Environmental Modelling & Software 62 (2014) 128–138

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

A text mining framework for advancing sustainability indicators

Samuel J. Rivera [a, *], Barbara S. Minsker [a], Daniel B. Work [a], Dan Roth [b]

[a] Department of Civil and Environmental Engineering, 205 N. Mathews Ave., University of Illinois at Urbana-Champaign, Urbana IL 61801, USA
[b] Department of Computer Science, 201 N. Goodwin Ave., University of Illinois at Urbana-Champaign, Urbana IL 61801, USA

ARTICLE INFO

Article history:
Received 16 October 2013
Received in revised form
18 July 2014
Accepted 15 August 2014
Available online

Keywords:
Sustainability indicators
Text mining
Informatics
Knowledge discovery

ABSTRACT

Assessing and tracking *sustainability indicators* (SI) is challenging because studies are often expensive and time consuming, the resulting indicators are difficult to track, and they usually have limited social input and acceptance, a critical element of sustainability. The central premise of this work is to explore the feasibility of identifying, tracking and reporting SI by analyzing unstructured digital news articles with text mining methods. Using San Mateo County, California, as a case study, a non-mutually exclusive supervised classification algorithm with natural language processing techniques is applied to analyze sustainability content in news articles and compare the results with SI reports created by Sustainable San Mateo County (SSMC) using traditional methods. Results showed that the text mining approach could identify all of the indicators highlighted as important in the reports and that the method has potential for identifying region-specific SI, as well as providing insights on the underlying causes of sustainability problems.

Software availability

Name of software: SI News Classifier
Developers: Samuel Rivera (srivera2@illinois.edu)
Contact address: 4129 Newmark Lab, MC-250,205 N. Mathews Ave. Urbana, IL 61801 USA
Availability and Online Documentation: Free download with installation manual and supporting material at GitHub account of the Environmental Informatics and Systems Analysis group at the University of Illinois at Urbana–Champaign (https://github.com/EISALab).
License: The University of Illinois/National Supercomputer Application Center (NCSA) Open Source License (http://opensource.org/licenses/NCSA).
Year first available: 2014
Software required: Matlab and Dataless Classification (developed by the Cognitive Computation Group led by Professor Daniel Roth at the University of Illinois at Urbana–Champaign and available at: http://cogcomp.cs.illinois.edu/page/software_view/Descartes.)
Programming language: Matlab
Program size: 72 KB

1. Introduction

1.1. Motivation

The planet's environmental challenges, economic instability, and finite resources have raised interest in assessing sustainability and measuring progress towards sustainable development (Farr, 2008; Rockström et al., 2009; Solomon et al., 2009; UNDESA, 2010). Over the last 20 years, *sustainability indicators* have emerged as the preferred method to track this progress, and to aid decision making towards sustainable development (Dahl, 2012).

Sustainability indicators are metrics that track the current state and evolution of these complex systems (Hammond et al., 1995; IISD, 2000), such as the number of people living in poverty or the health of endangered species. To be comprehensive, these indicators must address the political, economic, social, and environmental components of communities, and must be understood by all members of society (Innes and Booher, 2000; Dahl, 2012). Indicators are most effective when they are aligned with the values and concerns of the target audience (Dahl, 2012). Lastly, these indicators must be created at multiple scales of governance, including global, national, regional, and city scales in order to be effectively utilized within local cultural, social, political and economic characteristics of each community (Bossel, 1999; Innes and Booher, 2000; Gahin et al., 2003; Dahl, 2012).

To date, 895 initiatives exist worldwide to develop sustainability indicators ranging in scales from cities to global projects (IISD,

* Corresponding author. Tel.: +1 787 240 0044.
E-mail addresses: srivera2@illinois.edu, sammy.rivera14@gmail.com (S.J. Rivera), minsker@illinois.edu (B.S. Minsker), dbwork@illinois.edu (D.B. Work), danr@illinois.edu (D. Roth).

http://dx.doi.org/10.1016/j.envsoft.2014.08.016
1364-8152/

2013). However, not all of the methodologies and guidelines for developing and implementing sustainability indicators have been effective (Gahin et al., 2003; Krank and Wallbaum, 2011). Most indicator projects seek input from representative populations to decide which problems should be addressed (Innes and Booher, 2000; Yli-Viikari, 2009; Dahl, 2012), for example through surveys, public hearings, professional meetings, and other means. Unfortunately, these approaches often fail to achieve large-scale participation, and consequently may not be representative of the community's true values and concerns (Innes and Booher, 2000; Gahin et al., 2003; Adinyira et al., 2007; Yli-Viikari, 2009; Scerri and James, 2010; Krank and Wallbaum, 2011; Dahl, 2012; Moldan et al., 2012). Furthermore, collecting this input is often extremely time consuming and resource intensive, which ultimately leads to large latencies in the reported data (Innes and Booher, 2000; Gahin et al., 2003; Moldan et al., 2012).

These limitations contribute to the challenges associated with using information from sustainability indicators in environmental modeling and management. Approaches such as integrated environmental modeling (Laniak et al., 2013; Jakeman et al., 2008) and decision making, participatory modeling and socio-environmental modeling (Krueger et al., 2012; Voinov et al., 2014) require input on the state of sustainability indicators in order to inform the interactive decision making process. Additionally, it has been argued that the presentation of scientific evidence without an understanding of the preferences and values affecting the decision and actions of stakeholders will usually fall short of being effective (Laniak et al., 2013). It is necessary to have an understanding of the perceptions and values of stakeholders on sustainability issues to generate adequate (re)actions in the form of policies and management strategies (Krueger et al., 2012; Voinov et al., 2014).

The objective of this work is to begin addressing these limitations through the development of a new method to design and track sustainability measures using digital news media and recent progress in text mining. More specifically, this paper addresses the question of whether the news media contains relevant information that can enable fast identification, tracking, and reporting of sustainability indicators for a region. The hypothesis to be tested is that the unstructured data of news media can provide insight into sustainability problems within the cultural and contextual characteristics of a community, thereby also addressing the social component that has been underdeveloped in previous approaches.

1.2. Related work

Recent statistics show that half of the news across the world have been digitized, and are supplanting print and broadcast news (Leetaru, 2011). News media represents a “mediated public sphere” (Holt and Barkemeyer, 2012) with the potential to influence people's mindsets and create a feeling of worldwide connectedness by changing the public's level of awareness and attention to a specific issue (Szerszyski et al., 2000; Holt and Barkemeyer, 2012). Also, it has been argued that newspapers are the most important source of information used to spread scientific knowledge (Nelkin, 1995; Morse, 2008), and are very effective in placing topics in the public mind (Holt and Barkemeyer, 2012). Moreover, it has been shown that there is a causal relationship between thematic priorities of the media and the relevance of social problems in the population (Rogers et al., 1993). Furthermore, news articles contain far more than just factual details; they provide insights into the cultural context upon which they are written, a spatial and temporal component to the facts, and a window for forecasting many social behaviors using text mining techniques (Thøgersen, 2006; Tang et al., 2009; Leetaru, 2011; Michel et al., 2011).

To identify actionable information from the large volumes of unstructured digital news, efficient text mining tools are needed to process the data and extract trends. Recent studies suggest that analysis of text archives can generate new knowledge about the functioning of society (Leetaru, 2011). Moreover, text mining tools can detect the tone of news articles, which enables applications such as forecasting social behaviors ranging from ticket movie sales to stock market trends (Mishne and Glance, 2006; Tang et al., 2009; Leetaru, 2011; Michel et al., 2011). In one recent study, an analysis performed on the global news tone of a 30-year worldwide news archive demonstrated that this type of analysis could have forecasted the revolutions in Tunisia, Egypt, and Libya, including the removal of Egyptian President Mubarak (Leetaru, 2011). This suggests the use of text mining techniques to quantify and assess the social components of sustainability could hold promise.

Text mining techniques have previously been applied to measure some elements of sustainability. “Trends in Sustainability” is a Web application that searches for predetermined keywords related to different sustainability topics in 115 newspaper sources from 41 different countries (Barkemeyer et al., 2009). The result of the application is a display of the trend of the volume of news containing the keyword across time. The “Carbon Capture Report”, is a similar application that searches the social media (e.g. news, blogs, Twitter, and Youtube) for predetermined keywords to identify relevant articles. However, in the “Carbon Capture Report” the data are further processed by implementing *natural language processing* (NLP) techniques and a sentiment analysis that provide further information. The application displays a time series analysis of the volume of data with an overall tone (positive, neutral, negative) and activity of the topics and color coded world regions that indicate the magnitude of their contribution to the data, as related to that topic. A similar application is the Media Watch on Climate Change, a public Web portal that aggregates large archives of digital news and social media coverage on climate change and related issues (Scharl et al., 2013). Using an interactive dashboard the location, the frequency, and the sentiment of the information is displayed to stakeholders with the idea of increasing awareness and availability of environmental information.

1.3. Contributions

Previous sustainability studies using text mining have focused on tracking general trends in different sustainability and climate change topics at the national or global scale. This study focuses on demonstrating a faster method for identifying, tracking, and reporting of sustainability indicators specific to a region using news articles that can better incorporate society's values. Furthermore, the approaches taken in this study provide links between observed indicator trends and their underlying causes; previous indicator methods focus primarily on tracking the issues.

Additionally, this study develops a new methodology that combines a suite of text mining methods to more accurately classify sustainability articles given the limited training set of regional news articles and their non-mutually exclusive topic areas (e.g., an article on the effects of pesticides on water quality would be equally relevant to the water quality and pesticide indicators). The methodology is applied in San Mateo County, CA, to demonstrate feasibility of the approach. Future work is then recommended to extend the methodology to other types of data available on the Web (e.g., social media data and blogs).

2. Methodology

The tracking and extracting of information from sustainability related news articles is accomplished by integrating different classification approaches and NLP