Neural Networks 58 (2014) 38–49

Contents lists available at ScienceDirect

Neural Networks

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

2014 Special Issue

Modeling virtual organizations with Latent Dirichlet Allocation: A case for natural language processing

Alexander Gross [a], Dhiraj Murthy [b,*]

[a] University of Maine, United States

[b] Goldsmiths College, University of London, United Kingdom

ARTICLE INFO

Article history:
Available online 2 June 2014

Keywords:
Natural language processing
Latent Dirichlet Allocation
Big Data
Social media
Virtual organizations

ABSTRACT

This paper explores a variety of methods for applying the Latent Dirichlet Allocation (LDA) automated topic modeling algorithm to the modeling of the structure and behavior of virtual organizations found within modern social media and social networking environments. As the field of Big Data reveals, an increase in the scale of social data available presents new challenges which are not tackled by merely scaling up hardware and software. Rather, they necessitate new methods and, indeed, new areas of expertise. Natural language processing provides one such method. This paper applies LDA to the study of scientific virtual organizations whose members employ social technologies. Because of the vast data footprint in these virtual platforms, we found that natural language processing was needed to 'unlock' and render visible latent, previously unseen conversational connections across large textual corpora (spanning profiles, discussion threads, forums, and other social media incarnations). We introduce variants of LDA and ultimately make the argument that natural language processing is a critical interdisciplinary methodology to make better sense of social 'Big Data' and we were able to successfully model nested discussion topics from forums and blog posts using LDA. Importantly, we found that LDA can move us beyond the state-of-the-art in conventional Social Network Analysis techniques.

1. Introduction

In recent years, the Internet has undergone enormous transformations. From its inception as a framework for the interconnection of fragments of information from disparate locations through a vast network of hyperlinks, the Internet has evolved into a new medium of communication that almost seamlessly connects individuals with one another. Far from early critiques that the Internet separates and isolates people, it could now be argued that an important function of the Internet is in fact to keep people socially connected and even increase their social capital (Wellman et al., 2001). With sites including Facebook and Twitter amongst the most heavily used sites on the Internet, and even sites traditionally thought of as informational like Google relying heavily on social features, social technologies have become increasingly important to us. Indeed, the Internet has become primarily semantic, contextual, and social (Gruber, 2008). A key challenge now is to parse, understand, and visualize these online formations and spaces and their role in our lives.

Many social networking sites implement features which allow users to record and report their relationships with other users and groups (e.g. friending, liking, and following) (Boyd & Ellison, 2007). A whole research domain has arisen in recent years to quantitatively explore the kinds of networks defined by such user-reported relationships. But a complete and thorough analysis of online communities and organizations, which stops there, would be incomplete. These types of relationships only capture a small portion of the activity that defines any true online community. For example, virtual organizations, "collection[s] of geographically dispersed individuals, groups, organizational units – either belonging or not belonging to the same organization – or entire organizations that depend on electronic links in order to complete the production process" (Travica, 1997), involve complex social webs which depend on the development and maintenance of trust. And because virtual organizations are increasing mediated by social technologies, they often include virtual spaces or forums for fostering trust and completing tasks. However, an analysis of these communicative spaces alone is insufficient to the task of understanding the organizational and community-building potential provided by

* Corresponding author. Tel.: +44 207 717 3371.
E-mail address: d.murthy@gold.ac.uk (D. Murthy).

http://dx.doi.org/10.1016/j.neunet.2014.05.008
0893-6080/

modern virtual social formations. Understanding the human relationships within virtual organizations requires much more than capturing who interacts with whom and the topics about which people communicate. Rather, the formations of profiles and other facets of people's virtual presence are negotiated and constructed over time in complex ways. Understanding this activity within virtual organizations involves the study of many users, actions, connections, and communities taking place across a universe of diverse threads, discussions, groups, and micro-communities within a typical social platform and potentially even across multiple social technologies.

In the last decade, even as interactive social technologies were settling in as the dominant online paradigm, there was a severe lack of accepted research methods or even availability of information about fledgling online social activity. Even where rich information was available, robust computational analysis of that information was often intractable. This status quo often led to favoring approaches to network analysis focused on the analysis of reported relationships, as there was an established framework for the study of such structures borrowed from graph theory (Fombrun, 1982). These methods are and remain important to understanding social networks online. However, they can and should be augmented with other types of data and methods of analysis. Natural language processing provides one important avenue for this.

In an earlier issue of this Journal, Fabisch et al. (2013) argue that "Artificial intelligence is facing real world problems and thus machine learning problems become more and more complex". Part of this real world problem is the exponentially increasing volume of data that machine learning has to process. The debates around 'Big Data' challenges have appeared in this Journal as well as across the computer science literature more broadly (Kantardzic, 2011). A particular challenge for artificial intelligence in the context of Big Data is to ensure less – not more – effort is "shifted from human to machine" (Fabisch et al., 2013).

Of course, not all Big Data is equal. A major challenge facing the artificial intelligence community is machine learning with small chunks of text, such as text commonly found on online social networks, social media, and other online spaces. For example, Twitter data has a wide range of text quality, text length, and content types. Machine learning with 140-character tweets is possible and has been done by many (Pak & Paroubek, 2010), but the task remains rife with problems, especially when conventional machine learning algorithms are applied. As Cambria and White (2014) argue, "NLP research has evolved from the era of punch cards and batch processing [...] to the era of Google", and machine learning continues to evolve to adapt to complex settings, such as the analysis of social media data. One part of this evolution is the move from coarse to fine-grained analysis methods. As Cambria et al. (2013) point out in their discussion of opinion mining and sentiment analysis, early NLP methods often classified sentiment based on a whole document, whereas newer methods strive to analyze segment-level sentiment.

In this article, we describe the application of one of these more fine-grained NLP techniques – Latent Dirichlet Allocation (LDA) (Blei et al., 2003) – to information gathered from two prominent virtual communities of life scientists. LDA is a robust and versatile unsupervised topic modeling technique, originally developed to identify latent topics within a collection of text documents. It has shown great flexibility in being easily adapted to situations where objects in a collection are each associated with a unique set of exchangeable attributes (words, in the case of text documents). In general, the technique discovers latent topics of associated and co-occurring attributes within the collection. A latent topic has a probability distribution over words (as opposed to a strict list of words that are included in or excluded from the topic). Instead of simply determining an object's simple group membership, as is the case with many machine learning algorithms, LDA uses a mixture model that models each of its objects as drawing proportionally from each of a set of latent topics. These models provide a rich, almost genome-like structure for the comparison of objects, classifying each on the entire range of latent groupings. This article first introduces the context of our research question, then describes LDA and its variants before moving to our study methods and results.

2. Background

In this section, we will provide a brief background to the LDA method for topic modeling and highlight some of the opportunities that exist in the application of LDA to understanding collective behavior and latent, normally unseen network structures that can be discovered and explored from aggregated communication (in our case, drawn from virtual communities and organizations). Of particular interest to our research are the patterns and networks of latent behavior and communication that help to understand and illuminate the collective activity of scientific social networking sites and particularly the virtual organizations that develop from them.

Social media and social networking technologies have become ubiquitous in our social lives. However, they are also increasingly pervasive in organizational settings. For example, corporate internal social media systems such as HP's WaterCooler (Brzozowski, 2009) and IBM's Beehive (Geyer et al., 2008) confirm the utility of social technologies to organizational innovation, collaboration and general knowledge sharing. Individual discussion threads and even small clusters of interactions on these platforms can be readily analyzed, but it is not easy or straightforward to do this on a much larger scale. As the field of Big Data reveals, an increase in the scale of social data available cannot be effectively managed by merely scaling up hardware and software, but creates new challenges which necessitate new methods and, indeed, new areas of expertise (Kaisler et al., 2013). Our project is particularly interested in the study of virtual organizations mediated by social technologies.

2.1. Virtual organizations

Virtual organizations (VOs) are organizations or enterprises not tied to a singular physical locality (i.e. a specific lab or work place), and are a product of changes in global economic, social, and political systems. A useful working definition of VOs is provided by Travica (1997) who views them as manifesting themselves as a "collection of geographically dispersed individuals, groups, organizational units – either belonging or not belonging to the same organization – or entire organizations that depend on electronic links in order to complete the production process". The work of Mowshowitz (1997) and Travica (1997), though useful in defining elements of VOs and mapping their history, does not offer a general articulation of what constitutes a virtual organization. Indeed, VOs are conceptualized differently in different contexts. The VOs form, disband, and re-configure as required for the task. A VO in this context is a virtual collection of geographically disparate team members brought together to solve a particular problem/task or accomplish a specific goal. Ultimately, in global virtual teams, the 'grid' is distributed human resources connected together through collaborative new media technologies to work together as a VO. In this way, VOs share with offline organizations a purpose of organizing individuals towards a common cause. But, with the exponential increases in textual data being produced with