Geoderma 334 (2019) 49–54

Contents lists available at ScienceDirect

Geoderma

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

Natural language indexing for pedoinformatics

John Furey*, Austin Davis, Jennifer Seiter-Moser

Environmental Laboratory, U.S. Army Engineer Research and Development Center, 3909 Halls Ferry Rd., Vicksburg, MS 39180, United States of America

ARTICLE INFO

Handling Editor: A.B. McBratney

Keywords:
Soil science
Classification
Taxonomy
Databases
Text mining

ABSTRACT

The multiple schema for the classification of soils rely on differing criteria but the major soil science systems, including the United States Department of Agriculture (USDA) and the international harmonized World Reference Base for Soil Resources soil classification systems, are primarily based on inferred pedogenesis. Largely these classifications are compiled from individual observations of soil characteristics within soil profiles, and the vast majority of this pedologic information is contained in non-quantitative text descriptions. We present initial text mining analyses of parsed text in the digitally available USDA soil taxonomy documentation and the Soil Survey Geographic database. Previous research has shown that latent information structure can be extracted from scientific literature using Natural Language Processing techniques, and we show that this latent information can be used to expedite query performance by using syntactic elements and part-of-speech tags as indices. Technical vocabulary often poses a text mining challenge due to the rarity of its diction in the broader context. We introduce an extension to the common English vocabulary that allows for nearly-complete indexing of USDA Soil Series Descriptions.

1. Introduction

Soil science can be considered settled science for basic agricultural applications (U.S. Department of Agriculture, 1951), but soil scientists have long struggled to extend their pedological techniques towards other soil functions (Wilding and Lin, 2006; Hartemink, 2006). Churchman (2010) applauded the increasing numbers of diverse publications involving the word “soil” in recent years, but reiterated that soil science per se deals with the formation and properties of soil (Brady and Weil, 2002). A fundamental observation is that soil formation is a complex set of processes with multiple factors (Jenny, 1941), and hence soil science terminology has to be complex to be useful (Bridges, 1997; Krasilnikov et al., 2009). The multiple time scales and the geospatial diversity of soil formation processes make it especially challenging to integrate them into a unified field of study (Baveye et al., 2011; Lin, 2011; Richter and Yaalon, 2012).

1.1. Soil classification

One type of product of soil science *qua* science is the development of soil groupings based upon similarities resulting from inferred formation processes. Alternative soil classifications focus on other aspects besides formation, such as suitability for construction engineering purposes in the Unified Soil Classification System (ASTM, 2017). Typically soil science groupings are arranged hierarchically (Nachtergaele et al., 2002) in the form of ranked levels of classification. In this work the terms soil classification level and soil taxonomy level are used interchangeably. For example in the widely-used United States Department of Agriculture (USDA) system there are 12 soil orders at the highest taxonomic level, and tens of thousands of soil series at the lowest level (Soil Survey Staff, 2014) which continues to gradually develop (Kimble et al., 1999). However despite the emphasis on soil formation processes soil classification systems are in no way cladistic, which is to say that originating “parent” soils are not placed at higher classification levels than further developed “child” soils.

In fact, soil classifications are derived by soil scientists primarily from consolidation of data from soil survey reporting and mapping, meaning that someone went to the location of a soil, described the soil in situ, and classified it based on soil scientists' criteria. These criteria depend on aggregates of soil sampling reports and to a far lesser extent laboratory geochemical and physical analyses, even though literally millions of soil samples have been characterized in the laboratory (Smith et al., 2014; National Cooperative Soil Survey, 2018; Davis et al., 2018). This expert system approach, along with the fact that pedogenesis is so tied to geography (Schaetzl and Thompson, 2015; Zinck et al., 2016), makes it difficult to generalize across disparate regions and to directly compare soil surveys from widely different areas, despite the increasing availability of such data (Beaudette and

* Corresponding author.
E-mail address: john.s.furey@usace.army.mil (J. Furey).

https://doi.org/10.1016/j.geoderma.2018.07.050
Received 9 March 2018; Received in revised form 20 June 2018; Accepted 28 July 2018
0016-7061/ Published by Elsevier B.V.

O'Geen, 2009). Adding to the degree of difficulty of inter-comparison, many different regions (e.g. countries) use quite different and at times competing classification systems (Duchaufour, 1988; Hewitt, 1992; Zitong, 1994; Isbell, 1996; Soil Classification Working Group, 1998; Ferg, 2001; Shishov et al., 2001). The main purpose of standing up the World Reference Base for Soil Resources soil classification system (International Union of Soil Sciences Working Group, 2015) was to harmonize these and other differently structured soil classification systems, not to translate between them.

Since approximately the middle of the 20th century, soil scientists have increasingly turned to numerical techniques to make soil classification more quantitative (Bidwell and Hole, 1964). These techniques have grown to include methods such as pedometrics (McBratney et al., 2000), geostatistics (Lark, 2012), geomatics (Davis et al., 2018), and taxonomic distance (Minasny et al., 2009; Láng et al., 2013). The success of each of these highly supervised techniques depends on the judicious selection of appropriate measures to discriminate and classify soils. Perhaps a more general term to use is pedoinformatics, the application of informatics techniques to soils data (Wilson, 2012). Pedoinformatics has been recently applied in the form of unsupervised multivariate cluster analyses to predict contaminant degradation and sorption in varied soils (Chappell et al., 2016; Katseanes et al., 2016).

1.2. Scientific text mining

In contrast to quantitative numerical data, resulting for example from laboratory measurements amenable to regular statistical analyses, most of the published information concerning soils is in the form of non-quantitative text i.e. words. The words themselves are not suitable for statistics designed for numbers, but are well suited for word analyses including automated text mining techniques. Understandably, text mining was first approached systematically by library scientists (Deerwester et al., 1990). The techniques of general document text mining have become highly advanced in recent decades (Salton et al., 1994; Berry and Castellanos, 2007; Weiss et al., 2015). Notably, the publication of an efficient machine learning algorithm for text mining (Blei et al., 2003) led first to its main application in the analysis of general documents including newspaper archives (Wei and Croft, 2006), before being applied to technical and scientific documents (Blei and Lafferty, 2007).

Due principally to data availability, text mining continues to be applied primarily to analyses of general documents that use a common lexicon (Han et al., 2011; Liu et al., 2017), although methods of analyzing relationships between scientific documents, especially in authorship and references, are very actively researched (Bertin et al., 2013). The self-reflective world of scientometrics is beginning to see scientific text mining as a scientific activity in itself (Mayr and Scharnhorst, 2015). Since text mining is primarily taught in computer science departments, computer science literature often comprises a favorite target of analyses, and applications to internet text are especially scrutinized by computer scientists, and others, to discern social connections and shared intents as well as for other kinds of web analytics (Gupta and Lehal, 2009; Yu et al., 2010; Miner et al., 2012; Sun and Han, 2012; Kiritchenko et al., 2014; Wang and Han, 2015).

But by far the largest investment in scientific text mining in the past decade has involved the published biomedical literature. Biomedical terminology can be challengingly technical, and the analysis of biomedical word usage is its own field of study with its own journals (Shotton, 2010).The initial potential for biomedical discovery using unsupervised text mining methodology was recognized early (Jensen et al., 2006), and recent and ongoing work has been accomplishing much of the initial promise (El-Kishky et al., 2015; Ji et al., 2015; Gonzalez et al., 2016; Zhou and Fu, 2018). In fact the UK National Centre for Text Mining leverages its biomedical research support to further advance the study of scientific text mining methods (Brockmeier et al., 2017). To our knowledge, text mining pertaining specifically to soil science has not been previously published (Furey et al., 2017).

2. Methods

We chose to initially study the highly structured USDA soil taxonomy text documentation, including embedded user manuals and linked help files, and the text data entries in the National Cooperative Soil Survey (NCSS) characterization and Soil Survey Geographic (SSURGO) databases (National Cooperative Soil Survey, 2018; Soil Survey Staff, 2018). It would have been premature to instead start with analyses of the overall soil science literature, most of which exists in text forms that are unstructured or variously-structured including textbooks and journal articles. We obtained and prepared local copies of the NCSS and SSURGO databases on a workstation (Dell Precision 7810) running the Linux (Ubuntu 14.04) operating system. A local PostGIS (The PostGIS Development Group, 2018) implementation (PostGIS 2.1.10) housed the text data for custom database queries.

The Python (Python Core Team, 2018) language framework (Python 3.4.6) was used to execute the database queries and calls to Natural Language Processing (NLP) (Bird et al., 2009) modules from gensim (Rehurek and Sojka, 2010; Rehurek, 2018) for topic modelling (gensim 1.0.1) and from the Natural Language Toolkit (NLTK) (Bird, 2017) for word tagging (NLTK 3.2.5). Due to the complexity of the structures revealed by text mining of the database documentation files (Fig. 1), we determined that a more quickly productive focus of this initial investigation would be the set of USDA Soil Series Descriptions.

The local databases contained 16,599 uniquely named soil series as extant in the USDA taxonomy, and these text data were parsed to paragraphs operationally defined by whitespaces. A portion of one of the soil series entries is shown in Fig. 2. Many of the words in these entries are common English words as used in the general English lexicon (e.g. “somewhat”), and other words that are essentially solely soil science terminology (“superactive”), but many other words are common English words imbued with soil science significance by usage and context (“fine sand”). Taken together, a technical inventory of certain lexemes for soil science can be extended from common English using English morphemes and idiomatic expressions, and thereby generate a wordstock of soils-related terminology and technical parts of speech.

To initialize a training set to perform scientific text mining, we first ran 38 paragraphs randomly chosen from the soil series entries through the default NLTK stochastic part-of-speech tagger (the maxent_treebanck_pos_tagger model). As might have been expected, most of the words were not properly tagged on the first pass since the default tags had been developed for non-technical texts. For an example using just the default tags, the word “mixed” was unfailingly tagged as VBD (indicating a past tense verb) instead of its actual functioning as JJ (adjective). This improper tagging of ordinary words was undoubtedly due to the dissimilarity of usage and context between the soil series texts and the default tagger corpus, but certainly another large cause of the wrong tagging was the uniquely technical vocabulary. This failure to appropriately tag vocabulary in a technical context motivated the authors to refine and extend the part-of-speech tagset to enable technical word indexing.

We manually tagged the initial set of 38 paragraphs by an extension of the NLTK tagset using custom technical word tags, two being generically technical tags such as MM (related to measurements), and a dozen others being tags denoting specifically soil science technical usage such as SP (soil property) and PGM (physiographic modifier). Tagging technical words in this manner enabled direct lookup of specifically tagged technical words and groupings. We highlight the fact that a large fraction of the words in each soil series description involved properties of the location in which the soil was found, which we designated as physiographic words, while other words described properties of the soil itself. Table 1 exhibits the complete list of tags used to extend the NLTK tagset.