ISSN:2231-0711
www.ijcset.net

Advanced Analysis of Internet Text Mining

Visam Praveen , Gousiya Begum,
CSE Department,Mahatma Gandhi Institute Of Technology
Hyderabad, India
visampraveen2@yahoo.in

Abstract - **Large amounts of data have been collected routinely in the course of day-to-day management in business, administration, banking, the delivery of social and health services, environmental protection, security and in politics. Such data is primarily used for accounting and for management of the customer base. Manufacturers selling product on the internet often ask their customers to review their product. Such reviews provide information about these products. They are used by potential customers to find opinions of existing users before deciding to purchase that product. They can also be used by product manufacturers to identify the problems in their product. Unfortunately, the importance of reviews gives good incentives for Spam, which contains false positive or malicious negative opinions. Typically, management data sets are very large and constantly growing and contain a large number of complex features. While these data sets reflect properties of the managed subjects and relations, and are thus potentially of some use to their owner, they often have relatively low information density. One requires robust, simple and computationally efficient tools to extract information from such data sets. The development and understanding of such tools is the core business of data mining.Through this system we study this issue in the context of product reviews, which are opinion rich and are widely used by customers and product manufacturers. Review spam is quite different from internet spam and email spam and hence requires different detection techniques. Review spams can be broadly classified into three groups such as Untruthful opinion, reviews on Brands only and non-opinions. In this proposed system untruthful opinions are detected using Secure Hash Algorithm (SHA-256 bits) and reviews on brands only as well as non-reviews are identified using multilayer perceptron. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration.**

Keywords - **Opinion spam, review spam, fake reviews, review analysis**

1. INTRODUCTION

Frequent pattern mining is an important area of Data mining research. The frequent patterns are patterns (such as item sets, subsequences, or substructures) that appear in a data set frequently. For example, a set of items, such as milk and bread that appear frequently together in a transaction data set is a frequent item set. A subsequence, such as buying first a PC, then a digital camera, and then a memory card, if it occurs frequently in a shopping history database, is a (frequent) sequential pattern. With the rapid expansion of e-commerce, more and more products are sold on the internet, and more and more people are also buying products online. In order to enhance customer satisfaction and shopping experience, it 1has become a common practice to enable their customers to review or to express opinions on the product that they have purchased. With more and more common users becoming comfortable with the internet, an increasing number of people are writing reviews. Reviews are useful to both individual consumers and product manufacturers. For example, if one wants to buy a product, one typically goes to a merchant site to read some reviews of existing users of the product. If the reviews are most positive, one is very likely to buy that product. If the reviews are mostly negative, one will likely choose another product. Positive opinions can result in significant financial gains or fames for organizations and individuals. This gives good incentives for Review Spam.

Review spam is similar to Internet page spam. In the context of search, due to the economic and/or publicity value of the rank positions of a Internet page returned by the search engine, internet page spam is widespread .Internet page spam refers to the use of “illegitimate means” to boost the rank positions of some target pages in search engines .In context of reviews, the problem is similar, but also quite different. We collected reviews on various manufactured products. We discovered that spam activities are wide spread. For example we found a large number of duplicate and near-duplicate reviews written by the same reviewers on the same product or different products.

There are generally three types of spam reviews:

(i) Untruthful opinions: Those that deliberately mislead readers or opinion mining systems by giving undeserving positive reviews to some target objects in order to promote the object and/or by giving unjust or malicious negative reviews to some other objects in order to damage their reputations.

(ii) Reviews on brands only: Those that do not comment on the products in reviews but only the brands, the manufacturers and sellers of the products.

(iii) Non-opinions: Those that are non-reviews, which have two subtypes

(i) Advertisements

(ii)Other irrelevant reviews containing no opinions(e.g.: questions, answers, comments)

This system proposes some novel techniques to study review spam and spam detection. In general, spam detection can be regarded as a classification problem with two classes, spam and non-spam. For two types of spam reviews, we can detect them based on supervised learning because these two types of reviews are recognizable manually and thus training data can be labeled manually. The main task is to find a set of effective features for model building.

However, for the type of spam reviews, which we call untruthful opinions, manual labelling by simply reading reviews is very hard, if not impossible, because a spammer can carefully craft a spam review to promote a target product

or to damage the reputation of another product that is just like any innocent review. Thus, other ways have to be explored in order to find training examples for detecting possible untruthful spam reviews. In our analysis we found a large number of duplicate and near-duplicate reviews. They contain untruthful spam reviews because of the following types of duplicates:

1. Duplicates from different user ids on the same product.
2. Duplicates from same user ids on different products.
3. Duplicates from different user ids on different product.

Thus our review spam detection takes the following strategy; first, we detect duplicates and near-duplicates. We then detect spam reviews of type's brands only and non-opinions based on machine learning and manually labeled examples. Finally we detect untruthful opinion spam by exploiting the above three types of duplicates and other relevant information.

II. EXISTING SYSTEM WORKS

Analysis of on-line opinions became a popular research topic recently. Current studies are mainly focused on mining opinions in reviews and/or classify reviews as positive or negative based on the sentiments of the reviewers. The most extensively studied topic on spam is internet spam. The objective of internet spam is to make search engines to rank the target pages high in order to attract people to visit these pages. Internet spam can be categorized into two main types: content spam and link spam. Link spam is spam on hyperlinks, which does not exist in reviews as there is usually no links among them. Content spam tries to add irrelevant or remotely relevant words in target pages to fool search engines to rank the target pages high. Many researchers have studied this problem. Review spam is quite different. Spammers write undeserving positive reviews to promote their target object and/or malicious negative reviews to damage the reputation of some other target objects.

Another related research is email spam, which is also quite different form of review spam. Email spam usually refers to unsolicited commercial advertisements. Although exist, advertisements in reviews are not relatively easy to detect. Untruthful opinion spam is much harder to deal with.

Recent studies on spam also extended to recommender systems where they are called attacks. Although the objective of attacks to recommender systems are similar to review spam; their basic ideas are quite different. In recommender systems, a spammer injects some attack profiles to the system in order to get some products more (or less) frequently recommended. A profile is a set of ratings for a series of products. The recommender system uses the profiles to predict product rating of a single user or a group of users. The spammer does not usually see other users rating profiles .in the context of product reviews, there is no concept of profiles. Each review is only for a particular product, and is not used for any prediction. Also the reviewer can see all reviews for every product. Rating is only part of the review and another main part is review text. Spam is a much broader concept involving all types of objectionable activities. Introduces the problem of review spam, and categorised different types of spam reviews. However it did little study on detecting untruthful opinions. Gives some general discussions about spam review as well, but no computational study is reported.

III. PROPOSED SYSTEM BACKGROUND

3.1 Module description

3.1.1 Detection of duplicate reviews-by verifying review property:

Most of the reviews include the following properties:

< Reviewer name>, <Reviewer id>, <Date>, <Review body>

1) Author

The same author's reviews are likely duplicate reviews.

2) IP Address:

The reviews released from same IP address have the greater possibility to express duplicate reviews.

3) Time:

The smaller the time interval T (T=T2-T1), the more likely the duplicate reviews are.

By verifying (author, IP address, time) duplicate reviews can be identified to some extent.

3.1.2 Detection of duplicate reviews-by verifying review body

The above criteria (author, IP address, time) may not be accurate to judge similarity results. Therefore duplicate reviews are identified by calculating the similarity of review's main content.

3.1.2.1 Extraction content features from reviews

The content of reviews may be regarded as text document or a long character strings. Detecting duplicate reviews need to extract text blocks from reviews, calculate similarity of text blocks. If similarities are greater than the given threshold, we determine that the reviews are duplicate reviews. We utilize the character strings comparison to compute the similarities. We select some character strings called "the fingerprint" and map them into the hash table. In the hash table, one fingerprint will correspond to one hash value. Finally the same fingerprint numbers is counted. The similarity decision function is commonly used as the following:

$$S(a, b) = \frac{F(a) \cap F(b)}{F(a) \cup F(b)}$$

Where F (a) - fingerprint set, F (b)-fingerprint set, S (a,b)-similarity of set a&b.

3.1.2.2 Granularity

The detecting unit is called the text granularity (block).the thinnest granularity is just one character. The thickest granularity is all the main text. The thinner granularity often leads to low accuracy rates, whereas thickest granularity is only used to detect identical reviews. Using the various granularities we can obtain a better performance, however the time complexity is high. Therefore, we should select the appropriate granularity size.

We first delete the attributes (author Name, IP address and time T) of reviews. We take sentences as elemental detection units. In general, the reviewers like to use various

punctuations as the boundary symbols, (e.g., “too great!!”, “great ~” and so on.) Therefore, in this paper we regard the symbols besides carriage return as the sentences boundaries. We calculated the Hash value of each sentence, then places them in (sentence, sentence number, Hash value list), and uses the (1) to calculate the similarities. They are regards as duplicate reviews if their similarities value is above the threshold.

Taking the carriage return as division symbol of sections, we calculate the Hash values of each section, places them in(section, section number, Hash value list), and uses the (1) to calculate the similarities. They are regards as duplicate reviews if their similarities value is above the threshold.

3.1.2.3 Algorithm description:

We select SHA-2 algorithm as the method to detect duplicate reviews .In this algorithm 256 bit hash value is calculated. Detecting duplicate reviews need to extract text blocks from reviews, calculate similarity of text blocks. If similarities are greater than the given threshold, Threshold value is set by the developer. We determine that the reviews are duplicate reviews. We utilize the character strings comparison to compute the similarities. We select some character strings called “the fingerprint” and map them into the hash table. In the hash table, one fingerprint will correspond to one hash value. Finally the same fingerprint numbers is counted.

3.1.2.4 Multilayer Perceptron:

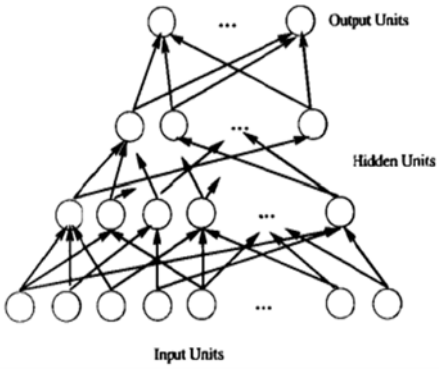
The Multi-layer perceptron is one of the most widely used types of neural networks. It is simple and based on solid mathematical grounds. Input quantities are processed through successive layers of "neurons". There is always an input layer, with a number of neurons equal to the number of variables of the problem, and an output layer, where the perceptron response is made available, with a number of neurons equal to the desired number of quantities computed from the inputs (very often only one). The layers in between are called "hidden" layers. With no hidden layer, the perceptron can only perform linear tasks (for example a linear discriminant analysis, which is already useful).

All problems which can be solved by a perceptron can be solved with only one hidden layer, but it is sometimes more efficient to use 2 hidden layers. Each neuron of a layer other than the input layer computes first a linear combination of the outputs of the neurons of the previous layer, plus a bias. The coefficients of the linear combinations plus the biases are called the weights. They are usually determined from examples to minimize, on the set of examples, the (Euclidian) norm of the desired output - net output vector. Neurons in the hidden layer then compute a non-linear function of their input.

Usually, the non-linear function is the sigmoid function y(x) = 1/(1+exp(-x))). The output neuron(s) has its output equal to the linear combination. Thus, a Multi-Layer Perceptron with 1 hidden layer basically performs a linear combination of sigmoid function of the inputs. A linear combination of sigmoids is useful because of 2 theorems:

1. A linear function of sigmoids can approximate any continuous function of 1 or more variable(s). This is useful to obtain a continuous function fitting a finite set of points when no underlying model is available.

2. Trained with a desired answer = 1 for signal and 0 for background, the approximated function is the probability of signal knowing the input values. This second theorem is the basic ground for all classification applications. ”



IV. EXPERIMENTAL FLOW

The below specified diagram explains the experimental flow of the complete spam filter.

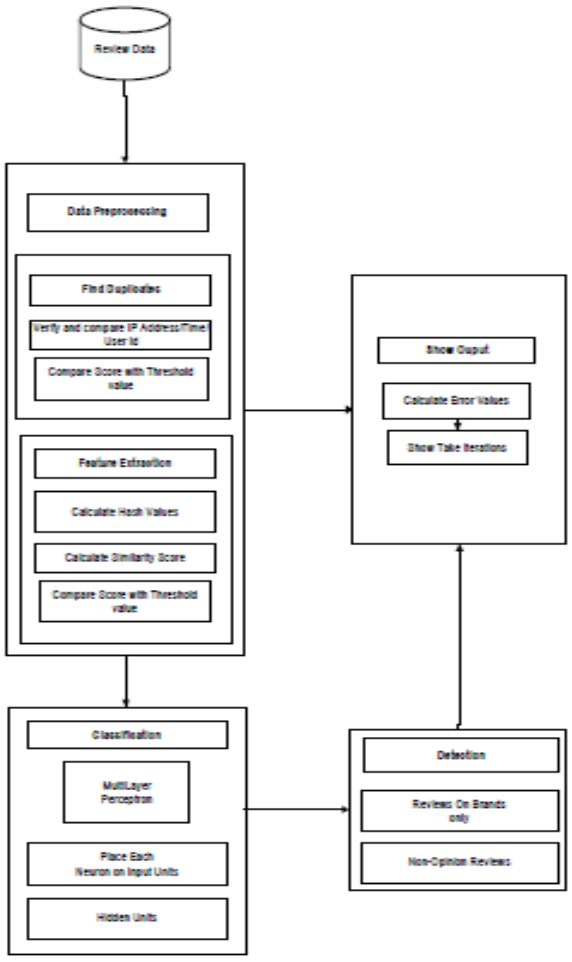


Fig 1:system architecture

V.EXPERIMENTAL RESULTS

We created a data set of about 400 reviews which includes both spammed and non spammed reviews. Through initial checking of IP address, time, name of reviewer about 20% of duplicates can be identified by setting a threshold value of time as 5ms. Duplicates from text are done by extracting text from the data set and calculate their hash values using SHA-2 algorithm and then finding the similarity score as explained in section 4. The advantage of using SHA-2 is security as well as efficiency. The duplicates identified through this process are stored in a database called spam database.

VI. CONCLUSIONS

A-M. Popescu and O. Etzioni [1] introduces OPINE, an unsupervised information extraction system which mines reviews in order to build a model of important product features, their evaluation by reviewers, and their relative quality across products. Compared to previous work, OPINE achieves 22% higher precision (with only 3% recall) on the feature extraction task. OPINE's novel use of relaxation labelling for finding the semantic orientation of words in context leads to strong performance on the task of finding opinion phrases and their polarity. The key component of OPINE described in this paper are the PMI feature extraction and the use of relaxation labelling in order to find the semantic orientation of potential opinion words.

Through this paper N. Jindal and B. Liu[2] studies the issue of review spam in the context of product reviews, which are opinion rich and are widely used by consumers and product manufacturers. In the past two years, several start-up companies also appeared which aggregate opinions from product reviews. It is thus high time to study spam in reviews. To the best of our knowledge, there is still no published study on this topic, although Internet spam and email spam have been investigated extensively. We will see that opinion spam is quite different from Internet spam and email spam, and thus requires different detection techniques. Based on the analysis of 5.8 million reviews and 2.14 million reviewers from amazon.com, we show that opinion spam in reviews is widespread. This paper analyses such spam activities and presents some novel techniques to detect them.

M. Hu & B. Liu.[3] Merchants selling products on the Internet often ask their customers to review the products that they have purchased and the associated services. As e-commerce is becoming more and more popular, the number of customer reviews that a product receives grows rapidly. For a popular product, the number of reviews can be in hundreds or even thousands. This makes it difficult for a potential customer to read them to make an informed decision on whether to purchase the product. It also makes it difficult for the manufacturer of the product to keep track and to manage customer opinions. For the manufacturer, there are additional difficulties because many merchant sites may sell the same product and the manufacturer normally produces many kinds of products. In this research, we aim to mine and to summarize all the customer reviews of a product. This summarization task is different from traditional text summarization because we only mine the features of the product on which the customers have expressed their opinions and whether the opinions are positive or negative. We do not summarize the reviews by selecting a subset or rewrite some of the original sentences from the reviews to capture the main points as in the classic text summarization. Our task is performed in three steps: (1) mining product features that have been commented on by customers; (2) identifying opinion sentences in each review and deciding whether each opinion sentence is positive or negative; (3) summarizing the results. This paper[4] proposes several novel techniques to perform these tasks. Our experimental results using reviews of a number of products sold online demonstrate the effectiveness of the techniques.

REFERENCES

1. Celebrating 40 years of the net, by Mark Ward, Technology correspondent, BBC News, 29 October 2009
2. "A Technical History of CYCLADES", Technical Histories of the Internet & other Network Protocols, Computer Science Department, University of Texas Austin
3. "The Cyclades Experience: Results and Impacts", Zimmermann, H., Proc. IFIP'77 Congress, Toronto, August 1977, pp. 465–469
4. A Chronicle of Merit's Early History, John Mulcahy, 1989, Merit Network, Ann Arbor, Michigan
5. "Roads and Crossroads of Internet History" by Gregory Gromov. 1995
6. Hafner, Katie (1998). Where Wizards Stay Up Late: The Origins Of The Internet. Simon & Schuster. ISBN 0-684-83267-4.
7. Ronda Hauben (2001). From the ARPANET to the Internet. Retrieved 28 May 2009.
8. "Events in British Telecomms History". Events in British TelecommsHistory. Archived from the original on 5 April 2003. Retrieved 25 November 2005.
9. Walter Willinger, Ramesh Govindan, Sugih Jamin, Vern Paxson, and Scott Shenker (2002). Scaling phenomena in the Internet, in Proceedings of the National Academy of Sciences, 99, suppl. 1, 2573–2580
10. Jesdanun, Anick (16 April 2007). "Internet Makeover? Some argue it's time". Seattletimes.nwsource.com. Retrieved 8 August 2011.
11. R. Cohen, K. Erez, D. ben-Avraham, S. Havlin (2000)."Resilience of the Internet to random breakdowns". Phys. Rev. Lett **85**: 4625.
12. R. Cohen, K. Erez, D. ben-Avraham, S. Havlin; Erez, K; Ben-Avraham, D; Havlin, S (2001). "Breakdown of the Internet under intentional attack". Phys. Rev. Lett **86** (16): 3682–5.DOI:10.1103/PhysRevLett.86.3682. PMID 11328053