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

Ana ESTEBAN-MALUENDA
Technical University of Madrid
Luis SAN PABLO MORENO
ArchiteXt Mining Project (MINECO-ERDF)

Ana ESTEBAN-MALUENDA

The ArchiteXt Mining project is funded by the Spanish Government, included in the National Program of Excellence. It started in 2016 and it will continue to be developed during the next three years. Before explaining the technical details of the way the project is organized, I will present how the project was born.

In the XXI century, it is scarcely possible to think about scientific research without considering the digital aspect. The world that we are living is more and more reliant on data. We can find more and more information, identify types of behavior and visualize information in a different scale. Some analytical tools like data warehousing and data mining have reached very important results in terms of the massive treatment of the information. Text Mining is a more specific technique that looks for patterns and tendencies in texts. This technique allows us to discover hidden knowledge so it is possible to respond to questions that were previously asked or to discover hidden patterns in a group of texts. This kind of analytical processes have been largely used in many scientific and humanistic disciplines with positive results. (Fig.1)

For example, the Italian priest Roberto Busa in 1946 began to build the Index Thomisticus as a tool for performing text searches within the corpus of Thomas Aquinas works. At the beginning Busa used basic tools like punched cards to make his analyses, however thanks to the strong development of these techniques the best results of this project have been obtained later in the end of the XX century. More recently many other projects have applied text mining tools. Many works of authorship identification of literary pieces are well-known, specially the analysis of William Shakespeare works which has achieved remarkable results.

In architecture, there were also several pioneers in digital analysis, like Juan Pablo Bonta who wrote the book American Architects and Texts published in 1996. (Fig.2)

In his research, Bonta worked with data cited in 380 texts about the American architecture since 1815. Today the vast work of Bonta that quantifies the projects of all those architects have been overtaken by tools like the Ngram Viewer by Google Books, which allows to visualize in few seconds a huge amount of cites of a certain author included in books of the XIX and XX centuries. These graphs show cites that were found in books during the XX century of the five Masters of Modern Architecture: Le Corbusier, Frank Lloyd Wright, Mies van der Rohe, Walter Gropius and Alvar Aalto. As you can see, the cites are growing with the time due to the fact that nowadays the publishing market is much more developed than in the past. Therefore, it seems that this tool isn't good to measure the real interest provoked by these authors in a certain moment. 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 answer is: yes.

The architecture of the XX and XXI 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 in the same way that it has been done for the last 50 years. We still need to go to the library and review page by page all these issues. In last decades, many architectural periodicals indexes have

been built in order to help researchers, but those indexes are usually incomplete and they don't include records about the minor texts, for example, the news section. The big quantity of information in periodicals hinders the researcher's understanding. It is necessary the help of computers in order to transform this large database into a readable format which can be easily computed.

Fig.1 The Jesuit Priest Roberto Busa shows the *Index Thomisticus* in the IBM stand of the Expo 58 in Brussels.

ArchiteXt Mining proposes the use of advanced techniques in data analysis for building tools for researchers that use periodicals for their work. Also, ArchiteXt Mining 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 it is a pilot project which hopes to explore a new way of carrying out research. It 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 tools possibilities with the well known area of a recent theme, which is the Spanish Architectural Periodicals of the 1940s, 1950s, 1960s and 1970s. The changes in the Spanish Architecture that happened during these years give a wide range of possibilities to explore.

At the beginning, we have already digitalised the journals of the Institute of Architects of Madrid, *Revista Nacional de Arquitectura* and *Arquitectura*, and the one of the Institute of the Architects of Barcelona, *Cuadernos de Arquitectura*. Apart from these, it will be completed these material with other important Spanish periodicals like *Hogar y Arquitectura* and *Nueva Forma*. The initial aim was to scan and digitize some European Periodicals as *L'Architecture d'Aujourd'hui*, from France, *The Architectural Review* and *Architectural Design*, from Great Britain, and *Domus* and *Casabella*, from Italy. Due to the reduced budget of the project, these European sources probably will not be included in the first phase. But this project thinks big and wants to branch out internationally, sharing information with other countries.

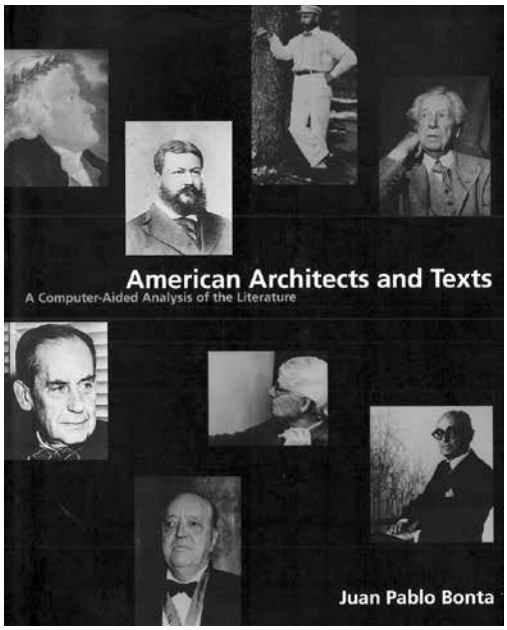

Fig.2 Cover of the book *American Architects and Texts* (1996), by Juan Pablo Bonta.

One issue that we would like to highlight in this project is that it is not a simple machine digitalization. The digitalization projects are very important since they bring the magazines closer to the researchers but in fact they do not change the need to look up information in journals page by page. Another very important aspect is the digitalization with OCR (Optical Character Recognition), as Hugo Segawa has used in the case of the *Acropole* magazine, that provide us the possibility to do searches of words or group of words. But we want to provide researchers with something more. A tool that not only helps them to save time, but it even serves them as a source of inspiration. So we proposed the creation of a database on Modern Architecture published in Spanish periodicals which will be accessible to the academic world. This will be more than the basic bibliographic information contained in the indexes that are already available. At the same time, we will start making in-depth analysis of the contents of articles applying the methodologies of Text Mining. The intention is to establish several patterns of similarities and differences not only between magazines but also between Spanish and foreign architecture. The quantitative analysis of the trends will be fundamental as well as the location of the main nodes of the reception and

admission 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 or continuation of the same. On the other hand, we aim at establishing rankings that indicate the importance of architects, buildings, critics and a considerable number of variables of interest for the researches. One of the stronger goals of this project is its potential growth. We wish to be the starting point to a worldwide project in which Text Mining becomes a really powerful analytical tool.

So our first task was the development of a biblio-thematic database. This is an initial classification according to the traditional formula done by the members of the research group. However, it provides a lot of information that is not included in traditional bibliographic indexes. Trying to explain the power of this tool, I propose you a very simple search in the contents of the journals of the Institute of Architects of Madrid - *Arquitectura* and *Revista Nacional de Arquitectura* (RNA) - and the one of the Institute of Barcelona - *Cuadernos de Arquitectura*. We can look, for example, the term "Japan" and the results are the following: 2 little mentions in RNA and 19 references in "Arquitectura". Just with the results of this search in the biblio-thematic database we can deduce very different conclusions. For example, the first one that we can deduce it's the very different interests of the two architectural notes ongoing in Spain (Madrid and Barcelona): the magazine of the Institute of Architecture of Barcelona did not publish anything referring to Japan during those 20 years, compared to those 21 mentions that were made from Madrid.

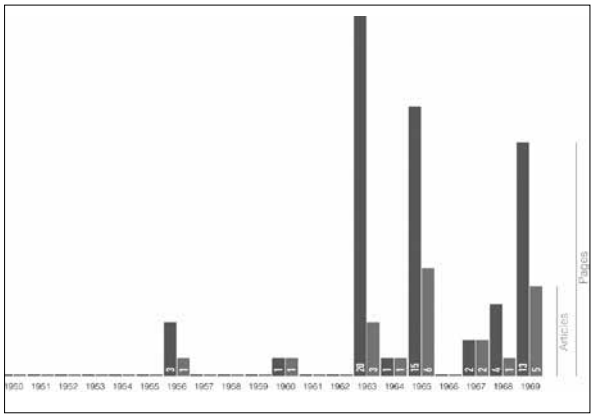

Fig.3 Distribution over time of the articles on Japan, published in the magazines of the Institutes of Architects of Madrid and Barcelona in the 1950s and 1960s.

On the other hand, we can see how articles are distributed over time and how the Spanish architects became interested in the Japanese architecture since the 1960s, particularly since 1963. And you could ask me: why? Because in that year a Spanish architectural exhibition was shown in Tokyo and that circumstance lead to the Japanese Journal *Kokusai kenchiku* to publish a monographic issue about the Spanish architecture.

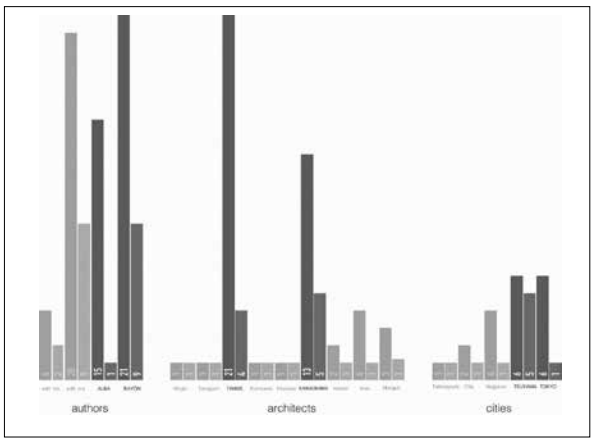

Fig.4 From left to right, graphs of the Spanish authors who wrote more articles on Japan; and the most popular Japanese architects and cites published in the magazines of the Institutes of Architects of Madrid and Barcelona in the 1950s and 1960s.

Most pages devoted to Japan in Spanish architecture 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, Tokyo and Tsuyama captured the Spanish 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 in Japan, but in addition to this reflection, this database aims to bring additional values.

Luis SAN PABLO

Our objective today is to show you how can we use the elemental unit of information in the data-

base ArchiteXt Mining, that is no more than the article.

This is the normal presentation of an article in our reviews of periodicals. We can see different pages full of texts and pictures and also with several draws, but for us this kind of information is quite complicated to process, as before prof. Segawa has already remarked. It is very complicated to us to work with these different formats of information. So we need to "clean" these articles and take only the texts to be processed automatically. For example, we have a Spanish article to be translated in English for the audience's comprehension. This is the material we are going to work with and, in a second step, we are going to put the material into a work sheet, no matter which kind of work sheet. We have to transpose the orientation of the text not because it is indispensable, but only for practical purposes. After that we see that there are certain words that are alone themselves and they don't have specifically meanings. They are not useful in text mining and they are called in data mining as "stop-words" - words that are very useful for connecting ideas and phrases in texts but for our purpose to analyze semantic significance of words are not useful, therefore we need to remove them. So this is our text when these words are removed, and as can you see it makes a lot of difference when we have a text with "stop-words". "Without stop-words" we have only names, substantives and verbs, which are the nucleus of the information. Now we can create a matrix of words with the ranking of frequencies of different words that appear in the text. On the left, we have our text and, inside the matrix, we have the words organized according to appearance frequencies: words with high frequencies in one side and, in the other side, words with low frequencies. It is not complicated to understand this organization, but probably it will be easier to understand with a graphic view the appearance of words in a text. This is the usual technic to approach texts in a digital work.

After this we would like to demonstrate three practical examples of how we can use these technologies and techniques and to start to work with examples with deeper analysis.

Text similarity studies

The first type of analysis possible to be executed is to create indexes of similarities between two texts. Here we have two different texts, both written by Spanish architects, the first was written by Antonio Fernández Alba in 1963 and the second was written by Mariano Bayón in 1968. There are five years of difference between them, and we are going to make an experiment to correlate them in terms of frequency matrix, not in terms of a text, trying to understand what it covers.

So, here we have the first frequency matrix of the text number one, that could be considered the DNA of the text, and here a matrix of the second one. If we consider only the common words of both texts and we put together only the common words in two columns, we get to this: $Sim(d1,d2) = x1y1+ \ldots +xNyN = \sum Ni=1\ xiyi$

This mathematical expression means that we have the same words in both texts and it shows the differences of frequency appearance of these words in text number 1 (T-1) and text number 2 (T-2). For example, the word "architecture" appears seven times in T-1 and four times in T-2. So we would like to propose you an index that could measure - in an objective way - how these texts are close in terms of significance. We are going to make a scalar product, a pondered scalar product of frequencies of many terms, as common words we have in our texts. For example, we would like to multiply for each different common word the frequencies of T-1 with the frequencies of T-2, obtaining a sort of terms that finally will be summarized for the analysis, in order to obtain an index like this: 0,436. If we express it in a percentage format, it means 43,6% of similarity. This is an objective criteria for researchers that we can provide them. And with these indexes they could have some criteria before reading if one text is quite similar to other. This is very important to save time. It is not the same to write a doctoral thesis in 10 years than writing it in 5 years. We will provide these tools for investigators and researchers to do similarity searches on all texts that we have stored in our database. This probably will become a powerful tool for them. Our aim is to give this information as fast as possible. Thus, we will calculate in advance the combinations of all the articles stored in the database two by two and obtain an index for any combination of text, in order to make it available to researchers and users under demand. Regardless of any combination users want to do, the fast output will be possible because these indexes will be calculated in advance.

Word correlations' studies

With word correlation analysis, it is possible to know how words are presented in a text together

with others. Usually, words show tendencies of how they can be grouped with others in a certain way. For example, in our texts the word "form" usually appears near to the word "world", "claim", "needs", "aspects" or "life". This graph shows the strength of relations between words and provide us a powerful tool to know concepts that are related together. In the second test, we have the word "urban", which is usually presented together with the words "studies", "terms", "structure" or "approach". For example, the year "1960" is clearly related closely to the word "urban". As I am not a specialist in the history of architecture, I don't know why it is. But with this clue, I can search and find that in 1960 Kenzo Tange proposed a large urbanistic scheme on Tokyo Bay. Thus, even for a non-specialist like myself, it becomes possible to realize that these two words are grouped together and make easy and clear to understand the relations that exist between words.

Studies of frequencies

There are different ways to present ideas of an article. For example, on the left we have the T-1 word-cloud and on the right the T-2 word-cloud.

Fig.5 Word-clouds that show the most common words in T-1 (Fernández Alba, 1963) and T-2 (Mariano Bayón, 1968).

In the first one, We can see the most common words in T-1 in larger fonts: "forms", "structure", "plastic" and "material". Thus, in 1963 Fernández Alba gave us the physical component of the architecture of Kenzo Tange. These words are followed by others like "new", "time" or "born", giving the idea that this concept of urban planning presented by Kenzo Tange is really interesting to write about as a novelty. On the other hand, in T-2 we also see the most common words: "system", "communication", "prospective" or "flow". These are more abstract concepts so, probably, the other author – Mariano Bayón - was willing to give us the idea of more abstract aspects of the Tange's architecture. But, even though both texts have 5 years difference between them, we can found common concepts like "human", "planning", "urbanism", or "city". So here is another tool that gives us different ways to extract information about articles.

Here I present another very interesting graph that shows how information is disposed in an article in time and space. Before, in other presentations, we have seen other projects that are structured in three parts: introduction, development of contents in the middle, and conclusions at the end. This is the classical way to organize a discourse, a speech, a presentation, a book, an article, whatever...

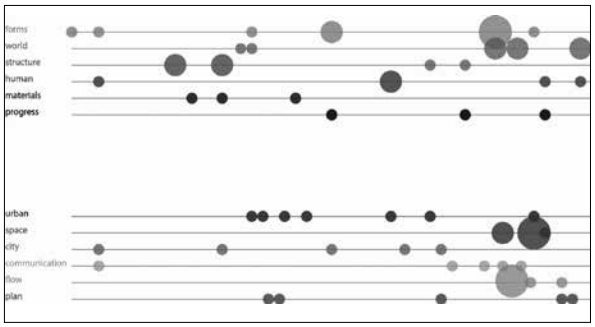

Fig.6 Bubble-graphs of the most common words in T-1 and T-2. They show how words are disposed in the two articles in time and space.

In T-1 the word "structure" is used to propose certain theories at the beginning. After that, in the middle of the text, Fernández Alba uses the words "forms" and "urban" to develop these ideas. At the end, in the conclusions, he writes the word "forms" to obtain or give power to the conclusion. In T-2 probably the conclusions are even more evident. For example, look at the density of bubbles in the final part, where the concepts are concentrated all together. The word "space" and "flow" are probably in the basis of these conclusions. On the other hand, the word "urban" doesn't appear with high frequency but appears quite persistently, and the same happens to the word "city". They are not treated as very important concepts but they are the substratum of the text because they appear all time. In contrast, the word "communication" is mainly used at the end. So probably with this kind of analysis, it is possible to develop different points of view that are not also appreciated in a first lecture of the article.

These are three practical examples that serve to explain how this technology can help the researchers. However, these techniques and technologies cannot substitute researchers because the brains of the investigators are much more powerful than any

of these tools. The main objective of creating these tools is to help researchers in getting into new types of results.

Finally, in the future we would like to reinforce the international component of this project but, at least for now, we must overcome certain limitations of budget. Thank you very much.

Questions and Comments

(Hugo SEGAWA) Thank you for your presentation, I learnt a lot. I once read a book about this kind of analysis of the speech of politicians. But I have a doubt, not as an analyzer, but as a writer. Usually, when I write a text, in my final revisions I tend to search for words that are repeated and I try to avoid the repetition of words, thus I try to find synonyms or other language supports in order to change the words but saying the same content. Thus, is there any analysis that could consider this kind of variation? Then, a second question. Of course, I do not know the methodology of your analysis, but I remember that Juan Pablo Bonta in his analysis of words, he well referred to studies of semiotics and linguistics. So, is there any possibility to link your methodology to semiotics? Or is there any possibility to analyze the frequency of terms that you presented from the point of view of linguistics?

(Luis SAN PABLO) In fact, we have shown very simple examples of how we can work with printed texts, but there are also other algorithms that avoid these two problems that you have remarked. The first, is that synonyms are different words that have similar meanings, obviously this happens in any language. But we also need to know that most of the languages have no more than 40,000 or 50,000 words. In average, a child can use about 500 words and an adult about 10,000 words, and if we use more than 10,000 words we are considered to be erudite or intellectual. So, it is possible to store all these words in a database, and we can group words with similar meaning, and finally we can associate these words with one concept. We are prepared to consider the first situation related to the use of synonyms. Then, the second issue. We have tools to consider words with the same etymological root: for example, architect, architecture, or architectural. It is possible to consider that similar roots imply in similar concepts and it is possible to generate automatic analysis of this type. It is not a perfect solution, if we go so fast, we will always forget some aspects in the process of analysis. But I prefer to have this kind of analysis – imperfect or inexact – to begin to work than having nothing.

(Adriana PICCININI) Your work is really interesting. You showed us the analysis in English. I was wondering if you have tried to make this analysis in Spanish? And if you have tried in Spanish what was the difference? How much do we get lost in translation?

(Luis SAN PABLO) This is a very interesting question. Obviously, the process must be done in only one language. I think that if we do it in English or Spanish the index will probably show little variation, most of the variation will probably appear due to different organization of words in the text. In fact, when two different texts speak about the same topic usually they have remarkable indexes. But we have not done this exercise yet, experimenting with more languages. This would be a challenging exercise. Most probably we are still not ready to do this in this moment. It is easier to do with some languages, for example, Spanish and Portuguese are not so different. If the vocabularies of the authors are not radically different, the indexes will be the similar and the trust can be guaranteed. There are other ways to test the similarity and avoid the search interference or noise brought in by certain "enemies", like style and the use of synonyms, that could be not beneficial for this technic.

(Gaia CARAMELLINO) I got really fascinated by your presentation because I am very far from all this. Since I am also working on magazines, I am very curious about how you can treat critical statements. Of course, because working on circulation and statements, if you work on a project or a figure, it is possible to map the circulation and transfer experiences. When you arrive to texts, it starts to be difficult. So enlarging the geographies of the observation, how can you treat these shifts of different languages? Which I think is an important question.

(Luis SAN PABLO) I think it is a question of criteria. Finally, I think English would become a common language for an international project because of the practical aspect, that we need to have everybody understanding each other. I know that it will not be beneficial for the Italian researchers, or the French, or the Spanish. This is obvious! But we need to accept this kind of minor disadvantage that might come with the "standardization" of the information.

(Gaia CARAMELLINO) The second one has to do with the analysis on a specific geography. If when you work on the recurrence and combination of words, is there a possibility to include the variable of time in this observation. For example, "urban artifact", we would use only urban, or only artifact, but it is a expression that after Aldo Rossi experiments gained a specific meaning and started to be used regularly. Other expressions, like "city region" have similar development.

(Luis SAN PABLO) The component of time is always difficult to be managed. It is complicated because the example we brought here treat texts in a fixed moment of time. But the evolution of one authors' style could be a very interesting project. It would be interesting to see how one author's style has been transformed from the beginning to the end. For example, Picasso had in the beginning of his career one style, which was somehow understandable, but he reaches the end of his career with paintings which are impossible to understand, with a series of transformations in style that show very different ways to express.

(Ana MALUENDA) I think to incorporate time in this project would come in a second step because we are now in the very beginning of the project. Right now, we are only working with words; in the next step, we plan to work with groups of words; then later, we would like to work with other countries and other continents. We would love to see this project grow and grow, but one step after the other, and right now we are dealing with words.

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).