Finding topics in email using formal concept analysis and fuzzy membership functions

Geng, Liqiang; Korba, Larry; Wang, Yunli; Wang, Xin; You, Yonghua

Publisher's version / Version de l'éditeur:
https://doi.org/10.1007/978-3-540-68825-9_11
Advances in Artificial Intelligence: 21st Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2008 Windsor, Canada, May 28-30, 2008 Proceedings, Lecture Notes in Computer Science; no. 5032, pp. 108-

National Research
Council Canada

Institute for
Information Technology

Conseil national
de recherches Canada

Institut de technologie
de l'information

NRC · CNRC

*Finding Topics in Email Using Formal Concept Analysis and Fuzzy Membership Functions **

Geng, L., Korba, L., Wang, Y., Wang, X., You, Y.
May 2008

* published at The 21st Canadian Conference on Artificial Intelligence (CAI 2008). Windsor, Ontario, Canada. May 28-30, 2008. NRC 50327.

Canada

Finding Topics in Email Using Formal Concept Analysis and Fuzzy Membership Functions

Liqiang Geng[1], Larry Korba[1], Yunli Wang[1], Xin Wang[2], Yonghua You[1]

[1]Institute of Information Technology, National Research Council of Canada
Fredericton, New Brunswick, Canada
{liqiang.geng, larry.korba, yunli.wang, yonghua.you}nrc-cnrc.gc.ca

[2]Department of Geomatics Engineering, University of Calgary
Calgary, Alberta, Canada
xcwang@ucalgary.ca

Abstract. In this paper, we present a method to identify topics in email messages. The formal concept analysis is adopted as a semantic analysis method to group emails containing the same keywords to concepts. The fuzzy membership functions are used to rank the concepts based on the features of the emails, such as the senders, recipients, time span, and frequency of emails in the concepts. The highly ranked concepts are then identified as email topics. Experimental results on the Enron email dataset illustrate the effectiveness of the method.

1 Introduction

Email is one of the most important communication and information exchange tools for modern organizations. It has greatly improved work efficiency. However, with the increasing use of email system, managing emails efficiently becomes an important issue. Identifying activities in emails is a technique to address this issue. At the individual level, identifying topics in emails can facilitate the access to the email messages for users. At the organization level, analyzing the emails in terms of topics can help discover the real workflow within that organization.

Some work has been conducted on identifying activities embodied in emails. The major techniques adopted to tackle this problem are data mining and machine learning methods, such as classification and clustering. Huang et al. proposed to use clustering methods to infer activities from emails based on the subjects and body of the emails [3]. This is one of the first work in this domain. Li et al. incorporated semantic analysis and named entity to email clustering [6]. Khoussainov and Kushmerick combined the relation identification and the speech act classification to improve the performance of email topic identification [4]. Dredze et al. proposed a method to classify emails into activities based on the people involved in the activities and the content of the email messages [1, 2]. Kushmerick and Lau tried to identify a more structured workflow from an email dataset of e-commerce transactions [5].

However, there are some limitations of the above-mentioned classification and clustering based topic identification methods. First, classification method requires that email messages be labeled before training the classification models, which is a not a trivial task, especially without domain knowledge. Secondly, both clustering and classification methods (except hierarchical clustering) produce a single partition on the data set, i.e., one email message can only belong to one class or cluster. Therefore, an email regarding both "trip" and "meeting" will be either assigned to topic "trip" or topic "meeting". Thirdly, the current work did not distinguish between a topic class and a topic instance. Topic class can be considered as an abstract topic, while a topic instance is a concrete topic. For example, topic class "trip" may contain many instances, one of which could be "Smith's trip to Toronto in May 2007". Classification and clustering methods may achieve good results on identifying different topic classes, but may not be able to distinguish between instances of the same topic class. Lastly, the granularity level of clustering methods, i.e., the number of clusters, is not easy to determine.

In this paper, we use the formal concept analysis (FCA) method to organize the groups of emails in a concept hierarchy. Since FCA groups data based on subsets of features, it has the potential to distinguish between the instances of the same topic class. Also FCA assigns an email message to multiple groups, which allows users to view an email from different perspectives. The groups of emails identified by FCA, which are called concepts, are potential topics. We then use fuzzy membership functions to rank and filter the concepts to find the topic instances. From now on, we will use topic and topic instance interchangeably throughout this paper.

The rest of the paper is organized as follows. In Section 2, we describe the method to identify topics from emails. In Section 3, we present the experimental results on Enron data set. In Section 4, we conclude the paper and discuss some future work.

2 Finding Topics in Emails

We first find concepts in emails using formal concept analysis and then rank them according to the likelihood that they each represent a single topic. We talk about the formal concept analysis in Sections 2.1. Then we present the email topic identification method in Section 2.2.

2.1 Formal Concept Analysis

Formal Concept Analysis (FCA) is a mathematical method for data analysis, knowledge representation, and knowledge visualization [7]. It is similar to clustering in terms of grouping similar objects together, but it generates much more clusters since it views objects from different perspectives, i.e., an object can belong to different clusters rather than belong to only one. For example, an ostrich can be classified as an animal that can not fly. It also can be classified as a bird.

The basic idea of FCA is to extract concepts consisting of similar objects and their common features/attributes from a data table and build a hierarchy according to the generality of the concepts.

The input of the FCA is a two-dimensional table called *formal context*. Each row in the table represents an object. Each column represents an attribute. If an object has an attribute, we put value 1 in the cell in the intersection in the table. Otherwise, we put 0 in the cell. Formally, a formal context is a triplet (O, A, R), where O represents the universe of objects, A represents the universe of attributes, and $R \subseteq O \times A$ is a binary relation between O and A. We define two mappings: f: 2^O->2^A and g: 2^A->2^O as follows. Given a set of objects $O_1 \subseteq O$, $f(O_1) = \{a \in A \mid \text{for any } o \in O_1, (o, a) \subseteq R\}$. Given a set of attributes $A_1 \subseteq A$, $g(A_1) = \{o \in O \mid \text{for any } a \in A_1, (o, a) \subseteq R\}$. The mapping f finds all the attributes shared by the objects in O_1, while the mapping g finds all the objects that share the attributes in A_1.

With a formal context and the two mappings being defined, FCA can extract formal concepts from the formal context. A *formal concept* is represented as a pair (E, I) such that $E \subseteq O$, $I \subseteq A$, $f(E) = I$, and $g(I) = E$. In other words, (E, I) is a concept if and only if the objects in E only share the attributes in I and the attributes in I are only shared by the objects in E. E and I are called *extent* and *intent* of the concept, respectively. Given a set of concepts, we define a partial order relation $\leq$. For two concepts (E_1, I_1) and (E_2, I_2), we say that $(E_1, I_1) \leq (E_2, I_2)$ if and only if $E_1 \subseteq E_2$ holds. Equivalently, we say $(E_1, I_1) \leq (E_2, I_2)$ if and only if $I_2 \subseteq I_1$ holds. (E_1, I_1) represents a more specific concept than (E_2, I_2), therefore (E_1, I_1) is called a *sub concept* of (E_2, I_2) and (E_2, I_2) a *super concept* of (E_1, I_1).

A complete lattice can be generated based on the relation $\leq$, with each node representing a concept and each arc representing a direct partial relation. This lattice is called *formal concept lattice* or *Galois lattice*.

2.2 Identifying Topics of Emails from Concepts

We first use the FCA to group emails into concepts based on their subjects or content. Preprocessing is needed to transform email corpus into a two dimensional table. Each column in the table represents a keyword and each row represents an email. Then the FCA identifies a set of concepts, which can be considered as potential topics, and generates the concept hierarchy. The user can explore the concept hierarchy to find the concepts which correspond to topics.

One of the advantages of FCA based topic detection is that it can find topics at different levels of granularity. The FCA can also assign an email to different concepts / topics. The disadvantage is that it may produce huge numbers of concepts for a large data set which may overwhelm the users. Also many concepts in a concept hierarchy may not correspond to a topic. For example, a concept regarding *trip* may involve several topics about different cases of trips rather than only one topic. Although some work on visualization has been done to assist the users to explore the concept hierarchy, it is still a burden for the users to find interesting and meaningful topics in the hierarchy. Therefore, besides the navigation functionality intrinsic to the FCA, we need more functionality to facilitate user's explorations and analysis.

In email messages, besides subjects and content, other information can be used as indicators to show how likely a concept really represents a topic. Here we consider three factors: the number of participants (senders and recipients) p, the time span t of the emails in the concept, and the frequency *t* of the emails in the concept. They are defined as follows.

$$p(C) = |\bigcup_{e \in C} (sender(e) \cup recipient(e))|,$$

$$t(C) = \max_{e \in C}(date(e)) - \min_{e \in C}(date(e)), \text{ and } f(C) = |C| \big/ t(C),$$

where *e* denotes an email message, *C* denotes a concept, and |*C*| denotes the number of emails in concept *C*.

We use fuzzy membership functions [8] to represent users' domain knowledge. In fuzzy logic, a fuzzy membership function represents how likely an object belongs to a set. The inputs of the function are the values of the attributes of the object. The output is a numeric value between 0 and 1 representing the degree that the object belongs to the set. The value 1 means that the object fully belongs to the fuzzy set. The value 0 means that the object does not belong to the fuzzy set. A value between 0 and 1 means that the object partially belongs to the fuzzy set, with higher value representing higher degree of belongingness.

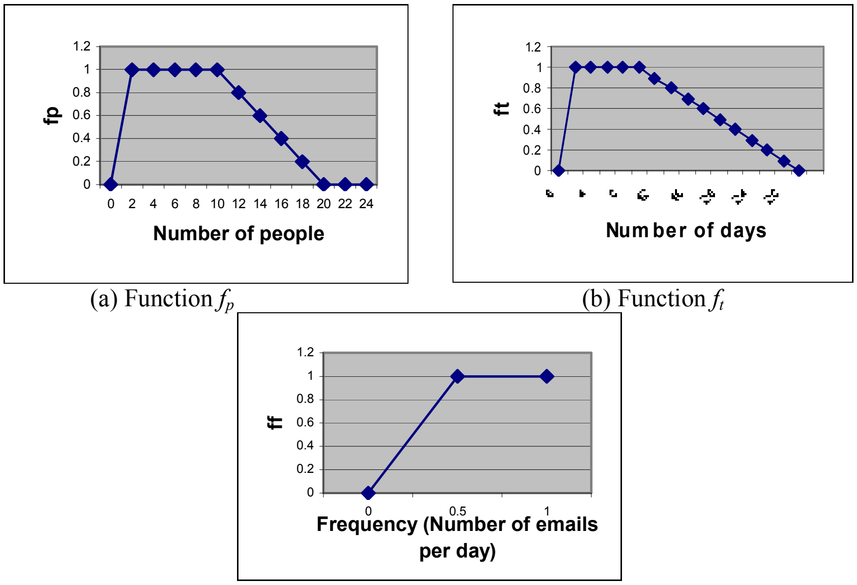
Intuitively, emails related to a topic should involve a certain group of people and should occur over a short period of time with high frequency, although values for these parameters may differ from application to application. We use these features as the input to the fuzzy membership functions. The output of the functions is the fuzzy values that represent the degree that the group of emails belongs to a topic. Figure 1 shows three examples of the fuzzy membership functions. Figure 1(a) describes the relationship between the degree of a topic and the number of people involved in the group of the emails. We denote the function as f_p. It says that if 2 to 10 people are involved in a group of emails from a concept, the membership degree that they belong to the same topic is 1. If there are 10 to 20 people involved, the degree of membership linearly decreases. If there are more than 20 people involved, the degree becomes 0. Similarly, we define the fuzzy membership functions for the time span (f_t) and frequency of emails (f_f) in Figures 1(b) and 1(c), respectively. In real applications, these functions should be defined by the domain experts.

Finally, we combine these three factors by multiplying the three fuzzy values to get the overall fuzzy value. For example, if a concept includes 2 emails, involves 5 people as senders or recipients, and time span is 6 days, then the final fuzzy value is 0.67 according to Figure 1.

By setting a threshold for the fuzzy values, we can identify the topics in the concepts.

3 Experiments

The experiments are implemented in Java and run on a PC with 3GHz CPU and 2G bytes memory. We chose Enron email data set for our experiments. Enron Email data set is a benchmark for research in fields like link analysis, social network analysis, fraud detection, and textual analysis. We worked on the cleaned version from [9], which contains 252,759 messages from 151 employees distributed in around 3000 user defined folders. We selected emails sent in the year 1999 (4760 emails) for our experiments.



(a) Function f_p

(b) Function f_t

(c) Function f_f

Fig. 1. Fuzzy membership functions for ranking the concepts

Table 1. Concepts regarding *trip*

Score	Intent	Number of Emails
1.0	london, trip	2
0.91	houston, trip	5
0.0	brazil, trip	9
0.0	trip, canadian	1
0.0	trip, ba	1
0.0	houston, trip, meet, request	1
0.0	trip	26

We extracted the words in the subjects of the emails and cleaned up the data in preprocessing, including removing the stop words and using the Porter algorithm to stem words. After these steps, 935 words were left. Then we set frequency threshold to 5 to prune the words that do not appear frequently enough. Finally we had 402 words left as features. We applied the FCA open source code from [10] and found 782 concepts.

We choose to look into the topics regarding *trip*, which are comprehensible to general readers. Seven concepts concerning *trip* were identified, which are shown in Table 1. We manually checked the concept *london trip* and found that it talks about a schedule for Tana's visit to London, which is a topic. The emails for concept *houston* *trip* include 5 emails, which involve two topics. The first four emails focus on one topic and the last one belongs to another topic. The concept *brazil trip* contains 9 emails involving three topics. The results show that the higher the fuzzy value is, the more likely the concept corresponds to one topic.

4 Conclusion and Future Work

We proposed a method based on formal concept analysis to find concepts (potential topics) in emails. To deal with the overwhelming number of the concepts produced by formal concept analysis, we used fuzzy membership functions based on the features of email, including timestamps, senders, and recipients, to rank the concepts according to the likelihood that they are topics. Preliminary experiments on Enron email dataset show the promising results.

In the future, we will do more comprehensive experiments on the whole Enron data set and on both subjects and content of emails. We will compare our method with other email classification and clustering methods. We will also improve our method by taking into account the explicit threading information in the emails when ranking the concepts, and looking for more effective method for fuzzy member function aggregation.

References

[1] Cselle G., Albrecht K., and Wattenhofer, R. BuzzTrack: topic detection and tracking in email. *Intelligent User Interfaces*. 190-197, 2007.

[2] Dredze, M., Lau, T.A., and Kushmerick, N. Automatically classifying emails into activities. *Intelligent User Interfaces*. 70-77, 2006.

[3] Huang, Y., Govindaraju, D., Mitchell, T.M., de Carvalho, V.R, and Cohen, W.W. Inferring ongoing activities of workstation users by clustering email. *Proceedings of the First Conference on Email and Anti-Spam*. Mountain View, California, USA, July, 2004.

[4] Khoussainov, R. and Kushmerick, N. Email task management: An iterative relational learning approach. *Proceedings of the Second Conference on Email and Anti-Spam.* Stanford University, California, USA, 2005.

[5] Kushmerick, N. and Lau, T.A. Automated email activity management: An unsupervised learning approach. *Proceedings of the 2005 International Conference on Intelligent User Interfaces*. 67-74, San Diego, California, 2005.

[6] Li, H., Shen, D., Zhang, B., Chen, A., and Yang, Q. Adding semantics to email clustering. *Proceedings of the 6th IEEE International Conference on Data Mining*. 938-942, Hong Kong, China, 2006.

[7] Wille, R. Restructuring lattice theory: an approach based on hierarchies of concepts. In Rival (Ed.), *Ordered Sets*. Reidel, Dordrecht-Boston, 445-470, 1982.

[8] Zadeh, L. Fuzzy sets. *Information and Control*, 8(3), 338-353, 1965.

[9] http://www.isi.edu/~adibi/Enron/Enron.htm.

[10] http://www.iro.umontreal.ca/~galicia/