A Complex Mining Process about Air Quality
Mihaela Juganaru-Mathieu, Silvia González Brambila

To cite this version:

HAL Id: emse-00870769
https://hal-emse.ccsd.cnrs.fr/emse-00870769
Submitted on 31 Oct 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
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.

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 briefly 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

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

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 governmental decisions.

Also air pollution data set are used as validation data set for innovative 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].

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

3. Collection

The collection that we are interested in is free and accessible on the Web site of Dirección de Monitoreo Atmosférico1 which have to supervise the metropitan area and the Mexico Valley2. The collection contains exclusively annual reports and these reports are classified in three categories:

- air quality (1994–2011)
- acid rains (1994–1999)

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
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 pairs of document in the aim to find common parts like bibliography, acknowledgements, technical description of physical measures. 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/
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 data mart 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
used can be also applied in these cases.

References


