Abstract

Evelina Di Corso
Doctoral Program in Computer and Control Engineering
XXXI Cycle

March 2019

Nowadays, large volumes of heterogeneous data are continuously collected at an ever-increasing rate in various modern applications, ranging from social networks (e.g. Twitter, Facebook), digital libraries (e.g. Wikipedia), smart city environments, Internet of Things (IoT) services and so on. In addition, we are in an age of data-intensive science and we are witnessing the unprecedented generation and sharing of large scientific datasets. Indeed, the pace of data generation has now far exceeded the pace of data analysis.

The analysis of these data collections is challenging, as it is a multi-step process in which the data scientists tackle the complex task of configuring the analytics system to transform data into actionable knowledge to effectively support the decision making process.

A plethora of algorithms are currently available for performing a given data analysis phase, but for each one the specific parameters have to be manually set and the obtained results validated by a domain expert. Moreover, real datasets are also characterised by an inherent sparseness and variable distributions, and their complexity increases with the data volume. Thus, a proper combination of different analytics algorithms should be defined to correctly model data under analysis. These activities are very time-consuming and require a lot of expertise to achieve the best trade-off between the quality of the result and the execution time. Innovative, scalable, and parameter-free solutions need to be devised to streamline the analytics process for large data collections.

The aim of this dissertation is to design and develop an automated data analytics engine to effectively and efficiently analyse large collections of textual data with minimal user intervention. Both parameter-free algorithms and self-assessment strategies have been proposed to suggest algorithms and specific parameter values for each step characterising the analytics pipeline. The proposed solutions have been tailored to textual corpora characterised by variable term distributions and different document lengths. Specifically, a new engine named ESCAPE (**E**nhanced **S**elf-tuning **C**haracterisation of document collections **A**fter **P**arameter **E**valuation) has been designed and developed. ESCAPE includes two different solutions to address document clustering and topic modelling. In each proposed solution, ad-hoc self-tuning strategies have been integrated to automatically configure the specific algorithm parameters, as well

as the inclusion of novel visualisation techniques and quality metrics to analyse the performances of the methodologies and help the domain experts to easily interpret the discovered knowledge. Specifically, ESCAPE exploits a data reduction phase computed through the Latent Semantic Analysis, before the exploitation of the partitional K-Means algorithm (named joint-approach) and the probabilistic Latent Dirichlet Allocation (named probabilistic approach). The former is based on dimensionality reduction of the document-term matrix representing each corpus, while the latter is based on learning a generative model of term distributions over topics. Both the joint-approach and the probabilistic model permit to find a lower dimensional representation for a set of documents with respect to the simple document by term matrix. Furthermore, ESCAPE includes several weighting strategies, which are able to measure term relevance in the same dataset by exploiting a local weighting schema (e.g. TF, LogTF) together with a global weighting schema (e.g. Entropy, IDF). Moreover, the outputs of the two methodologies are disjoint groups of documents with similar contents. To compare the results, different visualisation techniques have been integrated in ESCAPE to help the analyst in the interpretation of the ESCAPE results. The proposed engine has been tested through different real textual datasets characterised by a variable document length and a different lexical richness.

ESCAPE correctly identifies a good partition of a given corpus based on its main content, grouping the documents into well separated topics. Both the exploratory methodologies are able to split the corpora into well separated groups, both in terms of quality indices and easily-interpretable graphical representations. Through the joint-approach, based on a dimensionality algebraic phase before the application of the partitional K-Means algorithms, ESCAPE finds homogeneous partitions in terms of documents characterising each topic. In other words, this approach creates balanced clusters. Moreover, changing the weighting strategy, the end-user is able to partition the same dataset at different granularity levels. Specifically, the local weighting schema LogTF tends to find a small number of clusters. While, the local weighting schema TF is able to characterise the corpus by identifying also the hidden subtopics of interest. Moreover, the weighting schema TF-IDF is able to create more clusters characterising sub-topics related to the major category. On the other hand, the global weighting schema Entropy is able to find less clusters but with a larger cardinality, finding only the main relevant topic associated with each partition.

On the other side, the probabilistic model tends to find more heterogeneous clusters than the joint-approach. The probabilistic approach, exploiting the semantic similarity among the produced topics turned out to outperform the current used approach to find proper numbers of clusters. Indeed, ESCAPE is able to capture the effective cohesion level of the clusters, and then properly identify the optimal number of topics. The clusters found for all the corpora are well separated, especially for certain weighting schemas such as TF-IDF. However, with respect to the joint-approach, some weighting schemas lead to very poor results, such as the Entropy-based schemas.

Possible future extensions concern the integration of other (i) algebraic data

reduction algorithms, (ii) probabilistic topic modelling methods, and (iii) visualisation techniques. Furthermore, we are planning to introduce a semantic component able to support the analyst during the pre-processing phase (to reduce semantically correlated terms) and the post-processing phase (to help the analyst during the exploration of the results).