Information & Security

D. Lande, I. Subach, O. Puchkov & A. Soboliev, vol. 50, no. 1 (2020): 79-86
https://doi.org/10.11610/isij.5013

Published since 1998

ISSN 0861-5160 (print), ISSN 1314-2119 (online)

Research Article

A Clustering Method for Information Summarization and Modelling a Subject Domain

Dmytro Lande [1,2] (✉), *Ihor Subach* [2], *Olexander Puchkov* [2], *Artem Soboliev* [2]

[1] *Institute for Information Recording of the National Academy of Sciences of Ukraine, Kyiv, Ukraine, http://www.ipri.kiev.ua*

[2] *Institute of Special Communications and Information Protection, National Technical University "Igor Sikorsky Kyiv Polytechnic Institute," Ukraine, https://iszzi.kpi.ua*

ABSTRACT:

The article presents a discriminant cluster analysis method used to form real-time models of subject areas and digests based on automatic analysis of a large number of messages from social networks. It is based on estimating the discriminant value of terms. Cluster analysis, like the well-known LSA algorithm, provides a matrix representation of the data. The novelty is in using the most significant discriminant values as centroids to define clusters.

The algorithm is simplified; it does not involve referencing to the adjacency matrix, definition of eigenvectors. Its complexity is O(N2), where K is the number of clusters and N – the number of reference terms. If it is necessary to improve the quality of the proposed approach, the defined centroids can be transferred as input data for other known algorithms. Based on the above algorithm, toolkits for the formation of a language network and digests were developed and embedded in the "CyberAggregator" system, which provides accumulation, processing, summarization of data from social networks on cybersecurity issues.

ARTICLE INFO:

RECEIVED: 11 JUNE 2021
REVISED: 07 SEP 2021
ONLINE: 18 SEP 2021

KEYWORDS:

clustering method, information summarization, subject domain, words network, visualization, social media monitoring, CyberAggregator

Creative Commons BY-NC 4.0

✉ Corresponding Author: Tel.: +380674668917 ; E-mail: dwlande@gmail.com

Problem Statement

The analytical processing of large thematic document numbers includes such procedures as the formation of a terminological network, which can be regarded as a kind of "information portrait," a model of the subject's area, data aggregation on time, basis keywords, other parameters, automatic formation of digests, informational of issues, etc. Tasks, which require automatic grouping of such objects as subject area terms, subjects, documents, are, first of all, the tasks of forming the digests, forming the terminology network.

A terminological network consists of nodes, which are the terms, keywords of the subject area, and connections – meaningful or statistical connections between these terms. When such a network is built, there is a need for grouping messages that are similar in meaning. When the domain is known and has long been defined, classification algorithms by known classes are used. Otherwise, cluster analysis methods, the so-called teacherless machine learning methods, are used.

Selection of the most weighty, relevant to the user's information needs documents by information arrays is a complex and ambiguous task. A limited set of such documents arranged in an easy-to-understand form is usually called a digest. Digests are usually compiled by human analysts.

The automation of the above processes can be carried out based on various linguo-statistical algorithms, among which the well-known algorithms of cluster analysis, first of all K-means,[1] LSA,[2] algorithms based on text markers (critically depending on document language), network,[3] and hybrid ones.

The common problem of all existing algorithms is the long execution time, related to the problems of: 1) computational complexity; 2) the need to obtain qualitative models of subject areas or digests; and 3) evaluating the quality of the created products. Experts can evaluate the quality of the created digests, but in need to obtain real-time reporting documents, quality evaluation methods based on information theory (Jensen-Shannon divergence) [3] can be applied, which partially solve the third problem. The discriminant cluster analysis algorithm proposed in this paper provides a partial solution to problems 1) and 2) by taking into account the most significant terms as semantic markers of texts and almost linear computational complexity.

Purpose

The aim of the study was to create a fast discriminant method of cluster analysis, which should be used to form real-time networks of language (models of subject areas) and digests based on automatic analysis of large arrays of documents (messages) from social networks, based on account of the discriminant weight of the terms from these documents.

Grounding

The prerequisite for the creation of the method was that the authors have at their disposal full-text databases of documents collected from social networks using the CyberAggregator system.[4] Each of these documents is assigned so-called reference terms, a predetermined number of the most significant terms defined by the CHVG method.[5]

It is supposed that, on a request to the CyberAggregator system, the corresponding thematic information array of documents is formed, on the basis of which the network of terms should be constructed and automatically grouped by topics as a model of subject area and the most significant in some sense documents should be selected, whose content should constitute the digest.

It would seem that the frequent terms themselves, from the set of all reference terms included in the relevant documents, should be prominent as centroids for the clusters of the terminological networks or as markers for selecting documents for the digest. There is another requirement; namely, multiple term clusters, like the digest, should cover different main aspects of the content of the thematic information array; accordingly, individual documents as components of the digest should be as meaningfully different as possible. This property has to be taken into account by selecting the terms with the highest discriminant weight as the basis for clustering (cluster centroids).

To solve this problem, we propose an approach ideologically close to the well-known *TF-IDF* approach,[6] used in the theory of information retrieval. Accordingly, for this approach, the definition of the weight of each term can be given as the product of some non-decreasing function of the absolute frequency of its occurrence and the number of terms from the set *T*, which this term is not related to (not included in the same documents):

$$w_i = F(tf_i) \cdot (x+1).$$

It is assumed that both the entire set of terms from the array of relevant documents and some sample of the most frequent words can be considered as a term set. Therefore, in the above formula, one is added to the value.

As in the well-known *TF-IDF* approach, the term weight in our case acts as a kind of singularity measure or discriminant power. The most frequent terms have the greatest weight and are relatively weakly related to the others due to the close connection to the cluster of their nearest neighbors in the network.

Cluster network analysis, as a rule, solves the problem of two-criteria optimization, namely:

1) within each cluster K, elements (nodes) should be connected as much as possible, that is

$$\sum_{\substack{i \in K \\ j \in K}} a_{ij} \to \max.$$

Here and hereafter, a_{ij} it is some estimate of the relationship between elements with indexes i and j, which are included in the cluster K. a_{ij} can take non-negative values. This estimate can be calculated in different ways, for example, for words from documents as the number of documents simultaneously containing words with indexes i and j.

2) the connectivity between any separate different clusters, for example, K_p and K_q should be minimal, that is

$$\sum_{\substack{i\in K_p \\ j\in K_q}} a_{ij} \to \min.$$

Often, in the general case, the sums for all indices are estimated (N is the number of clusters):

$$\sum_{p=1}^{N}\sum_{\substack{i\in K_p \\ j\in K_p}} a_{ij} \to \max, \quad \sum_{p=1}^{N}\sum_{\substack{q=1 \\ q\neq p}}^{N}\sum_{\substack{i\in K_p \\ j\in K_q}} a_{ij} \to \min.$$

By association, in our case, the first factor in the formula for determining the weight w_i corresponds to the first requirement, since tf_i, for example, if it is of great importance, then it is connected by strong ties with a certain number of words (including from its own group). If the parameter x is of great importance, then the bonds are concentrated within their own group (cluster) – this corresponds to requirement 2.

We can assume that the individual terms with the highest discriminant weight (the number chosen in advance) will constitute the centroids for clustering or the basis for selecting the documents that will form the basis of the digest.

Algorithms

To select the terms that should form the basis for clustering or selecting documents to form a digest, the following steps must be carried out:

1. A thematic information array of documents corresponding to the information need of users is formed. It is considered that each document is prescribed in advance reference terms, the most weighty words from this document according to the CHVG algorithm.
2. Calculated the absolute frequency of occurrence of each reference term, selects a given number of the most significant reference terms.
3. A matrix of mutual occurrence of the selected reference terms in the documents is formed. The matrix element is the number of mutual occurrences of pairs of terms with indices and in the same documents. Obviously, the diagonal of this matrix corresponds to the absolute frequency of the corresponding terms.
4. Based on the generated matrix and the above formula (as a function of F the square root is chosen) for each word, its discriminant weight is calculated.
5. A given number of important by discriminant weight reference terms are selected as centroids for clustering or document selection markers for the digest.

After that, to determine the clusters in the terminology network, for each of the selected centroids a set of terms from the set that are most closely related to it according to the matrix data is determined.

In the future, you can improve the quality of clustering by considering certain pivotal terms as the initial centroids of the K-means algorithm. In accordance with this algorithm 1) at the first stage of clustering, a set of initial pivotal terms (centroids) is selected 2) on the basis of these terms, clusters are selected from a set of all terms - network nodes. 3) After the initial selection of clusters, a central element is calculated for each of them, which may not coincide with the previous centroid. After redefining the centroids, a new selection of clusters is performed, that is, go to step 2). This happens until the process stabilizes. The question of how the centroid is selected in step 3) in the case of a network of terms can be solved in different ways, for example, how the centroid of cluster K can be selected the term i that is most related to other terms ($i = \arg\max_{i \in K} \sum_{j \in K} a_{ij}$), the term with the highest PageRank value within the cluster (PageRank can also be seen as centrality in networks, including word networks [7]), or the term with the highest the value of the entered discriminant weight $(i = \arg\max_{i \in K} F(tf_i) \cdot (x+1))$.

Obviously, the given formula for determining the weight of the discriminant node-term is a simplification of the strict requirements of clustering and may require further development of the algorithm. At the same time, as practice shows, the given approach solves well the problem of clustering a network of terms within the framework of the CyberAgregator system.[4]

The CyberAgregator system is designed to collect, search and analyze information from social networks and websites on cybersecurity issues. Search and data aggregation in this system is provided by Elasticsearch tools. At the stage of primary information processing, concepts are extracted from messages, such as keywords, names of persons, names of firms, brands, toponyms. Among the possibilities provided by the CyberAgregator system, one can single out the tasks of forming digests and models of subject areas that correspond to user requests.

Figure 1 shows an example of a term network visualization from documents (posts) from social networks (Telegram, YouTube, Facebook, Reddit, Medium) and selected websites after the cluster analysis. The network is displayed using the JavaScript library D3.js.[8] Each node in the network contains links to a refined search by the appropriate term.

It should be noted that the CyberAgregator system does not group terms presented in different languages, so it is possible to form clusters of similar content in different languages (for example, nodes “intelligence,” “разведка,” “розвідка”).

One of the modes of CyberAgregator system functioning is the formation of digests which consist of the most important documents on various aspects of the investigated subject. When forming a digest, the required number of documents containing the terms with the highest weight, included in different word clusters, is used.

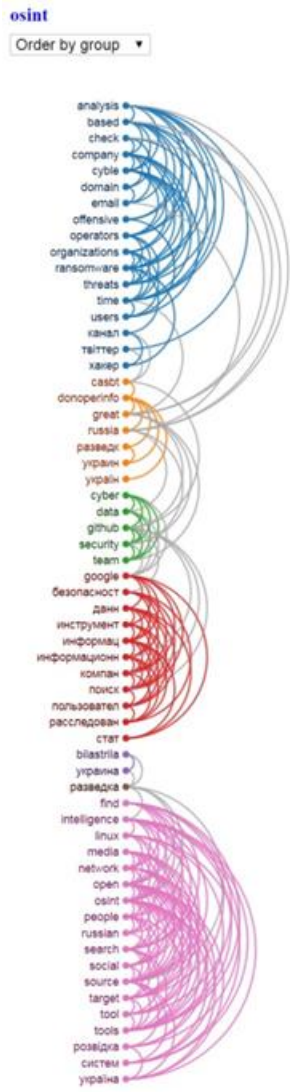

Figure 1: Visualization of the network of terms that correspond to the OSINT (Open Source Intelligence) theme.

Some additional empirical criteria are also applied to the selection of documents. They can be related to the length of these documents, their headings, to take into account signs of duplicate documents and filter some of the accounts, and so on.

Figure 2 shows an example of visualization of a digest generated from documents from social networks. The digest displays one of the analytical functions of the "CyberAggregator" system. Each digest message contains links to a search refined by the appropriate term (document selection marker).

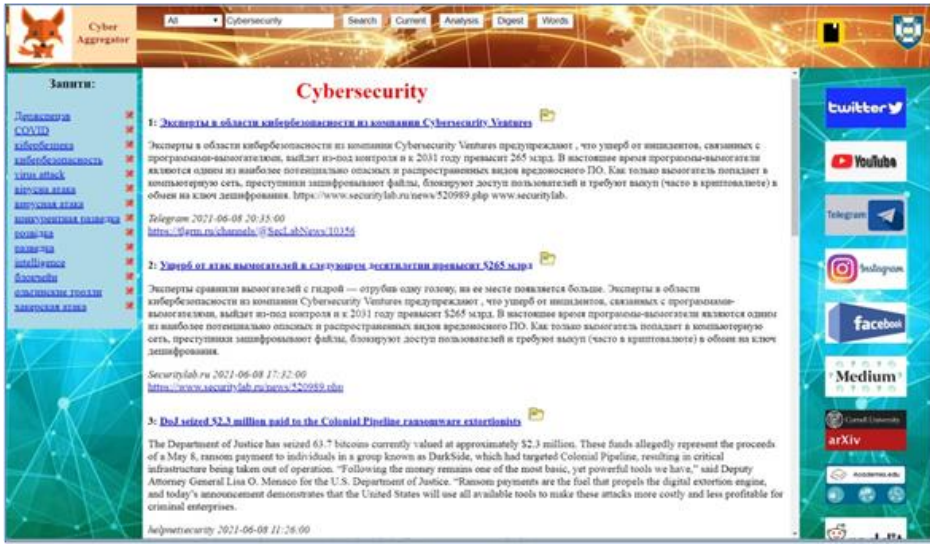

Figure 2: An example of a digest display that corresponds to the Cybersecurity theme.

Conclusions

The algorithm presented in this article involves working with predetermined reference timelines, which allow ensuring speed and quality. This algorithm, as well as the well-known algorithm LSA, is an algorithm of cluster analysis, which applies matrix representation of data.

The novelty of the algorithm lies in the approach (formula) to calculate the discriminant weight of the terms, the most significant of which act as centers to determine the clusters' centroids. The algorithm is simplified; it does not provide the reference of adjacency matrix, definition of eigenvectors, etc. The complexity of the above algorithm is O (N^2), where N is the number of reference terms. If it becomes necessary to improve the quality of the proposed approach, the defined centroids can be passed as initial data for other known algorithms, such as K-means.

In accordance with the above algorithm, a toolkit for the formation of the language network and digests has been developed, which are built into the system "CyberAggregator."

Acknowledgements

This research was supported by CyRADARS project (SPS G5286 "Cyber Rapid Analysis for Defense Awareness of Real-time Situation") in the frame of the NATO Science for Peace and Security program.

References

[1] Kristina Sinaga and Miin-Shen Yang, "Unsupervised K-Means Clustering Algorithm," *IEEE Access* 8 (2020): 80716-80727, https://doi.org/10.1109/ACCESS.2020.2988796.

[2] Said Salloum, Rehan Khan, and Khaled Shaalan, "A Survey of Semantic Analysis Approaches," In: Hassanien AE., Azar A., Gaber T., Oliva D., Tolba F. (eds) *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020), Advances in Intelligent Systems and Computing,* Vol. 1153 (Cham: Springer, 2020), https://doi.org/10.1007/978-3-030-44289-7_6.

[3] Dmitry Lande, Yang Zijiang, Zhu Shiwei, Guo Jianping, and Wei Moji, "Chinese legal information automatic summarization," in *Selected Papers of the XVIII International Scientific and Practical Conference on Information Technologies and Security (ITS 2018)*, *CEUR Workshop Proceedings (ceur-ws.org)*, Vol. 2318, (2018): 222-238.

[4] Dmytro Lande, Igor Subach, and Alexander Puchkov, **"**System of Analysis of Big Data from Social Media,**"** *Information & Security: An International Journal* 47, no. 1 (2020): 44-61, https://doi.org/10.11610/isij.4703.

[5] Dmytro Lande and Oleh Dmytrenko, **"**Methodology for Extracting of Key Words and Phrases and Building Directed Weighted Networks of Terms with Using Part-of-speech Tagging,**"** in *Selected Papers of the XX International Scientific and Practical Conference "Information Technologies and Security" (ITS 2020), CEUR Workshop Proceedings,* vol. 2859 (2020): 168-177.

[6] Rahim Khan, Yurong Qian, and Sajid Naeem, "Extractive based Text Summarization Using K-Means and TF-IDF," *I.J. Information Engineering and Electronic Business* 3 (2019): 33-44, https://doi.org/10.5815/ijieeb.2019.03.05.

[7] Sujatha Das Gollapalli and Xiao-li Li, **"**Using PageRank for Characterizing Topic Quality in LDA," *ICTIR '18: Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval*, September 2018, pp. 115–122, https://doi.org/10.1145/3234944.3234955.

[8] Scott Murray, *Interactive Data Visualization for the Web. An Introduction to Designing with D3* (O'Reilly Media, 2017), 472.

About the Authors

Dmytro **Lande**, Dr. of Sci. (tech), is Professor in the Institute for Information Recording of National Academy of Sciences of Ukraine, Kyiv, Head of Department; Institute of Special Communications and Information Protection of the National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute," Kyiv, Ukraine.

Igor **Subach**, Associate Professor, Dr. of Sci. (tech), is Department Head in the Institute of Special Communications and Information Protection of the National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute," Kyiv, Ukraine.

Olexander **Puchkov**, Professor, PhD, is the Director of the Institute of Special Communications and Information Protection of the National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute," Kyiv, Ukraine.

Artem **Soboliev**, PhD, Institute of Special Communications and Information Protection of the National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute," Kyiv, Ukraine, Senior Researcher.