All in a twitter: self-tuning strategies for a deeper understanding of a crisis tweet collection

Evelina Di Corso, Francesco Ventura and Tania Cerquitelli
Dipartimento di Automatica e Informatica
Politecnico di Torino
Torino, Italy
Email: {evelina.dicorso, francesco.ventura, tania.cerquitelli}@polito.it

***Abstract*—Natural disasters have become more frequent during the past 20 years due to significant climate changes. These natural events are hotly debated on social networks like Twitter and a huge amount of short text messages are continuously and promptly exchanged with personal opinions, descriptions of the natural events and their corresponding consequences. The analysis of these large and complex data could help policy-makers to better understand the event as well as to set priorities. However, the correct configuration of the tweet mining process is still challenging due to variable data distribution and the availability of a large number of algorithms with different specific parameters. The analyst need to perform a large number of experiments to identify the best configuration for the overall knowledge discovery process. Innovative, scalable, and parameter-free solutions need to be explored to streamline the analytics process. This paper presents an enhanced version of** PASTA **(a distributed self-tuning engine) applied to a crisis tweet collection to group a corpus of tweets into cohesive and well-separated clusters with minimal analyst intervention. Experimental results performed on real data collected during natural disasters show the effectiveness of** PASTA **in discovering interesting groups of correlated tweets without selecting neither the algorithms nor their parameters.**

***Keywords*-Text mining; Parameter-free technique; data weighting function; Big Data framework; Tweet analysis.**

I. Introduction

Geophysical disasters (e.g., earthquakes, tsunamis, volcanic eruptions) have become more frequent during the past 20 years [1] due to an increase in climate-related events (mainly floods and storms) that negatively impact calamities. During these events people write a lot of realtime messages on social networks to share with friends their personal emotions, descriptions of the natural calamity and its consequences, and how they perceive the critical issues. For example, Twitter enables users to share a lot of short messages. The analysis of these textual data could help policy-makers to better understand how people perceive these events and thus set priorities while ensuring the immediate availability of risk measurement results [2]. In general, the analysis of textual collections includes several methodologies: grouping documents with similar content [3], topic-modeling [4] and detection [5], document summarization [6], pattern analysis [7] and enrichment [8]. Furthermore, although a lot of research has been devoted to analyzing tweets [9], [5], it is still a challenging task due to the variable data distribution characterizing a tweet collection and the variety of analytics algorithms available with different specific parameters that need to be manually set by the analyst. Knowledge discovery is thus a multi-step process in which data analysts tackle the complex task of configuring the analytics system to transform the overload of tweets into actionable knowledge [10]. Innovative, scalable and self-tuning techniques need to be devised to streamline the analytics process and minimize user intervention in configuring knowledge discovery.

PASTA (PArameter-free Solutions for Textual Analysis) [3] is a distributed self-tuning engine able to cluster a corpus of documents into cohesive and well-separated groups with limited analyst intervention. PASTA includes two strategies to relieve end-users of the burden of selecting specific values for algorithm parameters. Specifically, it incorporates all the analytics blocks to suggest to the analyst a good configuration of the text mining process in terms of the best algorithm to apply in each analytics step and how to set its specific parameters. PASTA runs on the Apache Spark [11] framework, supporting parallel and scalable processing for analytics activities.

This paper presents an enhanced version of PASTA named e-PASTA which is tailored to tweet messages. The contribution of the paper is threefold, providing: (i) a new local weight to measure the importance of words in tweets, (ii) a new version of the algorithm to automatically configure Latent Semantic Indexing to reduce data dimensionality, and (iii) the characterization of the cluster set through the top-k frequent words to understand the main topic addressed by tweets grouped in each cluster. Validated on a crisis tweet collection, the preliminary results of e-PASTA show the effectiveness of PASTA in identifying a good tweet partition.

This paper is organized as follows. Section II briefly describes the PASTA engine and its main building components, while Section III presents the main new features of e-PASTA. The preliminary experiments performed on a crisis tweet collection are discussed in Section IV. Finally, Section

V draws conclusions and discusses the future directions of e-PASTA.

II. The PASTA engine

PASTA is a distributed self-tuning engine whose aim is to cluster collections of textual documents into correlated groups of documents, each one addressing a specific topic. PASTA includes automatic strategies to relieve the end-user of the burden of selecting proper values for the overall cluster analysis process. The PASTA architecture, reported in Figure 1, includes three main components: (i) *Textual data processing*, (ii) *Document modeling and transformation* and (iii) *Self-tuning textual data clustering*.

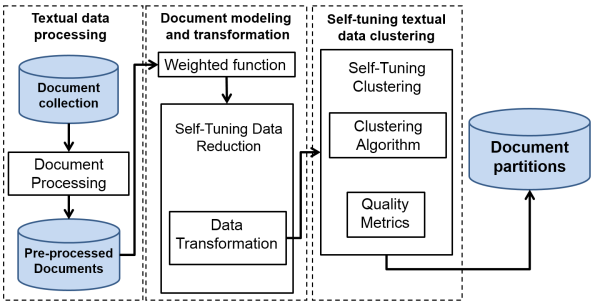

Figure 1: PASTA architecture

Textual data pre-processing is achieved through five steps which are performed sequentially as interrelated tasks. (1) *Document splitting*: Documents can be split into sentences, paragraphs, or analyzed in their entire content, according to the next analytics task; (2) *tokenization* breaks each document into a group of words named tokens within the same split; (3) *stopwords removal* eliminates the most common words in a language (e.g., articles, prepositions), which do not give any additional information; (4) *stemming* removes prefixes and suffixes to reduce each word to its base or root form; (5) *case normalization* converts the text to uppercase or lowercase. In the *document modeling and transformation* building block, PASTA investigates several suitable data weighting strategies to highlight the relevance of specific terms both in the document (local weight) and in the collection (global weight) [12]. The current implementation of PASTA includes two local weights (Term-Frequency (TF) and Logarithmic term frequency (Log)) and two global weights (Inverse Document Frequency (IDF) and Entropy (Entropy)). PASTA includes two new algorithms for the automatic configuration of the text mining process to minimize analyst intervention. To make the analysis more effectively tractable PASTA applies a data transformation method [13], i.e., Latent Semantic Indexing (LSI), to reduce the data dimensionality before a partitional clustering algorithm [14] (i.e., K-mean algorithm) is exploited. Specifically, the *ST-DaRe (Self-Tuning Data Reduction)* algorithm automatically selects three good values for the LSI algorithm to identify the best number of dimensions to analyze in the subsequent analytics step (for more details see Section III-B. Lastly, the *Self-tuning textual data clustering* component entails the discovering of groups of tweets with similar topics through the self-evaluation of the quality of the discovered clusters. To this aim, PASTA tests several configurations by varying the specific-algorithm parameter (i.e., number of desired clusters). These solutions are then compared through the computation of different quality indices (i.e., *Silhouette-based indices*) to measure the cohesion and separation of each cluster set. The top three configurations, which identify a good partition of the original collection, are selected. PASTA includes two variations of the standard Silhouette index [15] to evaluate the quality of the discovered cluster set: (i) the *purified silhouette index* (PS) (ii) and the *weighted distribution of the silhouette index* (WS). The PS index [3] disregards documents that appear in singleton clusters. Thus, the impact of these documents in the overall Silhouette index is reduced, while the WS index (assuming values in [0; 1]) [3] represents the percentage of documents in each positive bin properly weighted with an integer value $w \in [1; 10]$ (the highest weight is associated with the first bin [1 - 0.9] and so on) and normalized within the sum of weights. The higher the weighted silhouette index, the better the identified partition.

III. The e-PASTA engine

This section describes several improvements of the integrated PASTA architecture tailored to a tweet data collection. The new version is named e-PASTA.

A. An additional local weight

PASTA includes two local (Term-Frequency (TF) and Logarithmic term frequency (Log)) and two global (Inverse Document Frequency (IDF) and Entropy (Entropy)) weights to highlight the relevance of specific terms in the collection of documents. However, tweets are short messages of at most 140 characters or less. Thus, the number of times that a term occurs in a document (i.e., term frequency) is often equal to one: a meaningful word is unlikely to be repeated twice in a tweet. In this case, local weighting factor LogTF is equal to TF, and it is trivial demonstrated. Moreover, the only values that TF can assume for each term in a document are 0 (word does not appear in that tweet) or 1 (word does appear in that tweet). Thus, e-PASTA includes as only local weight *Boolean* measuring either the presence or the absence of each word in each tweet.

B. An improvement of the ST-DaRe algorithm

The ST-DaRe (Self-Tuning Data Reduction) algorithm in PASTA automatically selects three good values for the LSI parameter to identify a good number of dimensions to consider in the subsequent analytics step without losing significant information. The ST-DaRe algorithm uses three

parameters experimentally set (i.e., two thresholds and the singular value step) to analyze the variability of the singular value curve, i.e., plot of the singular values (in descending order) obtained through the SVD decomposition. The singular values are analysed in pairs using the predetermined singular value step defined by the end-user. Then, the marginal decrease in the curve is computed for each couple of singular values. If this decrease is comparable with either one of the two thresholds, or their average, the smallest singular value of the pair is selected as one of the three values. In e-PASTA we include an enhanced version of ST-DARE with only one input parameter to analyze the trend of singular values in terms of significance. The significance of each dimension is represented by the magnitude of the corresponding singular value. Insignificant dimensions represented by a low magnitude of singular values may represent noise in the data and should be disregarded in the subsequent analysis steps. Thus, we only consider the first 100 singular values for the analysis. Specifically, the mean and the standard deviation values of the magnitude of the first 100 singular values are computed and then a confidence interval is defined. The selected three-good values of the number of dimensions to consider for the next analytics steps are distributed along the curve: (i) the first is the singular value in correspondence of the mean position, (ii) the second is the singular value in correspondence of the mean plus the standard deviation position, and (iii) the last one is the singular value in correspondence of the mean position of the previous ones. Through this method the problem of the local optimality choice is overcome.

C. Explainability of cluster set

The best partitions of a tweet collection identified through e-PASTA are anonymous groups of tweets. To enhance the explainability of the cluster set, e-PASTA characterizes each found cluster via the top-k words, based on their frequency. To this aim, the top-k frequent itemsets (set of words in each cluster characterized by a frequency higher than a given threshold named support) are extracted through the FP-growth algorithm.

Weight	ExpId	LSI reduction parameter	Number of cluster	Purified Silhouette	Weighted Silhouette
Boolean-IDF	Exp_1	6	6	0.681	0.737
	Exp_2	8	6	0.583	0.632
	Exp_3	16	10	0.413	0.413
Boolean-Entropy	Exp_4	7	8	0.597	0.623
	Exp_5	9	7	0.575	0.649
	Exp_6	13	7	0.455	0.481

Table I: Experimental results obtained through e-PASTA

IV. Preliminary results

We experimentally validated e-PASTA on a crisis tweet collection [16] containing 60,005 tweets with 24,615 distinct words. Tweets are collected across 6 large events in 2012 and 2013[1]. Thus, the dataset includes 10,000 tweets for each natural disaster and each tweet is labeled with relatedness (i.e., "on-topic" or "off-topic").

The set of experiments have been designed to show *the effectiveness of e-*PASTA *in discovering a good Tweet partition*. To compare different configurations, we run e-PASTA once for each combination of weighting function (Boolean as local weight and IDF and Entropy as global weights) together with the LSI reduction method. Table I shows the top three solutions identified by e-PASTA for each weighting strategy and for each one the number of considered dimensions (LSI reduction factor), the identified number of clusters, and the quality indices as purified silhouette and weighted silhouette are reported. All selected partitions are good clusterization because both PS and WS always assume values in [0.4, 0.8). As shown in Table I the identified number of clusters found by e-PASTA has a different trend based on the weighting schema used. By increasing the number of dimensions selected through LSI after applying the Boolean-IDF weighting schema, the number of clusters found increases, while with Boolean-Entropy weighting schema, a reverse trend is noted. Moreover, the number of clusters tends to decrease, approaching the expected number of categories (i.e., 6). The Boolean-IDF weighting schema is useful when a more detailed analysis of the categories (disaster type) is of interest because some relevant subtopics within each category are identified, while Boolean-Entropy tends to find the macro-categories at a higher granularity level. Tables II and III show the detailed results of Exp_6 in Table I, enriching each tweet with its label (i.e., "on-topic" or "off-topic"). Specifically, we split tweets grouped in each cluster according to label value and category (disaster type). Table II shows the number of tweets for each cluster and category by only considering the subset of tweets with *off-topic* label, while Table III is related to *on-topic* label partition. The sum of all tweets in Tables II and III forms the complete dataset. $Cluster_6$ mainly includes off-topic tweets (about 80%) and 1,342 tweets related to Alberta Floods. The other clusters are very similar in numbers of tweets ($Cluster_1$=4,736, $Cluster_2$=8,149, $Cluster_3$=5,058, $Cluster_4$=5,873, $Cluster_5$=4,741 and $Cluster_7$=5,542[2]) and for each of them a main category/topic has been identified (bold numbers in Tables II and III). For example, $Cluster_2$ is mainly related to Floods (both Alberta and Queensland) while $Cluster_5$ describes the West Texas Explosion (see Table III). Although the number of clusters is close to the expected value (i.e., six categories), the found partition well separates the main topics. To better explain the cluster set e-PASTA also characterizes each cluster with the top-K

[1]2012 Sandy Hurricane, 2013 Boston Bombings, 2013 Oklahoma Tornado, 2013 West Texas Explosion, 2013 Alberta Floods and 2013 Queensland Floods

[2]These values are obtained by summing the total number of tweets in Tables II and III for each cluster

	Off-topic							Total number of tweets off-topic
	Cluster ID							
Category	1	2	3	4	5	6	7	
Alberta Floods	495	304	53	114	43	**3.777**	49	4,835
Boston Bombings	200	114	481	154	53	**3,309**	46	4,357
Oklahoma Tornado	285	155	35	604	25	**3,994**	52	5,150
Queensland Floods	252	414	34	141	29	**3,669**	68	4,607
Sandy Hurricane	137	118	34	117	25	**3,093**	343	3,867
West Texas Explosion	190	79	43	159	152	**4,068**	63	4,754
Tweets for each cluster	1,559	1,184	680	1,289	327	21,910	621	27,570

Table II: Detailed of Exp_6: Number of tweets for each cluster and category for off-topic label

	On-topic							Total number of tweets on-topic
	Cluster ID							
Category	1	2	3	4	5	6	7	
Alberta Floods	**1,715**	**1,837**	18	235	23	**1,342**	19	5.189
Boston Bombings	185	46	**4,155**	308	125	801	22	5,642
Oklahoma Tornado	762	41	6	**3,686**	11	307	11	4,824
Queensland Floods	187	**4,844**	3	40	8	258	60	5,400
Sandy Hurricane	132	146	8	95	21	931	**4,802**	6,135
West Texas Explosion	196	51	188	220	**4,226**	357	7	5,245
Tweets for each cluster	3,177	6,965	4,378	4,584	4,414	3,996	4,921	32,435

Table III: Detailed of Exp_6: Number of tweets for each cluster and category for on-topic label

Cluster 2		**Cluster 5**	
Frequent Item	**Support**	**Frequent Item**	**Support**
flood	0.758	explos	0.876
queensland	0.356	texa	0.767
australia	0.270	plant	0.466
water	0.140	fertil	0.382
crisi	0.103	west	0.327
alberta	0.047	waco	0.179

Table IV: Top 6 items extracted for $Cluster_2$ and $Cluster_3$

itemsets. Table IV shows the top-6 itemsets (composed of one word) found in $Cluster_2$ and $Cluster_5$ by decreasing support (frequency) values. These itemsets well describe the main topic addressed by tweets grouped in the corresponding cluster and they are in line with the results reported in Table III.

V. Conclusion and future work

This paper presents an enhanced version of a distributed self-tuning engine tailored to a crisis tweet collection. The aim of the proposed engine, e-PASTA, is to cluster collections of short messages into correlated groups of tweets addressing a similar topic. Preliminary results performed on a real dataset demonstrate the potential of e-PASTA to automatically configure the overall analytics process with minimal user intervention. One possible future direction of this research work is to extend e-PASTA to address the real-time analysis of streams of heterogeneous data to provide a more effective tool to extract near-real time knowledge items during crisis events.

Acknowledgements

The research leading to these results has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement No 700256 ("I-REACT" project).

References

[1] C. for Research on the Epidemiology of Disasters., "The human cost of natural disasters 2015, a global perspective. Available: http://reliefweb.int/sites/reliefweb.int/files/resources/. Last access on September 2017."

[2] Z. Wang, H. T. Vo, M. Salehi, L. I. Rusu, C. Reeves, and A. Phan, "A large-scale spatio-temporal data analytics system for wildfire risk management," in *Proceedings of the Fourth International ACM Workshop on Managing and Mining Enriched Geo-Spatial Data, Chicago, IL, USA, May 14, 2017*, 2017, pp. 4:1–4:6. [Online]. Available: http://doi.acm.org/10.1145/3080546.3080549

[3] E. Di Corso, T. Cerquitelli, and F. Ventura, "Self-tuning techniques for large scale cluster analysis on textual data collections," in *Proceedings of the 32nd Annual ACM Symposium on Applied Computing, Marrakesh, Morocco, April 3rd-7th, 2017*, pp. 771–776.

[4] G. Bruno, "Text mining and sentiment extraction in central bank documents," in *Big Data (Big Data), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1700–1708.

[5] L. Cagliero, T. Cerquitelli, P. Garza, and L. Grimaudo, "Twitter data analysis by means of strong flipping generalized itemsets," *Journal of Systems and Software*, vol. 94, pp. 16–29, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.jss.2014.03.060

[6] E. Baralis, L. Cagliero, A. Fiori, and P. Garza, "Mwi-sum: A multilingual summarizer based on frequent weighted itemsets," *ACM Trans. Inf. Syst.*, vol. 34, no. 1, pp. 5:1–5:35, Sep. 2015. [Online]. Available: http://doi.acm.org/10.1145/2809786

[7] X. Song, X. Wang, and X. Hu, "Semantic pattern mining for text mining," in *BigData*. IEEE, 2016, pp. 150–155.

[8] K. Holley, G. Sivakumar, and K. Kannan, "Enrichment patterns for big data," in *2014 IEEE International Congress on Big Data, Anchorage, AK, USA, June 27 - July 2, 2014*, 2014, pp. 796–799. [Online]. Available: https://doi.org/10.1109/BigData.Congress.2014.127

[9] X. Xiao, A. Attanasio, S. Chiusano, and T. Cerquitelli, "Twitter data laid almost bare: An insightful exploratory analyser," *Expert Syst. Appl.*, vol. 90, pp. 501–517, 2017.

[10] J. Han, "On the power of big data: Mining structures from massive, unstructured text data," in *BigData*. IEEE, 2016, p. 4.

[11] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. J. Franklin, S. Shenker, and I. Stoica, "Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing," in *NSDI'12*, 2012, pp. 2–2.

[12] P. Nakov, A. Popova, and P. Mateev, "Weight functions impact on LSA performance," in *EuroConference RANLP'2001 (Recent Advances in NLP*, 2001, pp. 187–193.

[13] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[14] Pang-Ning T. and Steinbach M. and Kumar V., *Introduction to Data Mining*. Addison-Wesley, 2006.

[15] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53 – 65, 1987.

[16] A. Olteanu, C. Castillo, F. Diaz, and S. Vieweg, "Crisislex: A lexicon for collecting and filtering microblogged communications in crises," in *International AAAI Conference on Web and Social Media*, 2014.