Learning About Health and Medicine from Internet Data

Elad Yom-Tov
Microsoft Research Israel
13 Shenkar st.
Herzeliya 46733, Israel
eladyt@microsoft.com

Ingemar Johansson Cox
University College London
Gower st.
London WC1E 6BT, U.K.
ingemar.cox@di.ku.dk

Vasileios Lampos
University College London
Gower st.
London WC1E 6BT, U.K.
v.lampos@ucl.ac.uk

ABSTRACT

Surveys show that around 70% of US Internet users consult the Internet when they require medical information. People seek this information using both traditional search engines and via social media. The information created using the search process offers an unprecedented opportunity for applications to monitor and improve the quality of life of people with a variety of medical conditions. In recent years, research in this area has addressed public-health questions such as the effect of media on development of anorexia, developed tools for measuring influenza rates and assessing drug safety, and examined the effects of health information on individual wellbeing. This tutorial will show how Internet data can facilitate medical research, providing an overview of the state-of-the-art in this area. During the tutorial we will discuss the information which can be gleaned from a variety of Internet data sources, including social media, search engines, and specialized medical websites. We will provide an overview of analysis methods used in recent literature, and show how results can be evaluated using publicly-available health information and online experimentation. Finally, we will discuss ethical and privacy issues and possible technological solutions. This tutorial is intended for researchers of user generated content who are interested in applying their knowledge to improve health and medicine.

Categories and Subject Descriptors

H.3 [**Information Search and Retrieval**]: Search process; I.2 [**Applications and Expert Systems**]: Medicine and science

Keywords

Medicine, Information Retrieval, Machine Learning, User-Generated Data

WSDM'15, February 2–6, 2015, Shanghai, China.
ACM 978-1-4503-3317-7/15/02.
http://dx.doi.org/10.1145/2684822.2697042

1. OVERVIEW

Pew surveys recently found that more than 70% of US Internet users consult the Internet when they require medical information [3]. The data created by people who thus seek information offers an unprecedented opportunity for applications to monitor and improve the quality of life of people with a variety of medical conditions.

For the past decade, such data has been used in a variety of ways to the betterment of medical research and practice. Internet data has proven useful over traditional tools in several areas: First, it is a more sensitive sensor than that afforded by medical records in diseases where only a minority of infected people seek medical attention [9]. Thus, tracking the rates of influenza-like illness through Internet searches [2, 4, 8] and social media [5, 1, 6] has been the focus of significant research.

Second, Internet data can provide information where the its collection in the physical world would be difficult, if not impossible. For example, the mental state of cancer patients in the first few days following diagnosis has been analyzed using web searches [7], and changes in mood prior to the appearance of abnormal mental states detected [13].

A third area where Internet data is advantageous is where traditional data gathering is dependent on people making associations between different events, which are not always easy to link. Thus, the identification of adverse drug reactions is dependent on the realization of a causal link between taking a drug and the appearance of an adverse reaction. Internet query logs have been used to form this association without the need for voluntary reporting, consequently discovering previously unknown adverse reactions [12].

Finally, Internet data is useful when the activity of interest is primarily web-driven. This includes the actual search for medical information, as well as when studying patient groups. Here, studies have shown the effect of information on users learning about health topics [10] and demonstrated the detrimental effect of well-intentioned interventions for anorexia patients [11].

This tutorial will show how Internet data can facilitate medical research, providing an overview of the state-of-the-art in this area. The tutorial will be roughly divided into three main parts. The first part will discuss areas where Internet data is useful for medical research, and show examples representing the state of the art in the field. The second part will give researchers tools to address medical questions using Internet data. The third part will discuss ethical questions and privacy issues associated with this type of research, and possible technological solutions to the latter.

During the tutorial we will discuss the information which can be gleaned from a variety of Internet data sources, including social media, search engines, and specialized medical websites. We will provide an overview of analysis methods used in recent literature, and show how results can be evaluated using publicly-available health information and online experimentation. Finally, we will discuss ethical and privacy issues and possible technological solutions. This tutorial is intended for researchers of user generated content who are interested in applying their knowledge to improve health and medicine.

Our goal is to familiarize participants with the state of the art of the field, to show common approaches and tools used to address health questions with Internet data, to demonstrate where Internet data is useful (and where it is not) and to discuss open questions in the field.

2. INSTRUCTORS

Elad Yom-Tov is a Senior Researcher at Microsoft Research. Before joining Microsoft he was with Yahoo Research, IBM Research, and Rafael. Dr. Yom-Tov studied at Tel-Aviv University and the Technion, Israel. He has published three books, over 80 papers (of which 3 were awarded prizes), and filed more than 30 patents (19 of which have been granted so far). He is a Senior Member of IEEE and held the title of Master Inventor while at IBM. He has published several papers demonstrating the use of Internet data for learning about medicine, including the detection of side effect of drugs from query logs, the effects of media on the development of anorexia, and the information needs of cancer patients, among others.
URL: `http://www.yom-tov.info`

Ingemar J. Cox is a professor in the Computer Science departments at the University of Copenhagen and University College London (UCL). He is a Fellow of the ACM and IEEE. He is an inventor or co-inventor on 37 patents. Currently his primary research interests are in information retrieval and the application of data mining methods to online resources for medical purposes. He is presently part of a large UK EPSRC Interdisciplinary Research Centre on Early Warning Sensing Systems for Infectious Diseases ("i-Sense"), where he leads research targeting influenza.
URL: `http://mediafutures.cs.ucl.ac.uk/people/IngemarCox`

Vasileios Lampos is a Research Associate at the Computer Science department of UCL and a visiting researcher at Google. Before joining UCL, Dr. Lampos has worked as a Research Associate in the Natural Language Processing (NLP) group at the University of Sheffield. He was awarded his Ph.D. from the University of Bristol under the supervision of Prof. Nello Cristianini. His research interests span from Statistical NLP applications to Computational Social Science; he has published several papers on case studies of Social Media analysis, including the theme of influenza-like illness rates modelling from Twitter data. Currently, he participates in the interdisciplinary EPSRC project "i-Sense", with the general aim to develop a variety of sensing systems for the timely diagnosis and prevention of infectious diseases.
URL: `http://www.lampos.net`

3. REFERENCES

[1] A. Culotta. Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the First Workshop on Social Media Analytics - SOMA '10*, pages 115–122. ACM Press, 2010.

[2] G. Eysenbach. Infodemiology: Tracking flu-related searches on the web for syndromic surveillance. In *AMIA 2006 Symposium Proceedings*, pages 244–248, 2006.

[3] S. Fox and M. Duggan. Health online 2013. *Health*, 2013.

[4] J. Ginsberg, M. Mohebbi, R. Patel, L. Brammer, M. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2009.

[5] V. Lampos and N. Cristianini. Tracking the flu pandemic by monitoring the social web. In *Cognitive Information Processing (CIP), 2010 2nd International Workshop on*, pages 411–416. IEEE, 2010.

[6] V. Lampos and N. Cristianini. Nowcasting Events from the Social Web with Statistical Learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4):1–22, 2012.

[7] Y. Ofran, O. Paltiel, D. Pelleg, J. M. Rowe, and E. Yom-Tov. Patterns of information-seeking for cancer on the internet: an analysis of real world data. *PloS one*, 7(9):e45921, 2012.

[8] P. M. Polgreen, Y. Chen, D. M. Pennock, and F. D. Nelson. Using internet searches for influenza surveillance. *Clinical Infectious Diseases*, 47:1443Ű8, 2008.

[9] E. Yom-Tov, D. Borsa, I. J. Cox, and R. A. McKendry. Detecting disease outbreaks in mass gatherings using internet data. *Journal of medical Internet research*, 16(6), 2014.

[10] E. Yom-Tov and L. Fernandez-Luque. Information is in the eye of the beholder: Seeking information on the MMR vaccine through an internet search engine. In *Proceedings of the 2014 conference of the American Medical Informatics Association*, 2014.

[11] E. Yom-Tov, L. Fernandez-Luque, I. Weber, and P. S. Crain. Pro-anorexia and pro-recovery photo sharing: A tale of two warring tribes. *J Med Internet Res*, 14(6):e151, Nov 2012.

[12] E. Yom-Tov and E. Gabrilovich. Postmarket drug surveillance without trial costs: discovery of adverse drug reactions through large-scale analysis of web search queries. *Journal of medical Internet research*, 15(6), 2013.

[13] E. Yom-Tov, R. W. White, and E. Horvitz. Seeking insights about cycling mood disorders via anonymized search logs. *Journal of medical Internet research*, 16(2), 2014.