

Chapter 45

An Unstructured Information Management Architecture Approach to Text Analytics of Cancer Blogs

Viju Raghupathi

City University of New York, USA

Wullianallur Raghupathi

Fordham University, USA

ABSTRACT

In this research the authors explore the potential of the Unstructured Information Management Architecture (UIMA) platform in text analytics of cancer blogs. The application is developed using the UIMA open source platform. They use the text analytics methods of categorization, clustering, taxonomic classification, and others to identify and analyze the patterns in cancer blog postings. The authors establish a comprehensive UIMA methodology for developing text analytics applications for the analysis of cancer blogs. Additional insights are extracted through the development of categories or keywords contained in the blogs, the development of a taxonomy and the examination of relationships among the categories. The application has the potential for generalizability and implementation with health content in other blogs and social media. It has the potential to provide insight and decision support for cancer management and to facilitate the efficient and relevant search for information on cancer.

1. INTRODUCTION

Over the last decade, the healthcare domain has seen an exponential increase in the use of Web 2.0 tools and social media such as blogs, wikis, podcasts, twitter feeds, vlogs (video blogs) and on-line journals that convey health-related information.

Such content management applications allow key stakeholders in the health care system - physicians, patients, hospitals, insurance companies, government, and others - to create and disseminate health information via the web (Bradley, 2013; Chen, 2013; McNickle, 2013). Patients, for example, are finding useful health-related information with

DOI: 10.4018/978-1-4666-8619-9.ch045

regard to diagnosis, treatment, and the management of diseases. Physicians also use such tools to search for information in the context of evidence-based medicine and to address patients' concerns and issues (Miller & Pole, 2010). Hospitals and other providers use these tools as "gateways" to the communities (Hardy, 2012; Kotenko, 2013; McNickle, 2013). Large repositories of unstructured textual data are emerging and growing rapidly, and health entities are examining the potential of text analytics and other methods to analyze the data and glean patterns and relationships. These are, in turn, used to gain insight for making informed health decisions and improve clinical outcomes (Konkel, 2013).

Health data, such as general patient profiles, clinical data, insurance data, and other medical data, are being created for various purposes, including regulatory compliance, public health policy analysis and research, and diagnosis and treatment (Mulins et al., 2006). Data may include both structured data (e.g., patient histories as records in a database) and unstructured data (e.g., audio/video clips, textual information such as in blogs or physician's notes) (Spangler & Kreulen, 2007). Text analytics is typically used to identify patterns and trends in the unstructured data (Popowich, 2005). These patterns can shed light on a wide range of issues such as drug reactions, side effects, treatment outcomes, personalized medical treatments and efficacy of drugs. For example, analytics revealed the association between the arthritis drug Vioxx and an increased risk of heart attack/stroke, resulting in the withdrawal of the drug from the market (Versel, 2011). Furthermore, the data can be used for health quality assurance and clinical management queries (Einbinder & Scully, 2002; Scully et al., 1997). Therefore, text analytic applications, although relatively recent in healthcare (Brosette et al., 1998; Downs & Wallace, 2000; Lee et al., 2011), have the potential to improve quality (e.g., reduce medical errors) and reduce the cost of healthcare delivery (e.g., via reduced re-admissions).

In this article we describe our ongoing research project, which uses Unstructured Information Management Architecture (UIMA) for the text analytics of cancer blogs. UIMA defines a framework for implementing systems for the analysis of unstructured data (Kano et al., 2009; Mack et al., 2004). Unlike in structured information where the meaning is expressed by the structure or the format of the data, in unstructured information, the meaning is not explicitly stated, and may need to be inferred. Examples of data that carry unstructured information include natural language text and data from audio or video sources. For example, an audio stream has a well-defined syntax and semantics for rendering the stream on an audio device, but its music score is not directly represented. UIMA offers the capability to analyze unstructured health content at various levels. At the individual level (document-level analysis), one can perform analysis and gain insight about a patient in longitudinal studies. At the group level (collection-level analysis), one can identify patterns in network behavior (e.g., assessing the influence within the social group), in a particular disease group, the community of participants in an HMO or hospital setting, or even in the global community of patients (ethnic stratification). The results of these analyses may be generalizable to other blogs and web content.

Health blogs in particular are rich with information for decision-making. Even though web crawlers and blog analysis software generate statistics related to blogs (such as the number of blogs or top ten blogs) they are not advanced or useful computationally to help with the analysis and understanding of the social networks that evolve in healthcare and medical blogs. Thus there is a critical need for sophisticated tools to fill this gap. Furthermore, to our knowledge, there are not many studies or applications in the text analytics of cancer blogs. This study attempts to fill this gap by analyzing cancer blogs.

The rest of the article is organized as follows: first, we give a background of the research by

17 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:

www.igi-global.com/chapter/an-unstructured-information-management-architecture-approach-to-text-analytics-of-cancer-blogs/137384

Related Content

A Systems Overview of Commercial Data Centers: Initial Energy and Cost Analysis

Sardar Khaliq Uzaman, Atta ur Rehman Khan, Junaid Shuja, Tahir Maqsood, Faisal Rehman and Saad Mustafa (2019). *International Journal of Information Technology and Web Engineering* (pp. 42-65).

www.irma-international.org/article/a-systems-overview-of-commercial-data-centers/217694

State of the Art Review of IIoT Communication Protocols

Shruthi H. Shetty, Ashwath Rao and Rathishchandra R. Gatti (2021). *Transforming the Internet of Things for Next-Generation Smart Systems* (pp. 37-48).

www.irma-international.org/chapter/state-of-the-art-review-of-iiot-communication-protocols/278623

Research on O2O Platform and Promotion Algorithm of Sports Venues Based on Deep Learning Technique

Kaiyan Han and Qin Wang (2018). *International Journal of Information Technology and Web Engineering* (pp. 73-84).

www.irma-international.org/article/research-on-o2o-platform-and-promotion-algorithm-of-sports-venues-based-on-deep-learning-technique/204360

Web 2.0 and E-Discovery

Bryan Kimes (2010). *Web Technologies: Concepts, Methodologies, Tools, and Applications* (pp. 177-187).

www.irma-international.org/chapter/web-discovery/37631

WDPMA: An MA-Based Model for Web Documents Prioritization

Santosh Kumar and Ravi Kumar (2021). *International Journal of Information Technology and Web Engineering* (pp. 1-24).

www.irma-international.org/article/wdpma/275731