A Visualization Tool of Probabilistic Models for Information Access Components

Lorenzo De Stefani, Giorgio Maria Di Nunzio, and Giorgio Vezzaro

Department of Information Engineering, University of Padua, Italy
{destefan,dinunzio,vezzarog}@dei.unipd.it

Abstract. Since massive collections of textual documents become more and more available in digital format, the organization and classification of these documents in Digital Library Management System (DLMS) becomes an important issue. Information access components of a DLMS, such as automatic categorization and retrieval components of digital objects, allow users to interact with the system in order to browse, explore, and retrieve resources from collections of objects. The demonstration presents a two-dimensional visualization tool of Naïve Bayes (NB) probabilistic models for Automated Text Categorization (ATC) and Information Retrieval (IR) useful to explore raw data and interpret results.

1 Introduction

DLMSs are an example of systems that manage collections of multi-media digitalized data and include components that perform the storage, access, retrieval, and analysis of the collections of data. The visualization of the results returned by these components may be a key point for: firstly, system designers during the process of raw data exploration; secondly, users to interpret results more clearly and possibly interact with them.

In this demonstration, we present a tool for the visualization of NB probabilistic models for information access components of a DLMS that represents digital objects on the two-dimensional space [2,1]. This tool demonstrates to be a valid visualization tool for understanding the relationships between categories of objects, and helps users to visually audit the classifier and identify suspicious training data. This model defines a direct relationship between the probability of an object given a category of interest and a point on a two-dimensional space. In this light, it is possible to graph entire collections of objects on a Cartesian plane, and to design algorithms that categorize and retrieve documents directly on this two-dimensional representation. The demonstration will applied to the task of automatic text classification and text retrieval.

2 The Two-Dimensional Representation of Probabilities

In the two-dimensional representation of documents, the equation of the ranking or the classification function has to be written in such a way that each coordinate

M. Agosti et al. (Eds.): ECDL 2009, LNCS 5714, pp. 481–482, 2009.

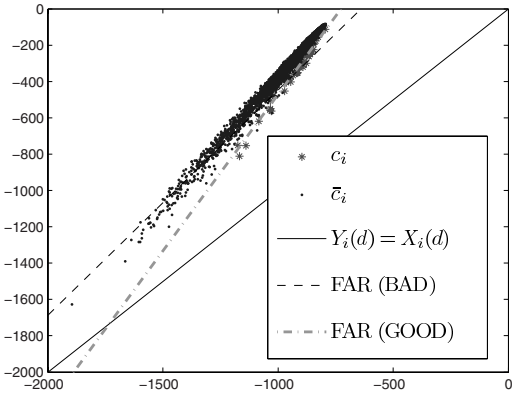

Fig. 1. An example of different separating lines for relevant and non relevant documents

of a document is the sum of two addends: a variable component $\mathrm{P}(d|c_i)$, the probability of a document given a category of interest, and a constant component $\mathrm{P}(c_i)$, the prior of the category of interest [1]. For example, in the case of NB models the equation becomes:

$$\underbrace{\log\left(\mathrm{P}(d|c_i)\right) + \log\left(\mathrm{P}(c_i)\right)}_{X_i(d)} > \underbrace{\log\left(\mathrm{P}(d|\bar{c}_i)\right) + \log\left(\mathrm{P}(\bar{c}_i)\right)}_{Y_i(d)}$$

In this demonstration we show the functionalities of the actual prototype on standard benchmark collections: how the ranking or classification functions are learned from the data as separating lines; how particular unbalanced distribution of documents can be corrected by means of parameter estimation; how the multivariate model and the multinomial model perform on different languages; how blind and/or explicit relevance feedback affect ranking list, and how the selection of relevant documents changes the shape of the clouds of relevant and non-relevant documents.

In Figure 1, a screen-shot of the main window of the visualization tool si shown. Relevant documents are plotted against non-relevant documents (respectively c_i and $\bar{c}_i$), three separating lines are shown to demonstrate how the same algorithm, the Focused Angular Algorithm (FAR), produces different separating lines according to different estimates of the parameters.

References

1. Di Nunzio, G.: Using Scatterplots to Understand and Improve Probabilistic Models for Text Categorization and Retrieval. International Journal of Approximate Reasoning (in press, 2009), http://dx.doi.org/10.1016/j.ijar.2009.01.002
2. Di Nunzio, G.M.: Visualization and Classification of Documents: A New Probabilistic Model to Automated Text Classification. Bulletin of the IEEE Technical Committee on Digital Libraries (IEEE-TCDL) 2(2) (2006)