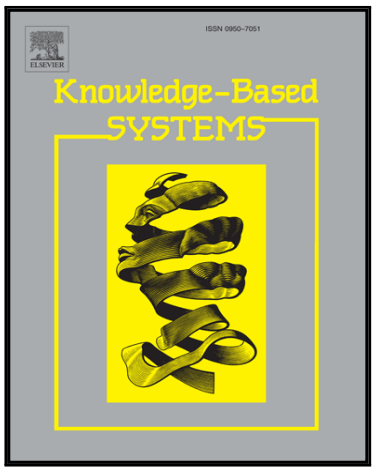

A Survey on Trends of Cross-media Topic Evolution Map

Houkui Zhou , Huimin Yu , Roland Hu , Junguo Hu

PII: S0950-7051(17)30127-2
DOI: 10.1016/j.knosys.2017.03.009
Reference: KNOSYS 3857

To appear in: *Knowledge-Based Systems*

Received date: 19 October 2016
Revised date: 22 February 2017
Accepted date: 7 March 2017

Please cite this article as: Houkui Zhou , Huimin Yu , Roland Hu , Junguo Hu , A Survey on Trends of Cross-media Topic Evolution Map, *Knowledge-Based Systems* (2017), doi: 10.1016/j.knosys.2017.03.009

A Survey on Trends of Cross-media Topic Evolution Map

Houkui Zhou [1,3,4],Huimin Yu[1,2,*], Roland Hu[1], Junguo Hu[3,4]

1Department of Information Science and Electronic Engineering, Zhejiang University,Hangzhou 310027,China

2 State Key Laboratory of CAD & CG,Hangzhou 310027,China

3 School of Information Engineering, ZheJiang A&F University, Linan 311300,China

4 Zhejiang Provincial Key Laboratory of Forestry Intelligent Monitoring and Information Technology, Linan 311300,China

Abstract: Rapid advancements in internet and social media technologies have made "information overload" a rampant and widespread problem. Complex subjects, histories, or issues break down into branches, side stories, and intertwining narratives; a "topic evolution map" can assist in joining together and clarifying these disparate parts of an unfamiliar territory. This paper reviews the extant research on topic evolution map based on text and cross-media corpora over the past decade. We first define a series of necessary terms, then go on to describe the traditional topic evolution map per 1) topic evolution over time, based on the probabilistic generative model, and 2) topic evolution from a non-probabilistic perspective. Next, we discuss the current state of research on topic evolution map based on the cross-media corpus, including some open questions and possible future research directions. The main contribution of this review is in its construction of an evolution map that can be used to visualize and integrate the extant studies on topic modeling – specifically in regards to cross-media research.

Keywords: Cross-media, topic evolution, topic map, probabilistic generative model

1 Introduction

The extensive development of modern information technology has created an information redundancy problem. There currently exists a wealth – an excess, even – of cross-media data having originated from heterogeneous and homogeneous media with multiple sources, such as news media websites, social media websites, photo/video sharing websites, mobile phones, video surveillance servers, and the Internet of Things. Extracting useful knowledge from cross-media big data on the network space is a challenging endeavor. Search engines are generally relied upon for accessing information, and efforts have even been made to create specialized search and retrieval tools (e.g., academic search and news search).

Search engines are indeed effective in retrieving knowledge, but output lists of search results that are unstructured and potentially unhelpful. Establishing a comprehensive and accurate understanding of the "big picture" when it comes to a given topic based on a large mass of text, speech, image, and video information can be nearly impossible. The "five Ws" (who, what, when, where, and why) are a common starting point in building connections between bits of information on a certain topic, as well as the topic's "evolution". Each of the first four W's is necessary to tell the whole "story" of a given topic or event; these answers can be readily extracted from the data available online, but what about the fifth W – the "why" of the topic?

The underlying cause (i.e., the "why") of a given topic is typically considered its most interesting characteristic; it is also the topic's most elusive aspect in terms of completing the topic evolution map. For this reason, the "why" tends to be relegated to an internal cause between different sub-topics in the map. The quantitative relation among sub-topics is highly appealing to the individual seeking the "cause and effect" of the topic. Knowledge is delivered by an array of data modalities (as opposed to a single medium) which represent the same semantics; this is known as

* Corresponding author Email:yhm2005@zju.edu.cn

Project supported by the National Key Basic Research Project of China (No. 2012CB316400)