Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 50 (2016) 436 – 441

www.elsevier.com/locate/procedia

26th CIRP Design Conference

Supervised Process of Un-structured Data Analysis for Knowledge Chaining

Matthieu Quantin[a,b,*], Benjamin Hervy[a], Florent Laroche[a], Alain Bernard[a]

[a]École Centrale de Nantes, IRCCyN UMR_CNRS_6597, Nantes, France
[b]Université de Nantes, CFV EA_1161, Nantes, France

[*] Corresponding Author *Tel.:* +33-2-40-37-69-51 *E-mail address:* matthieu.quantin@irccyn.ec-nantes.fr

Abstract

Along the product life-cycle, industrial processes generate massive digital assets containing precious information. Besides structured databases, written reports hold unstructured information hardly exploitable due to the lack of vocabulary and syntax standardization. In this paper we present a methodology and natural language processing approach to exploit these documents. Our method consists in providing connections based on supervised retrieval of domain-specific expressions. No prior document analysis are required to drive the algorithm. It underlines a scale of specificity in pattern visualization. This allows relevant and specific information extraction for feedback (e.g. design stage, after-sales service).

© 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license
(http://creativecommons.org/licenses/by-nc-nd/4.0/).
Peer-review under responsibility of the organizing committee of the 26th CIRP Design Conference

Keywords: Knowledge Management; unstructured data; design know-how; semantic information network; Natural Language Processing

1. Introduction

1.1. Context

Knowledge management during industrial processes implies massive digital documents and assets through information systems among the product lifecycle. These documents contain precious information for research and development improvements. Big data techniques are already used for knowledge discovery in order to tackle this issue on structured databases [1]. Besides structured databases of product or process data, there is also unstructured information. An estimate range between 50% and 80% of all the company's data is unstructured [2].

> "There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data". [3]

In this paper we focus on textual analysis of written information rather than massive and structured input data. Therefore, we deal with technical reports, testimonies, and any other written document or collection of documents. Then, we aim to work on knowledge management by chaining knowledge elements rather than data classification and knowledge discovery. The specificity of our work mainly lies on the nature of processed information: contrary to sensors or logs, human actors imply a strong variability in provided information.

1.2. Goals

Over a first phase, the main issue in the process of analyzing unstructured data for knowledge chaining is the domain specificity without any prior information. The process deals with documents that are not conforming to any vocabulary or syntax rules (such as customer reviews or technical notes for example). It underlines a domain-specific vocabulary without labeling or any pre-processing from the document's author. In addition, the process is mainly supervised. Thus, it keeps the results as close as possible from the author work habits and methods.
Over a second phase, our process aims at providing hypernavigation in corpora based on extracted information. It leads to two main outlines, and a third minor one:

1. Enabling different readings of the same corpus. New levels of reading are based on generic/specific keywords and expressions retrieved by the algorithm.
2. Providing understanding facilities through a located high precision decision-support analytic for unstructured textual corpora. The original text linearity of a single file, or the discontinuity of a corpus is outplayed by networking pieces of extracted information.
3. (minor outline) Offering a massive indexing system on server: anyone could upload a production, that is tagged and linked to other productions or any (other than text) tagged entity.

1.3. Usecase

Based on natural language processing techniques, we aim to demonstrate our proposition through a specific domain: technical history. Our work is inspired by the historians challenges.

2212-8271 © 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license
(http://creativecommons.org/licenses/by-nc-nd/4.0/).
Peer-review under responsibility of the organizing committee of the 26th CIRP Design Conference
doi:10.1016/j.procir.2016.04.123

Historical production, definitely human centered, discontinuous and multi-scaled is a perfect experimental field for raw text processing and knowledge chaining. For this article we also used a special corpus, the one from the CIRP paper annals, providing a tangible example for the CIRP community. We will discuss in the last part of this paper the interest of such work for industrial engineering community and how our proposal can be extended to other productions than of the provided use cases.
Thereafter, the term "corpus" will refer to the raw material input: a collection of written documents or a single one, that may contain internal and external references (pictures, quotation, links, etc).

2. Natural language processing for knowledge management

In research and industrial processes as in historical field, only external supports (mainly written) can capitalize the explicit information humans produce. This type of information is unstructured, and differs from any sensor's data output. In a classical conception of the DIKW[1] hierarchy [4], humans cannot produce "pure" data. As History is purely human produced, we have to deal with higher level agency: information to knowledge chaining, not data.

A piece of information is never independent, so the relationship that links to pieces is essential and is part of the knowledge. The aim of our work is to support the importance of these connections. This arise the ethical question, that won't be further discussed here, of the computer influence in (historical) knowledge establishment.

Three assumptions distinguish our work from other Text Mining tools for Knowledge Management:

1. Our work should not depend on any (human or not) data pre-processing step. Raw texts are our only reliable and expectable source of information. We aim to focus on unstructured data, massively produced in every business. This take us away from pure big data issue for decision making as described by the McKinsey Global Institute [5]. It also differs from text analytic oriented projects such as TXM [6] in french, or big data project with clustering for hierarchical structuring [7] in the same field of experiments.
2. Our previous works with historians taught us that information cannot be *a priori* compartmentalized in stringent entities [8]. Unlike biological [9] or manufacturing datamining, no exhaustive data classes exist. Our goal is not to "turn low level data into high-level and useful knowledge" [10], but to process already "compiled" data, i.e. information.
   A framework for text processing in (human) knowledge management should not be first semantically defined. Previously defined topics are at least narrowing and even irrelevant to a text. A POS-tagging or shallow parsing would not allow to match high pattern specificity, nor a core business vocabulary. So we aim to *extract* characteristic entities from a raw text and not to *recognize* already known ones. In this sense we deeply differ from "Named Entity Recognition for Classification" (NERC) processes such as [11] seeking to fill "the lack of publicly available labelled datasets" in french. In the same field of experiments, other works are focused on NERC and neglect the specificity of the input material [12]. Interpretation should be left for the human reading, this work try to stay as close as possible to the text.
3. The black box effect should be avoided. A fully supervised process is necessary for high quality results. We assume that 10 or 20% of mismatch is often not acceptable. Purely manual tagging is a laborious common practice, that hardly reach completeness, but builds high-quality tags. Our supervised process tries to get the best from the machine (completeness) and from the human (high-quality keywords). Moreover, this supervised step balances the machine responsibility (ethical).
   Once again we differ from a classical conception of supervision in NERC: "The main problem with Supervised Learning techniques is that a large amount of tagged data is needed to implement an effective system"[13], because in our case no training set is needed.

Regarding theses existing frameworks and tools, we state that our approach is complementary though significantly different. Complementary with the human "classical" approach of historians, but also with some other text mining tools. This option will be broached as a perspective in section 4.4.

3. Haruspex: a supervised process for knowledge discovery in un-structured data

To achieve our goal of compiling information produced under the form of written documents, we have designed a process based on natural language processing techniques. This process is implemented with Python3 programming language, with a basic GUI for end users. It takes the form of a simple software called Haruspex that takes a corpus as input and produces an undirected graph as output. The output graph is composed of documents or part of documents as nodes (with their metadata) and weighted keywords based relationships as edges. Users can supervise the entire process and have access to internal mechanisms such as the keyword extraction context.

3.1. Overview of the global process

The global process (see fig. 1) is composed of 4 main steps :

1. first step consists in processing input data files : format conversion, files concatenation, splitting, construction of "unit pages" that consists of file or part of file content.
2. second step runs Automatic Natural Acquisition of a Terminology [14] on the previously prepared pages.
3. third step maps the list of keywords from step 2 with their location in the different "pages" from step 1.
4. forth step builds weighted links between the related pages based on tf-idf (term frequency / inverse document frequency) indicator for each keyword.

Those four steps are detailed in the following sections.

[1] Data Information Knowledge Wisdom

Download English Version:

https://daneshyari.com/en/article/1698189

Download Persian Version:

https://daneshyari.com/article/1698189

Daneshyari.com