2009 Ninth International Conference on Intelligent Systems Design and Applications

A probabilistic method for text analysis

Fabio Clarizia, Massimo De Santo, Paolo Napoletano
Department of Information and Electrical Engineering
University of Salerno – Via Ponte Don Melillo – 84084 Fisciano (SA) – Italy
{fclarizia,desanto,pnapoletano}@unisa.it

Abstract

Textual materials are source of extremely valuable information, for which there must be a reflection on the techniques of analysis to be used to avoid subjective interpretations especially in the content. The Textual Analysis (TA), which makes use of statistical techniques, ensures the systematic exploration of the structure of the text (size, occurrence, etc.) and simultaneously the possibility to return at any time to the original text for the appropriate interpretations.

In this work we test a new technique based on a probabilistic model of language known in the literature as "topic model" for analyzing corpora of documents about electromagnetic pollution. The proposed method is able to reveal how the meaning of a document is distributed all along its spectrum (word-frequency) indicating that the real meaning of a document can be inferred following a multilevel analysis. Such analysis is carried out exploiting a new concept of ontology already used in literature and deeply explained here.

1. Introduction

The problem of the environmental impact of electromagnetic fields, the effects of exposure of the population, the tumors had a significant effect in Italy as in the rest of the World, where there have been sensationalist tones and alarms often unjustified.

Italy has always been characterized by a high-perceived risk of health effects resulting from exposure to electromagnetic fields. Installing radio base stations by the mobile operators in the area has increased the perception of risk.

Moreover, among the main causes of the widespread fear is caused by the spread of new and unconfirmed results of studies whose scientific validity is often questionable. The persistence of this situation makes imperative the need of correct information to the general public in taking care especially the scientific one.

This work will show how a probabilistic Text Analysis Tool, based on an extension of the topic model [1], [3], is able to create and to manipulate automatically ontologies from a corpora of documents extracted from the WHO (World Health Organization) one of the most important institution of electromagnetic risk communication. In this way, we can improve the quality and/or the objectivity of the interpretation of those documents and explore all occurrences of words, as already someone has made previously [4][5][6]. More details will be explained in the following sections together with experiments and charts to illustrate the goodness of this methodology.

2. Ontology Builder for TA

The *Ontology builder* is an automatic tool for construction of ontology based on the extension of the probabilistic topic model introduced in [1] and [2]. This method has been deeply illustrated in [3], next we will show the main idea behind it.

The original theory mainly asserts a semantic representation in which word meanings are represented in terms of a set of probabilistic topics z_i where the statistically independence among words w_i and the ``Bags of Words" (BoW) assumptions were made. The BoW assumption claims that a document can be considered as a feature vector where each element in the vector indicates the presence (or absence) of a word, where information on the position of that word within the document is completely lost. This model is generative and it allows us to solve several problems, including the word association problem, which is a fundamental for the automatic ontology building method. Such a problem was studied for demonstrating what is the role that the associative semantic structure of words plays in episodic memory. In the topic model, word association can be thought of as a problem of prediction. Given that a cue if presented, what new

978-0-7695-3872-3/09 $26.00 © 2009 IEEE
DOI 10.1109/ISDA.2009.26

IEEE computer society

words might occur next in that context? By analyzing those associations we can infer semantic relations among words, moreover by applying this method for automatic interpretation of a document, we can infer all the semantic relations among words contained in that document, as a result we could have a new representation of that document: what we call ontology. Assume we will write *P(z)* for the distribution over topics *z* in particular document and *P(w|z)* for the probability distribution over word *w* given topic *z*.
Each word w_i in a document (where the index refers to *i*th word token) is generated by first sampling a topic from the topic distribution, then choosing a word from the topic-word distribution. We write $P(z_i=j)$ as the probability that the *j*th topic was sampled for the *i*th word token, that indicates which topics are important for a particular document. More, we write $P(w_i \mid z_i = j)$ as the probability of word w_i under topic *j*, that indicates which words are important for which topic. The model specifies the following distribution over words within a document,

$$P(w_i) = \sum_{k=1}^{T} P(w_i \mid z_i = k) P(z_i = k), \quad (1)$$

where *T* is the number of Topics. In through the *topic model* we can build consistent relations between words measuring their degree of dependence, formally by computing joint probability between words,

$$P(w_i, w_j) = \sum_{k=1}^{T} P(w_i \mid z_i = k) P(w_j \mid z_j = k). \quad (2)$$

In this model, the multinomial distribution is drawn from a Dirichlet distribution, a standard probability distribution over multinomial. The results of LDA algorithm [2], obtained by running Gibbs sampling, are two matrixes:

1. The words-topics matrix Φ: it contains the probability that word *w* is assigned to topic *j*;
2. The topics-documents matrix Θ: contains the probability that a topic *j* is assigned to some word token within a document.

By comparing joint probability with probability of each random variable we can establishes how much two variables (words) are statistically dependent, in facts the hardness of such statistical dependence increases as mutual information measure increases, namely,

$$\rho(w_i, w_j) = \log\left| P(w_i, w_j) - P(w_i) P(w_j) \right|, \quad (3)$$

where $\rho \in [0,1]$, after a normalization procedure.
By selecting hard connections among existing all, for instance choosing a threshold for the mutual information measure, a graph for the words can be delivered. As a consequence, an ontology can be considered as set of pair of words each of them having its mutual informational value, see Figure 2.

3.1. Procedure for corpus analysis

Before start the real procedure we run the topic model on a set of corpora, all of those are on same topic, in order to learn the Φ and the Θ matrixes. Specifically we considered the following corpora:

1. WHO United Nations system;
2. COST European Cooperation in the field of Scientific and Technical Research;
3. EU European Union;
4. ICNIRP International Commission Non Ionizing Radiation Protection;
5. IEE Institution of Electrical Engineers – UK;

where the total number of documents is 23. After we decided to focus our studies on the WHO corpus, then we used those matrixes for building all the ontologies for this corpus. In the following we show how we analyze this corpus. The various steps to implement this process are given below:

I. Building the Ontology of the corpus with a certain, low, number of words and then by using a certain threshold;
II. As in the previous point but considering an average number of words;
III. Representation of the body through histograms generated according to the occurrence/frequency of words/tokens found in it;
IV. Splitting each bar of the histogram in a separate *n* sub-corpus so we could have *n* different *BoW (*the number *n* depends on how the words are distributed in the corpus);
V. Filtering range containing a number of words according to a given threshold.
VI. Ontology building for each *BoW* by using our method.

In our experiments, the previous algorithm is repeated twice with different set of threshold as indicated in point **V**: we first filter the range containing

a number of words less than 2 and next the range containing a number of words in excess of 5.

The reason why we filtered some range/bar of words (which are less than 2 and more then 5) is that, their presence may disturb the analysis of other frequency components. Of course to complete exhibition and then draw the appropriate considerations, the following will be given the words excluded from filtering. Before starting to present graphs we need to spent some words for the parameters which we used to tune the model. The two parameters are the *Words-Threshold* and *JointMatrix-Threshold*. The first threshold binds the process of generation of ontologies to consider only words with occurrence frequency above that value. The second is the ρ threshold discussed above.

4. Who is WHO?

WHO is the directing and coordinating authority for health within the United Nations system. It is responsible for providing leadership on global health matters, shaping the health research agenda, setting norms and standards, articulating evidence-based policy options, providing technical support to countries and monitoring and assessing health trends.

We selected 11 documents from collections of texts concerning this institutional and international organization indicating below a detailed list:

1. Establishing a dialogue on risk from electromagnetic fields
2. Electromagnetic fields and public health: cautionary policies
3. The international EMF Project. Fact sheet no. 181 - May 1998
4. Physical Properties and Effects on Biolog Systems. Fact sheet no. 182 - May 1998
5. Health effects of radiofrequency fields. Fact sheet no. 183 - May 1998
6. Public perception of EMF Risk. Fact sheet no. 184 - May 1998
7. Mobile telephones and their base stations. Fact sheet no. 193 - June 2000
8. Extremely low frequency fields and cancer. Fact sheet no. 263 - October 2001
9. Effects of EMF on the Environment - February 2005
10. Electromagnetic Hypersensibility. Fact sheet no. 296 - December 2005
11. Base stations and wireless technologies. Fact sheet no. 304 - May 2006

For better reflection of the reader, in *Figure 1* we show the histogram of that corpus, where we have in *x*-axis the occurrence of the words and on the *y*-axis we have the number of words of specific occurrence value.

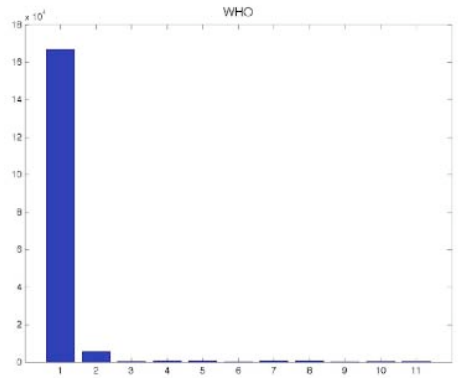
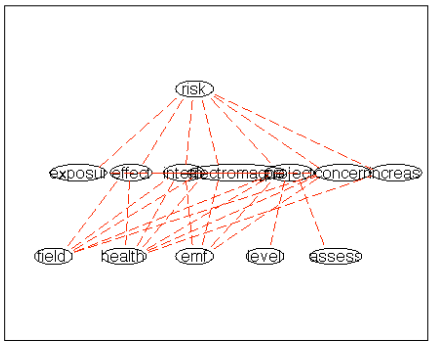


Figure 1. WHO Corpus - Number of words per documents.

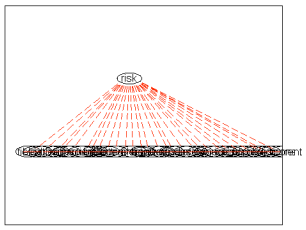


# token = 2322	*risk*	*assess*
Word-Th = 1	*field*	*project*
JM-Th = 0.593	*health*	*concern*
	exposur	*increas*
	emf	*level*
	effect	*intern*
	electromagnet	

Figure 2. First complete ontology WHO's Corpora with thirteen words

4.1. Analysis of the WHO corpus

Considering all the documents made available by this institution, we reported in Figure 1 the histogram, that is the corpus frequency spectrum. The first step of our procedure is represented in Figure 2 and 3, those figures are obtained by modulating appropriately the two thresholds. Number of token represents the number of different "forms" found in the documents.



# token = 2322	*risk*	*use*
Word-Th = 1	*field*	*studi*
JM-Th = 0.59735	*health*	*research*
	exposur	*assess*
base	*emf*	*low*
inform	*effect*	*project*
intern	*communic*	*concern*
electromagnet	*scientif*	*includ*
electr	*provid*	*possibl*
level	*sourc*	*protect*
frequenc	*respons*	*current*

Figure 3. Second complete ontology WHO's Corpora with twenty-nine words

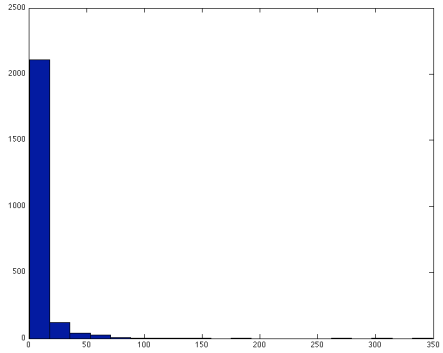


Figure 4. Words - Occurrences Distribution before filtering

In Figure 4 and 5 we can see the distributions of words before and after the two filtering step respectively.

It should be noted, moreover, how the intervals of analysis (showed in *Table 1*) become more narrowed (showed in *Table 2*) and therefore it lends us to a more detailed evaluation.

It is also useful to show the words eliminated by the filtering processes because of the leading contributions of the high frequency information, especially for the first filter (*Table 3*).

Following the procedure described above, in *Figure 6, 7,8* we show the ontology for each bar of the histogram reported in *Figure 5*.

The division of the corpus in bars, filtering and subsequently aggregating *BoW* of different sizes will inevitably lead to contain a few number of words (into histogram's tail) for which it is clearly inappropriate to create one ontology. In the *Table 4* e *Table 5* we show directly the contents of the four most important bars.

	Bar range's	# Token4Bar
I	range [1 18.4]	# token 2110
II	range [18.4 35.8]	# token 119
III	range [35.8 53.2]	# token 42
IV	range [53.2 70.6]	# token 27
V	range [70.6 88]	# token 8
VI	range [88 105.4]	# token 4
VII	range [105.4 122.8]	# token 3
VIII	range [122.8 140.2]	# token 2
IX	range [140.2 157.6]	# token 1
X	range [157.6 175]	# token 0
XI	range [175 192.4]	# token 1
XII	range [192.4 209.8]	# token 0
XIII	range [209.8 227.2]	# token 0
XIV	range [227.2 244.6]	# token 0
XV	range [244.6 262]	# token 0
XVI	range [262 279.4]	# token 1
XVII	range [279.4 296.8]	# token 0
XVIII	range [296.8 314.2]	# token 3
XIX	range [314.2 331.6]	# token 0
XX	range [331.6 349]	# token 1

Table 1. Tokens distribution before filtering in WHO's Corpora

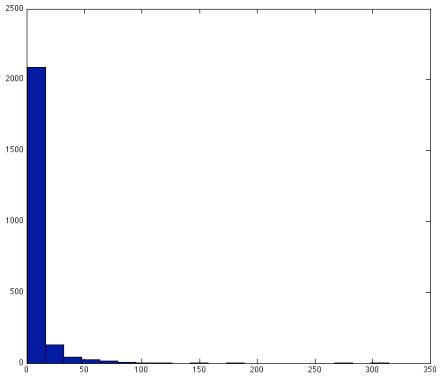


Figure 5. Words - Occurrences Distribution after filtering (six bars)

From the analysis of a single bar it is possible to obtain information about how the organization pay attention to some keywords. For instance we can note that words like "*medic*" is in the first bar and many *meaningful* keywords are contained in the 3th bar too as well (Table 4). Following the word "cancer", as showed in *Table 5* we can also find the links to others *meaningful* terms as "phone", "power" or "human". By looking at this kind of analysis we can argue that the proposed methodology can easily and automatically, highlight

meaningful concepts at different level of "*information* and it can support users to better analyze the real meaning carried out through a document. In the specific corpus we noticed that some important groups of words (in other terms concepts should carry out information about electromagnetic risk) occur in low levels (bars) of the spectrum.

	Bar range's	# Token4Bar
I	range [1 16.65]	# token 2086
II	range [16.65 32.3]	# token 131
III	range [32.3 47.95]	# token 45
IV	range [47.95 63.6]	# token 25
V	range [63.6 79.25]	# token 15
VI	range [79.25 94.9]	# token 5
VII	range [94.9 110.55]	# token 4
VIII	range [110.55 126.2]	# token 4
IX	range [126.2 141.85]	# token 0
X	range [141.85 157.5]	# token 1
XI	range [157.5 173.15]	# token 0
XII	range [173.15 188.8]	# token 1
XIII	range [188.8 204.45]	# token 0
XIV	range [204.45 220.1]	# token 0
XV	range [220.1 235.75]	# token 0
XVI	range [235.75 251.4]	# token 0
XVII	range [251.4 267.05]	# token 0
XVIII	range [267.05 282.7]	# token 1
XIX	range [282.7 298.35]	# token 0
XX	range [298.35 314]	# token 3

Table 2. Tokens distribution after filtering in WHO's Corpora

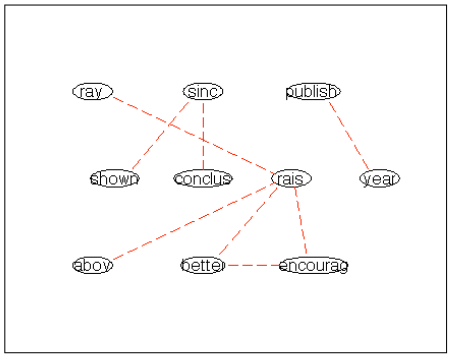
First Filter	Second Filter	
XX range:	VII range:	XII range:
risk	*electromagnet*	*effect*
	electr	XVIII
	mobil	range:
	level	*emf*
	VIII range:	XX range:
	base	*exposur*
	scientif	*Field*
	inform	*healt*
	intern	
	X range:	
	comunic	

Table 3. Words removed from filters

5. Conclusion and future work

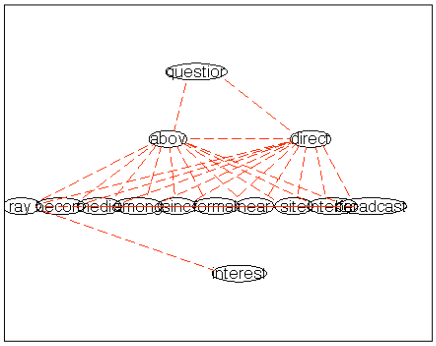
In this work we have adopted a methodology introduced in [1][2][3] to extract a synthetic and comprehensive representation of a corpus of documents taken from the WHO. We have shown, moreover, the ontologies representing these documents, taking care to extract the useful terms across the spectrum of available frequencies and we have made the appropriate considerations on communication skills of these documents.

Future work, in this direction, will characterize the documents presented by the most prestigious international organizations in the health's field with the methodology introduced in this paper so we can compare the information extracted and then we can make appropriate considerations.



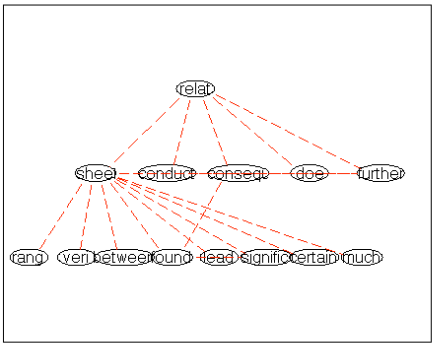
# token = 2086	*ray*	*conclus*
Word-Th = 1	*sinc*	*publish*
JM-Th = 0.586	*abov*	*rais*
	better	*year*
	encourag	*show*

Figure 6. First Bar's ontology WHO's Corpora with ten words



# token = 2086	*question*	*interfer*
Word-Th = 15	*ray*	*broadcast*
JM-Th = 0.48	*becom*	*abov*
	medic	*direct*
	among	*near*
	sinc	*site*
	formal	*interest*

Figure 7. First Bar's ontology WHO's Corpora with fourteen words



# token = 131 Word-Th = 15 JM-Th = 0.535	*relat* *sheet* *rang* *veri* *between* *conduct* *found*	*further* *signific* *certain* *much* *consequ* *doe* *lead*

Figure 8. Second Bar's ontology WHO's Corpora with fourteen words

3rd Bar			4th Bar
# token = 45			# token=25
time	*take*	*system*	*differ*
result	*ioniz*	*reduc*	*measur*
line	*biolog*	*manag*	*polici*
involv	*factor*	*communiti*	*increas*
standard	*precautionari*	*part*	*evid*
telephon	*govern*	*fact*	*radiat*
icnirp	*elf*	*evalu*	*percept*
environ	*uncertainti*	*radio*	*potenti*
high	*energi*	*associ*	*sourc*
group	*world*	*symptom*	*current*
general	*perceiv*	*make*	*advers*
caus	*report*	*one*	*stakehold*
hazard	*import*	*well*	*organ*
review	*present*	*industri*	*respons*
number	*develop*	*situat*	*issu*
			nation
			protect
			some
			technolog
			magnet
			establish
			provid
			need
			process
			peopl

Table 4. Bags of Word: III and IV Bar

5th Bar	6th Bar
# token = 15	# token = 5
guidelin	*use*
assess	*statio*
environment	*frequenc*
project	*research*
human	*studi*
low	
power	
individu	
possibl	
includ	
limit	
cancer	
phone	
concern	
decis	

Table 5. Bags of Word: V and VI Bar

Another different application from electromagnetic pollution could be interesting in the compression of information's field: if a big document can be represented with an ontology (simply some coupled word), we could think of storing only a few but meaningful words instead of the paper with considerable advantages of memory and time for retrieving it.

6. References

[1] T. L. Griffiths, M. Steyvers, J. B. T., 2007. "Topics in semantic representation", *Psychological Review*, 2007, 114 (2), 211–244

[2] Blei, D. M., Ng, A. Y., Jordan, M. I., "Latent dirichlet allocation". *Journal of Machine Learning Research 3*, 2003, (993–1022)

[3] Colace F., De Santo M., Napoletano P. (2008) A Note on Methodology for Designing Ontology Management Systems, *Symbiotic Relationships between Semantic Web and Knowledge Engineering*, AAAI Press

[4] Bisceglia, B., Valbonesi, S. (2008) People exposure to EMF in Italy. Monitoring network, risk perception and communication process. *30th Bioelectromagnetics Society Annual Meeting* June 8-12, San Diego, CA 480-482

[5] Bisceglia, B., Boumis, M. (2004) The environmental impact of RF electromagnetic fields in Italy: management of scientific and social aspects. *26th Bioelectromagnetics Society Annual Meeting* June 20 – 24, Washington, DC 288

[6] Bisceglia, B., Valbonesi, S. (2008) Analysis of textual data in significant documents. Some important words about electromagnetic pollution. *30th Bioelectromagnetics Society Annual Meeting* June 8-12, San Diego, CA 483-485