

Crossreads

Towards a rhizomatic narrative

Jaume Nualart

University of Canberra (AU), University of Barcelona
(CAT), and National ICT Australia (NICTA)
jaume.nualart@canberra.edu.au

Gabriela Ferraro

National ICT Australia (NICTA) and Australian
National University, Australia
gabriela.ferraro@nicta.com.au

Abstract

We present Crossreads, a manner to deconstruct linear narratives and to read texts in multiple orders. This is an ongoing work that studies data multiplicity, as well as textual visualization interfaces. The process starts with a text selection, then the text is segmented into small blocks, and a similarity among them is calculated, forming a network data set. Finally, a web interface allows the user to explore and read through the created network of texts.

Categories and Subject Descriptors H.5.4 [information Interfaces and Presentation]: Hypertext/Hypermedia: navigation

General Terms Experimentation

Keywords narrative, deconstruction, data multiplicity, visualization.

1. Introduction

Inspired by works like Rhizome [4] we present an ongoing project in the domain of information seeking and discovery called Crossreads. This work proposes an experimental way of reading texts, alternative to traditional linear reading. We propose to break the initial narrative line of a text by segmenting that text into small parts. Then we reorder the segments according to similarity scores, and, finally, we offer to the reader multiple paths to read the text. The aims of this project are to explore and study the effects when a reader processes fragmented information, as well as to analyze user activity and support reader's exploration with visualization techniques: the interaction text-reader, the text as a collection of segments, the most popular reading paths, and so on. At this stage of the project we cannot tell accurately what the benefits of crossreading texts are. According to some authors, learners naturally make connections between pieces of knowledge, and they are better able to retrieve and apply their knowledge when those connection are accurate and meaningful [1].

Crossreads outputs are a network data set where nodes are segments of text, and an interface that supports nonlinear reading in different ways.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Copyright © ACM [to be supplied]...\$15.00.
<http://dx.doi.org/10.1145/>

This project is part of a practice-led PhD project that has three main focus: the study of text visualization approaches, the multidisciplinary of the field, and the process of design and development tools. The presented approach, Crossreads, is one of the artefacts created as part of the PhD. The whole project includes a proposed new classification of text visualization tools [10], a visualization tool for single texts [9], and two visualization tools to explore and overview collections of texts [5], [8].

In the following sections, we discuss related work and influences. Then, we present the experiments done so far. Finally, we present some conclusions and directions for future work.

2. Related work

Several works in the past have explored the possibilities of breaking the linearity of a text. As mentioned in the introduction, the philosophers Deleuze and Guattari have described the rhizomatic structure of knowledge, which inspires this project too: "In a book, as in all things, there are lines of articulation, segmentarity, strata and territories; but also lines of flight, movement, deterritorialization and destratification". In the novel *Hopscotch* by J. Cortázar [3], the author proposes two reading order for the chapters; the text starts with: "In its own way this book is many books, but mostly it's two books". The Project Xanadu from 1960 [12] is considered the first hypertext project in the digital era, and it was a visionary definition of standards for the WWW that were mostly not included in the standard protocols. One of Xanadu's rules states: "Every document can consist of any number of parts each of which may be of any data type.". The open Xanadu project is accessible and like Crossreads, it encourages nonlinear navigation of text. The aim of Xanadu's demo is to demonstrate the possibilities of hypertext.

Those are the main examples that make us to investigate the effects of reading in alternative ways in combination with normal reading.

3. Experiments

Currently, two versions of Crossreads have been developed (I and II). Version I is part of an exhibition at Museum of Contemporary Art of Barcelona (MACBA), with texts in Catalan and Spanish by the artist Eugeni Bonet [6]. Version II uses texts in English by Domenico Quaranta about media art, compiled in the book "In Your Computer" [11], and it is accessible on line [7]. In both cases, the texts used are licensed under Creative Commons.

The creation of crossreads implies two different tasks: (i) data preparation and analysis. and (ii) interface design.

3.1 Data preparation and analysis

During the data preparation and analysis task, three main steps have been identified: data set selection, data segmentation, and similarity calculus.

3.1.1 Data set selection

So far, we have experimented using data sets from a single author and, more research should be done in order to propose text collections from multiple authors, topics, languages and other criteria.

The original data used for this experiment, in the two versions, have a particular feature: the data, i.e. the texts, are collections of documents opinion and critic documents compiled in books. We designed Crossreads respecting the original documents. The interface also enables linear reading of each document of the collection. In the future, we do not see any design problems if the original data is considered one single document.

3.1.2 Data segmentation

We have experimented with two segmentation approaches, with different benefices. In both versions (I and II) each document is divided into segments, where each segment consists of one or more existing paragraphs. The segments length is about seven hundred characters in total; which is an average of one minute of reading for an adult [13]. In Version I, segmentation is machine produced. In Version II, segmentation is human curated. Version I method is fast and able to process big collections. Version II method could be richer in terms of quality of the segments.

The reason for this two approaches is: this segmentation task is very subjective. A human expert could add a personal view to the segmentation (Version II). A machine produced segmentation (Version I) can accomplish well this task in terms of size of each segment, but it cannot be expected the richness of an expert. We wanted to compare both ways as part of the initial experimentation that will be validated by the user evaluation test.

3.1.3 Data similarity

For the similarity calculus between segments, thus to create the Crossreads network, we use the following of-the-shelf Natural Language Processing tools and techniques:

- Tokenization: words in the segments are separated by white space and punctuation characters.
- Stop word removal: standard stop word removal.
- Named Entity Recognition: identification and classification of Named Entities (NE) in each segment. We applied the OpenNLP Named Entity recogniser [2], which distilled four types of entities, Person, Location, Organization and Others.
- Similarity Calculus between segments.

The similarity between pairs of segments is calculated as the sum of the following factors,

$$Sim(i, j) = TokSim + EntitySim + NESim/3$$

where TokSim is the token cosine similarity between segments, which is a common vector based similarity measure. To calculate the similarity, the tokens of each segment are transformed into vectors and then the Euclidean cosine is used to determine the similarity between pairs of vectors; EntitySim is the sum of the NEs in each segment, normalized by the number of tokens in both segments; and NESim is the cosine similarity between NE. During this process, the similarity between different NE types (Person, Location, Organization and others) is calculated separately.

The similarity between the segments is calculated as follows. First, an arbitrary segment i is chosen and used to calculate the similarity between the segment i and the entire segment collection. Second, the segment with the highest similarity value score is set as the maximum similar segment of i . Since linear reading of a documents is enabled, in each iteration we decided to skip links to segments of the same document as segment i . Finally, we applied different constraints to Version I and II, which are:

Version I: in the crossreads network, each segment is linked to its most similar segment. The drawback of this approach is that links will have a wide range of similarity score, since in every iteration, the number of segments to compare with is smaller, and the possibility of finding a segment with a high similarity score decreases. The benefit it will not be orphan segments, i.e. all segments link to other segments, so the reader will always have the possibility to do crossreading.

Version II: each segment is linked to most similar one.. To avoid repetition of pairs, segments that has been already set as a maximum similarity segment during ten iterations are skipped. After ten iterations, the skipped segments are use again in the similarity calculus.

Again, both methods need to be evaluated by users.

3.2 Interface design and visualization

The interface has been designed to allow maximum comfort for the reader experience. It allows linear reading of the texts in combination with crossreading. A reader can choose any text of the collection to read. The collection of texts is presented in a time-line, and in a flat list with text-category filters. Both versions share a similar interface, with minimal differences according to the different data analysis and authors decisions.

In both cases there is a design principle: vertical navigation for linear reading of documents —using standard up and down arrows images—, and horizontal navigation for crossreading —using specially designed left and right arrows images.

Version I of the interface is adapted to the context of a museum exhibition. The interface has been assessed by a team of experts from MACBA including producers, curators and art historians. In this version, when the user reads a segment of text, there are two links to jump to —right and left— With the right link, the reader goes to its most similar segment. With the left link, the reader goes to its second most similar segment. Thus, both links offered the crossreads experience. In the links, it is announce the title of the document the links goes to, proving the user with some context before following the links. Version II of the interface has evolved offering link nuances. With the right link, the reader goes to its most similar segment. The link context is also showed as in Version I. Furthermore, the quality of the link is represented with an icon, which shows the similarity score between the current segment and the segment in the right link, as well as the token and entities similarity scores. With the left link, the reader jumps to a random segment of the collection.

4. Conclusions and future work

Crossreads proposes a novel way to explore a text collection, based on text segmentation, the textual similarity between the segmented pieces of texts, and a reader interface. For future work, a user evaluation is planned in order to assess: (i) the impact of the human and the automatic segmentation approaches in the crossreads experience, (ii) how the similarity among segments is interpreted by readers, and (iii) the effect of crossreading in the learning process.

Furthermore, future work will focus in discussing the conditions that a text must accomplish in order to suits crossreading, for exam-

ple, one or multiple authors, one or multiple genres, monolingual and/or multilingual collections, just to mention a few variables.

References

- [1] Susan A Ambrose, Michael W Bridges, Michele DiPietro, Marsha C Lovett, and Marie K Norman. *How learning works: Seven research-based principles for smart teaching*. John Wiley & Sons, 2010.
- [2] Jason Baldridge. The opennlp project. URL: <http://opennlp.apache.org/index.html>, (accessed 2 February 2012), 2005.
- [3] Julio Cortázar. Hopscotch (rayuela). *New York: Pantheon*, 1966.
- [4] Gilles Deleuze and Félix Guattari. Introduction: rhizome. *A thousand plateaus: Capitalism and schizophrenia*, pages 3–25, 1987.
- [5] Jaume Nualart. Area, 2013. URL <http://research.nualart.cat/area-ir/>. [Accessed: 2014-07-27. (Archived by WebCite at <http://www.webcitation.org/6RN89Vogt>)].
- [6] Jaume Nualart, Sonia Lopez, Gabriela Ferraro. Crossreads at MACBA, 2014. URL <http://www.macba.cat/eugenibonet/>. [Accessed: 2014-07-27. (Archived by WebCite at <http://www.webcitation.org/6RN8iHR48>)].
- [7] Jaume Nualart, Sonia Lopez, Gabriela Ferraro. Crossreads and D. Quaranta, 2014. URL <http://research.nualart.cat/crossreads>.
- [8] Jaume Nualart, Wray Buntine. Visference, 2013. URL <http://research.nualart.cat/visference/>. [Accessed: 2014-07-27. (Archived by WebCite at <http://www.webcitation.org/6RN7kqO7J>)].
- [9] J. Nualart and M Perez-Montoro. Texty, a visualization tool to aid selection of texts from search outputs. *Information Research*, 18(2), jun . ISSN 1368-1613.
- [10] J Nualart, Mario Pérez-Montoro, and Mitchell Whitelan. How we draw texts: a review of approaches to text visualization and exploration. *El Profesional de la Información*, 2014, vol. 23, num. 3, p. 221-235, 2014.
- [11] Domenico Quaranta. *In Your Computer*. Lulu. com, 2011.
- [12] Ted Nelson, et al. Project Xanadu, 1960. URL <http://xanadu.com/>. [Accessed: 2014-07-26. (Archived by WebCite at <http://www.webcitation.org/6RL0HzFo>)].
- [13] James R Williams. Guidelines for the use of multimedia in instruction. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 42, pages 1447–1451. SAGE Publications, 1998.