-semantic and lexical relations. This structure could be convenient to estimate words and sentences similarity. There're many measures of semantic similarity based on WordNet [2-8]. Some measures rely on WordNet structure to produce a numeric score that quantifies the degree to which two concepts are similar. According [17] measures which use information content values along with ontology structure are more accurate and provide greater correlation with human similarity judgements. That is why we use JCN similarity metric. It is a score denoting how similar two word senses are, based on the Information Content (IC) of the Least Common Subsumer [16]:

$$jcn(c_1, c_2) = \frac{1}{IC(c_1)+IC(c_2)-2*res(c_1,c_2)} \quad (1)$$

where $res(c_1, c_2)$ is Resnik measure of similarity and $IC(c)$ is information content value of concept [6].

Since we analyze news messages from social networks for the moment, so news usually are presented by one or few sentences. Our first step is to understand which messages are related to same theme. Thus sentences similarity estimation method should be proposed. We provide several steps for that method according [9,10, 15, 16]:

1) Sentence tokenization and stop-words removal. At this stage we represent each text message as token vector $\bar{v} = (v_1, .. v_n)$ consisting of words. We remove stop-words to avoid additional infelicities. [20]
2) Part-of-speech disambiguation. Each word is tagged by two tags: the first one indicated syntactic role of the word (object, subject e.t.c.) and the second one point at functional role (verb, noun e.t.c.). We estimate similarity between nouns aiming to reveal news similar by discussed theme.
3) Word stemming what means removing the common morphological and inflexional endings of words. This operation is especially useful in the field of information retrieval and increases accuracy [21].
4) Word Sense Disambiguation. In this stage we investigate which of word senses is more appreciate in current context. Lesk algorithm [22] could be used for that task. Word disambiguation is based on comparing of glossaries containing each word sense. The most probable sense is that one which is concluded in same glossary with the majority of other words in sentence. In [18] authors suggested adapted version of Lesk algorithm where they achieved more accuracy.
5) Compute sentences relatedness. This estimation is based on pair of words similarity according JCN metric. First similarity matrix have to be constructed. The matrix element $R_{i,j}$ is similarity estimation value between token v_i corresponding to first sentence and token w_j corresponding to second sentence. Similarity matrix could be examined as bipartite graph and sentence similarity computing task as computing a maximum total matching weight of this bipartite graph. Thus resulting similarity could be computed as average value

$$S = \frac{2\times Match(N,M)}{|N|+|M|} \quad (2)$$

where N, M are token vectors and Match(N,M) is token matching computed by Hungarian method [14]. This estimation takes into account the influence of each pair similarity value.

After we define sentences similarity computation method, we have to estimate similarity between all collected data and fetch out clusters of related messages.

4. Clustering data

Mass media news data from social networks has several features:

- news is presents as short text in 18 words in average;
- text corpora could contain hundreds of thousands of news and more. News set is always replenishing;
- the number of clusters in unknowns and it could vary in different moments of time;