ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

Data Mining Techniques for Online Communities

[1] Saloni Fathima, [2] A.Rajesh
[1][2] Faculty, Dept of CSE, KUCET

Abstract - **Data mining techniques can be applied to any type of old or new data, each of which can be best dealt with using specific technologies (not requiring all of them).In other words, data mining techniques are limited by data types, data set sizes, and task application environments. Each data set has its own suitable data mining solution. Data mining practitioners often face problems of the unavailability of all training data at the same time and the inability to process a large amount of data due to constraints such as lack of adequate system memory. Once older data mining techniques cannot be applied to new data types or if new data types cannot be converted to traditional data types, new data mining techniques will always need to be explored. The most popular and most basic form of data from the database, data warehouse, orderly data or sequence data, graphics data and text data. In other words, they are joint data, high-dimensional data, longitudinal data, streaming data, web data, numerical data, categorical data, or textual data.**

Keywords: - Abstractive summary, extractive summary, Keyword Extraction, Natural language processing, Text Summarization.

1. INTRODUCTION

Data mining is the discovery of a model in data, also known as exploratory data analysis, where useful, effective, unexpected, and understandable knowledge is found in the data. Some goals are the same as other sciences such as statistics, artificial intelligence, machine learning, and pattern recognition. In most cases, data mining is often seen as an algorithmic problem. Clustering, classification, association rules learning, anomaly detection, regression, and summarization are all part of the data mining task

2. THE BASIC METHOD OF MINING DATA MINING FOR ONLINE COMMUNITIES

2.1 Mining text data

Much of the information is stored in text, such as press releases, scientific papers, books, digital libraries, email messages, blogs and web pages. Therefore, to tap the online community is closely related to the mining of textual data, such as forecasting and analyzing hot topics in online forums [1] online hate group research [2] and so on.

2.2 Mining Web data

The World Wide Web is a huge, widely distributed global information center. It contains rich, dynamic information about structures with hypertext links and multimedia web content, hyperlink information, access and use of information, provides a wealth of resources for data mining. Web mining is the application of data mining technology to discover patterns, structures and knowledge from the Web. According to the analysis goal, Web mining can be divided into three main areas:

Web content mining, Web structure mining and Web usage mining.Web content mining analyzes Web content such as text, multimedia data, and structural data (in-page or linked pages) to understand Web content, scalability-based and information-rich keyword-based page search, entity / concept resolution, web page relevancy and rank assessment, web page content Summary of content, and other valuable information related to web search and analysis. Web Structure Mining uses graphs and network mining theories and methods to analyze nodes and link structures on the Web. Mining through the Web structure, you can Get membership in online community to be ready for categorizing your members. [2]

Web Usage Mining is extracting useful information from the server. It discovers patterns related to general or specific user groups and understands users' searches Modes, Trends and correlations Predict what users are searching on the Internet. This helps to improve search efficiency and effectiveness and also helps at the right time between different groups of users want to sell products or related information [3]

3. THE GENERAL PROCESS OF MINING AN

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

ONLINE COMMUNITY

3.1 Text, Web data acquisition and extraction

3.1.1 To use Web structure mining Sina online sports community data acquisition and extraction as an example [1] First, analyze the structure of Sina online sports community: online Sina sports community presents a root forum, twigs forum and the bottom of the page forum tree structure. Online community website data collection, extract User's personal letter interest, comments and more Data pre-set reason

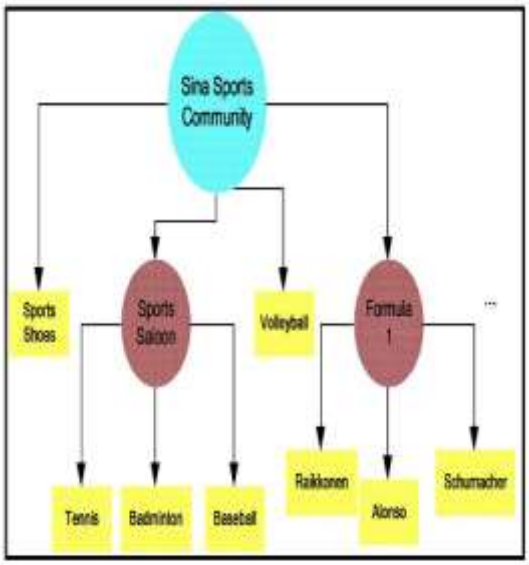

Fig.1 Mining online community

Figure 1 online community mining process Useful user Personal information, comment On and so on Category or poly class Get the corresponding in conclusion Opinion, preference analysis

According to the characteristics of this structure, the use of web crawler for the following steps:

Step 1. Manually create the SINA_ LEAFORUM _URLLIST table

Step 2. Create a table based on the SINA_ LEAFORUM _URLLIST table SINA_FORUM_URL: By parsing each leaf forum the first page to generate a series of top-ranked post URL and save

Step 3. Traverse the URL in the SINA_FORUM_URL table, climb down all the posts, respectively, into SINA_ FORUM_ POST and SINA_ RORUM_ COMMENT_ POST to extract the data.

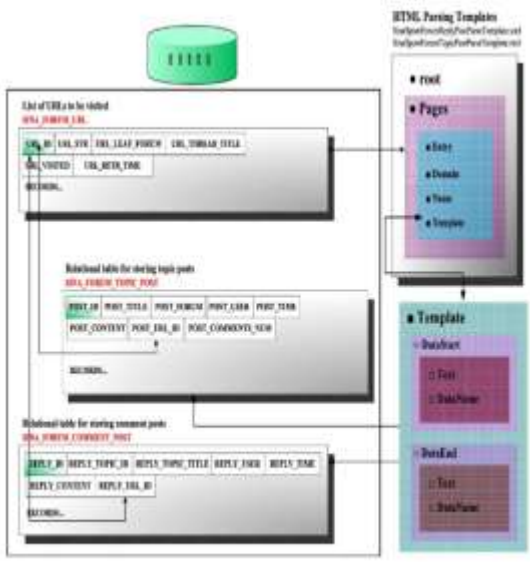

Fig.3 Parsing links in table SINA_FORUM_URL to generate tables SINA_FORUM_URL and SINA_FORUM_COMMENT_POST.

3.1.2 Different background data acquisition, extraction process differences

The above example, the direct use of XML format for data extraction, which requires basically the same format based on the post, but does not rule out the emergence different formats, this point in the data mining and analysis of the blog will be highlighted, because the format of different blogs will be significantly different. At this point, you need to use pattern matching and entity extraction techniques.

3.2 Data preprocessing

Today's real-world databases are highly susceptible to noise, missing values, and inconsistent data because databases are too large and mostly come from multiple Figure 2 Sina sports community tree structure Fig.3 Parsing links in table SINA_ FORUM_ URL to generate tables SINA_ FORUM_ URL and SINA_ FORUM_ COMMENT_ POST. Figure 3 parsing SINA_ FORUM_ URL in the link table SINA_FORUM_URL and table SINA_FORUM_COMMENT_POST kind of data source. Low quality data will result in low quality digging results. Therefore, the role of data preprocessing seems particularly important. The main steps of data preprocessing include: data cleansing, data integration,

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

data reduction and data transformation

3.2.1 Hottest forums as an example to analyze data cleaning

Data cleaning refers to the removal of noise data and irrelevant data in the data set. For different types of data, noise data, irrelevant data different definitions, cleaning methods also vary "number".

3.2.2 Take Amazon's Recommended System as an Example to Analyze Data Integration Process data Integration Consolidates data from multiple data sources into a consistent data store, such as a data warehouse Amazon's recommendation system based on user evaluation of the product, and control the quality of the evaluation will be directly related to the effectiveness of the recommended system, large amount of data, artificial screening unrealistic, and with subjective factors, so you will think of the machine learning skills to train the algorithm to automatically determine each, it is possible to impose the following two preconditions on the issue:

1. Pay attention to the same quality of the text of the original comment;

2 The same comment will show the same text quality Based on the above two points, we judge whether the above two points are suitable for the evaluation of "plagiarism". 3.3 Through the network analysis or classification

In order to forecast the hot spots in the forum, two kinds of machine learning methods are adopted in [1] : K-means and SVM. In order to classify the extracted network, three methods commonly used in topology research are adopted in [2] : average shortest path length shortest path length, clustering coefficient, and degree distribution. In order to mine the main content the user pays attention to, [8] uses computational entropy to capture the user's point of interest, defined for each user's directory The entropy of attention is as follows: iterates over 8 categories, and pi indicates the probability that the current user belongs to category i. Users in the more classification of comments, will get higher entropy and the lower the focus of attention. In terms of classification, the decision tree classifier [9] , the nearest neighbor classifier (the nearest neighbor classifier neighbor classifier [10] and support vector classifier [11] .

3.3.1 Classification of the basic concepts and common methods

Classification is an important form of data analysis that extracts models that characterize important data classes. This model is called classifier, predictive analysis class (from Scattered, disorderly) class label.

Data classification is a two-phase process that involves learning phases (building classifiers) and classification phases (classes that use models to predict given data Numbering). Common classification methods: decision tree classification, support vector machines Classification method of assessment:

First define four terms:

True Case / True Case (TP): Is the correct classification by the classifier ortho-tuples. Let TP be the number of real cases.

True Negative / True Negative (TN): Negative tuples correctly classified by the classifier. Let TN be the number of negative cases.

False Positives / False Positives (FP): is a negative tuple that is incorrectly labeled as a positive tuple. Let FP be the number of false positives.

False Negatives / False Negatives (FN): is a positive ancestor that is incorrectly labeled as a negative tuple. Let FN be the number of false negatives.

Measure	**Formula**
Accuracy, recognition rate	(TP + TN) / (P + N)
Error rate, misclassification rate	(FP + FN) / (P + N)
Sensitivity, real case rate, recall rate	TP / P
Accuracy	TN / N

Table 1 Mining online community

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

Receiver Operating Characteristic (ROC) curve is a useful visualizer for comparing two classification models with. The ROC curve shows the trade-off between the true rate of occurrence (TPR) and the false positive rate (FPR) for a given model. Given a test set and model, TPR is the proportion of orthodox correctly labeled by the model; and FPR is the proportion of negative tuples that the model incorrectly labels as positive. For the second type of problem, ROC.Table 1 online community mining process the curves allow us to examine the different parts of the test set and observe that the model correctly identifies the proportion and the model of the positive instance by mistake identifying the negative instance as positive tradeoffs between the proportions of the examples. The area under the ROC curve is a measure of the model preparation rate

3.3.2 Similarities and differences between clustering and classification

In the field of machine learning, classification is called supervised learning because, given the class label information, the learning algorithm is supervised because it is told class membership of training elements. Clustering is called unsupervised learning because no class label information is provided. For this reason, clustering is by observation rather than by example learning.

3.3.3 Take Digg as an example, analyze classification and evaluation process and result [8] In Digg, forecast popular points and social network analysis by mining user reviews In classifying user comments, a decision tree classifier [9] , a nearest neighbor classifier (the nearest neighbor classifier neighbor classifier [10] and support vector classifier [11], in which the decision tree classifier uses C4.5 decision Algorithm, denoted by DT; nearest neighbor classification algorithm using 9 neighbors classification method, denoted as 9-NN, support vector machine classification using linear and radial basis functions Kernel functions, denoted as SVM (L) and SVM (R), respectively. For k-class classification using SVM, we use the v-SVM regression method to estimate. Q_K means K-way classification accuracy, ROC means the area under the curve. F1 represents the precision and recall weighting average. The CC table is the correlation coefficient between the actual and the prediction when evaluating the regression results. The graph below shows that in the first 10 hours, the first 15 hours and all the data are predicted in (i) a 2 -class, (ii) a 6 -class, and (iii) a 14 -class evaluation result

Table 2 Evaluate of serveral classification, methods

	Ten Hours Data									
	K=2			K=6			K=14			K=∞
Method	ROC	F1	Q_2	ROC	F1	Q_6	ROC	F1	Q_14	CC
DT	0.83	0.80	0.80	0.72	0.63	0.62	0.64	0.41	0.41	-
9-NN	0.81	0.75	0.75	0.76	0.59	0.63	0.66	0.37	0.42	-
SVM (L)	**0.88**	**0.81**	**0.80**	0.74	0.63	0.63	0.63	0.44	0.42	**0.73**
SVM (R)	0.84	0.79	0.78	**0.79**	**0.66**	**0.64**	**0.70**	**0.46**	**0.45**	0.60
	Fifteen Hours Data									
	K=2			K=6			K=14			K=∞
Method	ROC	F1	Q_2	ROC	F1	Q_6	ROC	F1	Q_14	CC
DT	0.83	0.80	**0.80**	0.72	0.64	0.63	0.64	0.41	0.41	-
9-NN	0.81	0.75	0.76	0.76	0.59	0.64	0.66	0.37	0.42	-
SVM (L)	**0.89**	**0.82**	**0.80**	0.75	0.63	0.64	0.64	0.44	0.42	**0.75**
SVM (R)	0.84	0.79	0.78	**0.80**	**0.66**	**0.64**	**0.70**	**0.45**	**0.44**	0.61
	All Data									
	K=2			K=6			K=14			K=∞
Method	ROC	F1	Q_2	ROC	F1	Q_6	ROC	F1	Q_14	CC
DT	0.87	0.82	0.82	0.76	0.66	0.67	0.65	0.43	0.44	-
9-NN	0.85	0.79	0.79	0.80	0.63	0.66	0.69	0.38	0.43	-
SVM (L)	**0.91**	**0.84**	**0.83**	0.79	0.65	0.67	0.67	0.46	0.45	**0.80**
SVM (R)	0.86	0.81	0.80	**0.82**	**0.69**	**0.68**	**0.74**	**0.48**	**0.45**	0.64

Table 2 evaluation results of several classification methods

3.3.4 Taking online hotspot prediction as an example, we show the clustering process [8]

Algorithm: k-means. The k-means algorithm used for partitioning, where the center of each cluster is represented by the mean of all objects in the cluster enter: k: the number of clusters D: A dataset containing 31 leaf forums, each represented by V (i). V (i) consists of 5 elements: in the time frame the number of topic posts, the average number of responses for topic posts, the average opinion value for topic posts, the number of active posts for all topic posts, Negative posts in all topic posts. Respectively with NUM
(i), POS_PERC (i), and NEG_PERC (i). Express V (j) as follows:
Output: a collection of k clusters
method:

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

$$V^i(j) = \begin{pmatrix} NUM^i(j) \\ \overline{RESPONSE}^i(j) \\ \overline{SENTIMENT}^i(j) \\ POS\_PERC^i(j) \\ NEG\_PERC^i(j) \end{pmatrix}.$$

(1) Select k objects randomly from D as the initial cluster center
(2) Repeat
(3) According to the average of the objects in the cluster, each object is assigned to the most similar cluster
(4) Update the cluster mean, namely recalculate the average value of the objects in each cluster until no longer change
The following table shows the k-average clustering results when k is 5-7, respectively:

Table 3 Result of k-means for online forum hotspot

Time window	k=5	k=6	k=7
200701	1. Soccer Tycoons--Juventus 2. Basketball--Guangzhou Hengyuan 3. Outdoor activities 4. Soccer Tycoons--Milan International 5. Basketball--Yao Ming	1. Soccer Tycoons--Liverpool 2. Chinese Soccer--Cast About Chinese Football 3. Sports Saloon--Tennis 4. The Game of Go 5. Chinese Soccer--Shandong Luneng 6. Basketball--Yao Ming	1. Sports shoes 2. Soccer Tycoons--Juventus 3. Sports Saloon--Tennis 4. International Soccer--Italian Football League 5. The Game of Go 6. International Soccer--English Football League 7. Basketball--Yao Ming
200702	1. Sports shoes 2. Soccer Tycoons--Liverpool 3. Soccer Tycoons--Chelsea 4. International Soccer--English Football League 5. Chinese Soccer--Shandong Luneng	1. Sports shoes 2. International Soccer--Italian Football League 3. Chinese Soccer--Shandong Luneng 4. Outdoor activities 5. Basketball--Yao Ming 6. Chinese Soccer--China Super League of Football	1. Sports shoes 2. International Soccer--German Football League 3. Chinese Soccer--Shandong Luneng 4. Outdoor activities 5. Soccer Tycoons--Milan International 6. Basketball--Yao Ming 7. Chinese Soccer--China Super League of Football
200703	1. Sports shoes 2. Soccer Tycoons--Chelsea 3. Soccer Tycoons--Manchester United 4. Soccer Tycoons--FC Barcelona 5. Chinese Soccer--China Super League of Football	1. Sports shoes 2. Soccer Tycoons--Chelsea 3. The Game of Go 4. Chinese Soccer--Dalian Shide 5. Soccer Tycoons--Manchester United 6. Soccer Tycoons--FC Barcelona	1. Soccer Tycoons--Juventus 2. International Soccer--Spanish Football League 3. Sports Saloon--Billiard 4. Soccer Tycoons--Milan International 5. Soccer Tycoons--Manchester United 6. Basketball--Yao Ming 7. Chinese Soccer--China Super League of Football

Table 3 k-average results of online forum hot spots 4 The significance of data mining for online communities with the development of online communities, the huge number of users in the online community as well as the exposure of user personal information as well as bias, preferences and other information to data mining brings extremely rich resources, and data mining in this area can further promote social development, commercial prosperity, at the same time, to provide a guarantee for social stability.

CONCLUSION

This article analyzes the general process and method of mining online communities through multiple instances, and discusses the economic implications of mining online communities and Social Significance.

REFERENCES :

[1] Nan Li, Desheng Dash Wu. Using text mining and sentiment analysis for online forum hotpos detection and forecast,

[2] Michael Chau, Jennifer Xu. Mining communities and their relationship in blogs: A study of online hate groups. In: Int. J. Human-Computer Studies 65 (2007) 57-70.

[3] J. Han, M. Kamber and J. Pei. Data Mining: Concepts and Techniques., 3rd edition, Morgan Kaufmann, 2011. Yonezawa A. ABCL: An Object-Oriented Concurrent System. Cambridge: MIT Press, 1990.

[4] Cristian Danescu- Niculescu- Mizil, Yonezawa Gueorgi Kossinets, Jon Kleinberg, Lillian Lee. How Opinions are received by Online Communities: A Case Study on Amazon.com Helpfulness Votes.

[5] D. Sorokina, J. Gehrke, S. Warner, and P. Ginsparg. Plagiarism detection in arXiv.In Proc. ICDM, pages 1070-1075, 2006.

[6] Pedro Domingos, Matt Richardson. Mining the Network Value of Customers

[7] Ellen Spertus, Mehran Sahami. Evaluating Similarity Measures: A Large-Scale Study in the Orkut Social Network.

[8] Salman Jamali, HuzefaRangwala. Digging Digg: Comment Mining, Popularity Prediction, and Social Network Analysis. International Conference on Web Information Systems and Mining

[9] Geoffrey Webb. Decision tree grafting. In In IJCAI-97: Fifteenth International Joint Conference on Artificial Intelligence, pages 846-851. Morgan Kaufmann, 1997.

[10] David W. Aha, Dennis Kibler, and Marc K. Albert. Instance based learning algorithms. Machine Learning, 6 (1): 37-66, January 1991.

[11] Vladimir N. Vapnik. The Nature of Statistical Learning Theory. Springer Verlag, 1995.