Chapter 11
Role of Online Data from Search Engine and Social Media in Healthcare Informatics

M. Saqib Nawaz
Peking University, China

Raza Ul Mustafa
COMSATS Institute of IT, Sahiwal, Pakistan

M. Ikram Ullah Lali
University of Sargodha, Pakistan

ABSTRACT

Search engines and social media are two different online data sources where search engines can provide health related queries logs and Internet users' discuss their diseases, symptoms, causes, preventions and even suggest treatment by sharing their views, experiences and opinions on social media. This chapter hypothesizes that online data from Google and Twitter can provide vital first-hand healthcare information. An approach is provided for collecting twitter data by exploring contextual information gleaned from Google search queries logs. Furthermore, it is investigated that whether it is possible to use tweets to track, monitor and predict diseases, especially Influenza epidemics. Obtained results show that healthcare institutes and professional's uses social media to provide up-to date health related information and interact with public. Moreover, proposed approach is beneficial for extracting useful information regarding disease symptoms, side effects, medications and to track geographical location of epidemics affected area.

INTRODUCTION

Internet is now affecting and facilitating nearly every aspect of modern life, from healthcare and education to government and business. In past few years, healthcare organizations and professionals are using social media in order to promote, support and spread health related information and data for improving

both personal and community health practices (Househ, 2013; Chretien & Kind, 2013). Moreover, the younger generation also uses Internet especially social media for research and making health related decisions. In a survey (Health Fact Sheet, 2015), Internet users of approximately 72% checked online for information on health in past year. Online health seekers of about 77% used search engine. 11% said that they looked for health information at specialized health information site such as WebMD. 3% started their health related research at sites like Wikipedia and an additional 2% said that they used social networking sites such as Facebook and Twitter.

For health related applications and monitoring, the idea of using online data came from the estimation of Influenza incidence using logs of health related search engine queries (Eysenbach, 2006). Studies in (Ginsberg et al., 2009; Hulth et al., 2009; Palet et al., 2009; Achrekar et al, 2011, Aramaki et al., 2011; Polgreen et al., 2008) show that epidemics trend can be detected with information available on Web and there is a strong correlation in the frequency of online search queries and tweets with epidemics events (Xu et al., 2011). Hence, the behavior or pattern of when and how Internet users search may provide early indications or clues related to future concerns and expectations. For example, analysis conducted by Ettredge et al., (2005) on jobs and jobs opportunities related keywords searched by users over Internet has generated an accurate and useful statistics on the unemployment rate. These studies also suggested that people suffering from any kind of disease or health issues uses World Wide Web (WWW) to search for disease information. Logs of search queries (or terms) entered in search engines can provide valuable information on health related issues, especially the detection and monitoring of emerging epidemic diseases, as it is possible to track changes in the volumes of specific search queries.

However search query data is noisy, coarse and it does not provide any contextual information. People search for information related to health on search engine for various reasons, such as concern about oneself, friends or families. Some searches are done because of general interest that is usually initiated by a live event, news report or new scientific discovery (Signorini et al., 2011). Furthermore, errors discovered in Google Flu Trends serves as reminder that this big online data paradigm required further critical investigation and the development of more empirical methodologies for exploring the predictive utility of Internet data (Lazer et al., 2014a; Lazer et al., 2014b). One other limitation is that researchers and scientists do not have full access to search engines logs. Recently, social media data have been used effectively for disease surveillance as they contain contextual health information with diverse descriptions of health states. A study on "Twitter stream" (Twitter Study, 2009) revealed that despite high level of noise, a major proportion of Twitter message contain informative, links to the useful information and news content or spam and self-promotion.

In Natural Language Processing (NLP), opinion mining (also called sentiment analysis) is used for detection and extraction of subjective information from text documents (Liu & Zhang, 2012). Using sentiment analysis, one can find the overall contextual polarity about any topic in a document provided by its author. The challenging task in opinion mining is sentiment classification which is done by guessing opinion about anything such as book, movie, product, issues regarding politics and religion etc. These opinions can be in the form of sentence, document or feature, and the task is to label them as positive, negative or neutral. From developers and researchers point of view, various social media sites and search engine offers APIs (application programming interfaces) that can be used to collect data which later can be used for analysis. Moreover, we can combine these APIs to build our own applications. Therefore, opinion mining shares a strong fundament with the support of huge online data (Fang & Zhang, 2015).

For situation involving limited access to search logs, this chapter consider the use of search queries data to select tweets associated with health related information especially for Influenza epidemics. Main

Related Content

Avatar Control System Use Based on Bioinformatics in E-Business Research of Entrepreneurs
Natalia Rasskazova (2022). *International Journal of Applied Research in Bioinformatics (pp. 1-13).*
www.irma-international.org/article/avatar-control-system-use-based/290343

A Haplotype Analysis System for Genes Discovery of Common Diseases
Takashi Kido (2006). *Advanced Data Mining Technologies in Bioinformatics (pp. 214-230).*
www.irma-international.org/chapter/haplotype-analysis-system-genes-discovery/4253

Macromolecular Crystallographic Computing
Kostas Bethanis, Petros Giastas, Trias Thireouand Vassilis Atlamazoglou (2010). *Biocomputation and Biomedical Informatics: Case Studies and Applications (pp. 1-36).*
www.irma-international.org/chapter/macromolecular-crystallographic-computing/39601

Improving Prediction Accuracy via Subspace Modeling
Majid Masso (2012). *Computational Knowledge Discovery for Bioinformatics Research (pp. 33-48).*
www.irma-international.org/chapter/improving-prediction-accuracy-via-subspace/66703

Introduction to Emotional Chat Bots and the Basics of Bioinformatics
Svetlana Morkovina (2022). *International Journal of Applied Research in Bioinformatics (pp. 1-6).*
www.irma-international.org/article/introduction-emotional-chat-bots-basics/290345