Towards automated visualisation of scientic literature

Towards automated visualisation of scientific literature

Evelina Di Corso[1][0000−0002−3988−3512], Stefano Proto[1][0000−0002−8143−3611], Tania Cerquitelli[1][0000−0002−9039−6226], and Silvia Chiusano[2][0000−0002−5740−5004]

[1] Politecnico di Torino, Italy
Dipartimento di Automatica e Informatica
[2] Politecnico di Torino, Italy
Dipartimento Interateneo di Scienze, Progetto e Politiche del Territorio
name.surname@polito.it

Abstract. Nowadays, an exponential growth in biological data has been recorded, including both structured and unstructured data. One of the main computational and scientific challenges in the modern age is to extract useful information from unstructured textual corpora to effectively support the decision making process. Since the emergence of topic modelling, new and interesting approaches to compactly represent the content of a document collection have been proposed. However, the effective exploitation of the proposed strategies requires a lot of expertise. This paper presents a new scalable and exploratory data visualisation engine, named ACE-HEALTH (AutomatiC Exploration of textual collections for HEALTH-care), whose target is to easily analyse medical document collections through the Latent Dirichlet Allocation. To streamline the analytics process and enhance the effectiveness of data and knowledge exploration, a variety of data visualisation techniques have been integrated in the engine to provide navigable informative dashboards without requiring any a-priori knowledge on the analytics techniques. Preliminary results obtained on a real PubMed collection show the effectiveness of ACE-HEALTH in correctly capturing the high-level overview of textual medical collections through innovative visualisation techniques.

Keywords: Textual data visualisation · Topic modelling · Informative dashboards.

1 Introduction

In the recent years, we have been witnessing the exponential growth of biological data, including both structured and unstructured data. One of the main computational and scientific challenges in the modern age is to extract useful information from unstructured texts with minimal user intervention [4]. Their value is severely undermined by the inability to translate them into knowledge and, ultimately, actions [2].

This paper presents a new scalable and exploratory data visualisation engine, named ACE-HEALTH (AutomatiC Exploration of textual collections for HEALTH-care), whose target is to analyse medical document collections. The framework brings many different analytic techniques to help non-expert users to make sense of large medical data collections. The framework exploits the Latent Dirichlet Allocation [1], a generative probabilistic model, to divide a given corpus into correlated groups of documents with a similar topic. ACE-HEALTH includes several kinds of visualisations to show knowledge at different granularity levels to easily capture the high-level overview of textual medical collections, and drill-down the knowledge to the single document. We have experimentally evaluated the engine on a real textual medical collection. The performed experiments highlighted ACE-HEALTH's ability to autonomously identify homogeneous groups of medical documents and efficiently represent a manageable set of human readable results.

The next Sections are organised as follows. Section 2 presents ACE-HEALTH architecture and its main building components, Section 3 shows the tests run to assess ACE-HEALTH's performances and discusses the experimental results. Lastly, Section 4 draws the research conclusions and presents future works.

2 The ACE-HEALTH framework

ACE-HEALTH (AutomatiC Exploration of textual collections for HEALTH-care) has been tailored to analyse any medical textual collection. This new framework is able to automatically extract and graphically represent multiple knowledge items from textual collections, minimising the user intervention. ACE-HEALTH includes three main components: (i) *Data processing and characterisation*, (ii) *Self-tuning topic modelling*, and (iii) *Knowledge visualisation*.

2.1 Data processing and characterisation

In the data processing and characterisation, ACE-HEALTH computes five steps which are carried out conclusively. (1) *Document splitting*, in which documents can be split into paragraphs or analysed in their total content. (2) *Tokenisation* represents the process of segmenting a text into words. (3) *Case normalisation*, in which each word is converted completely to lower-case characters. (4) *Stemming* maps each word into its own root form, and (5) *Stopword removal*, which discards the words which do not bring any additional information (e.g. articles, prepositions). These steps lead to the Bag-Of-Words (BOW) representation.

Statistics definition and computation. To describe the data distribution [3] of each corpus, ACE-HEALTH computes a set of statistical features able to characterise the lexical richness of the corpus:

- *Avg frequency*: the average frequency of word occurrence in the corpus;
- *Max frequency*: the maximum frequency of word occurrence in the corpus;

- *Min frequency*: the minimum frequency of word occurrence in the corpus;
- *# documents*: the number of textual documents in the corpus;
- *# terms*: number of terms in the corpus, with repetitions;
- *Avg document length*: the average length of documents in the corpus;
- *Dictionary*: the number of different terms in the corpus, without repetition;
- *TTR*: the ratio between the Dictionary and # terms;
- *Hapax %*: the ratio between the number of terms with one occurrence in the whole corpus and the cardinality of the Dictionary;
- *Guiraud Index*: the ratio between the cardinality of the Dictionary and the square root of the number of tokens (*# terms*).

Weighting schemas. To highlight the relevance of specific words in the collections, ACE-HEALTH includes two weighting strategies. A weight is a positive real number associated with each term of the collection that quantifies its degree of importance. In literature, several weighting schemas have been proposed [8]. Each corpus is represented as a matrix, named *document-term* matrix, in which each row corresponds to a document of the collection and each column corresponds to a distinct term. Each weight is computed as the product of a local weight (which measures the relevance of each word in each document) and a global one (which measures the relevance of each word in the entire collection). In ACE-HEALTH, two local weights, *Term-Frequency* (TF) and *Logarithmic Term-Frequency* (LogTF), and the global weight *Inverse Document Frequency* (IDF) are integrated.

2.2 Self-tuning topic modelling

To cluster each collection into well-separated topics, ACE-HEALTH includes the Latent Dirichlet Allocation (LDA), an unsupervised generative probabilistic model for corpora [1]. Each document is described by a distribution of topics and each topic by a distribution of words. To generate the clusters, ACE-HEALTH selects the topic with the highest probability for each document. To automatically identify a suitable value for the number of topics, ACE-HEALTH includes the ToPIC-SIMILARITY strategy proposed in [9].

Given a lower and an upper bound for the number of clusters, a new LDA model is generated for each K value. For each partition, ToPIC-SIMILARITY computes three steps:

1. *topic characterisation*, to describe each topic with its most representative words;
2. *similarity computation*, to assess how the topics in the same partition are similar;
3. *K identification*, to find good clustering configurations to be proposed.

Steps 1) and 2) are repeated for each topic in every probabilistic model. For each weighting strategy, ACE-HEALTH reports to the analyst the top-3 values that

represent good quality partitions. ACE-HEALTH includes three quality metrics: (i) *Perplexity*, (ii) *Entropy*, and (iii) *Silhouette*. The perplexity [1] describes how well the probabilistic model depicts a sample. The lower the perplexity value, the better the model. The entropy [1] is defined as the amount of information in a transmitted message. The larger the entropy, the more uncertainty is contained in the message. Lastly, the Silhouette [10] measures the similarity of a document with respect to its own cluster (cohesion) compared to other clusters (separation). It assumes values in [-1, 1], where higher values indicate that the document is well matched to its own topic.

2.3 Knowledge visualisation

A dynamic and visual exploration of both data and information hidden in the scientific literature may significantly improve the knowledge understanding and its real exploitation. To this aim, ACE-HEALTH enriches the cluster set to provide two forms of human-readable knowledge: (i) *document-topic distribution* and (ii) *topic-term distribution*.

Document-topic distribution characterises the document distribution over the topics. It includes the exploitation of (i) *topic cohesion/separation* in terms of document distribution and (ii) *coarse-grained versus fine-grained* groups through the analysis of the impact of the different weighting schemas. (i) focuses on the characterisation of the documents distribution through *pie charts* and *t-Distributed Stochastic Neighbour Embedding* (t-SNE) [7]. t-SNE allows representing high-dimensional data into lower dimensional map. The colouring of the points reflects the assignment to a specific topic after the LDA model. (ii) carries out the analysis of the weights impact in terms of coarse versus fine grained groups, through the analysis of the correlation matrix. Documents belonging to the same macro category tend to be more similar to each other than those belonging to different ones.

Topic-term distribution describes the distribution over words for each latent topic. Specifically, ACE-HEALTH includes the characterisation of (i) *topic-term distribution* through the analysis of the top-k relevant words in terms of probabilities, and (ii) the *topic cohesion/separation* in terms of relevant words. For task (i), the most probable top-k terms are extracted from each topic and plotted using word-clouds [5]. The comparison of the word-clouds obtained is left to the human analyst judgement. For task (ii), we propose the graph representation. We introduce two types of nodes: *topic nodes*, which are green nodes one for each topic, and *term nodes*, which are pink nodes one for each distinct term. Then, for each topic we add an edge for each term linked with that topic. ACE-HEALTH extracts only the top-k most relevant words for each topic (the default value is 40). If a word appears in more than one topic, we colour that node in red. The framework computes the connectivity of the graph, since if a topic is characterised by words that are not used in other topics, that topic will be disconnected by the others.

3 Experimental results

Experimental validation has been designed to address two main issues: (i) the effectiveness of ACE-HEALTH in discovering good document partitions and (ii) the ability of visualising and making the information and the extracted knowledge easy to be interpreted at different detail levels by various non-expert analysts. ACE-HEALTH has been developed to be distributed and implemented in Python. All experiments have been performed on the BigData@PoliTO cluster[3] running Apache Spark 2.3.0. The virtual nodes, the driver and the executors, have a 7GB main memory and a quad-core processor each.

Features	**WH**	**WoH**
# documents	1,000	
Max frequency	775	
Min frequency	1.0	2.0
Avg frequency	15	18
Avg document length	3600	3469
# terms	3,600,153	3,469,305
Dictionary $\|V\|$	227,210	96,362
TTR	0.06	0.05
Hapax %	57.02	0
Guiraud Index	119.75	51.73

Table 1: Statistics for the PubMed collection

Dataset. We experimentally assessed ACE-HEALTH on a real collection of medical articles extracted from PubMed[4], which is the largest bio-medical literature database in the world. From this collection, we extract 1000 papers which statistics are reported in Table 1. The dataset contains documents characterised by a great lexical richness (defined by the *Guiraud index*). Moreover, removing Hapax (i.e. column WoH) does not change the main corpus features and statistics, but allows better LDA modelling. The number of expected categories is not a-priori known.

3.1 Performance

Table 2 includes a row for each K obtained using our proposed clustering methodology and the three quality metrics. For each weighting strategy, the best solution found by ACE-HEALTH is reported in bold. The number of clusters found by the TF-IDF schema is greater than the one obtained with LogTF-IDF. This means that the two weighting schemas characterise the same dataset in a different way.

[3] https://bigdata.polito.it/content/bigdata-cluster
[4] https://www.ncbi.nlm.nih.gov/pubmed/

Weight	K	Perplexity	Silhouette	Entropy
TF-IDF	3	9.715	0.285	0.208
	6	9.511	0.28	0.314
	8	**9.432**	**0.276**	**0.352**
LogTF-IDF	3	10.318	0.258	0.251
	4	10.229	0.283	0.293
	6	**10.164**	**0.301**	**0.301**

Table 2: Experimental results

ACE-HEALTH integrates the Adjusted Rand Index (ARI) [6] to compare the solutions obtained using the two weighting schemas. A larger ARI means a higher agreement between the two partitions. For the dataset, ARI is 0.487, which means that the partitions are different leading to different partitions.

To analyse the cardinalities of each cluster set, ACE-HEALTH integrates the pie chart. Figure 1 shows the document frequency for each weighting strategy. The TF-IDF finds a large number of clusters with respect to the LogTF-IDF.

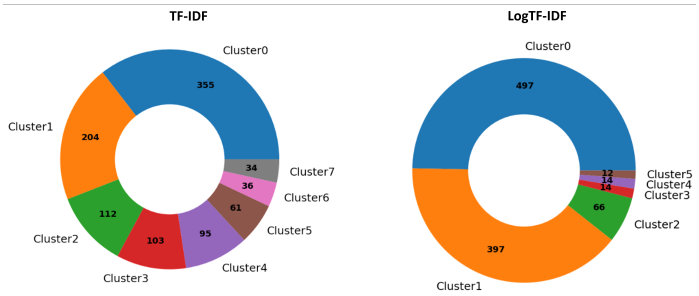

Fig. 1: Cardinality of cluster set though pie chart representation, weighting via TF-IDF weighting schema (Left) and LogTF-IDF weighting schema (Right)

3.2 Knowledge visualisation

ACE-HEALTH characterises the discovered topics through two types of interesting knowledge: *document distribution* and *topic-term distribution*. For the document distribution, two different visualisations are reported: (i) t-SNE and (ii) correlation matrix. For the topic-term distribution, ACE-HEALTH integrates (i) the word-clouds and (ii) the graph visualisation.

Document-topic distribution. To analyse the topic cohesion/separation in terms of documents distribution, ACE-HEALTH includes the t-SNE representation. Figure 2 (Top) reports the t-SNE representation, weighting via TF-IDF (Left) and LogTF-IDF (Right). Each colour point is based on the assignment to a specific topic. In both case the colours are well separated and not overlapped.

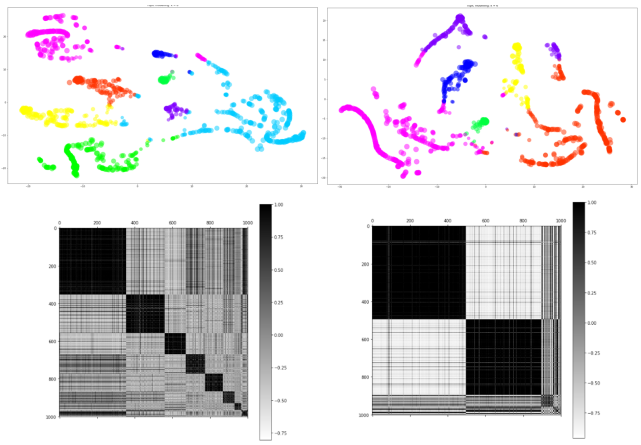

Fig. 2: t-SNE representation (Top) and Correlation matrix (Bottom), TF-IDF (Left) K=8 and LogTF-IDF (Right) K=6 weighting schemas.

To analyse the impact of the different weighting schemas, Figure 2 (Bottom) shows the correlation matrices. Documents belonging to the same topic tend to be more similar to each other than those belonging to different ones. The dark rectangles highlight that the topics are well-separated, although some correlations can be found by the analysis of topics with a lower cardinality.

Topic-term distribution. To interpret the content of each cluster ACE-HEALTH derives the word-clouds to analyse the top-k words in terms of probabilities and the graph representation to evaluate the topic cohesion/separation in terms of words.

As shown in Figure 3, Cluster0 includes documents concerning the analysis of problems related to shock or a feeling of anxiety. The most frequent words (i.e., those reported with a larger size), are related to terms such as *PTSD*, *paediatric*, *geriatric*, *adolescent* and *ASQ*. On the other side, Cluster7 (see Figure 3) includes documents addressing the cancer analysis. The fact that all the clusters are well separated is also confirmed by the graph representation shown in Figure 3 (Right). Specifically, it includes the graph representations using the top-40 words. The clusters are well-separated, and the graph is not connected.

4 Conclusion and future works

This paper presented ACE-HEALTH (AutomatiC Exploration of textual collections for HEALTH-care) a new scalable and exploratory data visualisation engine, whose target is to analyse textual medical collections through the LDA topic modelling and graphically represents the results using innovative visualisation techniques. As future directions, we are currently extending ACE-

Fig. 3: TF-IDF weighting schemas with K=8. Word-clouds of Cluster0 and Cluster7 (Left) and graph visualisation using the top-40 frequent words (Right)

HEALTH with (i) a querying engine able to assign topics to a new document using the generated LDA models.

Acknowledgements

The research leading to these results was partially funded by Project CANP, POR-FESR 2014-2020 - Technology Platform "Health and Wellness" - Piedmont Region, Italy.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)
2. Cerquitelli, T., Baralis, E., Morra, L., Chiusano, S.: Data mining for better healthcare: A path towards automated data analysis? In: 2016 IEEE 32nd International Conference on Data Engineering Workshops (ICDEW). pp. 60–63. IEEE (2016)
3. Cerquitelli, T., Di Corso, E., Ventura, F., Chiusano, S.: Data miners' little helper: data transformation activity cues for cluster analysis on document collections. In: Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics. p. 27. ACM (2017)
4. Di Corso, E., Cerquitelli, T., Ventura, F.: Self-tuning techniques for large scale cluster analysis on textual data collections. In: Proceedings of the Symposium on Applied Computing. pp. 771–776. ACM (2017)
5. Heimerl, F., Lohmann, S., Lange, S., Ertl, T.: Word cloud explorer: Text analytics based on word clouds. In: System Sciences (HICSS), 2014 47th Hawaii International Conference on. pp. 1833–1842. IEEE (2014)
6. Hubert, L., Arabie, P.: Comparing partitions. Journal of classification **2**(1) (1985)
7. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. Journal of Machine Learning Research **9**, 2579–2605 (2008)
8. Nakov, P., Popova, A., Mateev, P.: Weight functions impact on LSA performance. In: EuroConference RANLP'2001 (Recent Advances in NLP. pp. 187–193 (2001)
9. Proto, S., Di Corso, E., Ventura, F., Cerquitelli, T.: Useful ToPIC: Self-tuning strategies to enhance latent dirichlet allocation. In: 2018 IEEE International Congress on Big Data (BigData Congress). pp. 33–40. IEEE (2018)
10. Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics **20**, 53 – 65 (1987)