Applied Soft Computing 66 (2018) 506–511

Contents lists available at ScienceDirect

Applied Soft Computing

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

# Crowdsourcing based scientific issue tracking with topic analysis

Mucheol Kim [a], B.B. Gupta [b], Seunmin Rho [c,*]

[a] Department of Computer & Software Engineering, Wonkwang University, 53, 460, Iksan-Daero, Iksan, Jeonbuk, South Korea
[b] National Institute of Technology Kurukshetra, India
[c] Division of Media Software, Sungkyul University, 53, Sungkyuldaehak-ro, Manan-gu, Anyang, Gyeonggi-do, South Korea

ARTICLE INFO

*Article history:*
Received 31 July 2016
Received in revised form 7 September 2017
Accepted 14 September 2017
Available online 20 September 2017

*Keywords:*
Topic analysis
Scientific data analysis
Web technology
Big data
Information retrieval

ABSTRACT

With the advancement of web technologies, many people are participating in the information production and distribution process in the Web environment. In addition, many researchers have been interested in research on refining useful information using topic based recommendation system because the amount and complexity of web information is rapidly increasing. The proposed approach performs typical scientific data collection and then analyzes seed problem keywords using multi-level documents based on crowd sourcing. We then used the LDA algorithm to create a cluster of scientific themes to generate issue keywords that are responsive to the scientific trend issues. As a result, our approach suggests a methodology for recommending clusters of related issues when scientific issues are raised in each context.



## 1. Introduction

Due to the proliferation of smart devices and the growth of network technology, everyone could be accessing web data easily 1,36]. Then many people are often utilizing the web data with educational or personal purposes [2,3]. Furthermore everyone is taking an active part in the information production and distribution process, then the volume and complexity of web data are increasing rapidly [4,5]. Big data refers to structured or unstructured data which may not be collected or analyzed with the conventional methodologies [6]. In fact, big data technology is used in optimizing traffic volume and it is improving administrative efficiency. In big data era, however, there are too many data to handle by individuals. Therefore, many people are use the web data immediately, furthermore it is often disappeared without being used properly. Then, there are lots of data which are already meaning as they are or that can become valuable after being processed a little. In particular, the generated data from social network services contains emotional or communication information relating to users daily lives [7]. Since users share news through text messages, they are already valuable as information. Hence, a study aimed to analyze and manage big data in order to solve the social issues has become more important [8].

On the other hand, crowdsourcing is the emerging cultural concept including the process by which individuals are interested in a specific domain. Then they would generate not only knowledges, but also goods and software. It is also a new business model that harnesses creative and collaborative solutions in the web environment [9]. The concept of crowdsourcing may be defined as stated below:

Typical information is generated by everyone in web environments, then it could realize the collective intelligence [4]. It has given body to innovative technologies for learning, based on human resources in web 2.0 generation [10,11]. Although the web is not a learning technology by itself, it promotes a learning culture by various efforts of contributors [12]. Then, we could generate significant knowledge management model with the crowdsourcing concept. Especially, we can extract not only explicit contexts from various people, but also implicit knowledge from deduction with text mining and social network analysis [13,14].

Scientific data such as papers and reports are focused on suggesting solutions to various natural phenomena. However, in order to cope with real-time issues occurring in the real world, the amount and complexity of data make it difficult to use. Nonetheless, because of the characteristics of Vogue and reliability, scientific data is still attracting much attention as an information analysis tool for solving social problems. Therefore, in order to derive a solution method for various social phenomena, it is necessary to have an information analysis method that can provide information using scientific data that can cope with social issues. Therefore, in order to derive a solution method for various social phenomena, it is nec-

* Corresponding author.
*E-mail addresses:* mucheol.kim@gmail.com (M. Kim), gupta.brij@gmail.com (B.B. Gupta), smrho@sungkyul.edu (S. Rho).

https://doi.org/10.1016/j.asoc.2017.09.028
1568-4946/

essary to have an information analysis method that can provide information using scientific data that can cope with social issues.

This paper proposes a multi-layered information analysis method that reflects the crowd sourcing concept to generate focused topic group. The proposed method uses a combination of data generated for different methods and purposes in order to overcome the disadvantage that existing topic analysis methods generate unspecific topic groups for distributed topics.

In this paper, we collect science and social data from the web in real time. On the other hand, for the creation of focused topic group, science trend reports are collected and analyzed and used as seed data. Finally, we suggests the contextual issue tracking services when scientific issues are meeting to social phenomenon.

In this paper, Section 2 presents related works, and Section 3 describes the proposed scientific social issue tracking approach with topic analysis. In Section 4, experimental analysis and concluding remarks in Section 5.

## 2. Related work

Recently, many researchers are interested in extracting the context from typical web information. Some researches related to social network analysis are focused on the extracting the relationship between web documents [15]. They were utilizing the various properties such as personal information, relationships between contributors [16,17]. Kuada and Olesen, [18] proposed the provisioning and management approach based on collaborative strategy with social relationships in cloud computing services. [19] suggested the personalized recommender systems with social networks and collaborative filtering.

On the other hand, amount of topic analysis researches are applied to information retrieval and recommender systems. LDA (Latent Dirichlet Allocation) which is the issue extracting algorithm based on the probability model is used for extracting topics from social data and news data [20]. "Google Flu Trends" is the most favorite example of the topic analysis applications [21]. Some researches focused on the R&D issue tracking with scientific data which is related to research projects [22]. They should be dealing with advanced factors which are classifications of subject, technology, application for enhanced analysis results [23,24]. Xu et al. [25] suggested the candidate set for the proper opportunity of participating the R&D projects with their matchmaking algorithm. Wang et al. [26] proposed the social media recommendation approach with analyzing link structures between users and groups. Qian et al. [27] suggested the social event tracking and recommendation with event summary. Lessel et al. [28] proposed the issue tracking concept and its approach for deviations and disturbances in a manufacturing production environment.

## 3. Scientific social issue tracking with enhanced topic analysis

This section describes a crowdsourcing based topic analysis approach that can utilize the relationship between different level scientific data. First, we analyze seed data which is possible to identify the recent major issues periodically in diverse science and technology fields. It is written by governmental institutes or reputable researchers. Secondly, we analyze topic clusters using LDA algorithm. We clustered explicit or implicitly related terms from the seed data into issue keywords. In this process, we performed natural language processing as preprocessing for text analysis. Then it could be used to generate a word-document matrix. As a result, we could derive the topic analysis results and the contextual direction for decision support when the science problem was related to the social issue trend.

**Table 1**
The Concept of crowdsourcing.

| |
|---|
| Crowdsourcing represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined network of people in the form of an open call. This can take the form of peer production, but is also often undertaken by sole individuals. The crucial prerequisite is the use of the open call format and the large network of potential laborers. |

### 3.1. TF-IDF weighting for topic analysis

TF-IDF is the most common weighting method used to describe documents in the Vector Space Model. The TF-IDF function weights each vector component of each document on the following basis [29]. First, it incorporates the word frequency in documents. Thus, the more words appear in a document, it means that the word is more significant in the document. Meanwhile, IDF measures how infrequent a word is in a collection. It is estimated the degree of uniqueness through all documents. In this paper we assumed that each science and social news documents are consisted of typical terms which represents the identity for each document (Eq. (1)).

$$a_i = \{t_1, t_2, t_3, \ldots, t_n\} \tag{1}$$

TF means the frequency in each documents, then we calculated the generalized term frequency with maximum number of frequency in each documents.

$$TF(t_n, a_i) = \frac{ft_n, a_i}{\max(ft_n, d : t_n \in a_i)} \tag{2}$$

We also generated the IDF with traditional methodology which is counting the frequency of documents for each terms.

$$IDF(t_n, D) = \log \frac{N}{|\{a_i \in D : t_n \in a_i\}|} \tag{3}$$

As a result, we could deduct the TFIDF value with the harmony of TF and IDF for evaluating the term weight.

$$TFIDF(t_n, a_i, D) = TF(t_n, a_i) \times IDF(t_n, D) \tag{4}$$

### 3.2. Seed issue tracking with pilot analysis

This section explains the issue tracking process with pilot analysis. We collected the data-set from the 'Yearly Excellence National R&D 100′ in Korea, then we deducted various research issue keywords in each year [30]. They have selected by Korean governments and generated the case book with the excellent outcomes of research and technology through all research fields. We performed pilot analysis with the documents in selected research areas which are 'Life and Ocean Science' and 'ICT Technology' as you can see (Tables 1 and 2). We performed the crowdsourcing based pilot analysis with specific documents namely 'Yearly Excellence National R&D 100′ and it could be seed data-set of each specific research subjects.

In research fields related to 'Life and Ocean Science', there are typical scientific issues such as 'Cancer', 'Ion', 'Stem Cells'. Furthermore there are several significant research issues which are 'Graphene', 'Services', 'Mobile' in the documents related to ICT Technology. They could be the seed keywords for collecting the journal articles for tracking scientific research issues.

### 3.3. Topic modeling with LDA algorithms

In this paper, we generated topic models with LDA algorithms (Figs. 1–3) [31]. It could generate the typical clusters with Dirichlet distributions. It should be summarizing the documents quickly, and finding issue topics through all documents.

For LDA model, the number of topics $k$ have to be fixed a-priori. The LDA model assumes the following generative process for a