POLITECNICO DI TORINO

Dipartimento di Automatica e Informatica

PhD in Computer and Control Engineering

XXXI cycle

Supervisor

Prof. Tania Cerquitelli

Supporting decision making with self-learning methodologies

PhD Candidate: *Evelina Di Corso*

1. Introduction

Large volumes of heterogeneous data are being collected at an ever increasing rate in various modern applications. However, their exploitation is limited without effective approaches able to **automatically discover useful knowledge** from data collections **with limited user intervention**. Data-driven knowledge discovery requires users to interact with the system by tackling a variety of technical issues such as determining **optimal parameter settings** and finding **good quality models** without a-priori knowledge. Innovative and scalable data analytics methodologies are needed.

2. Goal

My PhD goal is the **design** and the **development** of a **new data analytics methodology** based on self-learning and self-tuning techniques. My research activities focus on automating data mining activities through the design of both **parameter-free algorithms** and **self-assessment strategies** able to provide good quality data models with **minimal user intervention**. The proposed solutions have been tailored to both unstructured data [1,2,3,4] (e.g., collections of documents**)** and structured data [5] (e.g., data produced through IoT devices).

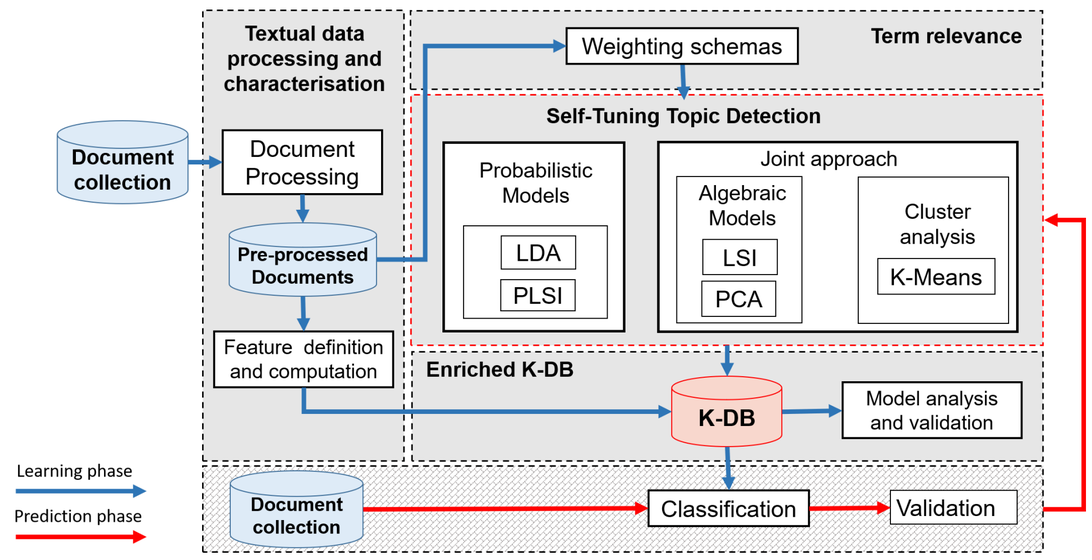

Fig.1: The ESCAPE Architecture

The proposed methodology, named **ESCAPE** (**E**nhanced **S**elf-tuning **C**haracterization of document collections **A**fter **P**arameter **E**valuation) performs different data analytics activities: (i) textual data processing and characterization [3]; (ii) term relevance weighting through different schemas (i.e., Fig.2 right); (iii) **self-tuning topic detection** through either **probabilistic models** (e.g., LDA) [1] or a **joint approach** based on algebraic models and cluster analysis, tailor to both large documents [2] and short social network messages [4]. ESCAPE is a project running on **Apache Spark** and has been validated on several collections of documents (e.g., Wikipedia datasets, Reuters Collection, Twitter).

Features	Values
# categories	5
Avg frequency	17
Max frequency	8,642
Min frequency	1
# documents	989
# terms	1,138,570
Dictionary	67,539
TTR	0.059
Hapax %	0.50
Guiraud Index	63.3

Weighting local schemas

Term Frequency
Logarithmic Term Frequency
Boolean

Weighting global schemas

Inverse Document Frequency
Entropy
TFglob

Fig.2: Feature computation (left) and proposed weighting schemas (right)

3. Results

Experimental validation has been designed to address two main issues: (i) the effectiveness of ESCAPE in discovering good document partitions, and (ii) the comparison with the state-of-the-art techniques. Here a Wikipedia dataset is discussed as a representative dataset whose features are reported in Fig.2 left. The quality metrics (e.g., perplexity, silhouette-based indices) computed for evaluating document partitions for each weighting strategy are reported for the joint approach (Table 1) and the LDA probabilistic model (Table 2). The ESCAPE partitions represent good quality models for the analysed dataset as shown in Figures 3 and 4.

Weight	K_{LSI}	$K_{K\text{-}Means}$	Accuracy	WA-F-Measure	PS	WS
TF-IDF	37	10	0.96	0.96	0.3	0.062
Log-IDF	19	7	0.95	0.95	0.3	0.068
TF-Entropy	38	10	0.91	0.91	0.3	0.056
Log-Entropy	21	5	0.99	0.99	0.4	0.076

Table 1: ESCAPE performance with the joint approach

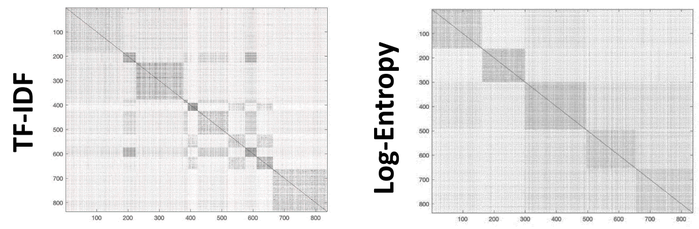

Fig. 3: Correlation Matrices

Weight	K	Perplexity	Silhouette	Entropy
TF-IDF	10	8.48	0.68	0.39
Log-IDF	8	9.18	0.67	0.32
TF-Entropy	5	9.07	0.76	0.28
Log-Entropy	7	9.88	0.84	0.17
Boolean-TF	5	6.46	0.66	0.48

Table 2: ESCAPE performance with the LDA model

Fig. 4: t-SNE (left), word-cloud (middle) and graph (right) representations

4. Conclusions

XXX includes ad-hoc auto-selection strategies to streamline the analytics process and off-load the parameter tuning from end-user. It features a distributed implementation in Apache Spark supporting parallel and scalable processing.

5. References

1. Proto, S., Di Corso, E., Ventura, F., Cerquitelli, T. Useful ToPIC: Self-Tuning Strategies to Enhance Latent Dirichlet Allocation. In *2018 IEEE BigData Congress.*
2. Di Corso, E., Cerquitelli, T., Ventura, F.. Self-tuning techniques for large scale cluster analysis on textual data collections. In *2017 ACM SAC.*
3. Cerquitelli, T., Di Corso, E., Ventura, F., Chiusano, S. (2017, June). Data miners' little helper: data transformation activity cues for cluster analyss on document collections. In *2017 ACM WIMS.*
4. Di Corso, E., Ventura, F., Cerquitelli, T.. All in a twitter: Self-tuning strategies for a deeper understanding of a crisis tweet collection. In *2017 IEEE Big Data.*
5. Di Corso, E., Cerquitelli, T., Apiletti, D.. METATECH: METeorological Data Analysis for Thermal Energy CHaracterization by Means of Self-Learning Transparent Models. Energies, 11(6).