Managing Educational Information on University Websites: A Proposal for Unibo.it

Federico Nanni

Department of Philosophy and Communication Studies, University of Bologna, Italy

federico.nanni8@unibo.it

Abstract. This article is focused on the complexity of finding and analyzing the totality of educational information shared by the University of Bologna on its website during the last twenty years. It specifically emphasizes some issues related to the use of the Wayback Machine, the most important international web archive, and the need for a different research tool which would guarantee more solid analyses of the corpus. This tool could initially be characterized by the use of standard Natural Language Processing techniques (such as tokenization, stop-words removing, parsing, etc.) but we also have to take into consideration more complex solutions, such as text mining analyses, WordNet integration and an ontological representation of knowledge. Thanks to approaches like the one here presented, future historians will be able to efficiently study the evolution of a website.

Keywords: web historiography, digital history, web archives, natural language processing, latent semantic analysis.

1 Introduction

It is difficult to deny that the World Wide Web is the most important technological innovation of the last Century. As we all know, it has not only exponentially improved our capacity of communicating with other people, but thanks to constant improvements, it has been offering year after year new ways of sharing information (websites, forums, Youtube, Twitter). As IBM remarked[1], in 2012 2.5 quintillion bytes of data were created each day, so much that 90% of the data in the World Wide Web today has been produced in the last two years alone. Therefore, it is self evident that being able to guarantee a precise and rapid access to all kinds of information available online is one of the most important tasks of these decades, and a robust interdisciplinary approach is needed.

However, the exponential increase of digital information is not the only problem which researchers interested in studying the Web have to face in this era. Since 30th April 1993[2], people have been creating websites and sharing information online. Facing the volatility of documents digitally born, web preservation has become another important task, which now involves National libraries, National archives and other no-profit privates organizations.

Taking all these reasons into account, my primary research objective is to emphasize both the difficulties and the potentialities that emerge when web historians[3] have to deal with documents that are digitally created. Specifically, I intend to remark the importance of Natural Language Processing and Text Mining techniques in studying these new sources.

---

[1] http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

[2] On this date CERN announced that the World Wide Web would be free for everyone, with no fees due.

[3] In my works I use the term "web historian" as Niels Brügger does in *Web History* (2010).

Contents of the document. In this article I will focus my attention on one specific case of study: how the University of Bologna has used its website as a platform to communicate educational information to its students, during the last twenty years. My interest will not be uniquely limited to a historical reconstructions of these materials, but I will also aim at performing a diachronic analysis of the changes in didactic at this university during the last two decades.

Even if my analysis is focused on a particular medium (the website) and a circumstantial kind of information (educational ones) related to a specific academic institution (the University of Bologna), it is important to notice that the same approach could be effective for other similar researches.

Given all the above points, the rest of the article is focused on three major tasks:

- First of all the "hypothetical corpus" I intend to recreate is described, and the importance of web archives for this specific task is remarked.

- Secondly a discussion on what kind of digital tools could be effective with a synchronic analysis of these materials is presented.

- Finally the importance, for historians, of developing new interdisciplinary competences and becoming able to deal with new born digital sources is emphasized .

2 Twenty years of educational information on www.unibo.it

Web historiography is a recent and challenging field of study, different from but closely related both to Internet Studies and Digital History, as Niels Brugger (2012) remarked. It involves a solid interdisciplinary approach in the creation of web archives and at the same time it requires specific tools in order to analyze the materials preserved. An active participation of historians is also crucial in dealing with more theoretical issues, such as the difference between web materials and archived web materials or the reliability of these preserved documents.

Therefore, I will take into consideration all these factors and I also intend to comment both on the digital preservation of materials and their automatic analysis.

2.1 The hypothetical corpus

My research objective is the study of how the University of Bologna has shared educational information with its students since 1993 through its website. Thus, the corpus I intend to rediscover includes all course programs offered online by the University during the last twenty years. To do so, I intend to use both the materials still present on Unibo.it and, if they are available, snapshots of its website preserved in the Internet Archive.

2.2 Piecing together the past

It is well known that Italy doesn't have a National Web Archive and, even if important preservation projects have been conducted since 2006[4], currently the only way of consulting Italian digital past is through international web archives.

The most important and extended one is the Wayback Marchine[5]: launched by the Internet Archive in 1996 it has preserved more than 240 billions URLs. This platform offers an enormous collection of snapshots of websites, in chronological order. Nowadays it is the only web archive which guarantees the preservations of almost two decades of Italian websites.

Unluckily, the University of Bologna's website is not preserved there (Fig. 1); supposedly it does not allow Internet Archive crawlers. Therefore, the only viable way of analyzing University of Bologna's digital past is by consulting materials still present on its website. Focusing on educational information, it has been offered on departments pages until academic year 2004/2005; however, because of an important website update in 2005, these information are often not available anymore[6].

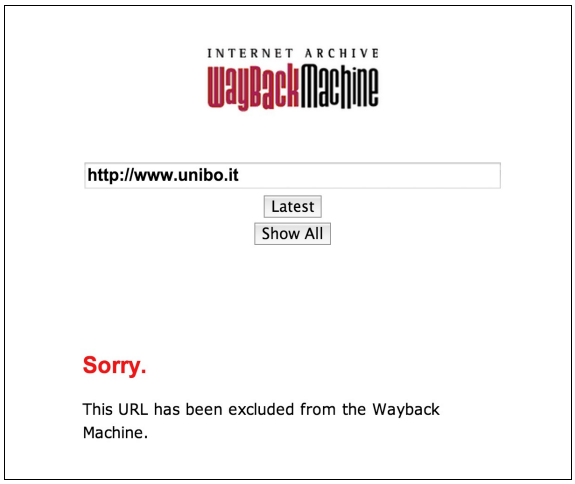

Fig. 1. University of Bologna is not preserved in the Internet Archive.

In conclusion, it is presently possible to have access only to educational information from academic year 2004/2005[7], but without a web archive we are not able to guarantee that the layout of-

[4] http://www.rinascimento-digitale.com/magazzinidigitali.phtml

[5] https://archive.org/web/

[6] For instance, this is the old version of the homepage of the Department of Classic and Medieval Philology: http://www2.classics.unibo.it/

[7] It is important to notice that this academic year is present both in the old version of the department websites and in the new ones.

fered today is the same as ten years ago and we cannot be certain of the preservation of educational materials linked from those pages.

2.3 From the "hypothetical corpus" to the real one

By analyzing this corpus it is evident that every year the University of Bologna offers online a massive amount of information related to its educational organization. As an example, let us focus on academic year 2012/2013: we have digital access to more than 6.300 different courses descriptions, characterized by a brief summary of the lessons, a bibliography, sometimes links to other educational materials, and the curriculum, research interests and a list of publications of the professor. If we multiply this corpus by ten years we obtain an enormous amount of non structured information which, were we able to analyze them, could give us a more complete understanding of the topic.

3 A possible approach for a synchronic analysis

Given the enormous size of the corpus previously mentioned, it is evident that in order to analyze this case of study an advanced search tool is needed: this could, for example, guarantee us to find specific courses by searching their main topic.

Nowadays this is not possible either using the generic search tool (`http://search.unibo.it/`) or the one dedicated to educational information (`http://www.unibo.it/it/didattica/insegnamenti`). Therefore, if we are interested in discovering which course is focused on "the historical importance of Charles Darwin", this is a piece of information difficult to retrieve with the system currently adopted.

3.1 Advantages given by NLP techniques

Given all these reasons, I believe that a viable solution could be a new research tool based on a Natural Language Processing approach. It will consider all the information listed in courses descriptions, on professors pages, and in publications and books abstracts as a single corpus. For instance, to improve the retrieval of educational information all courses programs will be tokenized; then all the stop-words and the other "useless words" (as "exam", "lessons", "program", etc.) will be eliminated. Next, a matrix will be created with the courses on the rows and the words selected on the columns and each word will be then represented as a vector <word, weight>, where the weight is determined by its recurrence and its position in the text.

After a training period exploiting users feedbacks and machine learning algorithms to improve the process of selecting the exact "word-weight" relationship, this tool will allow easier asking of structured query to the system while also receiving more precise information.

Another advantage of this solution could be the possibility of emphasizing similarities between courses, especially when they are from different departments. For instance with this approach we could easily discover that professor Fabio Vitali (Web Technologies) and professor Francesca Tomasi (Digital Humanities) both teach, from different points of view, "semantic web technologies", as the topic is present in both of their courses programs and in their publications.

3.2 Other problems, future solutions

It is clear that a straightforward comparison between two vectors could be an initial improvement for the retrieval of documents but this approach will not be able to extract the main topics present in each program, as the simple recurrence of a word is not enough[8]. Becoming capable of automatically extracting and representing knowledge from a textual document is one of the most difficult challenges that computational linguists have been faced with for more than fifty years. Certainly some progress has been made, as Google Knowedge Graph or IBM's Watson showed us[9], but we are still far from the complete comprehension of meaning from an unstructured text.

However, we could improve the approach previously described on three major aspects, following the advancements in the field. First of all it will be important to guarantee the integration of the tool with structured bases of knowledge, such as WordNet[10] and DbPedia[11], in order to help the system notice relationships between tokens which are synonyms and identify named entities.

Secondly it is evident that, with a bigger and bigger structured database, it will become vital to consider a more complex semantic structure of knowledge. Thus, creating an ontology representation of the educational activities offered at the University of Bologna would definitely help the planning of more complex analyses and information retrieval activities. A solution like this could be projected and initially experimented considering a limited number of courses, as the ones offered by the Department of Philosophy; here, another interesting option would be to automatically generate this ontology "from the data", with a bottom-up approach (Gangemi et al. 2013). Both methods will be considered.

To conclude, the last fundamental aspect that has to be taken into account is the use of more complex techniques for document clustering in order to discover relationships between different courses. Therefore topic models such as Latent Dirichlet Allocation (LDA) or the Pachinko Allocation Model (PAM) could definitely be appropriate for this purpose (Mendes, Antunes 2009 e Templeton 2011). For these reasons the software MALLET[12] will be a fundamental tool during this part of the analysis.

4 Conclusion: from synchronic analyses to diachronic ones

In this paper I underlined the complexities of reconstructing the digital corpus and developing a tool that could be able to automatically analyze the contents, offer more precise access to the information and also evaluate the proximity between two programs by their main topics.

---

[8] As Dan Cohen remarked here: `http://www.dancohen.org/blog/posts/its_about_russia`

[9] Here IBM's Watson project: `http://www.ibm.com/smarterplanet/us/en/ibmwatson/`
And here Google Knowledge Graph:
`http://www.google.com/insidesearch/features/search/knowledge.html`

[10] WordNet is a gigantic lexical database for the English language, but several versions for other languages (including Italian) have been realized: `http://wordnet.princeton.edu/`

[11] DbPedia is a project that aims at extracting structured information from Wikipedia and making them available on the Web: `http://dbpedia.org/`

[12] `http://mallet.cs.umass.edu/`

However it is evident that the approach previously described could also be an efficient solution for scholars who wants to study temporal changing of this new kind of sources.

4.1 Studying the web of the past

As early mentioned, ten years of courses programs are preserved on the website of the University of Bologna, starting from the academic year 2004/2005. Therefore, by implementing the NLP approach here presented to documents created in different years, it would become possible to emphasize correlations between past programs, even when they are taught by different professors, to describe the changes of theme in the same course during the last decade and to discover the increase or decrease of reference to a specific topic. Being able to extend the corpus by digitizing the course programs from previous years will definitely offer materials for way more complex analyses.

However, it is also important to note the fact that the documents here analyzed are not always a satisfying testimony of what the real course was (or is) about: they tend to be vague and, sometimes, an exact copy of the ones written the previous year, with no updated information. Therefore it would also be important to sensitize professors to the importance of these contents, which often represent the only reference for prospect students interested in the course.

4.2 New sources, new approaches but also new historians

Studying sources which are digitally born is about to drastically change historiography. New problems are arising, such as web preservation issues and the unstoppable increase of documents, but at the same time scholars are developing different solutions every day. As the web historian will be one of the researchers who will benefit the most from these innovations, I believe it is important that he will become more and more able to cooperate in this intensively interdisciplinary field of study.

5 Bibliography

- Brügger N. (2010). *Web History*, Peter Lang.
- Brügger N. (2012). *When the Present Web is Later the Past : Web Historiography, Digital History, and Internet Studies.* «Historical Social Research», Vol. 37, No. 4, 2012, pp. 102-117.
- Gangemi A. *et al.* (2013). *A Machine Reader for the Semantic Web*. International Semantic Web Conference (Posters & Demos), pp. 149-152.
- Mendes A. C., Antunes, C. (2009). *Pattern mining with natural language processing: An exploratory approach.* «Machine Learning and Data Mining in Pattern Recognition», Springer Berlin Heidelberg, pp. 266-279.
- Templeton C. (2011). *Topic Modeling in the Humanities: An Overview* URL= http://mith.umd.edu/topic-modeling-in-the-humanities-an-overview/ [Last visited 26.03.2014].