ArchiteXt Mining
—Taking advantage of Periodicals as an Architectural Data Base—

Ana ESTEBAN-MALUENDA[1]
[1] *Professor, Technical University of Madrid*

Luis SAN PABLO MORENO[2]
[2] *Data Scientist, BNP Paribas Spain*

Abstract

'ArchiteXt Mining. Spanish modern architecture through its texts (1939-1975)' is a research project funded by the Government of Spain through the 2015 Call for 'Excellence Projects' of the Ministry of Economy and Competitiveness.

This project aims to explore a new view point and look into the special features of Spanish Modern Architecture. Despite the success of the development of data analysis as a tool in different disciplines, the research on architectural theory has never made the most efficient use of these technologies. The Spanish and International circumstances of the development of the Modern Architecture has been scrutinized through qualitative research which established a shared theoretical ground. It is time to face a new in-depth research based on objective data. In order to obtain this, we propose the application of text mining techniques to take advantage of the best data source in the field: architectural magazines. Our objective is to offer a new vision of the transformation of the architectural production and how it is divulged through texts published on these magazines. In addition to this, we aim at creating a powerful database hosted in a public website for the scientific community. In this context, this project fulfills e-Research objectives: to make the computerization of research data easier, to support every stage of the data collection, and to contribute to the use of tools that allows big amount of data analysis.

As a first step, this project focuses on the Spanish case as a pilot for a bigger worldwide scale research. In particular, this stage will begin with the Spanish architecture magazines published during the Dictatorship period (1939-1975), when the cultural relations of the country at an international level were more difficult. We will be able to compare the important issues to Spanish architects with the other subjects of interest in European, American and Asian magazines, which will provide a new interpretation of the Spanish Architecture regarding the international panorama.

Keywords: Architectural Periodicals, Text Mining, Spanish Modern Architecture, Data Analysis.

Introduction

In the twenty-first century, it is scarcely possible to think about scientific research without considering the digital aspect. The world that we are living now is more and more reliant on data, in which we can find valuable information, identify patterns of behaviors and visualize information on a big scale. Some analytical tools, like Data Warehousing and Data Mining, have reached very important results in terms of the massive treatment of information.

Text Mining is a more specific technique that looks for patterns and tendencies in texts. It can discover hidden knowledge; so that it is possible to answer certain questions previously asked (descriptive models) or discover hidden patterns in a group of texts (predictive models).

These kinds of analytic processes have been largely used in many scientific and humanistic disciplines, with positive results. For example, let's think about the Jesuit priest Roberto Busa, who in 1946 began to build the Thomisticus Index; a tool for performing text searches within the Corpus of Aquinas's works.1

More recently, many other projects have been applied to Text Mining. There are well-known projects that supported the work of authorship identification.2 Especially, the analysis of William Shakespeare's texts has given remarkable results. Text Mining has demonstrated that certain texts attributed to Shakespeare, are far from resembling any other text written by him.

In architecture, there were also several pioneers in digital analysis, like Juan Pablo Bonta. He wrote the book, American Architect and Texts,3 published in 1996. In his research, Bonta worked with data cited in 380 texts about American architecture since 1815.

Contact Author 1: Ana Esteban-Maluenda, Professor,
Escuela Técnica Superior de Arquitectura,
Universidad Politécnica de Madrid, Avda. Juan de Herrera,
4. 28040-Madrid.
Tel: +34 913 364 232 Fax: +34 913 366 496
e-mail: ana.esteban.maluenda[a]upm.es

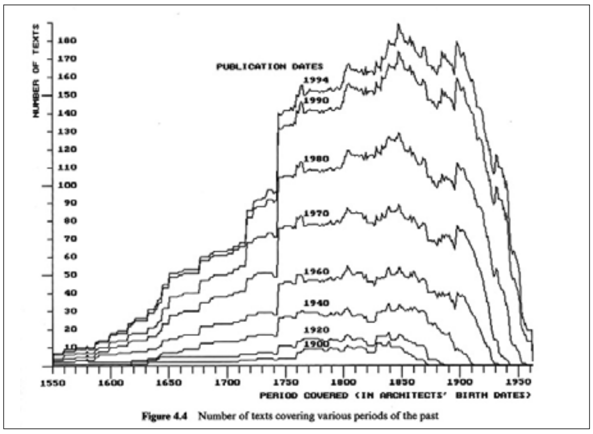

Fig. 1. Bonta studies. Number of text covering various periods of the past (Source: Bonta (1996), p. 53)

Today the vast work of Bonta that quantifies the cites of all those architects have been overtaken by tools like the Ngram Viewer by Google Books.4 Google Books allows us to visualize in a few seconds huge amounts of cites about a certain author included in books written during the nineteenth and twentieth centuries.

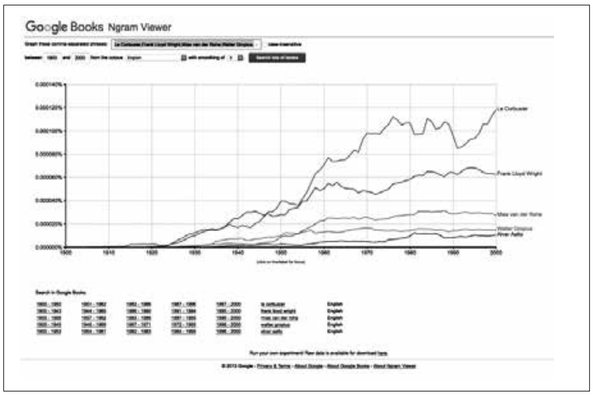

Fig. 2. Ngram Viewer. Number of cites: Le Corbusier, Frank Lloyd Wright, Mies van der Rohe, Walter Gropius and Alvar Aalto. (Source: https://books.google.com/ngrams)

With exceptions, the cites are growing with time due to the fact that nowadays the publishing market is much more developed than in the past.

So, if the development of the publishing industry ‘contaminates’ the quantifying of cites and makes this kind of analysis invalid, where can we obtain information about the evolution of Modern Architecture without accumulating these mistakes? Is it possible to use databases that can provide information about Modern Architecture at a certain point in time?

The architecture of twentieth and twenty-first centuries has an extraordinary database where the most important concepts, events and buildings have been registered: the architectural periodicals. Most researchers who are specialized in architecture use them as a source. Unfortunately, we are still doing this work the same way it has been done for the last fifty years. Meaning that we still need to go to libraries and review page by page all the issues.

In last decades, many architectural periodicals indexes have been built in order to help researchers. But those indexes are incomplete and usually they do not include records about the minor texts (i.e. news sections). The big quantity of information in periodicals hinders the researchers’ understanding. It is necessary the support from computers in order to transform this large database into a readable format which can be easily computed.

ArchiteXt Mining

The ArchiteXt Mining project (which is the acronym of Architectural Text Mining) proposes the use of advanced techniques in data analysis for building tools for researchers that uses periodicals for their work. Also, ArchiteXt Mining (AM) aims to be a collaborative tool which provides information at the same time as receiving it from users and researchers.

Another aspect to highlight is that ArchiteXt Mining is a pilot project which hopes to explore a new way of carrying out research. AM was born in Spain with the focus on Spanish architectural periodicals, but with the aspiration to grow into something larger. But before, we need to explore the tool’s possibilities in a limited area by the research team: the Spanish architectural periodicals of the 1940s, 1950s, 1960s and 1970s. AM will study the period of General Franco’s govern (1939-1975) in depth. It’s well-known that changes in Spanish architecture happened during these years giving a wide range of possibilities to explore.

At that time, the most important periodicals belonged to the Institutes of Architects of Madrid and Barcelona (Revista Nacional de Arquitectura and Cuadernos de Arquitectura). We have already scanned and digitalized both publications between 1939 and 1975. Apart from this, we intend to complete this material with other important Spanish periodicals, like Nueva Forma and Hogar y Arquitectura. The initial aim was also to scan and digitalize some European periodicals from the three most important spreading nodes of architectural news from that time: L’Architecture d’Aujourd’hui (France), The Architectural Review, Architectural Design (Great Britain), Domus and Casabella (Italy). Due to the reduced budget of the project, probably these European sources will not be included in the first phase. But, this project thinks big and wants to branch out internationally, sharing information and tendencies with other countries.

One issue that we like to highlight is that this project is not a simple magazine digitization. The digitization projects are very important since they bring the magazines closer to the researchers, but in fact they do not change the need to looking up information found in journals, page by page. We want to provide researchers with something more: a tool that not only helps them to save time, but it even serves as a source of inspiration.

In short, we propose the creation of a database on Modern Architecture published in the Spanish periodicals which will be accessible to the academic world. This will be more than the basic bibliographic information contained in the indexes which are already available.

At the same time, we will start making in-depth analysis of the contents of the articles applying the methodologies of Text Mining. We intend to establish several patterns and differences not only between the magazines, but also between different decades and between Spanish and foreign architecture. The quantitative analysis of the trends will be fundamental, as well as the location of the main nodes of reception and emission of news.

Another target of this project is to supply an objective list of texts that have set trends in Spanish architecture and those that, on the contrary, have been a mere reflection and continuation of the same. It is considered that this is a goal of great importance in the elaboration of future researches.

On the other hand, we aim to establish rankings that indicate the importance of architects, buildings, critics and a considerable number of variables of interest for the researchers.

But one of the strongest goals of this project is its potential for growth: we wish to be the starting point of a worldwide project in which Text Mining becomes a really powerful analytical tool.

Our first task is the elaboration of a biblio-thematic database. This is an initial classification according to the traditional formula. The research group covers the contents of the journals. However, it provides added value including sections that are not usually in traditional bibliographic indexes. The objective of this biblio-thematic database is to make this search easier and quicker. A web portal will allow the researchers to do cross-searching on all the articles, so that they will be able to look for certain terms simultaneously in all fields and records of the database.

Trying to understand the power of this tool, we will perform a very simple search in the two magazines of the Institute of Architects of Madrid and Barcelona: the term ‘Japan’.

The results are the following:

1. Two very brief mentions in Revista Nacional de Arquitectura in 1956 and 1957.

2. Nineteen references in Arquitectura between the years 1960 and 1969.

The first conclusion that can deduce is the very different interests of the two architectural nodes of the moment in Spain (Madrid and Barcelona). The magazine of the Institute of Architects of Barcelona did not publish anything referring to Japan during those twenty years, compared with the twenty-one mentions that were made from Madrid.

On the other hand, we could see how these articles are distributed over time: the Spanish architects became interested in Japanese architecture since the sixties, particularly since 1963. That year a Spanish architecture exhibition was shown in Tokyo and that circumstance lead the Japanese journal Kokusai Kenchiku to publish a monographic issue about Spanish architecture.

Most pages devoted to Japan in Spanish periodicals were written by Antonio Fernández Alba and Mariano Bayón; and the most popular Japanese architects were Kenzo Tange and Koji Kawashima. In terms of cities, it was Tokyo which captured the Spaniard’s attention throughout the whole decade.

In short, looking up the biblio-thematic database the researchers could have a global idea about the Spanish architects’ interest for Japan. But, in addition to the usual bibliographic data (date, title, author, journal, volume) and the thematic fields added by the research team, we aim to bring an additional value to the database: the searches by means of the study of the DNA strand of the text. We add the TMD (Text Matrix Document) to each article register, giving a statistical appearance to the traditional metadata.

For example, we can provide the most frequent words to the researcher. This is a very easy process, and it works really well, especially if those words are displayed in the form of word clouds.

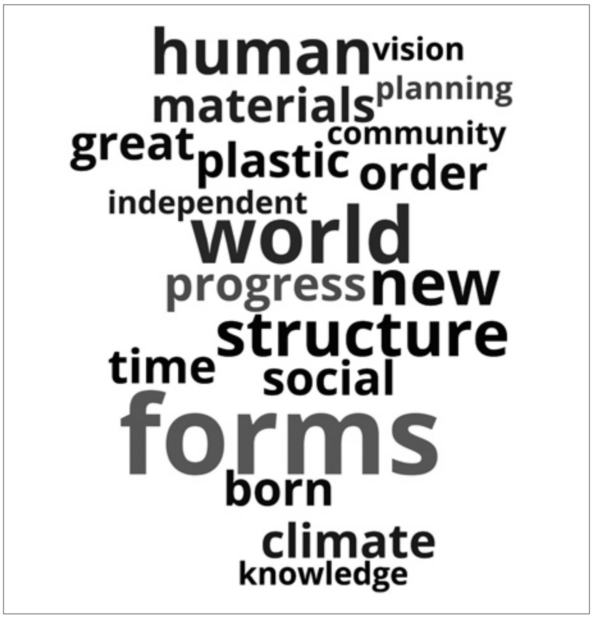

Fig. 3. Word cloud of the article: Antonio Fernández Alba, (1963) 'Kenzo Tange', Arquitectura, no. 60, December, p. 29-43. (Source: prepared by the authors)

The aim is to squeeze this information and be able to cross them between different papers obtaining new data, like term frequencies, similarity coefficients and correlational analysis.

Applications of Text Mining

Text similarity studies

Statistical methods allow us to overlap the digital footprint of two different texts if we have the corresponding TDMs from both of them stored in a database.

The information of two different texts can be crossed to obtain the similarity coefficient. Basically, is a frequency scalar product pondered from all the common words in both compared texts.

The coefficients for the articles can be calculated, the TDMs stored in our database and provide this information rapidly in real time, when the users are asking for words and specific terms on the website.

The first task is scanning them to transform the printed text into digital text.

Next, the texts are stored in a worksheet, placing each word in a cell. With this operation, we are able to work automatically with the texts and make operations with it, such as counting rows, words, aggregating values, measuring frequencies, etc.

Right after it is necessary to remove the stopwords, that is those words with no specific significance and elements that create background noise and interference in our analysis.

With the clean text, it is possible to build the TDM (Text Document Matrix) and store it in the data base, that now includes more information than the classical metadata of an article. Stored the TDM information by article, we can overlap them and make calculations. One of them is the similarity index (SIM), which provide us a quantitate value of how near two texts are in terms of significance. This indicator is directly proportional to the specific similarity between the two texts. The SIM coefficient has values between 0 (no coincidence) and 1 (total identity between texts).

Even more, it is possible to calculate - via TDM - the different SIMs of all articles stored in our database, what constitute a powerful tool to indicate how close two texts are in terms of significance.

This praxis saves researchers a lot of time, and puts them rapidly in the picture, selecting what they are only interested on.

Word correlation studies

This technique establishes word association patterns that commonly appear together and validate previous hypothesis. This skill gives the opportunity to make word associations, but much faster than we were able to in the past and giving new analytical components to our study.

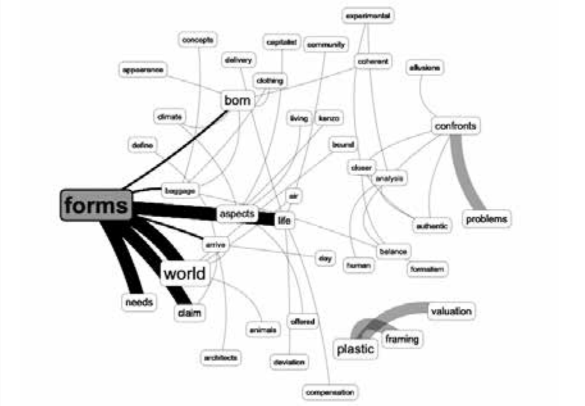

Fig. 4. Correlation study of 'forms', one of the most frequent words included in Fernández Alba (1963). (Source: prepared by the authors)

Studies of frequency

Delimiting the perimeter in advance is always beneficial. The database could be remarkable for researchers in terms of obtaining high word frequencies. With this first request, a list of principal articles can be obtained on a first approach. Completing this with coefficients of similarity - already stored in the database - the selection criteria will be stronger.

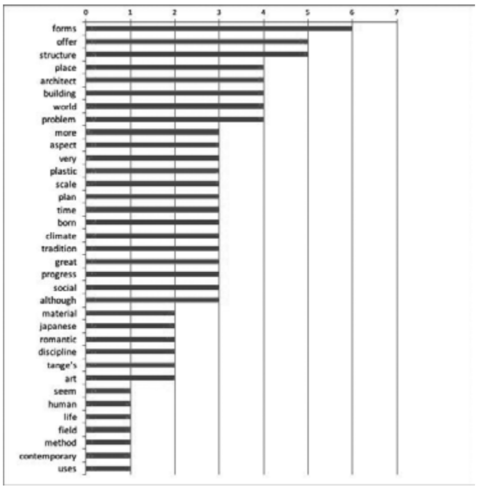

Fig. 5. Frequencies graph of the words included in Fernández Alba (1963). (Source: prepared by the authors)

Conclusions

Architectural periodicals include a great proportion of the events that occurred in the twentieth century and can therefore be considered the best database of Modern Architecture. However, the human brain cannot assimilate such information. To exploit it we need to digitalize it.

Text Mining techniques is unable to resolve the entire research process, but it can help in saving a lot of time. They can definitely contribute in locating new topics. If we add statistic treatment of information and string characters to the general data we have collected from the papers and articles and have included in a database, we can create a high-performance tool for text analysis.

ArchiteXt Mining is going to test this technology in the field of architecture but doesn't want to restrict itself to Spain or the decades of the fifties or sixties. It prefers to grow and turn into a global tool that will make easier the studies about the dissemination of the architectural trends and the exchange between countries and continents.

Acknowledgments

'ArchiteXt Mining. Spanish modern architecture through its texts (1939-1975)' HAR2015-65412-P (MINECO/ERDF) is a research project funded by the Government of Spain through the 2015 Call for 'Excellence Projects' of the Ministry of Economy and Competitiveness (MINECO) and the European Regional Development Fund (ERDF).

References

1)http://www.corpusthomisticum.org/it/index.age.

2)Philip Sallis, Subana Shanmuganathan, (2008) 'A Blended Text Mining Method for Authorship Authentication', in Analysis, Modelling & Simulation, 2008. AICMS 08. Second Asia International Conference on, IEEE Xplore Digital Library (http://ieeexplore.ieee.org/abstract/document/4530518/, search date: 03/10/2017).

3)Juan Pablo Bonta, (1996) American Architects and Texts: A Computer-Aided Analysis of the Literature, The MIT Press, Cambridge (Mass).

4)https://books.google.com/ngrams.