Information & Management 50 (2013) 540–552

Contents lists available at ScienceDirect

Information & Management

journal homepage: www.elsevier.com/locate/im

A research case study: Difficulties and recommendations when using a textual data mining tool

Abeer A. Al-Hassan [a,*], Faleh Alshameri [b], Edgar H. Sibley [c]

[a] Kuwait University, Kuwait
[b] Howard University, Washington, DC, USA
[c] University Professor and Eminent Scholar, Emeritus, George Mason University, USA

ARTICLE INFO

Article history:
Received 1 June 2012
Received in revised form 23 December 2012
Accepted 27 May 2013
Available online 26 July 2013

Keywords:
Corporate websites
Legal statements
Policy statements
Terms of Use
Textual data mining
Clustering
Industry classification
NAICS
SIC
Privacy statement

ABSTRACT

Although many interesting results have been reported by researchers using *numeric* data mining methods, there are still questions that need answering before *textual* data mining tools will be considered generally useful due to the effort needed to learn and use them.

In 2011, we generated a dataset from the legal statements (mainly privacy policy and terms of use) on the websites of 475 of the US Fortune 500 Companies and used it as input to see what we could detect about the organizational relationships between the companies by using a textual data mining tool. We hoped to find that the tool would cluster similar corporations into the same industrial sector, as validated by the company's self-reported North American Industry Classification System code (NAICS). Unfortunately, this proved only marginally successful, leading us to ask why and to pose our research question: What problems occur when a data-mining tool is used to analyze large textual datasets that are unstructured, complex, duplicative, and contain many homonyms and synonyms?

In analyzing our large dataset we learned a great deal about the problem and fortunately, after significant effort, determined how to "massage" the raw dataset to improve the process and learn how the tool can be better used in research situations. We also found that NAICS, as self-reported by companies, are of dubious value to a researcher—a matter briefly discussed.

1. Introduction

Electronic commerce is now an important part of national and international trade and thus more controls are needed to ensure effective website design and an efficient way of servicing customers. Today a person needing to buy a product will on his or her own behalf, or working as a purchasing agent for an organization, search the website of vendors to find a satisfactory and cost-effective product that is available and guaranteed by a vendor with whose products the buyer is familiar. However, as the electronic marketplace expands world-wide, the buyer needs to learn more about the organization and how it operates because the customer may live in a different country or be accessing the website of a small and relatively little-known company. Thus the material on a company website should be provided to satisfy the needs of worldwide customers whose search should be easy to perform; the data, of course, should be accurate and easy to understand. In our attempt to assess the "value" of a website, we decided to use a textual data mining tool. This led us to ask questions about the problems and potential of mining the contents of websites and to try to determine the difficulty in mining rather sparse and yet complex data. We therefore initially hoped to prove the following research hypothesis:

> The material on an organization's website discloses its sector of industry, where the industry is known by downloading the Corporation's self-defined NAICS (which is normally included on its website).

Since almost every corporation uses the website as a way to *advertize its wares*, we felt that mining the whole site to determine whether the clusters would form into sets of industries would prove too simple a task and that the result of such a research effort would be trivial. Instead, we chose to use a data mining tool on only part of the website: the *legal attachments statements*.[1] We therefore downloaded these parts of 475 of the US Fortune 500 company

* Corresponding author.
E-mail addresses: abeer@cba.edu.kw (A.A. Al-Hassan), fal-shameri@howard.edu (F. Alshameri), esibley@gmu.edu (E.H. Sibley).

0378-7206/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.im.2013.05.010

[1] These impose a legal obligation between the corporation and the website user and *vice versa*. They may appear in more than one section on different corporations' websites. Attachments or section such as a *privacy statement* and *terms of use* are the most common. But others, such as *terms and conditions*, *forward statements*, and links, cover the same topics but in no predefined order.

website attachments and their NAICS. Specifically, we used a data mining tool (CLUTO[2]) on the dataset consisting of all the available downloads,[3] hoping to find the results clumped into corporations considered to be in the same industry (i.e., performing business activities that have been categorized into easily understandable sectors, such as *the computer industry*). Governments and international bodies are interested in such categorization and the best known schemes today are the Standard Industry Code (SIC) and the North American Industry Classification System (NAICS).

Our attempt to determine the relationship between the legal attachment statements of FORTUNE 500 corporations and their (self-defined) industry code (NAICS) required some form of cluster analysis. At this point, we attempted to validate the results by checking to what extent the companies within a cluster had the same NAICS codes, and found that they did not perform as we expected. On examination of the NAICS, we realized that they did not seem to be what we expected—a surprising finding that led us to ask several questions about the process that a corporation takes to decide on its set of codes.

1.1. Our purpose and the research questions

We wished to determine the value of textual data-mining by clustering the datasets formed by downloading only the legal portions of the websites of major corporations in the hopes of finding that they would be grouped according to their industrial classification, as stated by their self-defined NAICS. This led to one major and one minor research question:

> Is it possible or reasonable to evaluate the effectiveness of the textual data mining process by finding how closely the clumps resulting from the use of the data mining tool on data downloaded from a corporate website is explained by the corporation's self-reported NAICS code?

And, because of our answer to this, it was necessary to add:

> What has to be done to the downloaded data to allow a tool to clump the data meaningfully?

1.2. The significance of our results

The results of our work on the major question led us to a discussion of how to reduce the time and effort expended in obtaining useful information using a textual data mining tool on a complex and unformatted set of downloaded data.

The second or minor question led us to further asking:

> What were the problems in stating a company's SICs or NAICS codes? And
> Are the data produced for international and local export/import analysis accurate (due to the lack of breakdown of the information delivered by individual corporations)?
> These two seemed important questions and led to us to consider them as questions for our next major research project.

1.3. The structure of the paper

In Section 2, we briefly discuss the portion of a typical website that deals with the legal aspects. This is followed (in Section 3) by a description of the NAICS coding system and a discussion of textual data mining (Section 4), leading to a discussion of our overall research methodology (Section 5). Section 6 provides an analysis of our results and Section 7 our conclusions. Our references and eight appendices complete the paper.

2. The contents of a corporate website

2.1. Legal issues affecting a corporate website

Most websites collect personal information from their visitors, and this gives rise to potential privacy infringement. Local, regional, and national governments have noted this and drafted laws to protect the individual; examples include the EU data protection law and US and its states' data privacy laws that attempt to protect individuals from the misuse of personal information. These regulate the collection, storage, use, and cross-border transfer of data until its final disposal [8].

Most Fortune 500 organizations use the US Federal Trade Commission's Fair Information Practice Principles (FIPP) [6] as a blueprint for their privacy policies. The FIPP has five core principles: Notice/Awareness, Choice/Consent, Access/Participation, Integrity/Security and Enforcement/Redress [7].

Many US laws, such as the Gramm–Leach–Bliley Act [3], which requires financial institutions to explain their information-sharing practices and ways of protecting sensitive information received from their customers, and the Health Insurance Portability and Accountability Act (HIPAA) [13], which addresses the storage and privacy of personal health data, have followed the FIPP core principles, as has the US Children's Online Privacy Protection Act (COPPA) [2], which requires websites to post clear rules on what, if any, information it collects from children who visit their site. However, there is no standard template for such legal issues, though most websites have similar parts and some may also have portions written to ensure compliance with the laws of states in which they do business.

2.2. Components of the privacy policy statement

The corporate privacy policies are explained in an attachment that says how the company protects the information that it collected from its visitors or potential customers. Its major purpose is to show any policies and practices when dealing with personal and private data that are collected from people and organizations, thereby making sure that they have a legal basis for their defense if sued for any release of private information, as well as giving all customers the right to decide whether to participate by providing their information or opt out of the process.

2.3. Components of the Terms of Use

Terms of Use are posted on a website to establish rules on operations that may be performed by the firm, its customers, and its partners. The website generally includes a statement of the service provided by its owner, used to disclaim any implied warranties [14] as well as a declaration of the site owner and visitors' rights and responsibilities. Most sites require that their users/visitors accept the terms of use before being allowed to access other parts of the site, a practice considered to be a valid contract that is legally binding. It is important to note that sites differ in the complexity of these terms, depending on the nature of the corporation; such as it being commercial or public.

[2] CLUTO is a free data clustering software package for clustering low and high dimensional datasets; it is owned by Karypis Lab at the University of Minnesota: www.cs.umn.edu/~karypis/cluto.

[3] In 2011, 25 of the Fortune 500 companies did not have websites.