

Apriori Algorithm base Model of Opinion Mining for Drug Review

M.Alekhy¹, B.Raghu Ram², B.Hanmanthu³

¹Kakatiya University, Kakatiya Institute of Technology and Science,
Yerragattu, Warangal-506 015, Telangana, India.
alekhyanetha57@gmail.com

²Kakatiya Institute of Technology and Science, Kakatiya University,
Yerragattu, Warangal-506 015, Telangana, India.
Raghu9b.naik@gmail.com

³Kakatiya Institute of Technology and Science, Kakatiya University,
Yerragattu, Warangal-506 015, Telangana, India.
bhcsekits@gmail.com

Abstract: *The growing trend of blogging the experiences can provide good feedback for patients to take decision about the consumption of drugs. In order to extract information from blogs about various drugs and symptoms is a challenging task to data mining research community. Applying opinion mining concepts for drug review can lead to a useful platform to drug users to decide to go with particular drug or not. We propose an Apriori base model form retrieving information about various drugs and symptoms. The proposed model is tested on WebMD blog reviews which shown the good results.*

Keywords: Opinion Mining, Apriori Algorithm, Drug Reviews.

1. Introduction

Opinion mining handles the retrieving of important and previously unknown information from a large amount of text opinions or reviews from various World Wide Web sources. In numerous times, solely an overall rating for a review cannot reflect the conditions of different features of a product or a service. Considering this many mining algorithms proposed from opinion review. The discussions and reviews in blogs considered as one of the important source for the opinion mining. Opinion mining was applied with many different type of the fields which include e-commerce, research communities, media, social networks and medical reviews.

A blog is a discussion or informational site published on the World Wide Web and consisting of discrete entries typically displayed in reverse chronological order. Until 2009 blogs were usually the work of a single individual, occasionally of a small group, and often covered a single subject. More recently multi-author blogs have developed, with posts written by large numbers of authors and professionally edited. MABs from newspapers, other media outlets, universities, think tanks, advocacy groups and similar institutions account for an increasing quantity of blog traffic. The rise of Twitter and other "microblogging" systems helps integrate MABs and single-author blogs into societal new streams. Blog can also be used as a verb, meaning to maintain or add content to a blog. The blogs and online reviews procedure is been extended to various fields including e-commerce, research media, social networks and medical reviews.

The drug reviews are currently growing trends in today's blogging trend. Unlike other reviews will mostly concentrate on price, ease of use and mostly side effects. Considering these reviews could provide effective information regarding different people experiences in using the drug. Which could be effective to other user to decide whether to use the drug or not. Considering this we propose effective apriori base model for reviewing the drug.

Apriori [1] is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets

determined by Apriori can be used to determine association rules which highlight general trends in the database.

Considering the effectiveness of the apriori algorithm in retrieving useful information we propose a apriori driven model for drug review. The review information can be used by the doctors, medical researcher and pharmacist to decide the effectiveness and site effects of the drug.

2. Related Work

It is very problematic in just identifying an particular aspect or an idea from a collection of words which are correlated by a class label. Many models like point wise mutual information (PMI)[2], information gain[3], association rules[4], class conditional probability of words etc are used for finding highly correlated words with class labels from a set of reviews which is a collection of written texts of words and class labels. All this models are failed and faced many problems because of not having any intuitive algorithm for classifying the words for getting an quick idea.

In the present context situations aspect-based opining mining is getting famous day by day. Frequency Based Approach [5] is used for mining high frequency noun phrases which will match the required specification or parameters of reviews and Relation Based Approach [6], [7] are used to recognize aspects by identifying aspect sentiment relation from reviews. But however both of these approaches are not used for drug reviews. Because this drug reviews are not mentioned by any author and this reviews and side effects are mentioned by patient is different from one person to another person. However classifying mined high frequency noun is a difficult task just based on those semantic meanings. But when we consider topic modeling it mainly focuses on co-occurrences of words in reviews and one merit of topic modeling is recognizing and classifying aspects are simultaneously performed.

For an collection of written text, Topic Modeling [8], [9], [10] is an well-known and best probabilistic approach. First based on the priority of probability of words of topic is stored in an order and high probability of words are semantically correlated manually. By this we can come to an decision state by using this topic modeling for a set of topics which are presented in reviews. Topic Sentiment Mixture

(TSM) [11], Joint Sentiment/Topic (JST) [12] model and Aspect and Sentiment Unification Model (ASUM)[13] are examples of topic modeling are mined and their associated sentiments. These types of problems are faced by topic modeling better rather than aspect opinion modeling.

3. Proposed Model

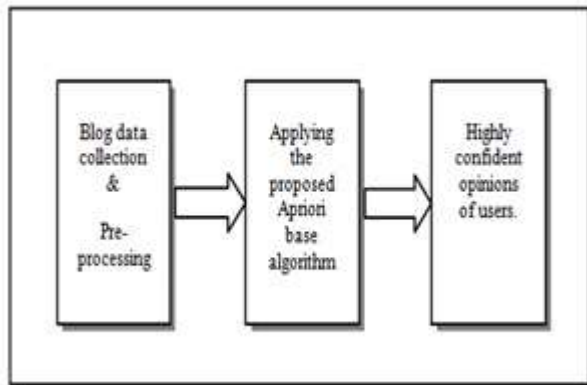


Fig.1: Model for apriori base opinion mining on drug review

The proposed model for opinion mining on drug review using apriori follows the three step procedure as shown in the figure.1. In first step it extract the information from the reliable blogs and the required data will be consolidated after which the data pre processing steps will be applied to process the data remove unnecessary and punctuation words. In second step the apriori algorithm will be applied on tested data to retrieve highly confident rules. Finally the highly confident rules will be extracted to show the results to need to take decision to go with or not.

When the index processing start, it parses each raw document and analyze its text content. The typical steps includes, Tokenize the document, Lowercase each word, Remove stop words, Stemming, Synonym handling. This can be done in two ways. Either expand the term to include its synonyms or reduce the term to a normalized synonym. At this point, the document is composed with multiple terms. $doc = [term1, term2 \dots]$. Optionally, terms can be further combined into n-grams. After that we count the term frequency of this document. The advantage of the preprocessing is that the unnecessary words will be extracted which leads to only use full information will be with the final text which leads to proper utilization of the terms by the data mining algorithm. Considering that the target of the work is to extract useful information from drug review we take care the medical terms will be given importance. In order to do so we supplied a list of terms which can be considered as the unimportant terms so that remaining terms will be considered useful terms which obviously include medically important terms.

After gathering useful terms by document clustering apriori algorithm will be applied to retrieve useful information from the text. The apriori algorithm is given in Algorithm.1. The apriori algorithm will take text as the input and perform mining operations by considering each term as a item and combination of items will lead to item set. In first pass it generates individual items and in next pass it will use the frequent item sets. The item set with minimum support will be considered as for next iterations. Finally the process will be repeated until the new item set with minimum support could be generated. Finally using the frequent item set the confident threshold rules will be generated.

Pass 1

1. Generate the candidate itemsets in C_1
2. Save the frequent itemsets in L_1

Pass k

1. Generate the candidate itemsets in C_k from the frequent itemsets in L_{k-1}
 1. Join $L_{k-1}p$ with $L_{k-1}q$, as follows:
insert **into** C_k

select $p.item_1, p.item_2, \dots, p.item_{k-1}, q.item_{k-1}$
from $L_{k-1}p, L_{k-1}q$

where $p.item_1 = q.item_1, \dots, p.item_{k-2} = q.item_{k-2}, p.item_{k-1} < q.item_{k-1}$

2. Generate all $(k-1)$ -subsets from the candidate itemsets in C_k
 3. Prune all candidate itemsets from C_k where some $(k-1)$ -subset of the candidate itemset is not in the frequent itemset L_{k-1}
2. Scan the transaction database to determine the support for each candidate itemset in C_k

Algorithm: Apriori algorithm for medical diagnosis

Finally obtained rule with highest confidence considered as opinion of the user about the drug. Different opinions rules can be considered with different confidence threshold as user with some specific medical problem may not active on the internet. Considering this the user with highest activity on blogs will be considered with high confidence threshold and users with low active at net will be considered for low confidence threshold.

Once the high frequency terms and their combination along with their confidence obtained then they finally used for retrieving the glanced the opinion mining of the drug. The opinion mining of the drug will be retrieved by mapping the obtained rules to prepared pervious set of expert rules. The obtained rules will first pruned using the confidence threshold issued by the user or the expert then the rules will be used for processing next step. At the next process the opinion mining rules prepared by the experts cross checked with the obtained high confidence rules. If the obtained high confidence rule is mapped with rule set of positive opinion rules of expert rule then the high confidence rule will be considered as positive else if the high confidence rule is mapped with rule set of negative opinion rules of expert rule then the high confidence rule will be considered as negative rule set. Finally out of retained all the high confidence rules will be rated as positive rule set or as negative rule set. At last the opinion of the drug will be decided by if it has high positive confidence rules more than low confidence rules then the drug opinion mining will be rated as positive reviewed drug else it will be rated as negative reviewed drug.

4. Evaluation

In order to testify the performance of model, our experiments utilized Pentium-dual core processors with 2.40GHz PCs with 512Mb main memory and windows OS. The three PCs are located in 100Mb LAN. We use the WebMD data sets obtained first the preprocessing, and then followed by a priori algorithm applied to find opinion mining with highest confidence.

In order to evaluate the working of the model first the expert opinion on the effectiveness of the different drugs collected from the WebMD dataset. Then the positive and negative key words about the drug collected and categorize then as positive as well negative rule set. Then the same drug opinion collected from WebMD and firstly preprocessed then proposed algorithmic model applied to collect the effectiveness of the drug. At final step of the application of the model the obtained high confidence rules of the drug of each medicine will applied to positive and negative key words and depends up on its rule set the final opinion of the rules was labeled. Then depends up on number of positive labeled rules of a medicine and negative labels of the rules will be collected. If the rule will more positive rules then the rule categorized as positive opinion rule else if the rule categorized as negative opinion rule. In order to evaluate the effectiveness of the rule the obtained opinion cross checked with published opinion of the WebMD website on the selected medicines. In comparisons the 90% of medicines ware success fully categorized as positive or negative medicine on par with opinion given in WebMD dataset. This proves the efficiency of the proposed model.

6. Conclusions

The review of the drug users in various blogs can be very useful to the new person who wants to take decision to go with the drug or not. Considering that the new apriori base model was proposed and evaluated in this article to retrieve use full information from the blogs. The evaluation results shown positive results to update model for further levels. The proposed model can be used by doctors also to decide the effectiveness of the doctors.

7. Acknowledgment

Our thanks to the management members and principal of Kakatiya Institute of Technology and Science-Warangal who have facilitated resources to read and compute in order to develop this model and narrate this article and our sincere thanks to Head of the Department Prof.P.Niranjan who encouraged us research and publish this paper.

References

- [1] C. Manning and H. Schütze, Foundations of Statistical Natural Language Processing. Cambridge, MA, USA: MIT Press, 1999.
- [2] T. Mitchell, Machine Learning. Boston, MA, USA: McGraw-Hill, 1997.
- [3] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in Proc. 20th Int. Conf. VLDB, San Francisco, CA, USA, 1994, pp. 487–499.
- [4] A.-M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," in Proc. Conf. Human Lang. Technol. Emp. Meth. NLP, Stroudsburg, PA, USA, 2005, pp. 339–346.
- [5] B. Liu, M. Hu, and J. Cheng, "Opinion observer: Analyzing and opinions on the web," in Proc. 14th Int. Conf. WWW, New York, NY, USA, 2005, pp. 342–351.
- [6] S. Baccianella, A. Esuli, and F. Sebastiani, "Multi-facet rating of product reviews," in Proc. 31st ECIR, Berlin,, Germany, 2009, pp. 461–472.
- [7] D. Blei and J. Lafferty, "Correlated topic models," in Proc. Adv. NIPS, 2006, pp. 147–154.
- [8] A. McCallum and X. Wang, "Topic and role discovery in social networks with experiments on enron and academic email," J. Artif. Intell. Res., vol. 30 no. 1, pp. 249–272, 2007.
- [9] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai, "Topic sentiment mixture: Modeling facets and opinions in weblogs," in Proc. 16th Int. Conf. WWW, New York, NY, USA, 2007, pp. 171–180.
- [10] Lin and Y. He, "Joint sentiment/topic model for sentiment analysis," in Proc. 18th ACM CIKM, New York, NY, USA, 2009, pp. 375–384.
- [11] Y. Jo and A. Oh, "Aspect and sentiment unification model for online review analysis," in Proc. 4th ACM Int. Conf. WSDM, New York, NY, USA, 2011, pp. 815–824.

Author Profile



M. Alekhya Currently persuing Master of Technology in Computer Science and engineering with specilization in Softwatre Engineering. Computer Science and Engineering Department, Kakatiya Institute of Technology & Science (KITS), Kakatiya University-Warangal. Telangana, India.



B. Raghuram obtained his Bachelor's degree in Computer Science and Engineering from JNT University of India. Then he obtained his Master's degree in Computer Science and Engineering from Pondicherry Central University Pondicherry, and he is also life member of ISTE. He is currently Assistant Professor of Computer Science and Engineering, Kakatiya Institute of Technology & Science (KITS), Kakatiya University-Warangal. His specializations include Data mining and Data warehousing, Databases and networking.



B. Hanmanthu obtained his Bachelor's degree in Computer Science and Engineering from JNT University of India. Then he obtained his Master's degree in Computer Science and Engineering with specialization Software Engineering from JNT University Hyderabad, and he is also life member of ISTE. He is currently Assistant Professor of Computer Science and Engineering, Kakatiya Institute of Technology & Science (KITS), Kakatiya University-Warangal. His specializations include Data mining and Data warehousing, Databases and networking.