Linnéuniversitetet
Kalmar Växjö

Master Thesis Project

An Application of Dimension Reduction for Intention Groups in Reddit

Author: Xuebo Sun, Yudan Wang
Supervisor: Morgan Ericsson
Examiner: Welf Löwe
Reader: Narges Khakpour
Semester: VT/HT 2016
Course Code: 4DV50E
Subject: Computer Science

Abstract

Reddit (www.reddit.com) is a social news platform for information sharing and exchanging. The amount of data, in terms of both observations and dimensions is enormous because a large number of users express all aspects of knowledge in their own lives by publishing the comments. While it's easy for a human being to understand the Reddit comments on an individual basis, it is a tremendous challenge to use a computer and extract insights from it. In this thesis, we seek one algorithmic driven approach to analyze both the unique Reddit data structure and the relations inside owners of comments by their similar features. We explore the various types of communications between two people with common characteristics and build a special communication model that characterizes the potential relationship between two users via their communication messages. We then seek a dimensionality reduction methodology that can merge users with similar behavior into same groups. Along the process, we develop computer program to collect data, define attributes based on the communication model and apply a rule-based group merging algorithm. We then evaluate the results to show the effectiveness of this methodology. Our results show reasonable success in producing user groups that have recognizable group characteristics and share similar intentions.

Keywords: Reddit, communication model, dimension reduction, similarity metric

Acknowledgements

Our master thesis project received guidance, ideas and strong support from our supervisor Dr. Morgan Ericsson. We are really grateful for his assistance. We also appreciate the help and encouragement from our thesis course manager Dr. Narges Khakpour. Many thanks are given to our close friends for their greetings and care inspiring us to move forward. We would also like to express our sincere gratitude to our families for their unwavering support and confidence in us during our most difficult time. We feel much stronger after completing the intense one year master program.

Contents

Chapter 1

Introduction

Reddit, a world-wide famous information sharing website, has become an ocean of information with the influx of a large number of data sources. Some research has been done on the popularity of the submitted content in the whole platform. The top popular text can reflect the hot ideas in the platform related to all users, which is a static result with isolated information units.

The fundamental idea of the Reddit platform rests upon an evolution process. It involves interactions among group of users. Users express their own opinions by submitting text of comments. Comments are organized under different topics – so called subreddit. Through this process, information is dynamically transferred from one to another and topics with more popularity can gain more weights due to more users following. One can thus explore the interconnection between the users via their communications and quantitatively measure user similarity based on the topics they share. Further one can apply certain dimension reduction algorithms to aggregate similar users into small groups. This will help both reduce the dimensionality of the data as well as identify the intent of different groups.

This chapter has following structure. First the background of the question proposed is introduced. Then the motivation and problems are discussed. In the later parts, research questions and contribution are described.

1.1 Background

The Internet makes it convenient for people to take part in global social networking. Large-scale network forum plays an important role in spreading information, in which people can participate the discussion of various topics and uncover valuable information or exchange their ideas and share emotional feeling with others. As the number of users increases rapidly, the amount of data in the digital world explodes and the network becomes much more complex. Being able to analyze large data sets has become a key basis of business competition, which underpins new waves of productivity growth and innovation [1]. It is helpful to figure out the underlying law of complex network and capture the essential context of information transfer. In a typical internet forum, users are often attracted by contents, topics or other users who they like to exchange ideas with. A group of like-minded people might gather together by the similarity between them. We are interested in getting different groups in which users from Reddit have the similar features and intentions. However, a large number of users bring large amount of information in many topics. To some extent, the data dimension processing need to be considered to make it available to measure the features of data in fewer group dimensions and highlight the group similarity [2, 3].

1.2 Motivation

The informative comment data from Reddit is the result of people from all over the world participating in the platform and interacting with each other. It seems that their operations appear everywhere in the forum. They read and post many

comments for sharing information and exchanging ideas. Many kinds of words represent different meanings. People's activities have their own purpose. Some of them like to communicate with certain group of people, or some of them prefer certain aspects of information content. It is useful to group people with similar intention. This could be done easily and naturally when not many people aggregate together in the club form in reality. However, in the wide digital world, only large amount of comment data is in front of us, which are related to a great number of users, comments and topics. The actions of a large number of users in Reddit platform construct a complex network. How is the information spreading? What is the communication topology of massive users? What is the main idea that they are talking about? In order to know the dynamics on this platform, it becomes necessary to process and analyze large-scale comment data from Reddit [4]. How could we analyze the data based on such great number of dimensions?

Real-world data, such as speech signals, digital photographs, or fMRI (functional MRI) scans, usually has high dimensionality [5]. In order to handle such real-world data adequately, its dimensionality needs to be reduced [5]. Traditional statistical methods break down partly because of the increase in the number of observations, but mostly because of the increase in the number of variables associated with each observation [6]. The dimension of the data is the number of variables that are measured on each observation. High-dimensional data sets present many mathematical challenges as well as some opportunities, and are bound to give rise to new theoretical developments [7]. What we are interested in is to build a data analysis model and use some algorithms to reduce data dimension [3, 8]. It's a big challenge to propose the hypothesis and to show how to build a model and evaluate its effectiveness [9].

1.3 Problem Formulation

Comment data are generated based on users' intent and are the results of the users' actions. Users publish comments in different topics to express their opinions, which are very flexible and diverse. They can cover almost all aspects of real life. Some of them might be potential threats to society. There might be inner relationship between the properties of the comments and the features of user groups. However, the number of the dimensions of data is enormous according to a great number of users, their comments and the topics of interest to them.

User's actions in Reddit could be summarized as:

(1) User could submit comments in different topics no matter if they subscribe them;
(2) User could answer other's comment by directly publishing its child comment;
(3) User could answer child comment by publishing the child comment of the same parent comment at the next position;
(4) Each comment has its own keywords that could represent its intention;
(5) Each comment could be scored by any user adding or reducing one karma score on it.

Users' actions in social network are not isolated. They are interrelated within certain group of users. The comments of these users might connect with each other in the data tree. The contents of the comments might be similar. Our analysis model will consider user's actions in all topics. To categorize the different types of user actions, we consider the following features:

(1) The number of comments related to certain pair of users;
(2) The number of users that have similar comments;
(3) The frequency of each word in each comment;
(4) The inverse frequency of documents that have certain word;
(5) The number of keywords in certain comment;
(6) The negative karma score of comment.

Our aim is to analyze complex data of Reddit and discover user groups with common intentions by modeling the features of data and defining the features for user groups. We also need to investigate previous research of dimension reduction and apply them to our model for reducing the data complexity of user group features [10, 11].

To fulfill the aim of our research, the following questions need to be answered:

RQ1.	What features of user could be used to compare different users?
RQ2.	What features of group could be used to aggregate similar users?
RQ3.	Which dimension reduction algorithm could be applied to our model?

People often behave independently. The characteristics of them are often different from each other. However, some people have similar features among plenty of characteristics, which could aggregate people into different groups. Each group is a study object and forms its own common preferences. The preferences could be reflected by the distribution of weight metrics in the group. The dimensions of each group are related to different users, different comment words and different topics. Several methods would be used to reduce the dimensions of group, like modeling or direct dimension reduction. Groups with evident intention and fewer dimensions are expected.

1.4 Contributions

As mentioned in the previous sections, the aims of our research in this thesis include methods and mechanisms of aggregating users into groups having small number of users; and some extent of similarity; and user communication network in the group; and the main content of intention; and the degree of negative impact. Along these research objectives, the main contributions of the thesis are shown below.

(1) A communication model is built to capture the feature of users. Some concepts are defined for the comment communication and its ways and types, and user communication. Some requirements are given to discover the valid communication between users;

(2) A proposal, based on the communication mode, is designed to group users, aggregate user groups with certain extent of similarity, and

highlight the main intention of the group and the important role of the users in the group. The information retrieval technique is used to extract keywords from the contents of the comments related to the group. The similarities of both users and keywords are computed between two groups. The rule-based dimension reduction algorithm is applied to merge similar groups based on the similarity computation. Similar users and keywords will result in getting high weight. Gephi Graph Visualization and Manipulation software is used to show the topology of the user communications in the group.

(3) A special approach is adopted to achieve the groups that have relatively high negative impact on society.

(4) A stand-alone application is developed to implement the above research design using Python programming language and SQLite database, and collect the result data to illustrate and evaluate the effectiveness of the design.

1.5 Outline

The report structure is as follow:
Chapter 2 reviews the existing research on the theory and technique that will be used in our research.

Chapter 3 focuses on what our methods are and how to build the communication model and apply it, along with the theory and techniques behind, and the methodology used in the research.

Chapter 4 describes the implementation of the methods in chapter 3, using python to as the main programming language for data transformation and algorithm execution.

Chapter 5 evaluates the grouping result by setting different thresholds, comparing various grouping results and offers intuition behind.

Chapter 6 concludes the above research and result, and summaries what we have done and the future work for our research.

Chapter 2

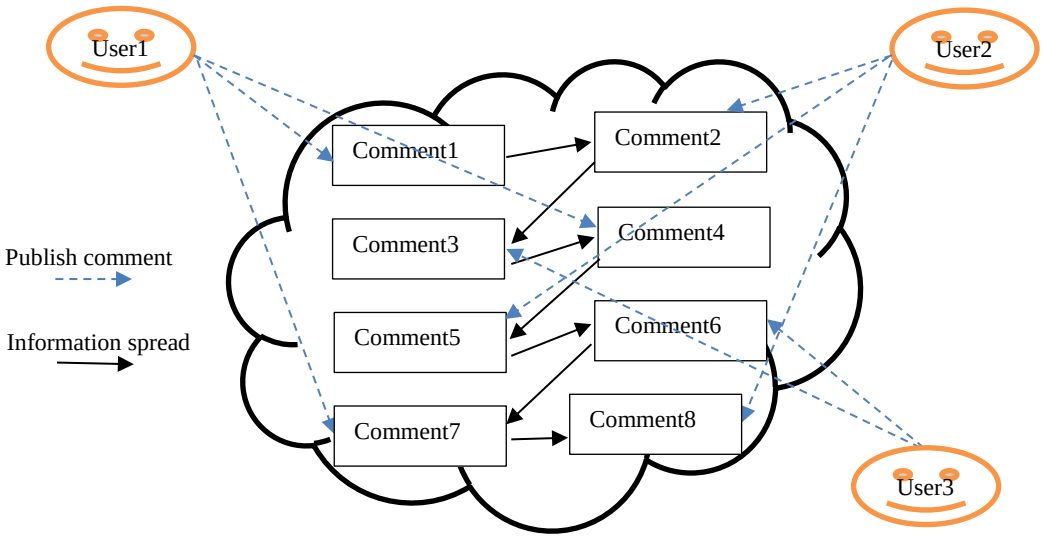
Background

The research related to this thesis focuses on social news website, Reddit research, key words algorithm, feature extraction and dimension reduction. Moreover, some kinds of techniques for dimension reduction are also discussed.

2.1 Social News Website

As the computer technology rapidly being developed, the Internet is becoming a good media for all kinds of information gathering and spreading because it provides a free and convenient way for people to share, browse and search for information. Many people are attracted to the Internet website by their interests in the rich content. They continue to create website content and exchange their ideas. Their interactions boost the development of network in this platform.

Social news website, like Reddit (www.reddit.com), is such a public and open network with a great number of active users publishing comments to share their ideas and emotions. The Web 2.0 technique makes the network self-organized. Each user has his or her personality viewpoint and interest. This means they have their own features to read information resources and express their opinions selectively. They choose to express opinions with specific populations, or on interesting topics. Small-scale network for each individual user grows differently according to their different interests. Their actions in the website are directional from their own perspective, which connect users with their characteristics. They can start new topics by posting new contents, or exchange ideas by posting comments. Actions from different users interact with each other, and different individual networks are intertwined together. A user can have many connections with different users by replying each other's comments. Such relationship is not as tight as that in the communication website [12]. He also can be involved in the formation of many opinions. This is the evolution of whole social network, the example of which is shown in Graph 2.1.



Graph 2.1 Example of social network evolution

The interactions on topics promote the information diffusion over the network. However, the impact of information is also spread along it. Some impact is positive and some is negative. In Reddit platform, this can be shown in a voting system [13]. Each comment can be “upvoted” or “downvoted” by all users. The scores of comments related to specific opinion can reflect its impact. The group of users around the opinion might have negative or positive impact on the society.

2.2 Reddit Research

2.2.1 Community in Reddit

Reddit is a type of online community where users vote on content [14]. It is also one of the most famous information sharing forum in North America. It is not a direct communication website, but a social news networking website, in which there are indirect communications in question and answer mode [15]. There are more than 11 million registered users in this platform and nearly 500 thousand classified topics (Subreddits). It has been increasingly popular with nearly 8 billion page views in the past one month [14].

Among the huge number of users, each user has his or her own perspective to be active in this platform. They have their own special actions according to their different interests. User’s actions may first focus on the rich content in the platform where they can get more useful information. Once the content is in user’s wide range of topic interests, he may become active and is likely to publish many reply comments. Usually, the reply comments from different users form the discussion around the same topic. Therefore, research on question and answer forum shows that users are connected with idea resources, which exerts the untapped capability of social network and promote opinion evolution [16]. The interactions between users might result in new opinions. The users having similar opinions will naturally gather together by their similar actions [17]. They will have more probability to exchange ideas together.

The content of the comments published in the platform plays an important role in users’ intention actions. Different users may express their ideas and emotions through comments. The aim of user’s actions needs to be induced from the content of the comments. Hence, it is necessary to analyze the main subject in each comment. It is better to assign weight to each word in a comment, showing the relative importance of the word.

Considering massive users and comments they published, in order to understand the procedure of opinion evolution and information transfer, and also highlight small number of users having common characteristics, it is necessary to calculate the similarity of the user actions and the content they published.

2.2.2 Voting System

Much of the traffic and discussion in Reddit is driven by its voting system. Because when users submit content to the forum, all users of the website can either upvote or downvote the submitted content. The difference between the

number of upvotes and the number of downvotes determines the karma score of the comment content [13]. The karma score value reflects the popularity of the content. Some research predicts comment's future popularity and the maximum score in its lifetime [13, 18]. The score can be positive or negative, which also means that the content has positive or negative impact. The impact of user's actions in the forum often can be quantitatively measured by adding up all impact of his corresponding comments.

In Reddit platform, the defined voting system can well reflect the popularity of comments. It is comprised of a set of rules. Voting need to be considered effective in accordance with these rules. The rules also decide how votes are added and accumulated to produce a final result. In this way, the choice of voters can be made between candidates in an election or on a policy referendum. There are two major classes of method, which are multiple-winner method and single-winner method [19].

In multiple-winner method, voters are more interested in the overall composition than exactly who get selected. Multiple-winner voting methods are basically formed through extending single-winner methods. Single-winner methods contain two types, which are ranked voting and rated voting. In ranked voting system, the candidates are ranked in preference order by each voter. For rated voting, each candidate option is given a score by voters.

Rated voting system exists in Reddit platform. Each option comment are scored or rated on a range by all voters, where the allowable ratings are -1 (downvote), 0 and 1 (upvote). The score results of each comment from all voters are cumulated together. If more voters choose -1 as ratings, the minus result can reflect the negative impact. On the other hand, if the result is positive and large, it means the comment is approved by more voters and has positive impact.

2.3 Keywords Algorithm

2.3.1 TF-IDF

There are some research interests in predicting Reddit post popularity, and the popularity is usually related to the content of the post [18, 26, 27]. The content is quite different so the problem is how to classify a Reddit post. How to compare the similarity of two posts? Some research use TF-IDF and cosine similarity algorithm. TF-IDF algorithm can indicate the word's relative importance to the text. It stands for term frequency–inverse document frequency, which is a numerical statistic that intended to reflect how important a word to a document in a collection. It is often used as a weighting factor in information retrieval and text mining. The TF-IDF value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the collection, which helps to adjust for the fact that some words appear more frequently in general.

TF-IDF assigns a weight to each word in a post, indicating the word's relative importance to the post [18]. From the following formula, let tf be the term frequency and idf be the inverse document frequency:

$$\text{tf (w, p)} = \frac{\#\ occurrences\ of\ w\ in\ p}{(\max \#\ occurrences\ for\ any\ word\ in\ p\ )}$$

where w is indicates words in a post, p indicates one post.

$$\text{idf (w,P)} = \log\left(\frac{|P|}{\#\ comments\ containing\ w}\right)$$

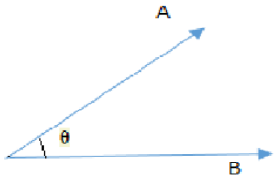
where w indicates words in a post and P indicates the total post collection.
So we can get the value of tf-idf by the following fomula:

tf-idf = tf (w, p) * idf (w, P)

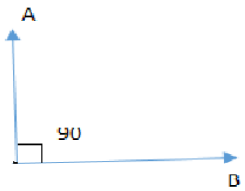
The idea behind applying tf-idf is to get difference between the relative important of words in the title and self-text features of posts. However, the result only showed mixed result with the inclusion of tf-idf [18].

2.3.2 Cosine Similarity

From some research, alternative TF-IDF approaches are more helpful for multinomial implementations. And it is mostly used combine with cosine similarity [16]. In order to compare the similarity between users' special action sets, cosine similarity is a good way to deal with this problem [16]. Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them [16]. The cosine of 0° is 1, and it is less than 1 for any other angle from 0° to 90°. We can get to know clearly from the following Figure 2.1, A and B shows two vectors, for example, $A= [A_1,A_2,\ldots A_n]$ and $B = [B_1,B_2,\ldots\ldots B_n]$.



(1) Two vectors have Similarity



(2) Two vectors are not similar

(3) Two vectors are same

Figure 2.1 the cosine similarity between two vectors

It is thus a judgment of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1, two vectors at 90° have a similarity of 0, and in between two vectors have certain similarity to some extent, independent of their magnitude. Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in. Similarity can be calculated by the formula:

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Where A_i and B_i are components of vector A and vector B respectively. For example, we have two vector A= [A_1,A_2,……A_n] and B = [B_1,B_2,……B_n], we can get the value of similarity according to the formula. The resulting similarity ranges from 0 indicating in correlation, to 1 meaning exactly the same, and in-between values indicating intermediate similarity or dissimilarity. This formula can be used not only for comparing between posts, but also for comparing the users from two groups or key words.

2.4 Feature Extraction

In order to get the proper information from a system, we need to analyze user actions in a system and extract the feature from the system. User action is always changeable according to their different preference. Some research considers variation of user actions and proposes an algorithm based on sequence of user action [17]. It helps to capture the characteristics of user's behavior. The first stage is data preprocessing. In this stage, using concept hierarchy and association rules to reduce the sparseness of data set. The user profile can then be created. In data mining stage, users can be grouped according to user profile. Data hierarchy refers to the systematic organization of data, which involves fields, records and so on. According to the whole data set, it can be divided into different small set. Association rules is mainly used for mining the relationship of users [17]. Two sets of data can be merged according to the association.

After getting the user actions, one needs to identify features from different user actions. So feature extraction is a good way to release it. In machine learning, pattern recognition and in image processing, feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations. Feature extraction is related to dimension reduction. When the input data to an algorithm is suspected to be complex and redundant, it can be transformed into a reduced set of features. This process is called feature selection. The selected features are expected to maintain the major relevant information from the input data, so that subsequent desired tasks can be performed by using this reduced representation instead of the complete initial data. Selection of a feature extraction method is probably the single most important factor in achieving high recognition performance in character recognition systems. Different feature extraction methods are designed for different representations of the characters. There are usually different features in one system. One research is dedicated to the extraction of musical features from audio files [20]. The different algorithms are decomposed into stages, integrating different variants proposed by alternative approaches. Normally, the different musical features extracted from the audio files are highly interdependent. Some features are based on the same initial computations. It is important to avoid redundant computation of these common components.

Feature selection is effective in reducing dimensionality, removing irrelevant information and improving result comprehensibility [21]. It's a challenge to many existing feature selection methods with respect to efficiency and effectiveness. Some researches introduce a predominant correlation and propose a fast filter method that can identify relevant features. Feature selection algorithms fall into two broad categories [21], one is filter model and the other one is wrapper model. Filter model relies on data to select some feature without involving any algorithm. Wrapper model requires one predetermined learning algorithm in feature selection [21]. When the number of features becomes very large, the filter model is usually chosen due to its computational efficiency. Some researchers try to combine the advantages of both models. They proposed an algorithm in a hybrid model in order to deal with high dimension data. In these algorithms, a good measure of feature that based on data characteristics it the first important, and they also consider cross validation which is exploited to decide a final best subset [21]. These research mainly based on combining filter and wrapper algorithms to achieve best possible performance [21]. The aim of research is to find a new feature selection algorithm that can reduce the irrelevant and redundant features.

Research shows that different feature selection algorithms can be further categorized into two groups with some filter model [17]. There are two algorithms including feature weighting algorithms and subset search algorithms, which is based on how they evaluate the features individually or through feature subsets. Feature weighting algorithms assign weight to features individually and rank them based on the target concept [17]. We select one feature if it's weighting value is greater than a threshold value. It's an easy way to select different desired features by setting different threshold. But some feature selection literature shows that sometimes features will all be selected even though they are highly correlated to each other [17]. Especially for the features which have high dimension data, there exists many redundant features, so pure relevance based feature weighting algorithms do not meet the feature selection very well. It's better to use feature weighting algorithms after reducing the dimensionality.

How to evaluate the goodness of features for a particular task? Research shows that a feature is good if it has more relationship with the class concept but not redundant to any of the other relevant features [17]. If the correlation between a feature and the class is high, it can be predicted by any of other relevant features, so this feature can be regarded as a good feature for classification. Subset search algorithms is guided by a certain evaluation measure which computes the goodness of each subset [17]. The research shows that the optimal subset is selected when the search stops. There is a hypothesis that if a feature is highly correlated to the class, and uncorrelated to each other, this feature is a good feature. But now existing subset search algorithms do not have the ability to deal with high dimensional data. How to overcome the problem of algorithms in both groups and meet the demand for feature selection for high dimension data? Some research tries to deal with this question and proposes a novel algorithms which can identify both irrelevant and redundant features.

2.5 Dimension Reduction

Data collection and storage capabilities play an important role in most sciences. Researchers work in different domains as engineering, biology, consumer transactions and so on [6]. Traditional statistical method is not effective anymore because the number of observations is too large. The dimension of the data is the number of variables that are measured on each observation. High dimensional database present many mathematical challenges. Researchers show some traditional and current state dimension reduction methods published in the statistics, signal processing and machine learning. In many cases, the researcher found one of the problems with high dimensional data set is not all the measured variables are important for the underlying phenomena of interest [6]. Some methods can build predictive models with high accuracy for high dimension data set. It is very interest that dimension reduction for original data has higher priority than any modeling of the data. However, for the area of social media websites, there is no related research work done with dimension reduction.

Dimension reduction is important in many domains, because it mitigates the curse of dimensionality and other undesired properties of high dimension spaces [5]. The function of dimension reduction is to reduce the transformation of high dimension data into a meaningful representation [5]. Some traditional research mainly uses linear techniques such as PCA, factor analysis and classical scaling. But these linear techniques can't adequately handle complex nonlinear data [5]. Nowadays a large of number nonlinear methods for dimensionality reduction have been proposed. Compare with the traditional linear techniques, the nonlinear techniques have the ability to deal with complex nonlinear data. The real world data are usually most nonlinear data. So it will have advantage to use nonlinear techniques to reduce the dimension of real word data. In order to get clear idea about what extent the performances of the various dimensionality reduction techniques differ on traditional techniques for dimension reduction and nonlinear techniques, some research analysis these two techniques and give a comparison. Even the comparison is very limited in some scope.

PCA is the most important linear dimensionality reduction technique. PCA is short for Principal Component Analysis, which is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. PCA is the best in the mean-square error sense and linear dimension reduction technique [6]. PCA seeks to reduce the dimension of the data by finding a few orthogonal linear combinations of the original variables with the largest variance [6]. PCA and classical scaling have been successfully applied in a large number of domains such as face recognition, coin classification and seismic series analysis. But PCA and classical scaling suffer from one drawback [5]. The size of covariance matrix is proportional to the dimensionality of the data points, so the computation of the eigenvectors might be infeasible for very high dimensional data [5].

Another linear method like linear discriminant analysis (LDA) is a generalization of Fisher's linear discriminant, a method used in statistics, pattern recognition and machine learning to find a linear combination of features that characterizes or separates two or more classes of objects or events. The resulting combination may

be used as a linear classifier, or more commonly, for dimension reduction before late classification. LDA is also closely related to principal component analysis (PCA) and factor analysis in that they both look for linear combinations of variables that best explain the data [22]. LDA explicitly attempts to model the difference between the classer of data.

PCA on the other hand does not take into account any difference in class, and factor analysis builds the feature combinations based on differences rather than similarities. Discriminant analysis is also different from factor analysis in that it is not an interdependence technique: a distinction between independent variables and dependent variables (also called criterion variables) must be made [22].

Compared with the linear dimension reduction algorithms, high correlation filter (HCF) [3] is more suitable for research the features that have relations. Normally, with high correlation filter algorithm, if two sets of data's variation tendency are similar, they would contain similar information. In order to get the similarity of two sets of data, we can calculate correlation coefficients for them using cosine similarity formula. When correlation coefficient exceeds the threshold that we defined, the two sets of data can be merged to one, in which way the dimensions of the data are reduced. Some research base on correlation analysis of feature, develop a procedure to select good feature for classification.

Chapter 3

Methodology

This chapter introduces the methodology applied to our research. First, the overview of the approach is given. Then, the unique properties of the original data are analyzed, based on which, a communication model is built and illustrated. Thereafter, some grouping metrics are defined for the application methods of the data model, and the process of the merging method is described, followed by the description of High Correlation Filter (HCF) [3] used for similarity aggregation. Subsequently, the expected results are prospected based on the methods. Lastly, the reliability and validity of the research are assessed, and the ethical impact is estimated.

3.1 Overview of Our Approach

In our research, we use quantitative methods to answer our research questions. Quantitative research is the systematic empirical investigation of observable phenomena via statistical, mathematical or computational techniques. It emphasizes numerical analysis, automating data transformation and model building via rigorous programing languages. The objective of quantitative research is to develop and employ mathematical models, theories and hypotheses pertaining to phenomena.

In Reddit platform, users worldwide publish their comments, sharing and exchanging all kinds of information. Each comment is mainly stored as a record with several informative fields. Our study object is the sample data from forum which is consisted of 150,429 comments. The quantitative method is used to analyze the raw data that were exported from the platform and stored separately in the form of zipped JSON data file. The comment attribute fields were captured in the file. Python scripts and SQLite database are used to extract specialty data from the file, process them and persist them step by step for further analysis. The followings are important data features used in our research:

- Subreddit: the topic of the Reddit platform in which certain category of comments are posted.
- Author: the user who publishes the comment.
- Id: the comment id.
- Parent_id: the parent id of the current comment.
- Body: the content of the comment.
- Created_utc: the time of day that the comment was published.
- Score: the karma score of comment which could be changed by user adding one or subtracting one.

The communication model will be defined on these comment data traits. The basic concepts related to comment communication are given in order to formalize the features and relationships between two pieces of comment data. And then, the communication in the user level is brought forward. Based on the communication model, a grouping proposal is designed to group and merge users into some number of groups with common intention. In this proposal, some metrics are defined to capture the features of user groups. Some techniques are used to compute the values of metrics. Through the information retrieval algorithm, TF-

IDF (Term Frequency-Inverse Document Frequency), the keywords of the comments related to the user groups are screened out from lots of text in the content of the comments. The user group can be represented by user vector and keyword vector. Therefore, cosine similarity algorithm is utilized to compute similarity between two vectors. The user and keyword similarities between two groups are calculated by cosine algorithm respectively. The user similarity refers to the extent to which the users from two groups are the same. The keyword similarity is the extent of the same keywords matching between two groups. General similarity will be computed based the user similarity and the keyword similarity. Then the way how to reduce dimension will be investigated to do similar result merging. Two user groups will be compared by their similarity and be merged through dimension reduction algorithm, High Correlation Filter (HCF). Furthermore, the similar users and keywords of the merged two user groups will get high weight, which indicates the important role of them in the new generated user group. In order to figure out the user groups having much negative impact on society, a special approach is adopted to assess the negative degree of each user group.

We conduct experiment on the sample data by implementing program that could fulfill the design of the grouping proposal. Numeric result data will be collected to evaluate and illustrate the effectiveness and meaning of data modeling and its application proposal, and compare the dimension quantity of user groups before and after data analysis.

3.2 Method Description

3.2.1 Proposal Overview

Our quantitative method is to probe informative data set where some valuable implication could be induced step by step. The following steps compose the algorithms of our proposal:

(1) Extracting feature data of comments from sample data set;
(2) Setting up communication model:
    a. Collecting comment communications;
        i. Between parent and child;
        ii. Between two neighbor brothers;
    b. Collecting user communications;
(3) Grouping users:
    a. The metrics of basic group;
        i. Users having valid communications;
        ii. Keywords;
        iii. Similar weights of users;
        iv. Similar weights of keywords;
        v. Negative karma score;
    b. Merging group by similarity between two groups.

3.2.2 Sample Data Property

There are 150,429 comments in the sample data. 24,913 users published these comments in 32 different topics. Total number of words that users mentioned in all comments is over five million. The dimensionality of this large data group is

relatively enormous when the valuable information is analyzed. If the data are naturally divided into relatively small groups by topics without considering the ideas of users, in each group, the average number of the users and that of the words are still large, which are 779 and over 160 thousand respectively. There ought to be something in common among small group of users, which could be figured out by studying the features of users.

From the structure of the sample data, some relationship between comments could be discerned. Each record of data represents one comment. One attribute of each comment is the ID of parent comment, which means the relationship among comments is like tree structure. The Subreddit topics are the roots. All comments are on the branch of the root or at the end of branch as leaves. Each comment is one-action result of the user in Reddit platform. The content of comments could reflect the information of user's communication.

The communication relationship between users could be reflected well in the comment tree. There would be small number of users who are the comments' authors, and have much probability of communication and similar intention words to communicate. Hence, the communication model could be built to embody the relationship of the comments, and more importantly, characterize communication between the authors of the comments.

3.2.3 Data Modeling

Based on previous analysis, our proposal is to firstly separate all users into individual groups, and then aggregate one-user group into a smaller number of groups related to fewer users and fewer words than those the sample data have. The main aim of users who behave on a social network is to share and exchange information. It is easy to know that two person are communicating and their topic when they are talking with each other. But Reddit forum is not a specialized chatting website. Many people publish information together in it. Therefore, it is imperative to study the way how they communicate, and the opinions they have. We need to build model for this problem.

Our analysis is that the communication between users happens through the interactions between comments published by users. The users have several kinds of actions related to comments in the forum, which is mentioned in chapter 1.3. The features of comments associated with user's actions are also shown in chapter 1.3. Thus, the interaction between comments is defined as comment communication, which is basic for user communication.

The communication model has three levels of definitions, which are comment communication (CC), user communication (UC) and valid user communication (VUC). The basis definitions are CC and UC. VUC will be defined as a metric to aggregate users.

Comment communication (CC) is the relationship between two comments, in which, one comment asks question and the other answers question. The question comment is generated earlier than the answer comment, which is according to people's way of questioning and answering in reality. There are two possible

ways of CC. One way is the communication between parent comment and direct child comment, and the other way is between neighbor brother comments which have the same parent comment and are neighbors among the child comments. The communication between parent and direct child can be seen directly through data feature, “Parent_id”. However, neighbor brother communication is a possible and habitual interaction between two users. This potential relationship could increase the cohesion between users.

CC is unidirectional between two comments. Thus, in each way, the comment usually has two types of CC with other comments. Question comment is generated earlier than answer comment. Being associated with different other comments, a comment could be either parent node or child node. If the comment is question node and the other is answer node, it means the comment has QA (Question-Answer) type of CC with the other comment. Conversely, if the comment is answer node and the other is question node, the comment has AQ (Answer-Question) type of CC with the other comment. For parent and child communication, QA type of CC is parent-child (PC), and AQ type of CC is child-parent (CP). For neighbor brother communication, QA type of CC is brother-question-answer (BQA), and AQ type of CC is brother-answer-question (BAQ). Hence, totally there are four types of CC.

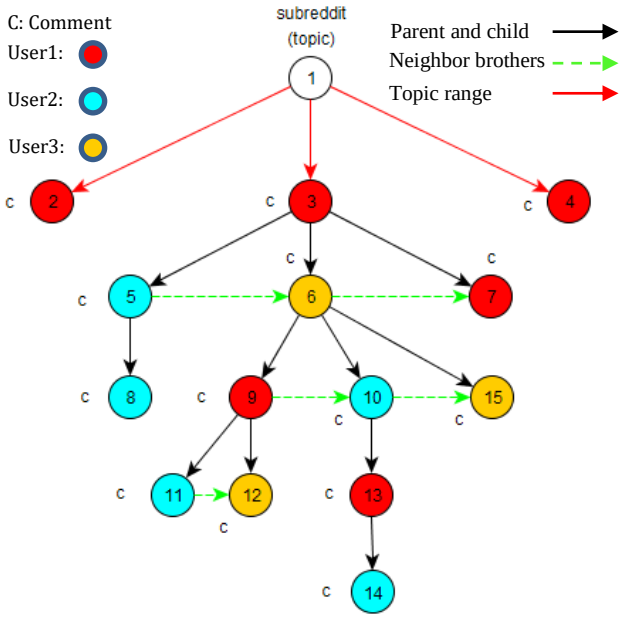


Figure 3.1 Comment tree example

Figure 3.1 is an example to illustrate the CC model in comment level. Each colored nodes are comments. Different colors means the comments belong to different users. Black arrows represent parent and child communication. For comment 3, it has PC communication with child comment 5. On the contrary, for comment 5, it has CP communication with parent comment 3. The neighbor brother communication cannot be seen directly, which is shown as green dotted arrows. For instance, comment 6 has three direct child comments, 9, 10 and 15. Comment 9 is the neighbor of comment 10, and the smaller number 9 means earlier and the one who asks question. So it has BQA communication with

comment 10. For comment 10, it has BAQ with comment 9 and BQA with comment 15. For comment 15, it has BAQ with comment 10.

Based on CC, user communication is defined for those who have CC in between. Some of the user communications in the forum are random, especially neighbor brother communication, and some are intentional. Not all user communications are meaningful. Therefore, valid user communication will be considered. In the next section, a metric will be defined as valid user communication by which similar users are grouped together in our later research.

3.2.4 Grouping Metrics

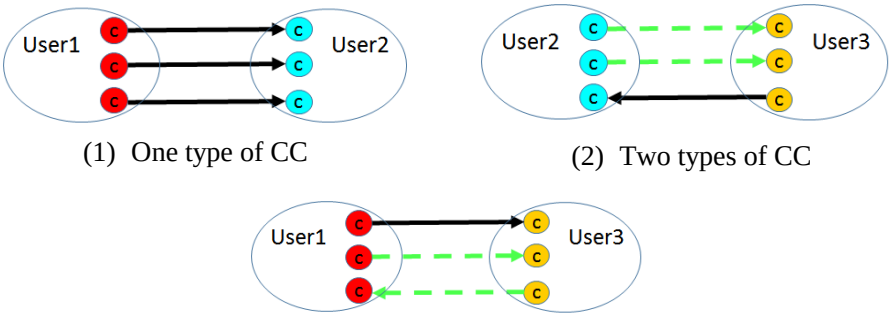
Based on the communication model, some metrics will be calculated to characterize a group of like-minded users. Valid user communication is one metric that is defined for the user to associate herself with other possible interactive users. Keywords of comment represent the main idea that user expresses in the submitted comment. General similarity is defined to measure the similarity between two user groups by comparing users in the two groups and their total keywords of the comments related to valid user communication in the two user groups. The similar weight of element is defined for both each user and each keyword measuring their importance in one user group. Negative karma score of user group is designed and calculated to show the extent of negative impact that one user group might have on the society. In the following paragraphs, these metrics will be explained in detail.

(1) Valid user communication (VUC) in user level
Within topic (Subreddit), if the user communication meets one of the following requirements, it will be valid user communication, and the two users become valid user communication pair, which is unit group introduced later. Here are the requirements of VUC between two users:

- ➢ One type of CC
  - Unidirectional parent-child communication (3 or more communications)
  - Unidirectional neighbor brother communication (3 or more communications)
- ➢ Two types of CC
  - Bidirectional parent-child communications (1 or more communication pairs)
  - Unidirectional parent-child communication and reverse neighbor brother communication (2 or more communication pairs)
  - Bidirectional neighbor brother communications (2 or more communication pairs)
- ➢ Three or four types of CC
  - All of them

Some thresholds are given in the requirements. If there is only one type of CC between two users, the number of CC should be equal to or larger than three. For two users having two types of CC in between, there are three cases. The first case is that there should be one or more pairs of bidirectional parent-child communication. The second case is that at least two pairs of mixed ways of bidirectional CC should exist. Two or more pairs of bidirectional brother

communication is the third case. Finally, all pairs of users that have three or four types of CC between them are valid.



(1) One type of CC

(2) Two types of CC

(3) Three types of CC

Figure 3.2 VUC examples

Based on Figure 3.1, Figure 3.2 illustrates VUC according to the requirements. Figure 3.2 (1) shows there is one type of comment communication between User1 and User2. User1 has three PC communications and user2 has three CP communications. Figure 3.2 (2) shows there are two types of comment communication between User2 and User3. User2 has two BQA communications and one CP communication. User3 has two BAQ communications and one PC communication. Figure 3.2 (3) shows there are three types of comment communication between User1 and User3. User1 has one PC communication, one BQA communication and one BAQ communication. User3 has one CP communication, one BAQ communication and one BQA communication.

These requirements are assumptions for data modeling. The actions of users in the forum could be reflected through them. If the result is not good, the thresholds could be modified to collect reasonable VUC data.

(2) Keywords of comment

The user communication property can be modeled by the relationship among comments. However, what users communicate with each other is also important. The content of comments is the basic information exchanged between users. All kinds of words constructing meaningful message are filled into users' comments. In one comment, each kind of word takes up certain percentage of the whole number of words. In all comments, each kind of word appears in certain number of different comments. TF-IDF (Term Frequency-Inverse Document Frequency) value is calculated for each unique word in the comment. The word that have larger value in certain comment can represent the content of the comment better.

TF-IDF algorithm is introduced in chapter 2. The TF value of the unique word in the comment is larger, which means the word takes up more percentage in the comment. If the IDF value of the unique word is larger, it means that the word appears less among all different comments and be special in its comment. Therefore, the product of larger TF value and larger IDF value of the unique word in the comment gets a larger TF-IDF value that is more representative for the comment.

After the unique words in the comment are sorted by their TF-IDF values, the keywords that can well represent the comment need to be screened out. A percentage threshold is used here. The assumption is made that keywords of the comment take up high 10 percent words of the comment with highest TF-IDF values. If the number of unique words in the comment is less than 10, the word that has highest TF-IDF value in the comment will be chosen as keyword. In this way, the large number of words in the sample data set is reduced to the small number of keywords which can reflect the content of comments well.

(3) General similarity between two groups

The behaviors of users in Reddit platform are characterized by VUC and keyword. VUC connects users who have communication. The keywords of the comments related to VUC pair are the feature of users. Similar entities in different user groups should be merged to reduce the complexity of study objects. Therefore, the users who have enough similar VUCs and keywords are aggregated into one group. User and keyword similarity between two groups will be considered respectively. The following paragraph will describe this process in detail. The importance is how to compute the similarity between two groups of users.

We use general similarity to compare two groups. The general similarity (GS) is calculated based on two independent metrics which are user similarity (US) and keyword similarity (KWS), given more weight to KWS, because keywords are more important to the intention of groups. The formula is below:

$$GS = (US + 2 * KWS)/3$$

The algorithm of Cosine Similarity is used to compute US and KWS respectively. Cosine operation is between the two vectors. User vector of group is defined for US computation. Keyword vector of group is defined for KWS computation. Therefore, GS value is computed, considering two aspects of similarity. The Cosine value between two vectors is larger, which means two vectors are more similar. If GS value of two groups is larger than a threshold, two groups are similar enough to be aggregated together.

Here shows the examples of calculating US and KWS. Group1 has user set {user1, user2, user3, user4} and Group2 has user set {user3, user4, user5}. The total user set of both groups is {user1, user2, user3, user4, user5}. Then, the user vector of Group1 is (1, 1, 1, 1, 0) and that of Group2 is (0, 0, 1, 1, 1). So the cosine value of two vectors is US value. KWS is computed in the similar way. Group1 has keyword set {“A”, “B”, “C”, “D”, “E”} and Group2 has keyword set {“B”, “C”, “E”, “F”, “G”}. The total keyword set of both groups is {“A”, “B”, “C”, “D”, “E”, “F”, “G”}. The keyword vectors of Group1 and Group2 are (1, 1, 1, 1, 1, 0, 0) and (0, 1, 1, 0, 1, 1, 1) respectively. Therefore, the cosine value of these two vectors is KWS value.

(4) The similar weight of elements

The group is related to similar users and similar words. Some users and words are more important in the final group because they appear as similar elements in more original groups. When these groups are merged into one group, similar elements will be repeatedly compared more times, which means that they are more important. Each element is initially given similar weight 1. When two groups are merged together, the similar weights of the elements in the intersection of two

groups will be added up. Hence, in the final group, the elements that have larger similar weight will be more important in the group.

(5) Negative karma score of group
Users are aggregated together by similarity. Some groups of users have much negative impact. In order to measure the extent of impact, the karma score of the comment is taken into account. Group is consisted of users who have similar user communication. The user communication happens between the comments. Thus, karma score can be accumulated from comments for each group. The negative impact can be amplified by only accumulating the karma score of the comments which are negative. In the case that comment has positive karma score, only 0 is added into the sum of negative karma score of group, which mean that the positive score will not be considered when assessing negative impact of group.

3.2.5 Grouping and Merging Process

In order to obtain intention groups, users with similar intention need to be aggregated together. Based on the communication model, users are initially split into unit groups which are VUC pairs. For example, Figure 3.3 (1) shows different pairs of users related to user1. Because these VUC pairs are related to the same users, which is the similar part automatically, they are merged into one basic group in Figure 3.3 (2). Then basic groups enclose users between which there are communications. Different basic groups have similarity in the aspects of users and keywords. The same users might have similar communication. The communications in the group might be related to the same keywords. From then on, the GS value between basic groups is calculated to compare their similarity. If two groups are similar enough, they will be merged into one group, which is shown in Figure 3.3 (3).

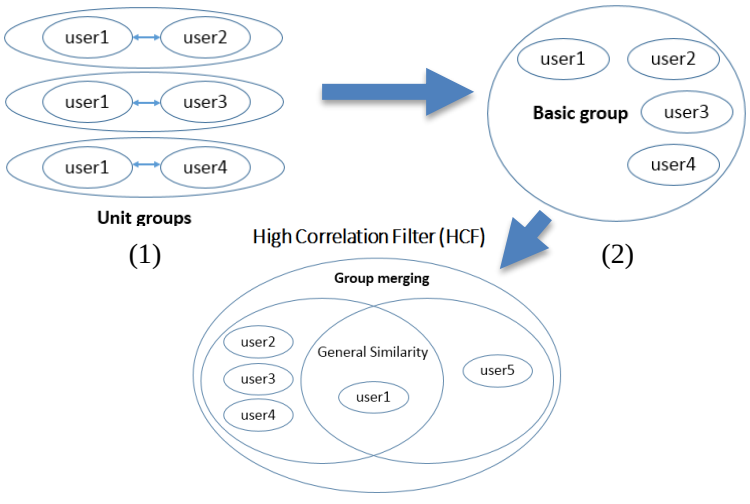


Figure 3.3 Merging process

3.2.6 High Correlation Filter (HCF)

HCF is an algorithm for dimension reduction [3]. Feature vectors of groups having similar trends are likely to carry similar information. The correlation

coefficient is calculated as GS between two vectors. A pair of groups with GS value higher than a threshold is merged into one groups. As mentioned previously in Figure 3.3 (3), basic groups will be merged into less number of groups using HCF algorithm. The number of the groups can measure the complexity of the comment data set. Each group is one dimension of the whole data set. As the number of groups decreases, the dimensions of the data set are reduced.

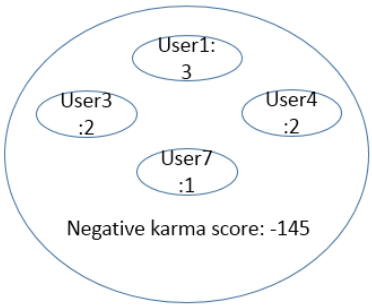


Figure 3.4 Expected result

3.2.7 Expected Results

Previous methods aggregate users into different groups regarding the similarity of the user communication and the keywords between groups. Each group has a negative karma score. The lower it is, the more negative impact the group has. The groups having lower negative karma score are expected.

When the negative groups are found, the importance of the elements (users or keywords) in the group should be considered. The similar weight of the element can reflect the importance well. The larger value the weight is, the more important the element is. For example, in Figure 3.4, the karma score of the group is -145. In this group, there are 4 user elements. User1 is most important because it has largest similar weight, which is 3. The second important elements are user3 and user4, weights of which are both 2.

3.3 Reliability and Validity

Reliability means if the research result is objective and can be obtained by other researcher using the same method. Our research is quantitative analysis by building data model. There are some threshold parameters according to the modeling assumption. The result data was collected by the programming implementation. The program has been run several times by changing or recovering threshold parameters. The collected data were the same. Therefore, the reliability is not a problem in our research.

Validity means if the research result is reasonable. Our research applies to the specific platform. To some extent, the result can measure the expectation of the effectiveness of data analysis modeling. Several thresholds in the research are changeable, which is up to people who have different point of view on communication and similarity. Hence, the validity of our research is based on the assumptions that have been made.

3.4 Ethical Considerations

Our research will not leak any privacy of users. The raw data is from Reddit database, and the data display comment ID, parent ID, user name or some other contents that have no relations with real user privacy. User name is a nickname which is a symbol of user, but not real name.

Chapter 4

Implementation Details

4.1 Technology Toolsets

In order to verify the effectiveness of data modeling, a stand-alone application needs to be developed to process data. We choose a combination of Python and SQLite database to build a full data transformation, algorithm execution and analysis system. This chapter first provides a high level summary of the tools used and then discusses the details of the algorithm.

4.1.1 Python Language

Python is a widely used high-level, interpreted programming language [23]. The aim of the language is to make code more readable and to express ideas in fewer lines of code than in other famous language [24] like C++ or Java. The features of Python are dynamic typing, automatic memory management and having a large and comprehensive standard library.

4.1.2 SQLite Database

SQLite is a relational database management system. Comparing with many other database management systems, it is relatively easier to use since it's not a client-server database engine. SQLite is a popular choice as an embedded database software for local storage in application software. SQLite is used here for its easy configuration and its easy backup through copying the individual database file.

4.1.3 Gephi Graph Visualization and Manipulation software

Gephi is an open-source software package for network analysis and visualization, which is written in Java on the NetBeans platform [25]. The software can be download from its official website – gehpi.org. One of its applications is to analyze social network. Social data connections can be easily created to map community organizations and small-world networks. Node can be defined according to different needs. Edge is defined as connection between two nodes. The network graph can be generated automatically based on nodes and edges. The connections in the graph can be exported as CSV formatted file. Reversely, CSV formatted file can also be created and imported into Gephi software. We will use database to analyze user social relation in different groups and the data will be the query results. Therefore, the results can be converted to CSV file and imported into Gephi platform. The user connections in group can be shown as graph intuitively.

4.2 Description of Data Transformation and Algorithm

4.2.1 Overview of Data Processing

Figure 4.1 provides the overview of our system. There are six major components. The input to the system is sample data from Reddit platform in zipped JSON

format with one-month span. The data importing and subsequent parsing of the JSON file is implemented in Python, which contains several steps and interacts directly with SQLite database. When the data processing application finishes running, all processed data are stored in the database. Then the SQL scripts are used to obtain statistical results from the previously generated data for assessing the effectiveness of the methods. Some of the results are shown in Excel diagram, and the others are shown in Gephi visualization software.

The data processing consists of several methods and are applied in sequence. At each step, the intermediate results need to be persisted in the database for further processing.

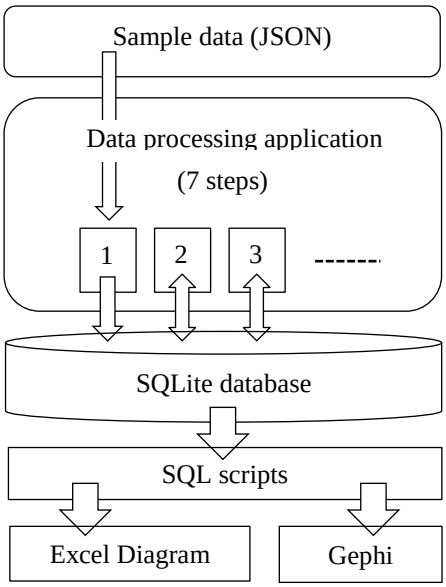


Figure 4.1 Architecture of implementation components

Figure 4.2 shows the multiple steps for the data extraction and the subsequent merging algorithm. In this pipeline, each step is implemented as a Python procedure. The output of one step will be the input for the next phase.

To assist the caching of intermediate data, several database tables are defined. They are listed in Table 4.1.

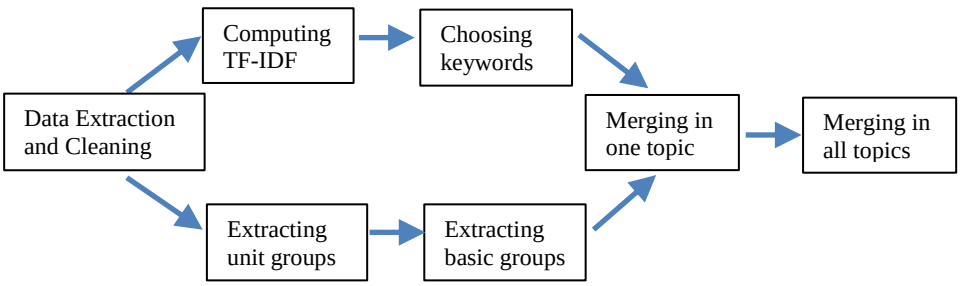


Figure 4.2 The steps of data processing

Table name	Description	
COMMUNICATION	**Source**	Data Extraction and Cleaning
	Destination	Unit Group
	All possible communications between two comments.	
WORDSEXIST WORDTF WORDIDF	**Source**	Data Extraction and Cleaning
	Destination	TF-IDF
	All identified words of comments; The TF and IDF values of each kind of word.	
WORDTFIDF	**Source**	TF-IDF
	Destination	Keywords
	The TF-IDF values of all kinds of words	
WORDTFIDFTOP	**Source**	Keywords
	Destination	Merging in one topic
	All keywords which have highest values in each comment and will be used when merging groups in one topic.	
GROUPUNIT	**Source**	Unit group
	Destination	Basic group
	All possible valid user communications (VUC) between two users.	
GROUPBASIS	**Source**	Basic group
	Destination	Merging in one topic
	All user groups in which one user has VUC with other users.	
GROUPMERGE_WITHINTOPIC_USER_WEIGHT GROUPMERGE_WITHINTOPIC_WORD_WEIGHT GROUPMERGE_WITHINTOPIC_SCORE	**Source**	Merging in one topic
	Destination	Merging in all topics
	The result of merging basic groups within one topic, in which all possible similar users in one group are aggregated together; Three kinds of metric data stored in three tables.	
GROUPMERGE_ALLTOPIC	The relationship between previous group number and final group number.	

Table 4.1 the description of main database tables. They serve as caching points for more efficient operation.

4.2.2 Data Extraction

Data model defines some metrics, like keywords, valid user communication, karma score etc. The metric related data, such as words of comment and comment communication, need to be extracted from the sample data. In order to manage data in context, the noisy and missing data need to be cleansed and the integrity of the data needs to be ensured.

4.2.2.1 Data Cleaning

There are several kinds of problem data. One problem is that the user names of some comments are deleted. Actually, the following data processes only care about user IDs generated according to the user names. Since some deleted users have same root comments, they are given the same user IDs as root comment. The other problem is that there are no comment data for some parent comment IDs that some comments have. This is because comments data are intercepted over time and some data associations are missing at the beginning of the intercepted data. In this case, the user IDs are set to an invalid number 0 for these parent comments that only have a comment ID. It means that the communications related to the invalid user IDs are invalid and will not be considered.

4.2.2.2 Extracting Communication Data

These are important metric related data. Two ways of communication data are obtained in different steps. When each comment is extracted, parent-child communication is stored based on the relationship between comment ID and its parent comment ID. In the meantime, a mapping dictionary is made between parent comment ID and all children ID. After all comment data are extracted, neighbor brother communication data are stored based on the mapping dictionary. Given a comment ID, all children ID are found through looking up the dictionary and sorted in ascending order. Between two neighbor child comments, the one with smaller ID is published earlier and regarded as question node, and the other one with larger ID is later as answer node.

Table 4.2 shows typical structure of a piece of comment communication data.

Field	Description
SID	Topic (Subreddit) ID
CID	ID of comment being observed
CUID	User ID of observed comment
CSCOREN	Negative karma score of observed comment
CCSCOREN	Negative karma score of communication comment
CCUID	User ID of communication comment
CCID	ID of communication comment
CCRELATION	The role of communication comment. (parent, brother, child)
QATYPE	The node type of observed comment (Question node, Answer node)

Table 4.2 the structure of comment communication data

4.2.2.3 Extracting Comment Words

The body of each comment has content filled with words. The whole body can be regarded as a string. The words will be identified by reading characters of the string one by one, and be associated with corresponding comment in the table.

4.2.3 Choosing keywords based on TF-IDF

The number of each kind of word in the comment and the total number of words in each comment are computed by querying table in database. The ratio of two

numbers is TF value of the word of certain comment. IDF value of the word is the result of dividing the total number of comments by the number of comments having the word. The product of TF and IDF value of the word is stored as TF-IDF value of the word, which is related to comment ID, user ID and topic ID.

Keywords are the certain percentage of words that have highest TF-IDF value in the comment. In our research, 10 is specified as the percentage, which means top 10 percent words are keywords.

4.2.4 Extracting Unit Groups

The phase is to screen out valid user communication from all comment communication data. Within one topic, the number of each type of comment communication between two users is counted out, and stored in a mapping dictionary related to each type. For each pair of users in one topic, they might have 1, 2, 3 or 4 types of comment communication between them. If the number of comment communication meets the requirements of VUC defined before, the two users have VUC between them, which will be stored in the table. A mapping dictionary is used here to store the number of each type of comment communication between users and verify the requirements of VUC.

These valid pairs of users are given initial group numbers, regarded as unit group. In the meantime, the negative karma scores of comments are accumulated to the valid user pair.

4.2.5 Extracting Basic Groups

Some user communicates with different users in different unit groups. Such user acts as core role among a group of users. The program sorted user pairs and merged all pairs having the same user into basic group.

4.2.6 Merging within Topic

This phase is to aggregate all possible similar basic groups within one topic. Five kinds of set dictionaries are built to do group merging, which are user ID, user ID with similar weight, keyword, keyword with similar weight and negative karma score of group. User ID and keyword set dictionaries are used to calculate similarity respectively. The set dictionary with weight are used to accumulate weight values of the same elements, which can reflect the importance of the elements. The dictionary of negative karma score is used to sum the total score of each group. The pseudo code of merging algorithm is shown in Algorithm 1 in Appendix.

4.2.7 Merging in All Topics

This phase has no topic limitation. The merging happens in all topics. The aim is to create relationship between final group ID and previous group ID. There are only two set dictionaries, user ID and keywords, which are used to compute similarity respectively. Once two similar groups are merged, new relationship between two previous group ID and new group ID will be created. The loop

algorithm is similar with the one in previous section. When the number of groups does not change any more, the final group IDs will be connected to previous group IDs, and stored into the database table.

Chapter 5

Results Analysis and Evaluation

The aim of this chapter is to analyze the result data of experiment from different perspective and evaluate the result to show significance that the data modeling and methods can bring.

5.1 Valid User Communication (VUC)

The communication model defines two ways of comment communication (CC). Between two users, there are four types of unidirectional CC, which are PC, CP, BQA, and BAQ. To generate valid user communication (VUC) data, some threshold values are set as assumption. Table 5.1 shows the distribution of VUC. Total number of VUC is 29095. Each requirement for VUC has a number of corresponding VUCs. The majority is bi-directional parent-child communication. VUC that have three or four types of CC occupies a large part, either. The parts of unidirectional CC, two-way reverse CC and neighbor brother CC are small, but not none.

		CC Types	**CC Threshold**	**VUC**
Number of CC types between two users	1	PC	3	127
		CP	3	127
		BQA	3	28
		BAQ	3	28
	2	(PC, CP)	(1, 1)	19931
		(PC, BAQ)	(1, 2)	69
		(CP, BQA)	(1, 2)	69
		(BQA, BAQ)	(2, 2)	458
	3	*	(1, 1, 1)	6118
	4	*	(1, 1, 1, 1)	2140

CC Types: PC (Parent-Child); CP (Child-Parent); BQA (Brothers-Question-Answer); BAQ (Brothers-Answer-Question)

CC: Comment Communication; *: All possible combination of three types or four types

Table 5.1 Valid user communication data

5.2 Dimension Reduction and Comparison

One goal of dimension reduction is to aggregate users into fewer groups in which users have common features. Figure 5.1 shows the variation of group number in different merging phases. At the beginning, each user is considered as an individual group. After the merging step, the number of users is reduced. However, these users have communication. Then, there is a big rise of number of unit groups because one user may be in several unit groups. At the phase of basic group, all unit groups related to the same user are merged into one basic group.

The group number decreases dramatically. The following step is merging groups within topic. The general similarity is calculated for making the decision of merging group. The threshold in this case is not less than 0.4, which means two groups are similar enough to be merged if their similarity is equal to or larger than 0.4. The last phase is merging in all topics. Two groups in different topics are merged together if their general similarity is not less than the threshold of this phase which is 0.3. The selection of the two threshold values will be described in section 5.3.

The group number is reduced from 24913 at the beginning to 4465 as the application result of communication modeling. Each group is the study object that is meaningful because of user similarity and keyword similarity.

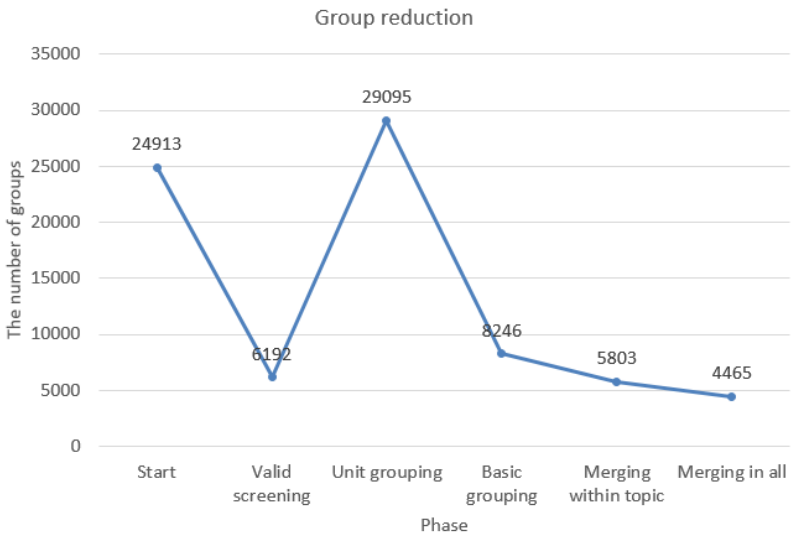


Figure 5.1 Group reduction

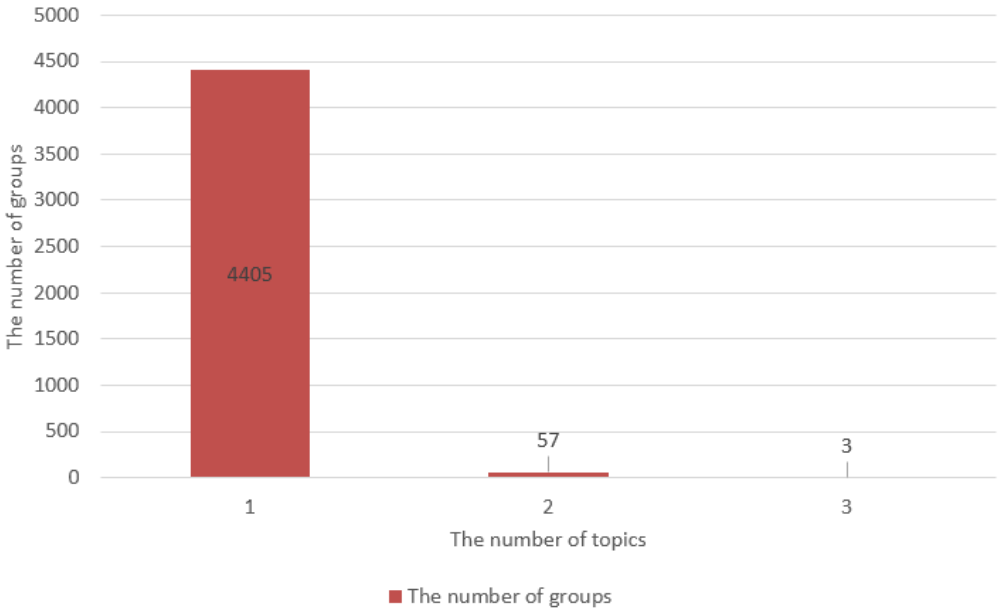


Figure 5.2 Number of topics related to one group

The number of related topics in one group ranges from 1 to 3, which is shown in Figure 5.2. There are 4405 groups, each of which has one related topic. 57 groups have two related topics. Three groups have three related topics.

The average user dimension of groups is shown in Figure 5.3. The number of users in initial data set is 24913. At the phase of unit group after the application of communication model, the average number falls to a minimum 2. After group merging, the average number goes up to about 6 because more similar users are aggregated into one group.

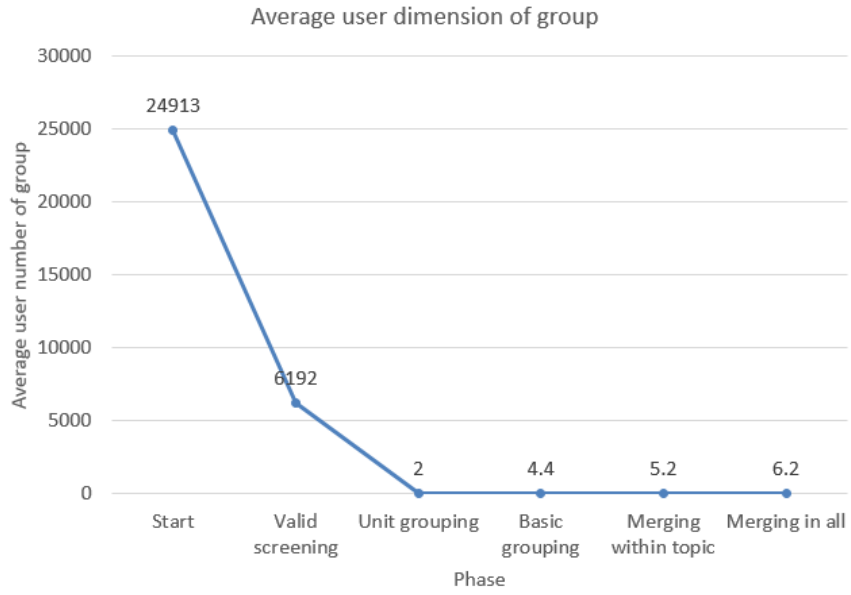


Figure 5.3 Average user dimension of groups

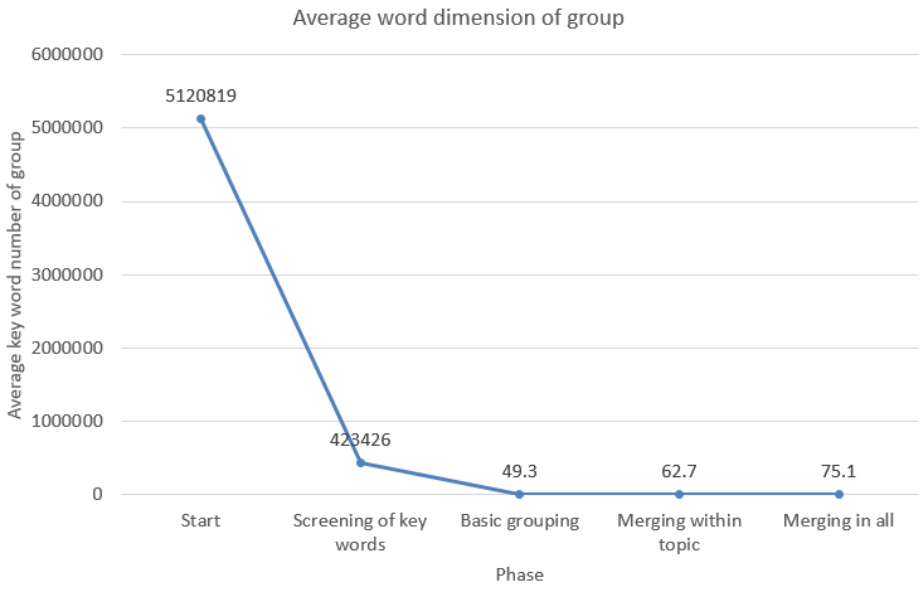


Figure 5.4 Average word dimension of groups

Figure 5.4 shows the average number of keywords in one group. The variation is similar with that of the average number of users in one group. The average number falls dramatically to about 49 after basic groups are generated. After group merging, the average number rises to over 75, which means that more users bring more intent into the group.

Table 5.2 and Table 5.3 compare the result with initial data set and natural topic group. The dimensionality is reduced dramatically. Some groups have more than one related topic, which means that similar users in one group have different preferences.

Study object	Users	Keywords	Topics
Initial data set	24913	5120819	32
Grouping result	6	75	1-3

Table 5.2 Contrast with initial data set

Study object	Users	Keywords	Topics
Natural topic group	779	160026	1
Grouping result	6	75	1-3

Table 5.3 Contrast with natural topic group

5.3 Merging Threshold Adjustment

Users are aggregated into unit groups, and then basic groups by the communication model. There are more probabilities of communication between users in each basic group. However, some metrics, mentioned in Chapter 3, could be used to characterize each of these groups, which are similar communicative users, and similar communication content, that is, the same keywords. Percentages could be computed for both kinds of similarity respectively. Based on this, general similarity is calculated, given more weight to the content of communication. The critical step is to compare the similarity of groups and merge two groups whose similarity meets the threshold requirement according to HCF dimension reduction algorithm. In Figure 5.5 and Figure 5.6, different merging results are shown when the threshold of general similarity is changed.

Figure 5.5 shows the result of merging within topic. The green line is the initial number of groups which is actually the total number of basic groups. It is constant when different merging results are generated. The larger threshold of similarity means that two groups need to be more similar to merge. Hence, there should be more similar users and more similar keywords. As the threshold gets larger, the curve goes up towards the number of basic group. More groups are kept unchanged and merging happens less times. When the general similarity threshold is 0.6, the number of groups is only reduced to 7186 which is close to the initial number. When the threshold is changed to 0.25, the lower number of groups, 3548, is generated. However, this seemingly small number of group scenario loses much similarity in user communication and the kind of information they exchange. Thus, middle general similarity 0.4 is chosen for the consideration of both similarity and group merging.

The number changing process during merging in all topics is shown in Figure 5.6. The green line is the initial number of groups in this process, which is the result of merging within topic. The red variation curve in this figure is similar with the one in Figure 5.5. When the threshold of general similarity is 0.4, merging happens rarely. The red line and green line nearly converge at this point. At the other end of the curve, 0.2 general similarity threshold reduces the number of groups into

2585, but for the same reason in Figure 5.5, this result also loses similarity of user communication and what kind of information they exchange. Therefore, the middle threshold 0.3 is chosen to get appropriate merging result.

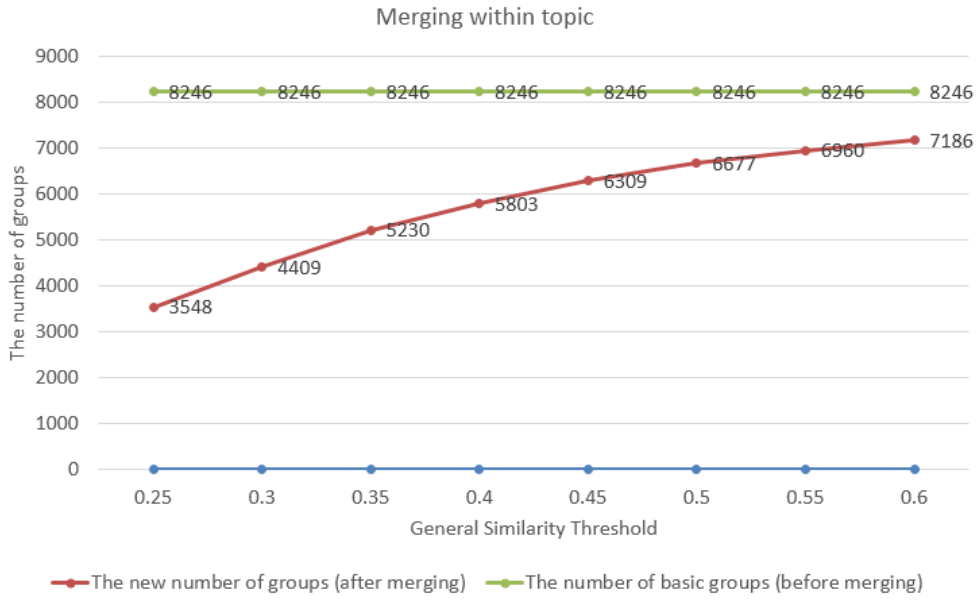


Figure 5.5 Merging within topic

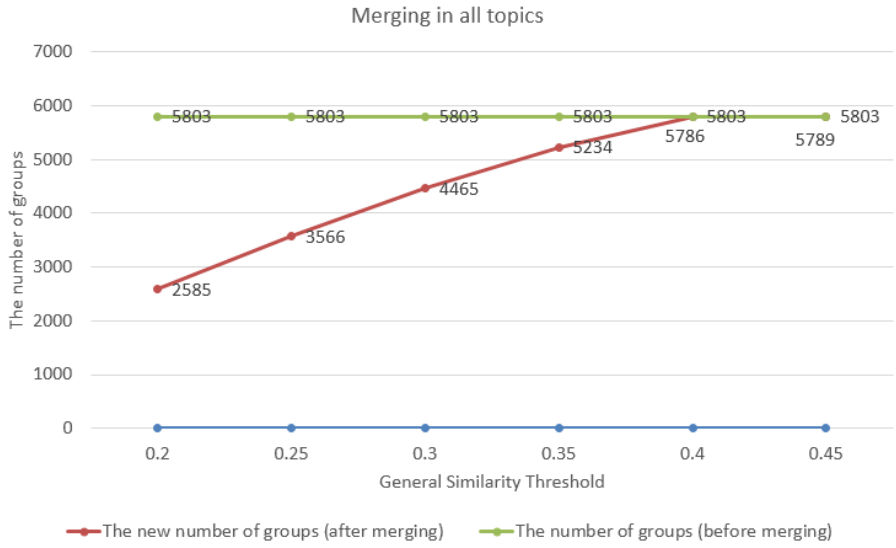


Figure 5.6 Merging in all topics

5.4 Negative Score Groups

In order to get certain groups having negative impact on the society, the negative karma scores in each group are accumulated to amplify the negative impact. Each comment has three kinds of scores, which are negative, zero and positive. The communication between two users is consisted of several pairs of communication comments. The negative score of group begins to be accumulated at the time that the unit group is formed from the valid communication user pair. Zero will be added if the score of certain comment is not negative. In the following steps, each time when several groups are merged into one group, their negative scores are

added together. Thus, at the end of merging in all topics, all groups have a score, negative or zero. The lower score has more negative impact on society.

Table 5.4 shows 30 groups whose negative scores are lower than threshold -200. To some degree, they could be considered as the groups having negative impact on society. The score range is from -7831 to -201. The lowest score -7831 is much lower than other scores. It means that Group 21699 is the one that the research expects most. However, different people would associate negative impact with different degrees of negative score. Based on the following score results, a new threshold could be given to choose a new set of negative impact groups. For example, if the threshold is -300, 18 groups the scores of which are lower than -300 will be chosen, which means fewer groups having negative impact on society are chosen when the condition becomes harsh.

Sequence	Group ID	Score	Sequence	Group ID	Score	Sequence	Group ID	Score
1	21699	-7831	11	25191	-390	21	23597	-276
2	23406	-1232	12	25349	-358	22	21262	-270
3	22929	-832	13	19897	-355	23	22070	-267
4	24443	-602	14	10334	-346	24	23377	-261
5	22231	-538	15	12021	-344	25	19876	-255
6	24014	-519	16	10122	-328	26	25934	-227
7	21414	-478	17	13991	-322	27	19583	-220
8	10351	-469	18	25681	-313	28	19488	-212
9	12164	-467	19	22815	-288	29	18338	-203
10	10717	-393	20	12533	-279	30	19544	-201

Table 5.4 Negative score groups

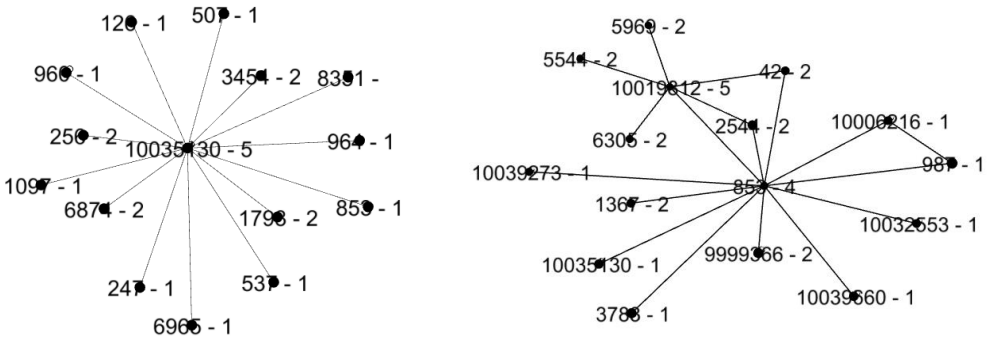
5.5 The Feature Distribution of Intention Group

The result of the research is a set of intention groups. There are two important features in each group, which are user communication and the information they exchange with each other. Because the weights of similar users and keywords are accumulated when merging happens, the larger weights means that the corresponding users or keywords are more important.

The user communication graphs of two groups are shown in Graph 5.1. The graph is generated by Gephi Graph Visualization and Manipulation software. Through SQL statement, all of the communication in certain group is collected into a set of records each of which has a pair of user id. All records are imported into Gephi by a formatted CSV file, and then a communication graph could be generated. Each node is labeled with user id and user weight.

The graph of Group 21699 is a star topology, and the graph of Group 22929 is a networking topology. In Group 21699, the weight of user 10035130 is 5, which means that this user is similar in most merged groups and very important. The importance of this user node is also reflected from its core position in the graph. The other user nodes with weight 2, like user 256, 3454, 6874 and 1798, are also important.

In the graph of Group 22929, the communication happens in more pairs of users. Both user 10019812 and 853 are most important with the weight 5 and 4 respectively. The other users with weight 2 in different positions of the networking topology also show their importance. In this group, more users communicate with each other and the interactions are more active.



(1) Group 21699 (2) Group 22929

Graph 5.1 Two communication graphs of users with weight

Table 5.5 lists all keywords in Group 21699 the weights of which are larger than 1. The words having weight values in top 5 are “homosexuals”, “include”, “marriage”, “redefine” and “redefined”. To some extent, these keywords could represent the intention of this group in general. The other keywords in the table are also relatively important.

Group ID	Keyword	Weight	Keyword	Weight	Keyword	Weight
21699	billy	2	goat	2	process	2
	binding	2	goats	2	propagandizing	2
	bless	2	guardianship	2	proposition	2
	calling	2	homosexuals	5	redefine	4
	canada	2	illustration	2	redefined	3
	cannot	2	include	5	redefining	2
	ceremony	2	involving	2	reilly	2
	churches	2	legal	2	relationship	2
	consent	2	legalize	2	spews	2
	constitutionality	2	lou	2	tolerance	2
	decomposition	2	marriage	5	vote	2
	defacto	2	marriages	2	ways	2
	deleted	2	marry	2	whichever	2
	entity	2	person	2	whites	2
	gave	2	pftl	2	why	2
	give	2	possession	2		

Table 5.5 Keywords with weight in group 21699

5.6 Results Discussion

Our research is to design a special application of dimension reduction to discover the potential intention group in which people have similar features and preference. A major objective is to gain certain intuition on the clustering effect of users as well as the topics themselves. Given the limited time and resource, we choose an approach based on a careful analysis of the detailed communication types between different users and further adopt a simple rule-based group merging algorithm.

We have built an end-to-end data exploratory process to help gain intuition from the Reddit dataset. One can further extend the intuition obtained from this research and combine with more advanced machine learning clustering algorithms.

Our results demonstrates that a) our communication model is reasonably effective and sensible in terms of filtering the relevant interactions between the users; b) a simple rule-based merging algorithm based on a weighted cosine-similarity measure works satisfactorily for grouping users with similar intent and thus achieve the goal of first level dimensional reduction.

Compared with other researches on question and answer forum [15], the requirements of our communication model are very close to people's actual social habit. One-time communication between two users' comments, shown as black edge in Graph 3.1, can mean that there is random talk between them. Table 5.1 shows that there exist certain times of conversations between two users according to our model, which can be regarded as formal social conversation and be used to construct user's communication network. The communication between brotherhood users can well reflect the flexibility and diversity of social habit. Furthermore, social conversations make people connect with each other. The connections can draw people who have common interests together, which is shown in the communication topologies of Graph 5.1 (1) and (2). The result shows that the complexity of study object is reduced dramatically, comparing Table 5.1 and 5.2. Massive users in the original data are grouped into small groups with obvious number of users by the similarity among them, for example, Group 22929 in Graph 5.1 (2). The features of this small group can be understood reasonably clearly.

The features can be embodied in several aspects: the connections among users, the importance of users and the importance of keywords. The importance of the element in the result is assessed through the weight of the element explained in Chapter 5. The weight is a concrete value that can objectively measure the degree of importance, which is calculated from the actual data. In Graph 5.1, the users who have higher weight can be quickly known as the key roles of ideas exchanging. In Table 5.5, it is clear to see the important words with higher weight in the group. Some other researches have been done previously to predict the popularity of the content in the topic [13, 18]. However, it is more meaningful to know the popular content in the certain user group, especially those who have potential threat to the society. Additionally, the popular keywords in one group are mainly about the common interests and intention of users in the group.

Moreover, the result of our research can also show the dynamics of information transfer in the network. For example, in Graph 5.1(2), to some extent, user 5969 and user 9999366 can exchange ideas indirectly through some consecutive connections between them. The possible path of information transfer could be "5969←→10019812←→2544←→853←→ 9999366".

One drawback of our system is that the efficiency in the data collection is not very good. The program had scalability issue. The sample data had only 150,429 comment records. It took about 22 minutes to finish grouping data and get the result. When we ran program on a larger data set having one month span, the program became very slow and even had no response. This larger data had

323,925,284 comment records. It took 150 minutes to finish extracting data from original data, that is, first step, while previously the small number of sample data only took 1.5 minutes. There is 100-fold difference between them. When we continued to do further data processing, the program began to have no response. We analyzed the problem. This was mainly because of the way how we use the relational database. The bottleneck was the performance of database operation when the resource was occupied exclusively and all operations were serial. It is better to use distributed concurrent programming to support big data processing.

Putting our work in the general context of clustering and dimensional reduction literature (chapter 2), we feel our problem probably should have been modeled as a community detection problem on a network graph. Namely, once we've defined the communication model and the similarity metric, we have essentially built a unipartite communication network with the nodes being the users, edges their communications and the weights the similarity score. Merging users of different groups corresponds to community detection on this graph.

Ref [29] reviews many different approaches for solving this kind of problem. In general, community detection on graphs focuses on finding clustered nodes (sub-graphs) that have high density of edges within the sub-graphs and low density of edges between them. Key elements in the concept of community include:
a) measures of intra-cluster density and inter-cluster density, which are defined as the edge densities within and between communities. b) measures of the adjacency between the vertices -- similarity scores between nodes. The evaluation of partitions can be done via a quality function, with the most popular one being that of Newman and Girvan's [30]. It is calculated by comparing the actual sub-graph density of edges against a null model (random graph)'s edge density. A large positive modularity value indicates good partitions. It can be expressed in terms of the intra-cluster and inter-cluster densities.

In our approach, we utilize the same concept of 'cosine similarity score' to evaluate the vertices' adjacency. However not enough exploration and analysis was done regarding the intra or inter-cluster edge densities. This makes it hard to draw conclusions regarding the optimality of our solution.

As a data exploratory tool, our rule-based group merging algorithm has the advantage of being single-pass and hence having a linear complexity in the number of nodes (with a multiplication factor to compute the similarity scores).
It is probably closest to the Agglomerative Algorithm in the Hierarchical Clustering method category, in which one starts from the basic nodes and iteratively merging them if the similarity is sufficiently high. This kind of method has the advantage that one does not need to specify the number and size of the clusters. However it also has the issues that a) a tendency to create very sparse clusters: vertices with only one neighbor are classified as separate clusters. b) results can be sensitive to the similarity measures. As a next step, it'll be very interesting to probe deeper the algorithm performance from these perspectives.

Chapter 6

Conclusions

Reddit comment data is a big information pool filled with large amount resources and ideas published by users from all over the world. The characteristics of the comments have been analyzed through our data modeling. A communication model was built to capture the features of users. Based on this, similar users were aggregated into different intention groups by some defined special rules and practical algorithms. The Python scripting programs were developed to implement our proposal, process data and collect experiment data. The result shows the effectiveness of our data modeling and design. Users were grouped together by their similarity. In some groups with negative karma score, the communication topology shows the importance of different users when exchanging information. Furthermore, there are some primary keywords that can well represent the intention of the group.

However, some important information in comments is deleted. If these data are included clearly, the intention of group will be more meaningful. Moreover, some parent comment data are missing which might have important communication and intention clues to aggregate users and make the result more meaningful as well.

The amount of Reddit comment data is increasing exponentially. Our sample data is chosen for experiment, which is only a very small part of data tsunami [1]. Given more time in the future, we would like to do a head-to-head comparison between a standard network clustering algorithm and our rule-based approach. We would also like to improve the efficiency of the program, make the data processing part more scalable to support the analysis of big data. Additionally, it would be a good idea to add more metrics into our data analysis. For instance, the activity level of users, which can reflect and predict the effect of users on the current event, could help to find group of users who are active in terms of information sharing.

References

[1] A. Labrinidis and H. V. Jagadish, ”Challenges and Opportunities with Big Data,” *VLDB Endowment,* vol.5(12), nr VLDB Endowment, pp. 2032-2033, 2012.

[2] S. Lafon and A. B. Lee, ”Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 28(9), nr IEEE Computer Society, pp. 1393-1403, 2006.

[3] R. Silipo, ”KDnuggets,” KDnuggets, 7 May 2015. [Online]. Available: http://www.kdnuggets.com/2015/05/7-methods-data-dimensionality-reduction.html. [Used 10 February 2016].

[4] F. Xiong and Y. Liu, ”Empirical Analysis and Modeling of Users’ Topic Interests in Online Forums,” *PloS One,* vol. 7(12), nr Public Library of Science, pp. 1-7, 2012.

[5] L. Van Der Maaten, E. Postma and J. Van Den Herik, ”Dimensionality reduction: A comparative review,” *Journal of Machine Learning Research,* vol. 10, nr Microtome Publishing, pp. 1-41, 2009.

[6] I. K. Fodor, ”A survey of dimension reduction techniques,” 9 May 2002. [Online]. Available: https://e-reports-ext.llnl.gov/pdf/240921.pdf. [Used 15 March 2016].

[7] D. L. Donoho, ”High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality,” 8 August 2000. [Online]. Available: http://statweb.stanford.edu/~donoho/Lectures/AMS2000/Curses.pdf. [Used 23 March 2016].

[8] M. Partridge and R. Calvo, ”Fast dimensionality reduction and Simple PCA,” *Intelligent Data Analysis,* vol. 2(3), nr IOS Press, pp. 292-298, 1997.

[9] N. Kambhatla and T. K. Leen, ”Dimension reduction by local principal component analysis,” *Neural Computation,* vol. 9(7), nr MIT Press, pp. 1493-1516, 1997.

[10] D. Niu, J. G. Dy and M. I. Jordan, ”Dimensionality Reduction for Spectral Clustering,” in *AISTATS,* vol.15, pp. 552-560, Fort Lauderdale, 2011.

[11] Z. Zhang and M. I. Jordan, ”Latent Variable Models for Dimensionality Reduction,” in *AISTATS,* vol.5, pp. 655-662, Clearwater Beach, 2009.

[12] Na Ni and Yaodong Li, "User Interests Modeling in Online Forums", ASONAM, 2012, 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2012, pp. 708-709, doi:10.1109/ASONAM.2012.122.

[13] A. Terentiev and A. Tempest, ”CS 229 Machine Learning, Final Projects, Autumn 2014,” 29 October 2014. [Online]. Available: http://cs229.stanford.edu/proj2014/Terentiev%20Tempest,Predicting%20Reddit%20Post%20Popularity%20ViaInitial%20Commentary.pdf. [Used 15 March 2016].

[14] Reddit, ”About Reddit,” Reddit, 10 March 2016. [Online]. Available: https://www.reddit.com/about/. [Used 10 March 2016].

[15] S. Budalakoti, D. DeAngelis and K. Suzanne Barber, ”Expertise Modeling

and Recommendation in Online Question and Answer Forums," in *Computational Science and Engineering*, vol.4, pp. 481-488, Miami, 2009.

[16] F. Rahutomo, T. Kitasuka and M. Aritsugi, "Test collection recycling for semantic text similarity," in *Information Integration and Web-based Applications & Services,* pp. 286-289, Bali, 2012.

[17] W. Wang and Y. Liu, "Recommendation algorithm based on customer behavior locus," *Computer Engineering and Applications,* vol. 09, nr Publishing House of Journal of Computer Engineering and Applications, pp. 35-38, 2006.

[18] J. Segall and A. Zamoshchin, "CS 229 Machine Learning, Final Projects, Autumn 2012," 29 October 2012. [Online]. Available: http://cs229.stanford.edu/proj2012/ZamoshchinSegall-PredictingRedditPostPopularity.pdf. [Used 15 March 2016].

[19] N. Barrot, L. Gourves, J. Lang, J. Monnot and B. Ries, "Possible Winners in Approval Voting," in *Springer-Verlag Lecture Notes in Artificial Intelligence 8176,* pp. 57-70, Brussels, 2013.

[20] O. Lartillot, "A Matlab toolbox for musical feature extraction from audio," in *International Conference on Digital Audio Effects,* pp. 237-244, Bordeaux, 2007.

[21] L. Yu and H. Liu, "Feature Selection for high-dimensional data: a fast corelation-based filter solution," in *Proceedings of the 19th International Conference on Machine Learning (ICML 2003),* pp. 856–863, Washington D.C, 2003.

[22] A. M. Martinez and A. C. Kak, "PCA versus IDA," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 23(2), nr IEEE Computer Society, pp. 228-233, 2001.

[23] L. Mark, Learning Python, Sebastopol: O'Reilly Media Press, 2009.

[24] S. McConnell, Code Complete, Redmond: Microsoft Press, 2004.

[25] M. Bastian, S. Heymann and M. Jacomy, "Gephi: An Open Source Software for Exploring and Manipulating Networks," in *Third International AAAI Conference on Weblogs and Social Media,* vol.8, pp. 361-362, San Jose, 2009.

[26] J. Martin, "Recovering subreddit structure from comments," 9 December 2015. [Online]. Available: http://cs.unc.edu/~jamesml/790-final-report.pdf. [Used 17 March 2016].

[27] D. Soni, "Reddit Karma Analytics," 9 January 2016. [Online]. Available: https://analyzekarma.herokuapp.com/. [Used 23 March 2016].

[28] U. Von Luxburg, "A Tutorial on Spectral Clustering," *Statistics and Computing,* vol. 17(4), nr Springer International Publishing AG, pp. 395 - 416, 2007.

[29] S. Fortunato, "Community detection in graphs," *Physics Reports,* vol. 486, nr Elsevier, pp. 75-174, 2010.

[30] M. Girvan and M. Newman, "Community structure in social and biological networks," in *Proceedings of the National Academy of Sciences,* vol. 99(12), pp. 7821-7826, 2002.

Appendix

Algorithm 1 – Group Merging Algorithm. This is our key rule-based algorithm used in the clustering step. Given merging *threshold* setting, the algorithm iteratively loops over each topic. Within each topic, it has an inner loop that computes the similarity score between each user pair and the comment content. If the total similarity score exceeds the threshold, then merge the users into one group. The algorithm continues the process until no valid merging events are left.

```
mergingGroup = intialGroup;
groupQuantityChangingFlag = true;
while (groupQuantityChangingFlag) {
        groupQuantityChangingFlag = false;
        newGroup = {};
        for each topic {
                groupIDSet =  initialGroupIDSet;
                chosenGroupIDSet = {};
                for each groupIDA in (groupNumberSet- chosenGroupNumberSet) {
                        chosenGroupIDSet.add(groupIDA);
                        mergingFlag = false;
                        for each groupIDB in (groupNumberSet- chosenGroupNumberSet) {
                                us = userSimilarity(groupA, groupB);
                                kws = wordSimilarity(groupA, groupB);
                                gs = (us + 2 * kws) / 3;
                                if (gs >= threshold) {
                                        groupA = merging(groupA, groupB);
                                        newGroup.add(groupA);
                                        chosenGroupIDSet.add(groupIDB);
                                        mergingFlag = true;
                                }
                        }
                        if not mergingFlag {
                                newGroup.add(groupA);
                        }
                }
        }

        for each topic {
                if mergingGroup.len() != newGroup.len() {
                        groupQuantityChangingFlag = true;
                        mergingGroup = newGroup;
                        break;
                }
        }
}
```