SeaLab Advanced Information Retrieval

Fabio Sangiacomo, Alessio Leoncini, Sergio Decherchi, Paolo Gastaldo and Rodolfo Zunino
SeaLab, DIBE, University of Genoa, Italy
Email: {fabio.sangiacomo, alessio.leoncini, sergio.decherchi, paolo.gastaldo, rodolfo.zunino}@unige.it

***Abstract*—Information Retrieval is a well established interdisciplinary topic in which machine learning, computational linguistic, computer programming and data mining merge together. SLAIR stands for SeaLab Advanced Information Retrieval and is an efficient software architecture that embeds these issues in a unique framework. SLAIR is expandable both from the data format and algorithm point of view. A pluggable notion of distance between documents drives the subsequent clustering/classification machinery; moreover SLAIR is explicitly designed to manage large scale text mining problems. The demo will be focused on the versatility of the framework; the main goal is to show how the different metrics provided by SLAIR can enhance clustering/classification ability and eventually lead to different views of the underlying textual data.**

***Keywords*-Text Clustering; Hybrid Metric; Semantic Representation; WordNet; Kernel K-Means; Support Vector Machines;**

I. Introduction

The SeaLab Advanced Information Retrieval (SLAIR) is a software framework devoted to the management of large masses of documents for clustering and classification purposes [1]. The term document here means any source of data that can act as a conveyor of information; hence it is not confined to textual format but may also include pictorial information such as images and video, as well as audio information and raw numerical data. SLAIR is designed and optimized for handling large scale data mining problems; in principle the capacity of SLAIR is bounded by the available space on the file system. SLAIR normalizes input documents into an internal representation and applies metric objects to compute distances between pair of documents. The distance measure is the computational core of SLAIR and from its definition depend all the subsequent steps and algorithmic performances.

II. Algorithmic Engine and Semantic Based Metric

A characterizing feature of the clustering engine [1] is the use of a document-distance that takes into account a conventional content-based similarity metric, a stylistic similarity criterion and a semantic representation of the documents. In the current framework the vector space model is used and also the positional information of the terms is maintained in a data structure. The Kernel K-Means algorithm supports the unsupervised grouping process for the present framework while the Support Vector Machine supports classification tasks. All the semantic elaboration of SLAIR is based on WordNet semantic network, which has been employed for its wide diffusion, reliability and generality [2]. An extension to other European languages has been implemented by using EuroWordNet databases [3], which have been properly re-factored for intensive queries.

III. SLAIR Software Architecture

SLAIR is entirely written in C++ and is a fully stand-alone framework of classes/functions; its core functions do not require neither proprietary / third-party libraries nor Standard Template Libraries, in order to maximize portability and platform independency. In addition SLAIR is designed to simplify the plugging of external libraries.

IV. Outlook of Demonstration

The exposure aims to underline the versatility of SLAIR and the ability of the the framework to exploit different notions of distance between documents to drive the clustering/classification process.
The demonstration will show that one can switch between different metrics by simply editing a row in the configuration file. Three different paradigms are provided. A pure frequency metric returns the Euclidean distance between the representing vectors of different documents. The stylistic distance uses a double description for each document that adds to the previous approach information about words positional distribution into the text. Finally a semantic metric based on WordNet can be chosen.
The demo will indeed discuss the set up of other important operating parameters. SLAIR can either try to automatically detect the language of a text corpus or use the dictionary manually set by the user.
Before the learning phase, SLAIR performs a number of tasks typical of any Natural Language Processing system. It's well known that a crucial help follows by the data representation. For this reason SLAIR uses the Porter's stemming algorithm and the morphological functions provided by WordNet. Then *Stop* and *Common Words* removals, wipe out low informative lemmas, returning a better vector space for any processed document of the corpus.

An essential feature provided by the framework is the capability of adopting three different strategies for the management of the kernel dot product matrix. The user can choose between a RAM-based approach, an ON-THE-FLY mode and a HD-based strategy. As a major result, the framework can be tuned to properly fit the available computer hardware configuration.
All the experiments will be carried on the standard Newsgroup dataset: the documents are collected from 20 different newsgroups.
There are approximately one thousand messages from each of the twenty newsgroups. Some couples of these sets are chosen for the demonstration.
Figure 1 shows a preview of the GUI that is currently under development, and that will be used during the demonstration.

A. Clustering

The first part of the demonstration focuses on the clustering functionalities of SLAIR, which exploits a hierarchical clustering technique followed by a calibration process. The demo will present and discuss different experiments involving different setups. In particular, the presentation will address the three main steps: the training process, the calibration process and the classification of unseen input documents.
For each experimental run, first the setup of the configuration file will be presented. Then, the outcomes of the training process will be shown and analyzed. SLAIR gives as output a textual file containing the hierarchy of the computed clusters; for each group the following information are provided: group ID and depth in the hierarchy, number of contained documents, cluster name built in the calibration process, number of errors in respect to input folder organization, assigned cluster label.
In addition, the framework exploits WordNet in the calibration process of each document and cluster; as a result, its possible to get a summarized description for all the data and groups that briefly outlines the main treated topics.The calibration process starts from the leaf nodes of the tree and continue with a bottom-up strategy; indeed, the process uses a smart strategy that avoids frequent words not to dominate the calibration process.
The exposure will finally deal with the crucial issue of classifying an unseen document. This process allows ascribing the candidate group membership for the input document, among a set of alternative clusters defined in the training phase. Indeed, it will be shown that SLAIR allows the user to easily evaluate the accuracy performance of the clustering and calibration algorithm. If the input data are organized in subfolders, the name of the directories are directly interpreted as the labels of the categories to be used as tags for the corresponding documents. As a consequence, classification accuracy on the trained corpus is computed automatically.

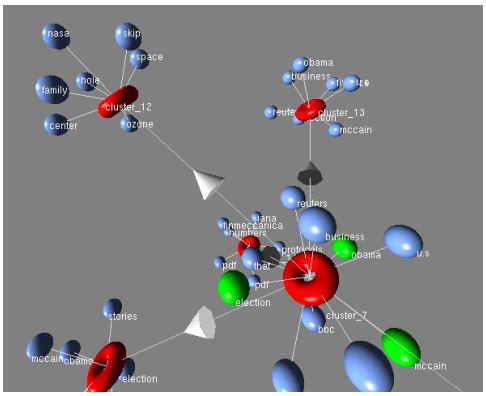

Figure 1. A sample view of the GUI

B. Classification

The second part of the demonstration deals with the classification functionalities of SLAIR. The framework uses a Support Vector Machine to provide a supervised classification tool as an alternative to the unsupervised approach supported by the clustering functionality.
This part of the demonstration will indeed analyze the performance of the different distance metrics when used for classification purposes. To allow a fair comparison with the experimental sessions discussed in the first part, the Newsgroup data will again act as a test bench.
The exposure will involve different experiments addressing various problem setups. Indeed, the discussion will analyze the outcomes of the classification process and compare them with those provided by the clustering engine, to illustrate the different role that the two functionalities may have in the text mining process.

V. Conclusion

SLAIR framework is a versatile tool to mine textual data both with a supervised or unsupervised strategy. The core of the system is the metric computation task and, as shown, is a highly configurable aspect of the architecture. The final goal of this demo is to show how different metrics mean different view on the underlying textual data; depending on the dataset, semantic or frequency or stylistic metric, or a combination of them, can be more suitable to represent data.

References

[1] S. Decherchi, P. Gastaldo, and R. Zunino, "K-means clustering for content based document management in intelligence," in *Advances In Artificial Intelligence for Privacy Protection and Security*, A. Solanas and A. M. Bellesté, Eds. World Scientific Publishing, 2009.

[2] G. A. Miller, "Wordnet: A lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[3] P. Vossen, *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers, 1998.