A Complex Mining Process about Air Quality

Mihaela Juganaru-Mathieu[1] **, Silvia González Brambila**[2]

[1]Ecole Nationale Supérieure des Mines, Saint Etienne, France
[2]Universidad Autónoma Metropolitana, Azcapotzalco, México D.F, México
mathieu@emse.fr, sgb@correo.azc.uam.mx

Abstract

In this paper we present a mining project about extracting knowledge from public documents concerning air pollution. Our collection contains annual reports about air quality, acid rains, climatological conditions in the large area of Mexico City. These reports contain reliable data and are generated by the Department of Environment, they are in a printable format (.pdf file) with number of pages, table of content, textual information, numerical information in tables, images. For a human being it is impossible to read the whole collection during a relatively short period (a few days or weeks) and understand the content of them. An automatic box of tools able to extract knowledge, to quick retrieve important term, to answer some exact questions about precise climate parameters would be an important help for lecturers.

We will describe our project based upon a text and data mining process; the aims of the complex process are extract frequent temporal pattern, to extract association rules, to integrate also some information retrieval simple tools. In parallel, some data mining techniques will be used to detect the same types of data presented in every report and then to extract a numerical datamart containing climatological data structured by month, year, geographical area. The datamart will be analyzed also. The main steps of our mining process are: preparing documents (cleaning, removing images, table of contents, footnotes), transforming in structured document (in a XML format with a precise DTD), indexing, various algorithms and methods of mining, visualising results and validating knowledge.

We think also that our methodology will concern also other collections of the same category : reliable data and informations presented in huge periodical reports.

Keywords : Knowledge Discovery, Text Mining, Data Mining, XML format

1. Introduction

Today many public administrations and gouvernements of all the countries are doing an important effort to inform the citizens about a large range of society subjects as health, infrastructure, education, pollution. These administrations and gouvernements are often publishing periodical reports about interest subjects. All the informations have a high quality and are published on the Web in a printable/readable format. Very often the same information (for example, the numbers of some diseases in a geographical area, information structured by age and gender of patients) is presented by month, trimester or year and in a recurrent manner (the lecturer will found the informations about 2005, 2006 and so one in every different report).

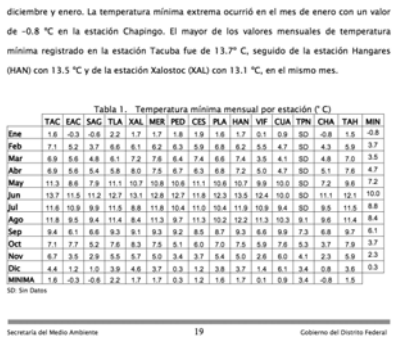
For an human who is not a specialist it becomes impossible to read all the reports about a time period (10 years, for example) and also to be able to extract himself the knowledge (for example, during the last x years the quality of y expressed by the parameter z was constantly increasing). Even a specialist of the domain (health, transportation, pollution) have to do an enormous effort to correct understand all the information and to extract correct knowledge. For the two categories of lecturers, specialist and nonspecialist, a computer aid would be very helpful.

Our project was motivated firstly by founding a complete and very well done collection of annual reports about the pollution, climate and air quality in Mexico City and secondly by some "subjective" observations : the air quality is better now than 10-15 years ago and there is not any acid rain.

So the main idea of this project is to develop a complete knowledge extraction and data mining process which is applied to this collection in the aim to offer for all lecturers powerful tools to explore the documents of the collection and the extracted knowledge and information. A first version of this project was presented in [9].

Some related works about mining in air pollution area are brievely presented in the following section. In section 3 we will present the collection of documents and its properties; section 4 contains the description of the whole knowledge extraction application which is based on data and text mining, the application is

composed mainly of three modules : integration and indexing, mining, visualization - information retrieval. Section 5 is dedicated for some important details about the cleaning, integration and indexing process, because during this iterative process the used indexes are fulfilled and the two other modules will use these indexes and by a feedback the mining processes will extract useful data integrated in indexes too. In the last section we will finish with the perspectives of our project and with a conclusion about the possible generalization of our work to other collections of printable documents in other domain.



diciembre y enero. La temperatura mínima extrema ocurrió en el mes de enero con un valor de -0.8 °C en la estación Chapingo. El mayor de los valores mensuales de temperatura mínima registrado en la estación Tacuba fue de 13.7° C, seguido de la estación Hangares (HAN) con 13.5 °C y de la estación Xalostoc (XAL) con 13.1 °C, en el mismo mes.

Tabla 1. Temperatura mínima mensual por estación (° C)

	TAC	EAC	SAG	TLA	XAL	MER	PED	CES	PLA	HAN	VIF	CUA	TPN	CHA	TAH	MIN
Ene	1.6	-0.3	-0.6	2.2	1.7	1.7	1.8	1.9	1.6	1.7	0.1	0.9	SD	-0.8	1.5	-0.8
Feb	7.1	5.2	3.7	6.6	6.1	6.2	6.3	5.9	6.8	6.2	5.5	4.7	SD	4.3	5.9	3.7
Mar	6.9	5.6	4.8	6.1	7.2	7.6	6.4	7.4	6.6	7.4	3.5	4.1	SD	4.8	7.0	3.5
Abr	6.9	5.6	5.4	5.8	8.0	7.5	6.7	6.3	6.8	7.2	5.0	4.7	SD	5.1	7.6	4.7
May	11.3	8.6	7.9	11.1	10.7	10.8	10.6	11.1	10.6	10.7	9.9	10.0	SD	7.2	9.6	7.2
Jun	13.7	11.5	11.2	12.7	13.1	12.8	12.7	11.8	12.3	13.5	12.4	10.0	SD	11.1	12.1	10.0
Jul	11.6	10.9	9.9	11.5	8.8	11.8	10.4	11.0	10.4	11.9	10.9	9.4	SD	9.5	11.5	8.8
Ago	11.8	9.5	9.4	11.4	8.4	11.3	9.7	11.3	10.2	12.2	11.3	10.3	9.1	9.6	11.4	8.4
Sep	9.4	6.1	6.6	9.3	9.1	9.3	9.2	8.5	8.7	9.3	8.6	9.9	7.3	6.8	9.7	6.1
Oct	7.1	7.7	5.2	7.6	8.3	7.5	5.1	6.0	7.0	7.5	5.9	7.6	5.3	3.7	7.9	3.7
Nov	6.7	3.5	2.9	5.5	5.7	5.0	3.4	3.7	5.4	5.0	2.6	6.0	4.1	2.3	5.9	2.3
Dic	4.4	1.2	1.0	3.9	4.6	3.7	0.3	1.2	3.8	3.7	1.4	6.1	3.4	0.8	3.6	0.3
MINIMA	1.6	-0.3	-0.6	2.2	1.7	1.7	0.3	1.2	1.6	1.7	0.1	0.9	3.4	-0.8	1.5	

SD: Sin Datos

Secretaría del Medio Ambiente 19 Gobierno del Distrito Federal

Fig. 1: A small piece of an air quality report : we can see textual information with a lot of abbreviations and a table

2. Related works

Around the world various scientific projects have been done about mining in the air pollution domain. All of these works treat data and not text and very often the mining process comes after other types of treatments.

Richards, Ma and al. in [15] and in [11] have presented a long and ambitious project in London, their project is based on a high throughput sensors distributed architecture and high grid computing abilities. Their first objective was to generate a real-time monitoring and mapping of the whole city of London, the second objective was to analyze and to mine the collected data sets. The mining processes used SOM and hierarchical methods, even distributed hierarchical algorithms. Lia and Shue in [10] had presented a complete application used in Taiwan and the most analyzed indicator was suspended particulate PM10. The clustering methods and SOM were done and the clusters obtained were useful for gouvernemental decisions.

Also air pollution data set are used as validation data set for innovant data mining techniques. Sahu and Baker [16] had worked on ozone pollution in New York state and had implemented a Bayesian model. Also Temiyasathit and al. [18] had used a collection on air quality in Dallas Fort Worth (DFW) area in a aim to realize multiscale and functional data analysis and also have purposed a spatial predictor based on kriging. Soft association rules were obtained from a quality air data set from Kuala Lumpur in [8].

We can see that all these mining works were done on detailed data sets, containing all or part of the evolution of the main polluants as carbon monoxide (CO), lead, nitrogen dioxide (NO_2), ozone (O_3). Also sometime the concentration of carbon dioxide (CO_2) are took into account. The mining process has broken predictive results, descriptions (as the daily highest pollutions in the air at London in [15]) or association rules indicating when it possible to arise is some risks.

3. Collection

The collection that we are interested in is free and accessible on the Web site of Dirección de Monitoreo Atmosférico[1] which have to supervise the metropolitan area and the Mexico Valley[2]. The collection contains exclusively annual reports and these reports are classified in three categories:

- air quality (1994–2011)
- acid rains (1994–1999)
- climatological (2001-2006)

For the previous information there is also the global report on air quality for the time period 1986–1992.

Every document can be extract in a printable format (.pdf file) and it contains a table of contents, figures, different colours of text, bibliography, large textual parts, tables with numerical values (see the figures 1, 3, 4). The size of every document is from 1Mbyte to 5 Mbyte and the number of pages are from 20 to 45. The size of the whole collection is around 80 Mbyte.

4. Complete Process of Knowledge Extraction and Mining

Our project is now in work and it is composed of three main parts:

- integration and indexing
- mining processes
- retrieval and visualization information and knowledge

The figure 2 shows the structure of our final knowledge extraction and data mining application. Its structure

[1] a service of the Department of Environment in charge with atmospheric monitoring

[2] The Web site is : `http://www.calidadaire.df.gob.mx/calidadaire/index.php?opcion=2&opcioninfoproductos=12`

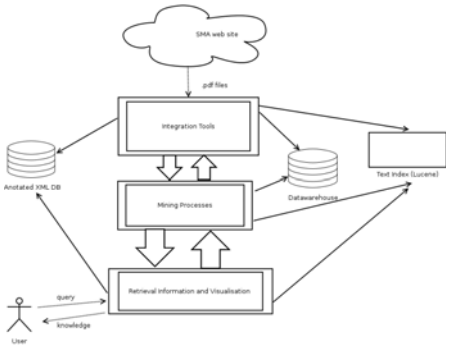


Fig. 2: The structure of our knowledge extraction and mining application

is modular and each module is able to connect directly with the previous modules and with the databases and the indexes.

The initial format of the documents is not adapted for our task, so we converted .pdf format into .ebook format and than with XSLT processor in a more useful .XML format.

The documents collected will be transformed in various manner and three indexes will be fulfilled:

- an XML database (implemented with eXist[3]) and an upper built mechanism allowing annotation
- a textual index (implemented with Lucene[4])
- a datawarehouse which contains the almost numerical values extract from document contents.

Various methods of XML mining [5], text mining and data mining will be implemented. We hope obtain frequent pattern, clusters of parts of documents, clusters of the datas, frequent fuzzy associations rules. Some methods of mining will be also used to be able to find and to extract interesting data or the localization of interesting images.

The last module "Visualization and information retrieval" is design to offer to the user the possibility to much more understand the knowledge extracted, to easy navigate into the information presented in the whole collection, and also to retrieve documents (one or more) with a precise information like "the mean Pb concentration in Azcapotzalco area in 2010" and also to extract the aggregated information like "variation of Pb concentration during the last 6 years".

5. Integration and indexing documents

After downloading all the documents, we have to covert all of them into a XML format using conversion software (like Adobe) and than to construct a XSLT processor [7] to transform XML ebook format into an other one more appropriate and able to differentiate the structure of the document. The paper [4] who describe a .pdf to .xml conversion was very helpful. The choice of XML [1] as central format of document representation is justified by the large capacities to present the XML document and parts of it, to preserve the structure of initial document, to allow easily future changes.

On the other hand the rich structure of the documents must be preserved because it is interesting to exploit for the text mining process [2].

An other exploration of the collection is to realize **xdiff**-like [19] comparison between paires of document in the aim to find common parts like bibliography, acknowledgements, technical description of physical mesures. These common parts will be indicated with annotations inside .xml document and during visualization these parts will be indicated with a different style.

The indexing part will be for :

- textual content and this step will be done with Lucene [12]
- numerical values and all numerical values will be stored in DBMS and a great part in datamarts with a model of hypercube [6]
- XML documents and their successive versions will be stored en a XML database (eXist).

The textual indexation is the first one of the indexing step. We use TreeTagger [17] for stemming, but we deplore the absence of an ontology about the environnement and, more particular, an ontology or some other knowledge representations about air pollution [13] for Spanish.

A problem we have met was the large number of abbreviations (see the figure 1 for example) The extraction of all the abbreviations will be done semi manual and using text mining like in [3] and [14].

We also planned to rapidly integrate all the `.xml` versions of document (initial version, version with annotation for images localization or for abbreviation changes) into eXist XML database. We will also do a large tool set to annotate the documents, to extract some interesting parts, to display friendly various parts of documents, these tools will be realized using XPath, XQuery and XSLT [7]. This database will be used for the XML mining step and also the database may be used by the visualization and information retrieval module.

In the aim to extract numerical values, we also need to detect the context of a value. If the datas are in a table like in figures 1 and 3 we have to translate the abbreviations, to fix the year of publication and to explore the caption of the figure or the text around the

[3] `http://exist-db.org/exist/`
[4] `http://lucene.apache.org`

Entidad	Estación	Clave	Horas que exceden la NOM de 1 h (NOM=0.110 ppm)	Indicador 8 h (NOM= 0.080 ppm, quinto máximo)	Cumple la NOM
Distrito Federal	Lagunilla	LAG	80	0.096	No
	Tacuba	TAC	188	0.118	No
	Azcapotzalco	AZC	145	0.116	No
	Merced	MER	95	0.099	No
	Pedregal	PED	291	0.126	No
	Cerro de la Estrella	CES	44	0.093	No
	Plateros	PLA	187	0.112	No
	UAM Iztapalapa	UIZ	149	0.112	No
	Taxqueña	TAX	15	—	Datos insuficientes
	Cuajimalpa	CUA	93	0.108	No
	Tlalpan	TPN	162	0.115	No
	Tláhuac	TAH	130	0.110	No
	Santa Úrsula	SUR	211	—	Datos insuficientes
	Coyoacán	COY	236	0.118	No

Fig. 3: A table inside a report with numerical values. A such table is present with actualized information in every annual report

reference to this table. We have also to use a data mining for a pattern search for detecting parameter tables which repeat in every year report. We have the objective to explore (extract) at least 75% of numerical values. These values will be saved in a relational database and a set of them will be organized as datamart in hypercube organized by month, year, area, parameter [6]. A similar approach as in [8] about extracting some association rules can be also possible, but we need the aid of an expert able to indicate as some thresholds of parameters like $CO_2 \geq 0.02$ or $O_3 \geq 0.007$. Our complete collection contain the values some polluants that are aggregate by month and by geographical area and also some climatic informations as temperature, wind, humidity.

An other type of information we can indexe is images. For an user it would be useful to see the carts like in the figure 4 for every year and to present the same information for a large time period. For this step we also need to do a step of text mining to discover the context.

This module must be able too to integrate and indexe the most recent published report without reconsider the whole collection.

6. Future work and conclusion

Our project is, surely, very ambitious. It started a few time ago and only a small part was done, this part is in the integration module. It is very possible that various unknown problems may arise and the planed methods may not be the best one. A particular care was accorded to the first module in the aim to be able to reconsider, if need, the mining processes.

On the other hand we are thinking that in other fields a lot of collections of the same type (big technical periodical reports for the people information) are published now on the web and the methodology we

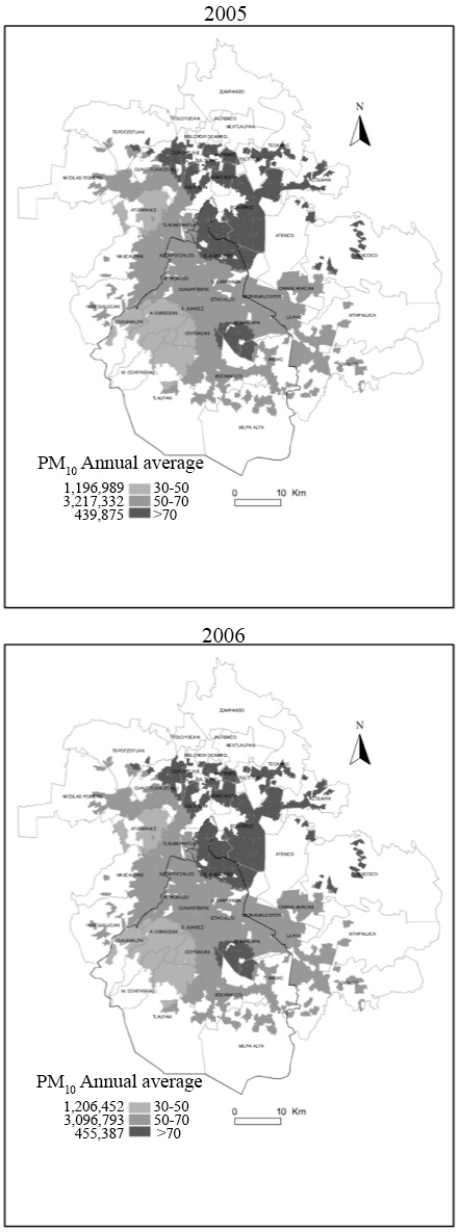


Fig. 4: Two carts presented into document with the evolution of a pollution parameter

used can be also applied in these cases.

References

[1] Serge Abiteboul, Peter Buneman, and Dan Suciu. Data on the Web: From Relations to Semistructured Data and XML. Morgan Kaufmann, 2001.

[2] Eugene Agichtein and Venkatesh Ganti. Mining reference tables for automatic text segmentation. In Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '04, pages 20–29, New York, NY, USA, 2004. ACM.

[3] Hiroko Ao and Toshihisa Takagi. Alice: An algorithm to extract abbreviations from medline. Journal of the American Medical Informatics Association, 12(5):576 – 586, 2005.

[4] Hervé Déjean and Jean-Luc Meunier. A system for converting pdf documents into structured xml format. In Horst Bunke and A.Lawrence Spitz, editors, Document Analysis Systems VII, volume 3872 of Lecture Notes in Computer Science, pages 129–140. Springer Berlin Heidelberg, 2006.

[5] Norbert Fuhr, Jaap Kamps, Mounia Lalmas, and Andrew Trotman, editors. Focused Access to XML Documents: 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2007, volume 4862. Springer, 2008.

[6] Jiawei Han and Micheline Kamber. Data mining: concepts and techniques. Morgan Kaufmann, 2006.

[7] Elliotte Rusty Harold and W. Scott Means. XML in a Nutshell, Third Edition. O'Reilly Media, 2004.

[8] Tutut Herawan and Mustafa Mat Deris. A soft set approach for association rules mining. Knowledge-Based Systems, 24:186–195, 2011.

[9] Mihaela Juganaru-Mathieu and Silvia Gonzalez Brambila. Projet d'exploration et extraction de connaissances sur la pollution de l'air depuis une collection de documents publiques. In STIC & Environnement 2011, pages 327–332, Paris, 2011. Presses de l'Ecole des mines.

[10] Sheng-Tun Lia and Li-Yen Shue. Data mining to aid policy making in air pollution management. Expert Systems with Applications, 27:331–340, 2004.

[11] Yajie Ma, Mark Richards, Moustafa Ghanem, Yike Guo, and John Hassard. Air Pollution Monitoring and Mining Based on Sensor Grid in London. Sensors, 8:3601–3623, 2008.

[12] Michael McCandless, Erik Hatcher, and Otis Gospodnetic. Lucene in Action. Second edition. Manning, 2010.

[13] Carolina Medina-Ramírez. La web semántica en el medio ambiente: necesidad de una ontología de dominio. 2009. `http://www.infotec.com.mx/work/models/infotec/`.

[14] Youngja Park and Roy J Byrd. Hybrid text mining for finding abbreviations and their definitions. In Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing, pages 126–133, 2001.

[15] M. Richards, M. Ghanem, M. Osmond, Y. Guo, and J. Hassard. Grid-based analysis of air pollution data. Ecological Modelling, (194):274–286, 2006.

[16] S.K. Sahu and K.S. Bakar. A comparison of Bayesian models for daily ozone concentration levels. Statistical Methodology, 9:144–157, 2012.

[17] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In Proceedings of International Conference on New Methods in Language Processing, volume 12, pages 44–49. Manchester, UK, 1994.

[18] Chivalai Temiyasathit, Seoung Bum Kim, and Sun-Kyoung Park. Spatial prediction of ozone concentration profiles. Computational Statistics and Data Analysis, 53:3892–3906, 2009.

[19] Yuan Wang, David J DeWitt, and J-Y Cai. X-Diff: An effective change detection algorithm for XML documents. In Proceedings. 19th International Conference on Data Engineering, 2003, pages 519–530. IEEE, 2003.