A FCA-based Framework for Discovering Hidden Knowledge from Twitter Content

Jeong-Dong Kim\(^1\) and Suk-Hyung Hwang\(^2\)

\(^1\)Department of Computer Science and Engineering, Korea University, Anam-Dong Sungbuk-Gu, Seoul, Republic of Korea
\(^2\)Department of Computer Science & Engineering, Sunmoon University, Kalsong-Ri, Tangjeong-Myoon, Asan-Si, Chungnam, Republic of Korea

\(^1\)kjd4u@korea.ac.kr, \(^2\)shwang@sunmoon.ac.kr

Abstract

Web data mining is a hot research topic that is a technique used to crawl through various web resources to collect required information. As the importance of such web data mining has been recognized, extensive studies have been conducted actively to analyze the data in a Social Networking Service (SNS). In a SNS, a large amount of data, which has a variety of characteristics, is generated through voluntary participation of users, which is also called “big social data”. Big social data can identify not only content registered on the web but also the relations of the friends of users. In this paper, we introduce Formal Concept Analysis (FCA) as the basis for a practical and well-founded methodological approach for web data analysis which identifies conceptual structures among data sets. As well as, we propose a framework for discovering hidden knowledge by using polarity from Twitter contents. Additionally, we show the experiments that demonstrate how our framework can be applied for knowledge discovery.

Keywords: Data Mining, Opinion Mining, Formal Concept Analysis, Social Networks, Knowledge Discovery

1. Introduction

Web data mining is a hot research topic that is a technique used to crawl through various web resources to collect required information. This collected information is used to gain more knowledge and based on the findings and analysis of the information make predictions as to what would be the best choice and the right approach to more toward on a particular issue [1].

As the importance of such web data mining has been recognized, extensive studies have been conducted actively to analyze the data in a Social Networking Service (SNS) [2-5]. A SNS is a platform to build social networks or social relations among users that is able to generated and shared a large volume of information by real time [6]. Most SNS are web-based and provide means for users to interact over the internet. As SNS continue to increase, more contents are created by users. Therefore, In a SNS, a large amount of data, which has a variety of characteristics, is generated through voluntary participation of users, which is also called “Big Social Data” [7].

Big social data can identify not only content registered on the web but also the relations of the friends of users [1, 5]. That is, big social data created in a SNS can be divided largely into “relational information between people” and “contents created by users”. In particular, “relational information between people”, which is the unique feature of a SNS, is highly appropriate information for personalized service, which can be utilized in search and recommendation services by means of the information. Thus, data generated via a SNS contains richer information than existing web data, which requires more complex data analysis process [1].
As a result, the SNS contents (such as big data) has emerged as a new issue, and also when the user's social network is available, the preferences of the user's related people can be utilized to assist in obtaining the user's preferences, assuming closely related people have similar interests. This is the main assumption when user interests and preferences are predicted based on the preferences of similar persons [6].

In a mountain of social contents, it is becoming more difficult for user to identify contents users are interacted in. To solve this problem, we introduce Formal Concept Analysis (FCA) as the basis for a practical and well founded methodological approach for web data analysis which identifies conceptual structures among data sets [8, 14]. The FCA is a method mainly used for the analysis of data, i.e., for investigating and processing explicitly given information. The FCA classifies data based on the ordinary set into concept units which consists of objects and attributes that those objects have commonly. That is, FCA extracts formal concept from a given data table, grasps conceptual structures between concepts, and constructs conceptual hierarchy. The FCA has been applied to various domains, such as medicine, bioinformatics, social science, data mining, ontology and software engineering, and others.

In this paper, we propose a framework for discovering hidden knowledge by using polarity from social contents. Additionally, we show some experiments that demonstrate how our framework can be applied for knowledge discovery.

2. Framework

A proposed framework in this paper consists of two mining modules largely. The first module is to analyze polarity of web data using the SentiWordNet [9] and the second module, which is a FCA-based analysis module, is to discover new knowledge through polarity of web data analyzed earlier.

2.1. Polarity Analysis

A polarity analysis, we can use the sentiWordNet [9] that is a lexical resource for opinion mining that is associated to three sentiment scores in which each WordNet synset [10]. In other words, the method relies on training a set of ternary classifiers, each of them capable of deciding whether a synset is Positive, or Negative, or Objective. In address, the SentiWordNet is to aid opinion mining by providing an off the shelf lexical resource that provides a granular level of opinion tags for a large set of words [15].

The SentiWordNet method relies of deciding whether a synset is Positive, or Negative, or Objective. Each ternary classifier differs from the other in the training wet used to train it and in the learning device used to train it, thus producing different classification results of the WordNet synsets [10, 11].

Each ternary classifier is generated using the semi-supervised method described in Esuli and Sebastiani [12]. A semi-supervised method is a learning process whereby only a small subset $L \subseteq T_r$ of the training data $T_r$ have been manually labelled. In origin the training data in $U \equiv T_r - L$ were instead unlabelled; it is the process itself that has labelled them, automatically, by using $L$ (with the possible addition of other publicly available resources) as input.

[Definition 1] The SentiWordNet method defines $L$ as the union of three seed (i.e. training) sets $L_p$, $L_n$ and $L_o$ of known Positive, Negative and Objective synsets, respectively.

$L_p$ and $L_n$ are two small sets, which we have defined by manually selecting the intended synsets for 14 “paradigmatic” Positive and Negative terms (e.g. the Positive term good, nice, excellent, positive, fortunate, correct, and superior; the Negative term bad, nasty, poor, negative, unfortunate, wrong, and inferior) which were used as seed terms in Turney and Littman [13]. Definition 1 have approached the problem of determining the orientation of terms by bootstrapping from a pair of two minimal sets of seed terms.
2.2. Formal Concept Analysis (FCA)

The FCA is a method mainly used for the analysis of data, *i.e.*, for investigating and processing explicitly given information. Such data are structured into units which are formal abstractions of concepts of human thought allowing meaningful comprehensible interpretation. FCA was introduced as a mathematical theory modeling the concept ‘concept’ in terms of lattice theory [8, 14]. This approach arose independently of ontologies, resulting in a different formalization of concepts. The FCA consists formal context, formal concept, and concept lattice.

2.2.1. Formal Context: FCA starts with a formal context that is comprised of a set of objects, a set of attributes and a relation describing which objects possess which attributes. In the formal definition, the set of objects is denoted by \( O \), and the set of attributes is denoted by \( A \).

[Definition 2] A formal context is a triple \((O, A, R)\), where \( O \) is a set of objects and \( A \) is a set of attributes, and \( R \subseteq O \times A \) is a binary relation between \( O \) and \( A \). In order to express that an object \( o \) is in a relation with an attribute \( a \), we write \((o, a) \in R\) and read it as the object \( o \) has the attribute \( a \).

Table 1 shows a formal context of some animals that is based on the set of objects \( O \) and the set of their attributes \( A \) as follows:

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Objects</th>
<th>four-legged</th>
<th>land-living</th>
<th>water-living</th>
<th>livestock</th>
<th>pet</th>
</tr>
</thead>
<tbody>
<tr>
<td>set-turtle</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dog</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>cat</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>cow</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
</tbody>
</table>

2.2.2. Formal Context: The central notion of FCA is the formal concept. Objects from a context share a set of common attributes and vice versa. Concepts are pairs of objects and attributes which are mapped into each other by the Galois connection. The set of all concepts of the context \( C = (O, A, R) \) is denoted by \( B(C) \) or \( B(O, A, R) \), *i.e.*, \( B(C) = \{ (X, Y) \in 2^O \times 2^A \mid X = \text{extent}(Y) \land Y = \text{intent}(X) \} \). For example, all concepts of the context of Table 1 are as follows: \( \{ \text{sea-turtle, dog, cat, cow} \}, \{ \text{four-legged} \}, \{ \text{dog, cat, cow} \}, \{ \text{four-legged, land-living} \}, \{ \text{dog, cat} \}, \{ \text{four-legged, land-living, pet} \}, \{ \text{cow} \}, \{ \text{four-legged, land-living} \} \).
The set of formal concepts is organized the partial ordering relation $\leq$ to be read as is a sub-concept of as follows:

**[Definition 4]** For a formal context $C = (O, A, R)$ and two concepts $c_1 = (O_1, A_1)$, $c_2 = (O_2, A_2) \in B(C)$ the sub-concept/super-concept relation is given by $(O_1, A_1) \subseteq (O_2, A_2) \iff O_1 \subseteq O_2 \land A_1 \supseteq A_2$.

### 2.2.3. Formal Concept Lattice:
A relationship shows that the dualism exists between attributes and objects of concepts. A concept $c_1 = (O_1, A_1)$ is a sub-concept of concept $c_2 = (O_2, A_2)$ iff the set of its objects is a subset of the objects of $c_2$. Or an equivalent expression is iff the set of its attributes is a superset of the attributes of $c_2$. That is, a sub-concept contains fewer objects and more attributes than its super-concept. The set of all formal concepts of a context $C$ with the sub-concept/super-concept relation is always a complete lattice, called the (formal) concept lattice of $C$ and denoted by $L := (B(C), \leq)$. Figure 1 shows the concept lattice for the Animal context of Table 1.

![Concept Lattice for the Animal Context of Table 1](image)

A concept lattice (a concept hierarchy) can be represented graphically using line diagrams (such as Hasse diagrams). These structures are composed of nodes and links. Each node represents a concept with its associated intentional description. The links connecting nodes represent the sub-concept/super-concept relation among them. This relation indicates that the parent's extension is a superset of each child's extension. Attributes propagate along the edges to the bottom of the diagram and dually objects propagate to the top of the diagram. More abstract or general nodes occur higher in the hierarchy, whereas more specific ones occur at lower levels. Herein, we can summarize the above considerations as a brief algorithm to construct concept lattice in Algorithm 1.
Algorithm 1. Generate Concepts and Build Concept Lattice

1: **INPUT**: a formal context \( C := (O, A, R) \)
2: **OUTPUT**: concept Lattice \( L := (B(C), E_{\leq}) \)
3: for all \( o \in O \) do
4: \( B(C) \leftarrow B(C) \cup (\text{extent}(\text{intent}(o)), \text{intent}(o)) \);
5: end for
6: for all \( c \in B(C) \) do
7: for all \( o \in (O - \text{extent}(c)) \) do
8: \( X \leftarrow \text{extent}(c) \cup \{o\} \);
9: if \( (\text{extent}(\text{intent}(X)), \text{intent}(X)) \notin B(C) \) then
10: \( B(C) \leftarrow B(C) \cup (\text{extent}(\text{intent}(X)), \text{intent}(X)) \);
11: end if
12: end for
13: for all \( c_1 \in B(C) \) do
14: for all \( c_2 \in B(C) - \{c_1\} \) do
15: if \( (c_1 \leq c_2) \land (\exists \, c_3 \in B(C) - \{c_1, c_2\} \mid (c_1 \leq c_3) \land (c_2 \leq c_3)) \) then
16: \( E_{\leq} \leftarrow E_{\leq} \cup \{(c_1, c_2)\} \);
17: end if
18: end for
19: end for

2.3. Proposed Framework

In this paper, the proposed framework consists of the following three regions: 1) Crawling Web Resources, 2) Opinion & FCA Analysis, and 3) Knowledge Discovery. Figure 2 shows the conceptual model of the newly proposed framework for discovering hidden knowledge.

- **Crawling Web Resources**: In this region, the actual social dataset (such as social relations and social contents) are collecting and saving. And we make our own Opinion Word Dictionary (OWD) expanding the SentiWordNet.

- **Opinion and FCA Analysis**: This region defines the PA and FCA. The PA is generated by retrieved contents through given terms for each domain. Retrieved contents are analyzed its polarity using Opinion Word Dictionary. Section 2.1, defined and its explained more detail.
• **Knowledge Discovery**: This region refers to the area where the existing FCA is extended to discover. To begin with, the FCA verifies the formal context, formal concept, and formal concept lattice for discovering hidden knowledge. Section 2.2, defined and its explained more detail.

3. **Experiment and Result**

In this section, we describe crawled dataset and experiments set in Twitter and then show the result of experiment for knowledge discovering.

3.1. **Experiment**

In the experiment, we use crawled contents of Korean Twitter dataset and user relation data during one month and ranging from July 1, 2012 to July 31, 2012. Whole data size of contents is about 93.1GB, and data size of user is 12.3GB. We saved contents data using JSON data style and user data is saved by simple text style.

First of all, we crawled contents related to smartphone domain such as Iphone, Galaxy, Optimus, Vega, Blackberry, and HTC. Then we find all user who has contents in our crawled data. Among them, we experiment for 100 active users who has many contents among 21,640 users related to smartphone domains. The Table 2 summarize the experiments dataset.

<table>
<thead>
<tr>
<th>Table 2. Dataset of Twitter Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Timescale</td>
</tr>
<tr>
<td>Domain</td>
</tr>
<tr>
<td>Smartphone</td>
</tr>
<tr>
<td>Items</td>
</tr>
<tr>
<td>Iphone, Galaxy, Optimus, Vega, Blackberry, HTC</td>
</tr>
<tr>
<td>Number of contents</td>
</tr>
<tr>
<td>259,176</td>
</tr>
<tr>
<td>All Users</td>
</tr>
<tr>
<td>21,640</td>
</tr>
<tr>
<td>Active Users</td>
</tr>
<tr>
<td>10</td>
</tr>
</tbody>
</table>

Twitter contents are analyzed by its polarity using OWD. We make our own OWD expanding the SentiWordNet. Using SentiWordNet, we find representative Korean vocabulary representing Positive and Negative, then add and modify vocabulary for smartphone domain. The OWD consists of 384 positive words and 509 negative words. In the twitter dataset, content whose opinion is ambiguous was classified into “Neutral”. Thus, twitter contents are classified into positive, negative, and neutral in this experiment. Table 3 shows a part of the OWD for Korean polarity analysis in smartphone domain.

The present experiment was conducted with regard to 100 active users who mentioned smartphone-related terms the most in the twitter contents. However, Table 4 showed a "formal context” created with regard to only five users. The reason for this is because “concept lattice” created with regard to 100 users could generate too many concepts and attributes, resulting in too much complexity of diagram which cannot be identifiable with human’s eyes. The overall experiment results of clustering and classifying for hidden knowledge are summarized as shown in Figure 4.

<table>
<thead>
<tr>
<th>Table 3. OWD for Smartphone Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polarity</td>
</tr>
<tr>
<td>Positive</td>
</tr>
<tr>
<td>간편하다, 감동받다, 강렬하다, 갖고 싶다, 가능하다, 좋다, 편하다, 잘되다, 건고하다, 고급스럽다, and so on</td>
</tr>
<tr>
<td>384</td>
</tr>
<tr>
<td>Negative</td>
</tr>
<tr>
<td>느리다, 좋지 않다, 어렵다, 귀찮다, 나쁘다, 오작동된다, 까다롭다, 깨끗하지 않다, 까지지 안한다, and so on</td>
</tr>
<tr>
<td>509</td>
</tr>
</tbody>
</table>
Table 4. Formal Context of Twitter Dataset

<table>
<thead>
<tr>
<th>Terms</th>
<th>Users</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>163299788</td>
</tr>
<tr>
<td>Iphone</td>
<td>x</td>
</tr>
<tr>
<td>Galaxy</td>
<td>x</td>
</tr>
<tr>
<td>Optimus</td>
<td>+</td>
</tr>
<tr>
<td>Vega</td>
<td>+</td>
</tr>
<tr>
<td>Blackberry</td>
<td>+</td>
</tr>
<tr>
<td>HTC</td>
<td>+</td>
</tr>
</tbody>
</table>

In Table 4, the analysis terms by using the formal context is shown of Twitter dataset. The meaning of the symbols (such as (+: positive), (−: negative), and (N: neutral)) signified a polarity from Twitter terms in their relations. Hasse diagrams, figure 3 shows the concept lattice for the Twitter dataset context of Table 4. According to the Figure 3, we can easily discover the hidden knowledge.

3.2. Result

In this paper, formal contexts were generated as “users: concepts” and the polarity of their attributes in the twitter contents were analyzed thereby extracting an Association Rules (AR) from the concept lattice created earlier. “users:concepts” of five persons were experimented thereby discovering 32 ARs in total: among them, only 19 ARs with 100% confidence, two ARs with 80%, three ARs with 75%, four ARs with 67%, three ARs with 50%, and one AR with 0% confidence rate. In this paper, knowledge discovery is defined as ARs with 100% confidence among the ARs extracted from the concept lattice.

Figure 4 shows 19 ARs experimented with regard to “users:concepts” of five persons. No. 2 AR in Figure 4 shows that the number of objects that satisfy the {Iphone+, IphoneN, Galaxy+, Galaxy-, GalaxyN, Optimus-, OptimusN, VegaN, Blackberry+, BlackberryN, HTCN} attributes is one {163299788} while object {163299788} must have {Iphone-} attribute. That is, No. 2 AR in Figure 4 means that every object(user: 163299788) possessing the attributes {Iphone+, IphoneN, Galaxy+, Galaxy-, GalaxyN, Optimus-, OptimusN, VegaN, Blackberry+, BlackberryN, HTCN} also has attribute {Iphone-}. For one more example, No. 18 AR in Figure 4 showed that the number of objects that have {IphoneN Galaxy+ GalaxyN HTCN} attributes is three while objects {163299788, 282578719, 612435148} have attribute {VegaN} mandatorily.
Figure 4. Results of Knowledge Discovery from Table 4: Association Rules

Figure 5 shows the number of “Concepts” that can be created when users are 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100, and the result of AR extraction with 100% confidence.

Copyright © 2016 SERSC
In Figure 5, the number of "Concepts" that can be created with 10 users is 36, in which 30 ARs were discovered. In addition, 621 "Concepts" were generated when the number of users was 100 while the number of ARs extracted with 100% confidence was 121.

4. Conclusions

This paper presented the framework for discovering hidden knowledge which identifies conceptual structures among social contents such as Twitter. The proposed framework consists of two mining modules: The first module is to analyze polarity of web data using the extending SentiWordNet (Opinion Word Dictionary) and the second module, which is a FCA-based analysis module, is to discover new knowledge through polarity of web data analyzed earlier. A key feature of proposed framework is that supports clustering and extracting hidden knowledge using the polarity of terms from social relations among users. Additionally, we showed the experiments that demonstrate how our framework can be applied for knowledge discovery from Twitter dataset. Next, we plan to improvise the experiment to realize various domain.

Acknowledgments

This research was supported by the basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education, Science and Technology (2011-0025588), and Next-Generation Information Computing Development Program through the NRF funded by the Ministry of Science, ICT & Future Planning (2012M3C4A7033346).

References


Authors

**Jeong-Dong Kim** received his MS degree in Computer Science and a PhD degree in Computer Engineering, in 2008 and 2012, respectively, from Korea University, Seoul, Korea, where he is currently Research Professor in the Department of Computer Science and Engineering. His research interests include learning management systems, metadata-based integration, Semantic Web, ontology, social computing, healthcare analysis and access control.

**Suk-Hyung Hwang** received a BS degree in Computer Science from Kangwon National University, Korea, in 1991, and ME and PhD degrees in Information and Computer Science from Osaka University, Japan, in 1994 and 1997, respectively. He is currently Professor in the Department of Computer Science and Engineering, Sun Moon University, Korea. His research interests include object-oriented analysis, design and programming, ontology, semantic web, and formal concept analysis, *etc.*