A Visualization Tool of Probabilistic Models for Information Access Components*

Giorgio Maria Di Nunzio
Dept. of Information Engineering
University of Padua
Via Gradenigo 6/a, 35131
Padua, Italy
dinunzio@dei.unipd.it

ABSTRACT

An effective graphic interface is a key tool to improve the fruition of the results retrieved by an Information Retrieval (IR) system. In this work, we describe a two-dimensional interface that represents the documents ranked on a Cartesian space and allows the user to interact with the documents in order to improve the results of the search engine. Results are classified and ranked according to the best separating line of the two classes of documents: relevant and non relevant documents. Mathematical tools such as least squares distances are used to train the supervised algorithm that finds the separating and ranking lines.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering, Relevance feedback, Retrieval models, Search process*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Graphical user interfaces (GUI)*

General Terms

Algorithms, Design, Experimentation

Keywords

Information Visualization, Machine Learning, Naïve Bayes Models, Relevance Feedback

1. INTRODUCTION

Visualization is the process of transforming data, information, and knowledge into graphic presentations to support tasks such as data analysis and information exploration. The definition of a spatial structure for information visualization is challenging because data in an information space may be multi-faceted, relationships of data are interwoven and are complicated. Moreover, the definition of such a space means a complex process of extracting displayable attributes from objects, organizing the information, projecting objects onto the structure, and synthesizing search features, objects and object relationships into the visual space [5].

The introduction of visualization environments may add cognitive processes to the user who needs to understand and learn the characteristics of the new environment and interact with them to get the best from the system. In fact, the aim of visualization environments, as external representation of the world of interest, is to reduce the amount of cognitive effort required to solve informationally equivalent problems [4]. In particular, an IR system should provide users an environment in which they can exploit their skills to maximize their cognitive abilities. The visualization of an IR system is nothing but a process that transforms invisible abstract data and their semantic relationships in a visible collection on a display in order to find the user information need more easily.

In this paper, we present the design and implementation a tool for the visualization of Naïve Bayes (NB) probabilistic models for information access components that represents digital objects on the two-dimensional space [2, 3, 1]. The demonstration will applied to the task of automatic text classification and text retrieval.

2. DESIGN

The model which upholds the visualization tool 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. This tool demonstrates to be a valid visualization tool also for understanding the relationships between categories of objects.

The design of the two-dimensional visualization tool follows two main requirements:

- for end-user, the interface should give the opportunity to define the query with simple or advanced options, and to express judgements for the documents retrieved which will be used to re-rank documents;
- for researchers, the interface should display the decisions taken by the search engine in terms of separating line and explain how the relevance feedback given by the user affects the list of ranked documents.

The interface offers the possibility to write free text queries, as any other search engine, or load predefined queries; predefined queries are used for research purposes and recreates

*This is an extended abstract of [1]

Appears in the Proceedings of the 1st Italian Information Retrieval Workshop (IIR'10), January 27–28, 2010, Padova, Italy.
http://ims.dei.unipd.it/websites/iir10/index.html

the environment of evaluation tasks organized by campaign such as TREC[1] or CLEF[2].

The interface associate each document of the collection to a point in the two-dimensional space according to a probabilistic algorithm: the abscissa reflects how much the document is relevant to the query, the ordinate reflects how much the document is not relevant to the query. The pair of numbers gives an indication of the fraction of relevance for that particular document given the query, this pair is plotted on a frame and the relative position of this point with respect to the other documents in the collection determines its position in the list of ranked documents.

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 of a document is the sum of two addends: a variable component $\mathrm{P}(d|c_i)$, the probability of a document d given a category of interest c_i, and a constant component $\mathrm{P}(c_i)$, the prior of the category of interest c_i [3] For example, in the case of NB models the equation becomes:

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

When the inequality holds, the document is considered an element of category c_i. If c_i and $\bar{c}_i$ are considered respectively the set of relevant documents and the set of non relevant documents, we can divide the collection of documents in these two sets; if we are only interested in the ranking of documents, we can compute the list of retrieved documents by combining the two components into one *relevance weight*.

Documents can be classified or ranked differently according to the Focused Angular Region algorithm which computes the best separating (or ranking) line by means of regression techniques and least squares orthogonal, and vertical, distances. Information about the categories of documents are collected during the interaction of the user with the interface; in particular, the relevance judgements that the user expresses for the documents are used to re-compute the probabilities and train the algorithm (details of this supervised algorithm are given in [3]). This part can be done automatically by selecting in the interface the option "Blind relevance feedback", which takes the first n documents of the current list of documents and set them as relevant.

3. RESULTS AND OPEN QUESTIONS

This visualization tool was tested on standard benchmark collections and a demonstration was presented at [1] in order to answer the following research questions: how well 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.

During the interaction with the system, new questions and new research ideas were collected about advances types of interaction: changing the estimated probability of terms directly; smoothing parameters in order to *see* how the clouds of points move in the space and how the performance changes accordingly; drawing the clouds of points incrementally, highlighting the contribution of each term to understand which terms better discriminate the two sets of points.

[1]http://trec.nist.gov/

[2]http://www.clef-campaign.org

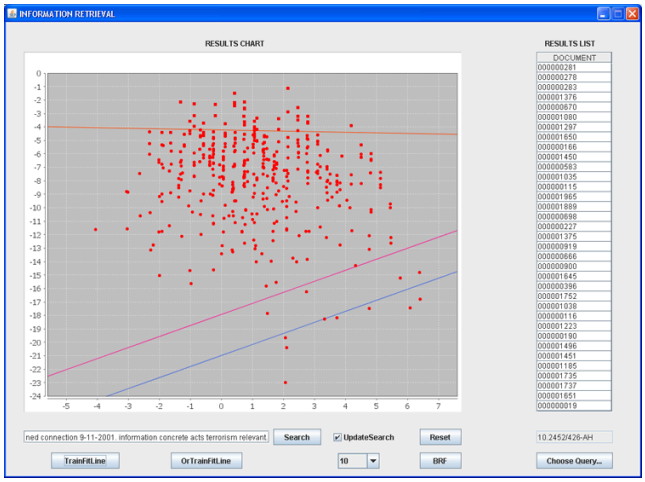

Figure 1: An example of the interface used by researchers.

In Figure 1, a screen-shot of the main window of the visualization tool si shown. The example shows the interface used by researchers. The different separating lines are calculated for a blind relevance feedback of 10 documents: the category of relevant documents in blue, the category of non relevant documents in red, the best separating line in purple. The list of retrieved documents is presented on the right. The user can choose to select a document, read it, and judge it as relevant or non relevant. This information is stored and used to train the supervised algorithm when the user selects the "update search" box.

4. REFERENCES

[1] L. De Stefani, G. M. Di Nunzio, and G. Vezzaro. A visualization tool of probabilistic models for information access components. In *Proceedings of Research and Advanced Technology for Digital Libraries (ECDL 2009)*, Corfu, Greece, September/October 2009. LNCS, Springer.

[2] G. M. Di Nunzio. 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.

[3] G. M. D. Nunzio. Using Scatterplots to Understand and Improve Probabilistic Models for Text Categorization and Retrieval. *Journal of Approximate Reasoning*, 50(7):945–956, July 2009. http://dx.doi.org/10.1016/j.ijar.2009.01.002.

[4] M. Scaife, M. Scaife, Y. Rogers, and Y. Rogers. External cognition: how do graphical representations work? *International Journal of Human-Computer Studies*, 45:185–213, 1996.

[5] J. Zhang. *Visualization for Information Retrieval*, volume 23 of *The Information Retrieval Series*. Springer, 2008. ISBN: 978-3-540-75147-2.