A Software System for Topic Extraction and Document Classification

Davide Magatti and Fabio Stella
Department of Informatics, Systems and Communications
Universita' degli Studi di Milano-Bicocca
Milan, Italy
Email: $\{magatti, stella\}$ *@disco.unimib.it*

Marco Faini
DocFlow Italia S.p.A.
Centro Direzionale Milanofiori, Strada 4 Palazzo Q8
20089 Rozzano, Italy
Email: marco.faini@docflow.it

***Abstract*—A software system for topic extraction and automatic document classification is presented. Given a set of documents, the system automatically extracts the mentioned topics and assists the user to select their optimal number. The user-validated topics are exploited to build a model for multi-label document classification. While topic extraction is performed by using an optimized implementation of the Latent Dirichlet Allocation model, multi-label document classification is performed by using a specialized version of the Multi-Net Naive Bayes model. The performance of the system is investigated by using 10,056 documents retrieved from the WEB through a set of queries formed by exploiting the Italian Google Directory. This dataset is used for topic extraction while an independent dataset, consisting of 1,012 elements labeled by humans, is used to evaluate the performance of the Multi-Net Naive Bayes model. The results are satisfactory, with precision being consistently better than recall for the labels associated with the four most frequent topics.**

***Keywords*-topic extraction; text mining; classification.**

I. Introduction

The continuously increasing amount of text available on the WEB, news wires, forums and chat lines, business company intranets, personal computers, e-mails and elsewhere is overwhelming [1]. Information, is switching from *useful* to *troublesome*. Indeed, it is becoming more and more evident that while the amount of data is rapidly increasing, our capability to process information remains constant. This trend strongly limits the extraction of valuable knowledge from text and thus drastically reduces the competitive advantage we can gain. Search engines have exacerbated such a problem by dramatically increasing the amount of text available in a matter of a few key strokes. In this paper the authors describe a software prototype for *topic extraction* and *multi-label document classification*. The prototype implements a conceptual model which processes a corpus of textual data to discover which topics are mentioned, and exploits them to learn a supervised multi-label classification model.

The rest of the paper is organized as follows; Section 2 introduces text mining, topic extraction and text categorization. Section 3 describes the main components of the software prototype along with their functionalities. Section 4 is devoted to numerical experiments. Finally, Section 5 presents conclusions and discusses further developments.

II. Text Mining

Text Mining (TM) [1], [2] is an emerging research area which aims to solve the problem of *information overload*. Typical tasks of TM are: text categorization, document clustering and organization, and information extraction.

Probabilistic Topic Extraction (PTE) analyzes the content of documents to discover the *topics* mentioned in a document collection. A variety of models has been proposed, in the specialized literature [3], [4], [5], which share the same rationale, i.e. a document is a mixture of topics. To describe how the PTE model works, let $P(z)$ be the probability distribution over topics z, $P(w|z)$ be the probability distribution over words w given topic z. The *topic-word distribution* $P(w|z)$ specifies the weight to thematically related words. A document is assumed to be formed as follows: the i^{th} word w_i is generated by first extracting a sample from the *topic distribution* $P(z)$, then by choosing a word from $P(w|z)$. We let $P(z_i = j)$ be the probability that the j^{th} topic was sampled for the i^{th} word token while $P(w_i|z_i = j)$ is the probability of word w_i under topic j. Therefore, PTE induces the following distribution over words within a document: $P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j) P(z_i = j)$ where T is the number of topics. Hofmann [4] proposed the probabilistic Latent Semantic Indexing (pLSI) method which makes no assumptions about how the mixture weights, i.e. $P(z_i = j)$, are generated. Blei et al. [5] improved the generalizability of pLSI to new documents by introducing a Dirichlet prior, with hyperparameter α, on $P(z_i = j)$, thus originating the Latent Dirichlet Allocation (LDA) model. Griffiths and Steyvers [6] extended the LDA model with a Dirichlet prior, with hyperparameter β, also to $P(w_i|z_i = j)$ to smooth the word distribution in every topic.

Document classification, the task of classifying natural language documents into a predefined set of semantic categories, has become one of the key methods for organizing online information. It is commonly referred to as text categorization and represents a building block of several applications such as: web pages categorization, newswire filtering and automatic routing of incoming messages at call centers, ...

978-0-7695-3801-3/09 $26.00 © 2009 IEEE
DOI 10.1109/WI-IAT.2009.49

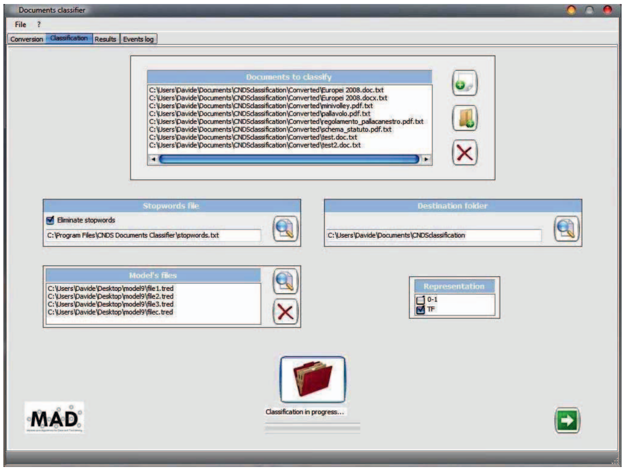
Text categorization is distinguished into binary, multi-class and multi-label settings. In the binary setting there are exactly two classes, i.e. *relevant* and *non relevant*, *spam* and *non spam* or *sport* and *non sport*. Some classification tasks require more than two classes, i.e. an e-mail routing agent at a call center needs to forward an incoming message to the right operator, depending on the specific nature of the message contents. Such cases belong to the multi-class setting where documents can be labeled with exactly one out of K classes. Finally, in the multi-label setting there is not a one-to-one correspondence between class and document. In such a setting, each document can belong to many, exactly one or no category at all. Several works have extensively studied the Naive Bayes model for text categorization [7], [8]. However, these pure Naive Bayes classifier models consider a document as a binary feature vector, and so they cannot utilize the term frequencies in a document, resulting in poor performances. The multinomial Naive Bayes text classifier has been shown to be an effective alternative to the basic Naive Bayes model by a number of researchers [9], [10], [11]. Recently Kim et al. [12] revisited the naive Bayes framework and proposed a Poisson Naive Bayes model for text classification with a statistical feature weighting method.

III. The Software System

The software system described in this paper consists of three main components; namely Text Pre-processor, Topic Extractor and Multi-label Classifier. These components, have been integrated to deploy a software system working on Windows XP and Vista operating systems.

Text Pre-processor. This software component implements functionalities devoted to document pre-processing and document corpus representation. It offers stopwords removal, different word stemming options and several filters to exclude those words which are judged to be too frequent/rare within a document and/or across the document corpus. This software component exploits a general purpose Italian vocabulary to obtain the *word-document matrix* following the *bag-of-words* model. Furthermore, the following document representations are allowed: binary, term frequency and term frequency inverse document frequency. The following document formats are valid inputs for the software system: *pdf*, *word* and *txt*.

Topic Extractor. This software component offers topic extraction and topic selection functionalities. The *topic extraction functionality* is implemented through a customized version of the LDA model [6]. LDA learning is obtained by using the Gibbs sampling algorithm which has been implemented in the C++ programming language. The *topic selection functionality* assists the user in choosing the optimal number of topics to be retained. Topic selection is implemented through a hierarchical clustering procedure based on the symmetrized Kullback Liebler distance between topic distributions. Each retained topic $z = j$ is summarized through the estimate of its prior probability $P(z = j)$, a sorted list of its most frequent words w, together with the estimate of their conditional probabilities of occurrence given the topic, i.e. the value of the conditional probability $P(w|z = j)$.



C:\Users\Davide\Documents\CNDSclassification\Converted\Europei 2008.doc.txt
C:\Users\Davide\Documents\CNDSclassification\Converted\Europei 2008.docx.txt
C:\Users\Davide\Documents\CNDSclassification\Converted\minivolley.pdf.txt
C:\Users\Davide\Documents\CNDSclassification\Converted\pallavolo.pdf.txt
C:\Users\Davide\Documents\CNDSclassification\Converted\regolamento_pallacanestro.pdf.txt
C:\Users\Davide\Documents\CNDSclassification\Converted\schema_statuto.pdf.txt
C:\Users\Davide\Documents\CNDSclassification\Converted\test.doc.txt
C:\Users\Davide\Documents\CNDSclassification\Converted\test2.doc.txt

☑ Eliminate stopwords
C:\Program Files\CNDS Documents Classifier\stopwords.txt

C:\Users\Davide\Documents\CNDSclassification

C:\Users\Davide\Desktop\model9\file1.tred
C:\Users\Davide\Desktop\model9\file2.tred
C:\Users\Davide\Desktop\model9\file3.tred
C:\Users\Davide\Desktop\model9\filec.tred

□ 0-1
☑ TF

Classification in progress...

Figure 1. MNPNB classifier: labeling GUI.

Multi-label Classifier. Implements a supervised multi-label classification model. This model exploits the output from the **Topic Extractor** software component. It uses a customized version of the Multi-Net Poisson Naive Bayes (MNPNB) model [12]. The MNPNB model allows to select the following *bag-of-words* representations: binary, term frequency and term frequency inverse document frequency. Each new document, represented according to the bag-of-words model, is automatically labeled depending on the user specified value for the posterior threshold. This software component has been implemented in C# and is available through a dedicated GUI (Figure 1) as well as a WEB service.

IV. The Italian Google Directory

The performance of the software system is investigated using a document corpus collected by exploiting the structure of topics offered by the *Italian Google Directory* (gDir)[1]. This topic structure relies on the Open Directory Project (DMOZ) which manages the largest human-edited directory available on the web. The topics associated with the first level of gDir are the following: Acquisti, Affari, Arte, Casa, Computer, Consultazione, Giochi, Notizie, Regionale, Salute, Scienza, Societa, Sport, Tempo Libero. Each first level topic is associated with a second and third level sub-topic structure summarized in a list of words.

[1] http://www.google.com/Top/World/Italiano/.

A. Document Corpus

The document corpus has been collected by submitting a set of 273 queries to the Google search engine. Each query contains a pair of words, randomly selected from the union of the lists of words associated to second and third level sub-topic structures. The Google search engine PDF filter has been used to ensure that only pdf format files are retrieved. The random query process retrieved 14,037 documents, each associated with one or more gDir first level topics.

B. Text Pre-processing

The document corpus has been submitted to the **Text Pre-processor** software component. Therefore, PDF files have been transformed to plain text, submitted to stopwords removal and to word stemming. Furthermore, size-based file selection has been applied to include only those PDF files with the size between 2 and 400 KB. The obtained document corpus consists of 10,056 documents ($D = 10,056$) while the global vocabulary consists of 48,750 word tokens ($W = 48,750$). The document corpus is represented by means of a *word-document matrix* WD consisting of $W \times D$ (term frequency) elements. The WD matrix is inputed to the **Topic Extractor** component.

C. Topic Extraction

The **Topic Extractor** has been invoked with the following learning parameters; 12 topics ($T = 12$), alpha prior equal to 1.67 ($\alpha = 1.67$), which implements the $\alpha = \frac{20}{T}$ rule cited in [5], beta prior equal to 0.01 ($\beta = 0.04$), which implements the $\beta = \frac{200}{W}$ rule cited in [5]. The Gibbs sampling procedure has been run 100 times (500 sampling iterations) with different initial conditions and different initialization seeds. The 12 topics extracted from the **Topic Extractor** software component have been summarized through their corresponding 100 most frequent words. Among the extracted topics, the four most interesting ones have been manually labeled as follows;

- SALUTE (medicine and health),
- COMPUTER (information and comm. technologies),
- TEMPO LIBERO (travels and vacations),
- AMMINISTRAZIONE (bureaucracy, public services).

The structure of the four topics, is described in Table I and Table II. Each topic is associated with an estimate of its prior probability (top of Tables). Furthermore, each topic is summarized in a *words list*, whose 15 most frequent words are reported in Table I and Table II. It is worthwhile to mention that for each pair of words w_i and topics j, an estimate of the conditional probability of the word w_i given the topic j, i.e. $P(w_i|j)$, is provided.

Table I
SALUTE AND COMPUTER.

SALUTE	0.0787
cellule	0.0032
emissioni	0.0029
nutrizione	0.0028
molecolare	0.0026
proteine	0.0022
dieta	0.0022
climatici	0.0021
foreste	0.0021
cancro	0.0021
aids	0.0020
disturbi	0.0020
infermiere	0.0019
cibi	0.0019
tumori	0.0019
veterinaria	0.0018

COMPUTER	0.0696
blog	0.0063
google	0.0035
linux	0.0034
copyright	0.0033
wireless	0.0030
source	0.0029
access	0.0028
client	0.0027
multimedia	0.0027
hacker	0.0026
password	0.0026
giornalismo	0.0025
browser	0.0023
provider	0.0022
telecom	0.0022

Table II
TEMPO LIBERO AND AMMINISTRAZIONE.

TEMPO LIBERO	0.0884
sconto	0.0077
aeroporto	0.0038
salone	0.0035
spiaggia	0.0028
lago	0.0028
colazione	0.0026
albergo	0.0025
vacanza	0.0025
piscina	0.0024
vini	0.0023
bagni	0.0023
voli	0.0021
pensione	0.0021
biglietto	0.0020
notti	0.0020

AMMINISTRAZIONE	0.1021
locazione	0.0021
federale	0.0021
direttivo	0.0021
finanze	0.0020
versamento	0.0020
lire	0.0019
commi	0.0019
prescrizioni	0.0018
vietato	0.0018
contrattuale	0.0018
richiedente	0.0018
utilizzatore	0.0017
agevolazioni	0.0017
contabile	0.0017
appalto	0.0017

D. Multi-Label Classification

The performance of the software system has been estimated by submitting a new document corpus to the **Multi-label Classifier**. This document corpus has been collected by using the same random querying procedure described in subsection IV-A. Its documents have been manually labeled, according to the 12 first level gDir topics, independently by three humans. The labeled document corpus consists of 1,012 documents. In detail, 478 documents are singly labeled, 457 are doubly labeled, while only 77 are associated with three labels. The **Multi-label Classifier**, queried by using the binary document representation and by setting a posterior threshold equal to 0.5, achieves an *accuracy* equal to 73%, which can be considered satisfactory. The estimates of precision and recall for the four selected topics: namely COMPUTER, SALUTE, AFFARI and TEMPO LIBERO, are reported in Table III. The best result is achieved for the topic AMMINISTRAZIONE, where the precision equals 92%, i.e. if the **Multi-label Classifier** labels a document with the label AMMINISTRAZIONE, then the labeling is wrong with respect to the manual labeling with probability 0.08. Furthermore, the recall equals 79% which means that the documents manually labeled with AMMINISTRAZIONE are correctly labeled by the **Multi-label Classifier** with probability 0.79. The topic COMPUTER achieves a precision value equal to 85% which is slightly lower than that achieved for AMMINISTRAZIONE. However, the achieved recall value drops from 79% of AMMINISTRAZIONE to 59%. The topic TEMPO LIBERO achieves a precision value equal to 78%, i.e. slightly lower than those achieved for COMPUTER, while the achieved recall value is equal to 44%. Thus, the achieved recall value is significantly lower than that achieved for COMPUTER. Finally, the topic SALUTE achieves performances comparable to those of TEMPO LIBERO. Indeed, the precision equals 76%, while the recall equals 41%. Around 57% of documents manually labelled with TEMPO LIBERO and/or with SALUTE, are not identified as such by the **Multi-label Classifier**. It is worthwhile to notice that for each label the achieved recall value is consistently lower than the precision value. A possible explanation for this behavior is as follows; the manual labeling procedure is both complex and ambiguous; it could label documents by using a broader meaning for each topic. Therefore, it is expected somewhat that automatic document classification could not achieve excellent performance with respect to both precision and recall. It is expected that the 'purer' a topic is, the better the performance achieved for the automatic documents classification task will be. However, it is important to keep in mind the difficulty of the considered labeling task, together with the fact that human labeling of documents can result in ambiguous and contradictory label assignment.

Table III
PRECISION/RECALL.

	SAL.	COM.	TEL.	AMM.
Precision	76	85	78	92
Recall	41	59	44	79

V. CONCLUSIONS AND FUTURE WORK

In this paper a software system for topic extraction and document classification has been described. The software system assists the user in correctly discovering which main topics are mentioned in a document collection. The discovered topic structure, after being user-validated, is used to implement an automatic document classifier. This model suggests labels to be used for each new document submitted by the user. It is important to mention that each document is not restricted to receive a single label but can be labeled with more topics. Furthermore, each topic for a given document is associated with a probability value that informs the user about the fitting of the topic to the considered document. This feature offers an important opportunity to the user who can sort his/her document collection in descending order of probability for each topic.

However, it must be clearly stated that many improvements can be achieved by taking into account specific user requirements. This aspect is under investigation, and particular attention is being dedicated to non-parametric models for the discovery of hierarchical topic structures. Finally, the interplay between topic taxonomies and topic extraction algorithms offers an interesting research direction to explore.

ACKNOWLEDGMENT

The authors would like to thank the anonymous referees for their valuable comments. Special thanks go to Gabriella Pasi for her suggestions.

REFERENCES

[1] R. Feldman and J. Sanger, *The Text Mining Handbook*. New York: Cambridge University Press, 2007.

[2] M. W. Berry and M. Castellanos, *Survey of Text Mining II: clustering, classification and retrieval*. London: Springer, 2008.

[3] T. L. Griffiths and M. Steyvers, "A probabilistic approach to semantic representation," in *Proceedings of the Twenty-Fourth Annual Conference of Cognitive Science Society*, G. W. and C. Schunn, Eds., 2002, pp. 381–386.

[4] T. Hofmann, "Probabilistic latent semantic analysis," in *In Proc. of Uncertainty in Artificial Intelligence, UAI 1999*, 1999, pp. 289–296.

[5] D. M. Blei, N. Andrew, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, January 2003.

[6] T. L. Griffiths and M. Steyvers, "Finding scientific topics." *Proc Natl Acad Sci U S A*, vol. 101 Suppl 1, pp. 5228–5235, April 2004.

[7] A. Mccallum and K. Nigam, "A comparison of event models for naive bayes text classification," in *In AAAI-98 Workshop on Learning for Text Categorization*. AAAI Press, 1998, pp. 41–48.

[8] D. D. Lewis, "Naive (bayes) at forty: The independence assumption in information retrieval," in *ECML '98: Proceedings of the 10th European Conference on Machine Learning*. London, UK: Springer-Verlag, 1998, pp. 4–15.

[9] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, "Inductive learning algorithms and representations for text categorization," in *CIKM '98: Proceedings of the seventh international conference on Information and knowledge management*. New York, NY, USA: ACM Press, 1998, pp. 148–155.

[10] Y. Yang and X. Liu, "A re-examination of text categorization methods," in *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM Press, 1999, pp. 42–49.

[11] A. K. McCallum and T. Mitchell, "Text classification from labeled and unlabeled documents using em," in *Machine Learning*, 2000, pp. 103–134.

[12] S.-B. Kim, K.-S. Han, H.-C. Rim, and S. H. Myaeng, "Some effective techniques for naive bayes text classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 11, pp. 1457–1466, 2006.