Information Processing and Management 52 (2016) 430–445

Contents lists available at ScienceDirect

Information Processing and Management

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

INFORMATION PROCESSING & MANAGEMENT

An International Journal

Semantic search for public opinions on urban affairs: A probabilistic topic modeling-based approach

Baojun Ma[a], Nan Zhang[b,*], Guannan Liu[c], Liangqiang Li[d], Hua Yuan[d]

[a] School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing 100876, PR China
[b] School of Public Policy and Management, Tsinghua University, Beijing 100084, PR China
[c] School of Economics and Management, Tsinghua University, Beijing 100084, PR China
[d] School of Management and Economics, University of Electronic Science and Technology of China, Chengdu 611731, PR China

ARTICLE INFO

Article history:
Received 26 December 2014
Revised 12 September 2015
Accepted 14 October 2015
Available online 14 November 2015

Keywords:
Probabilistic topic modeling
Public opinions
Big data analysis
Semantic search
Latent Dirichlet allocation (LDA)

ABSTRACT

The explosion of online user-generated content (UGC) and the development of big data analysis provide a new opportunity and challenge to understand and respond to public opinions in the G2C e-government context. To better understand semantic searching of public comments on an online platform for citizens' opinions about urban affairs issues, this paper proposed an approach based on the latent Dirichlet allocation (LDA), a probabilistic topic modeling method, and designed a practical system to provide users—municipal administrators of B-city—with satisfying searching results and the longitudinal changing curves of related topics. The system is developed to respond to actual demand from B-city's local government, and the user evaluation experiment results show that a system based on the LDA method could provide information that is more helpful to relevant staff members. Municipal administrators could better understand citizens' online comments based on the proposed semantic search approach and could improve their decision-making process by considering public opinions.

1. Introduction

The spread of Web 2.0 applications enables a new model for the use of the Internet (Burke, 2009). Users act not only as websites' visitors but also as content creators (Ingawale, Dutta, Roy, & Seetharaman, 2013). User generated content (UGC) significantly enriches the amount of online information and makes it more difficult for users to comprehensively understand information through regular reading behavior (Li et al., 2012; Zhu, Mo, Wang, & Lu, 2011). However, it is obvious that to appreciate the significance of UGC, any use of the knowledge underlying UGC must be developed based on a comprehension of its content (Williams, Wiele, Iwaarden, & Eldridge, 2010).

With respect to public administrators, the Web 2.0 environment with multiple exchanges provides a vital opportunity for enhancing interactions between the government and citizens (Horton, 2006). The explosion of online UGC and the development of big data analysis provides a new opportunity and challenge to understand and respond to public opinions (Rhoda & Norman, 2013; Yu-Che & Tsui-Chuan, 2014). In particular, when confronted with urban residents whose problems are comparatively homogeneous, local governments attempt to establish interactive platforms to collect citizens' opinions and recommendations from various perspectives to serve as a foundation for evaluating the performance of the government and to guide future policy

* Corresponding author. Tel.: +86 10 62772746; fax: +86 10 62772746.
E-mail addresses: mabaojun@bupt.edu.cn (B. Ma), nanzhang@mail.tsinghua.edu.cn (N. Zhang), guannliu@gmail.com (G. Liu), langmalee@gmail.com (L. Li), yuanhua@uestc.edu.cn (H. Yuan).

http://dx.doi.org/10.1016/j.ipm.2015.10.004
0306-4573/© 2015 Elsevier Ltd. All rights reserved.

adjustment (Hong, 2013). However, as citizens' enthusiasm for voicing their opinions on the Internet grows, how to understand the information both timely and effectively becomes more significant, which is directly related to whether issues that concern administrators can be addressed and whether the corresponding feedback could be timely offered to citizens (Linders, 2012). Semantic analysis and semantic search technology are likely to become effective tools for assisting the public administrators in the rapid and precise positioning of mass text information.

This study focuses on a platform for acquiring online citizen opinions on urban public affairs issues. B-city has been honored as one of China's international metropolises. To reinforce communication between municipal administrators and citizens, to listen to citizens' opinions and suggestions on urban public affairs in a timely fashion, and to facilitate public participation in urban construction and development, in 2005 B-city's municipal government installed an opinion acquisition module related to urban public affairs onto its official website; the city now receives around 30,000 text messages per year.

Confronted with this non-structured text information, it is difficult to use simple statistical methods and traditional data processing tools to help officials better understand those comments. Before the start of the study, face to average more than 100 comments of daily public feedbacks, the office of the website, only sorts responsibility to the different departments manually, and oversees the feedbacks every day, without any historical data analysis tools. When sometimes need to summarize a period of public opinions, or analyze deeply around one case or one area, the office can only conduct keywords retrieval then manual read and summarize the retrieved results. Obviously, the current manual method is inefficient, even invalid for those tasks which unable to provide accurate keywords initially. Semantic analysis and semantic search could play significant roles in solving this problem.

The paper describes a basic idea for designing a semantic search tool directed to special demands: a framework composed of two sets of procedures—namely, a user search process and a probabilistic topic modeling process—is proposed based on the systemization of literature related to semantic search and semantic analysis to characterize the actual requirements of the online citizen opinion platform. In other words, the foreground procedure facilitates obtaining keywords from search input, provides auxiliary keywords to help searchers determine the theme (when necessary), and derives not only search results but also longitudinal changing curves. The background procedure is a process of preprocessing subject clustering for the comment data based on a probabilistic topic modeling approach called Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003). Each theme that is generated based on probability subject modeling is appropriately viewed as the basic message block that waits constantly to be searched and invoked by the front-end search flow in the database. This study tries to make two aspects of contributions: (1) For this specific practice scenario mentioned in the paper, we provide a set of feasible solution to help the office search the public comments history data according to changeful requirements. In particular, the solution of the study is based on semantics, rather than merely based on keywords, which make the system "smarter" and could adapt to more complex requirements; (2) For academic area, we show a semantic search approach based on the LDA method. Compare to existing research involves the field of semantic modeling (Misra, Yvon, Cappé, & Jose, 2011; Xu, Zhang, & Wang, 2015), we try to show that the LDA could also play a role in the semantic search, and provide some key details of the techniques process including coordination between real-time search and non-real-time LDA calculation, keywords matching and suggestions in the user search process, determination of the number of latent topics.

To validate the usefulness and effectiveness of our proposed semantic search method based on LDA, we conduct a user evaluation experiment to compare with a baseline method, the Keywords-Matching approach (KM). The paper also presents the implications of the results for municipal management.

2. Brief literature review of semantic search

The idea of semantic search, which is understood as searching by meanings rather than literal strings of word and which aims to solve the limitations of keyword-based search models, has been the focus of a wide body of research in both the Semantic Web (SW) and the Information Retrieval (IR) communities (Fernández et al., 2011). An important aspect of semantic search approaches is that almost all of them use conceptual representations of content beyond mere keywords, and many of them also attempt to provide conceptual representations of user needs, as a method of enhancing traditional mainstream keyword-based search technologies.

Early in 2005, Mäkelä (2005) described five of the most-used methodologies in semantic search: (1) Resource Description Framework (RDF) Path Traversal; (2) keyword to concept mapping; (3) graph patterns; (4) logics; and (5) fuzzy concepts, fuzzy relations, and fuzzy logics. Then, Mangold (2007) surveyed and compared 22 different semantic search approaches or projects based on seven dimensions or criteria: architecture, coupling, transparency, user context, query modification, ontology structure and ontology technology. Thereafter, Dong, Hussain, and Chang (2008) conducted a brief survey based on a list of semantic search technologies from six categories: semantic search engines, semantic search methods, hybrid semantic search engines, semantic XML search engines, semantic ontology search engines and semantic multimedia search engines.

As described above, the core of semantic search technologies is the type and use of semantic knowledge representation. Accordingly, most of the semantic search methods in the literature can be distinguished according to the following three categories:

(1) *Linguistic conceptualization approaches* are based on light conceptualizations (usually considering few types of relationships among concepts) and low information specificity levels. For instance, early in 1998, Word Net was used to enhance search performance by considering the semantic relationships among words or concepts (Mandala, Takenobu, & Hozumi, 1998). Urbain, Goharian, and Frieder (2008) have explored unsupervised learning techniques for extracting semantic