

Digital Cancer Surveillance

Georgia D. Tourassi
Health Data Sciences Institute
Oak Ridge National Laboratory
Oak Ridge, TN, USA
tourassig@ornl.gov

Hong-Jun Yoon
Health Data Sciences Institute
Oak Ridge National Laboratory
Oak Ridge, TN, USA
yoonh@ornl.gov

Abstract— Web data mining has emerged in different domains as a powerful approach to harvest knowledge of unprecedented quantity, comprehensiveness, and diversity. The use of informatics technologies to make the most of online sources is known as cyber-informatics. We propose this concept to support the national cancer surveillance program. Such effort requires the development of novel cyber-informatics algorithms and tools to (i) automatically search disparate online sources for retrieving and integrating related web contents, and to (ii) effectively synthesize this information to automate hypotheses generation and accelerate knowledge acquisition. Developing and supporting such ecosystem will not only modernize the cancer surveillance program but it could also accelerate our understanding of individual differences in people's exposome and their effect on disease prevention, onset, progression, treatment, and survival. We will present representative case studies where we leverage online obituaries as the primary source of information. Obituaries contain important information about the deceased person's age, place of death, cause of death, family status, and even employment. Using advanced web crawling technology to automatically collect obituaries and death notices from openly available webpages of newspapers and funeral homes, a large corpus of cancer patients can be collected for further analysis. Then, using tailored natural language processing and information extraction algorithms the obituary text can be analyzed to derive spatiotemporal cancer mortality patterns and socioeconomic disparities patterns. Due to the considerable computational demands of the text parsing stage, we used the Titan supercomputer of the Oak Ridge Leadership Computing Facility. Using breast and lung cancer as use cases, our study findings were consistent with official cancer surveillance reports. Our studies suggest that non-traditional, openly available online sources can be a valuable information source for epidemiological discovery and validation.

Keywords—*web mining; natural language processing; digital epidemiology; public health informatics; cancer surveillance*