Filtering Web Pages by Sensitive Mining Approach

M.Sreedevi, A.Sowmya Kaveri, Deepak.V, K.Venkatesh , D.Sravan

Department of Information Technology,KLUniversity,Vijayawada,Andhra Pradesh, India.

Summary

Many abnormal topics or remarks on the world wide web may like crime, violence etc may disturb the public morality and cause social unrest. Most traditional methods filter a page as long as it contains a keyword in a predefined blacklist. Such methods cannot provide a quantitative measure of how sensitive the content is. In this paper, we propose a utility-based Web content sensitivity mining approach. Utility is viewed as the measure of how sensitive a page is. It allows the Internet regulators to take different operations according to different sensitivity values. We apply our approach on a real-world Web dataset. By varying the sensitive values of the keywords, different sets of high sensitivity keywords were discovered.

General terms

sensitive mining , filtering

Key words:

sensitive content, flexibility ,utility mining, sensitivity mining .

1. Introduction

With the emergence and prosperity of the Web as a media, a number of problems also arise. For instance, many people disseminate violence, conduct fraud, spread rumors, abuse, threaten, or promote superstition through the Web, especially through discussion boards. We refer the above annoying behaviors as sensitive content in this paper. It is an important task for the Internet regulators to identify the sensitive pages, trace the corresponding IP addresses, and even block the IPs when necessary. We call it Web content sensitivity monitoring system. With the help of such a system, the Internet regulators or Department of Security will be able to know the abnormal remarks or behaviors early and take appropriate operations, thus making the Internet a good, moral and upstanding place.

In this paper, we propose a utility-based Web content sensitivity mining approach. Utility is a measure of how useful or important an itemset is [i]. In the context of Web content, we can view the keyword set sensitivity as the utility of a keyword set. Web content sensitivity is defined as the total sensitivity of the high sensitivity keyword sets it contains, measuring how sensitive or abnormal its content is. Objective value is the keyword frequency in a given page; subjective value is assigned by the Internet regulators to express their sensitivity preference. Utility-based Web content sensitivity mining approach is to find all the Web pages whose sensitivity exceeds a user-defined threshold. It shows several desired properties over other existing methods:

- Level of sensitivity. By calculating the sensitivity of each Web page, users are allowed to rank the pages in the order of sensitivity. Department of Security may adopt different regulations under different circumstances.
- Flexibility. In real life, the definition of “sensitive” changes over time. In our approach, the sensitivity table is determined by the Internet regulators or experts, independent from the database. Thus, it is easy to update the sensitivity of each page regularly.
- Dimensionality reduction. The number of terms/ dimensions of each page is so high that makes typical text mining algorithms perform poorly. Our approach can reduce the dimensionality drastically and provide an understandable description of the discovered rules. In this paper, we apply our proposed approach on a real-world dataset. We compare our high sensitivity pages with the ones discovered by a frequent itemset- based method.

The rest of this paper is organized as follows. Section 2 overviews the related work. In Section 3, we introduce utility mining, the basis of our proposed approach. In Section 4, we present our proposed approach. Section 5 presents the experimental results and we summarize our work in Section 6.

2. Related work

Sentiment classification is a technique to judge a message as positive or negative (favorable or unfavorable) on a given topic. Traditional classification models have achieved acceptable accuracy [2]. Knowledge-based methods [3] pre-select a set of seed words or phrases to determine the semantic orientation of the content.

Sentiment analysis was used to test the relationship between Internet financial message boards and the behavior of the stock market to find a strong correlation between posts and volume of stock.

Public opinion monitoring systems [4, 5] can reflect the overall public opinion tendency, and provide real- time reports on hot topics and new topics at different time granularities. Most Web content monitoring systems, like anti- virus software, spam filter, Web filter, determine Web page normal or abnormal by checking if its IP

Manuscript received March 5, 2013
Manuscript revised March 20, 2013

address, or title, or content word matches a record in a predefined blacklist, no matter its content is positive or negative. Other systems use Bayesian classifiers to predict a new coming page as normal or abnormal .

3. Utility Mining

Since utility mining is the basis of our proposed approach, we would like to start with the formal definition of utility mining model [l].

- I = {i_1, i_2 …, i_m} is a set of items.
- D = {T_1, T_2, …, T_n} is a transaction database where each transaction T_i E D is a subset of I.
- $o(i_p, T_q)$, *objective value*, represents the value of item i_p in transaction T_q.
- $s(i_p)$, *significance value*, is assigned by a user to express his preference. It reflects the importance of an item, independent of transactions. $s(i_p)$ is greater than $s(i_q)$ if the user prefers item i_p to i_q.
- $u(i_p, T_q)$, *utility function*, is defined as $o(i_p, T_q)$ x $s(i_p)$, where $o(i_p, T_q)$, is the value of item i_p in transaction

u(X), utility of an itemset X, is defined as

$$\underset{T_q \text{ E} D \text{A} X \subset T_q}{2} u(X, T_q)$$

- An itemset X is a *high utility itemset* if $u(X)$ E, where $X \subset I$ and E is the user specified minimum utility threshold, otherwise, it is a *low utility itemset*.

Utility mining is to find the complete set of high utility itemsets, HU = {Xl X I, $u(X)$ ▬ ⁞ }.

$\underset{i_p \text{E} X}{2}$, where X = {i_1, i_2, …, i_k} is a T_q, and $s(i_p)$ is the unit profit of item i_p. $u(X, T_q)$, *utility of an itemset X in* transaction T_q, is defined as $\underset{i_p \text{E} X}{2} u(i_p, T_q)$, where X = {i_1, i_2, …, i_k} is a k-

itemset, $X \subset T_q$ and l k m.

4. Web content sensitivity mining approach

4.1 Keyword set sensitivity

We introduce the concept of utility into the problem of Web content mining. We propose a new term, *keyword set sensitivity*, defined as follows:

- I = {i_1, i_2,…, i_m} is a set of keywords in a predefined blacklist.
- D = {D_1, D_2, …, D_n} is a database where D_i is the **space vector** representation of a Web page.
- $o(i_p, D_q)$, *objective value*, represents the number of occurrences of keyword i_p in Web page D_q.
- $s(i_p)$, *sensitivity value*, is assigned by the Internet regulators to express the sensitivity level of a keyword. $s(i_p)$ is greater than $s(i_q)$ if i_p is more sensitive or abnormal than i_q. It is independent of D.
- $u(i_p, D_q)$, *utility function*, is defined as $o(i_p, D_q)$ x $s(i_p)$.
- $s(X, D_q)$, *sensitivity of a keyword set X in page* D_q, is defined as

k-keyword set, $X \subset D_q$ and l k m.

- $S(X)$, *keyword set sensitivity*, is defined as

$$\underset{D_q \text{ E} D \text{A} X \subset D_q}{2} s(X, D_q)$$

A keyword set X is a high sensitivity keyword set if $S(X)$ E$_l$, where $X \subset I$ and E$_1$ is the *keyword-set- sensitivity* threshold; otherwise, it is a low sensitivity keyword set. The concept of high sensitivity keyword set can help the Internet regulators understand the most influential or abnormal words or word sets on the Web.

We apply *phasefinder algorithm* to find all the high sensitivity keyword sets. This algorithm guarantees that the complete set of high sensitivity keyword sets will be identified efficiently.

```
Input: Web Page URL
          Phrase
   Output: Availability of phrase in web page
   Begin
        1. Enter the URL of webpage from which phrase
        p is to be searched there by extracting the
        required webpage.
        2. Enter phrase p to be searched
        3. Parsing starts from the beginning of page
    if webpage W is not empty
     read character from input file into array
      if the string matches phrase p
      then phrase p is available in webpage
      else
     phrase p is not available
   else
    web page is empty
```

Figure 1. Pseudo code of phrase finder algorithm

The algorithm parses the whole webpage from top to bottom and searches for the word.

4.2 Web content sensitivity mining

Based on the concept of keyword set sensitivity, we propose another concept, *Web content sensitivity*. It measures how sensitive or abnormal the content of a page is. *Buffer* stores the maximal high sensitivity keyword sets in p. For each Web page p, $WS(p)$, is the sum of the sensitivity of all the maximal high sensitivity keyword sets.

Web content sensitivity mining is to find out all the Web pages whose content sensitivity exceeds a *page-sensitivity threshold* ⁞$_2$.

Let's use Table 1 as an example. In Table 1 (a), the number in each cell represents the word frequency. *s*({B}, D$_7$) is 8, *S*({B}) is 36, and *S*({D, E}) = *s*({D, E},D$_2$) + *s*({D, E},D$_4$) + *s*({D, E},D$_7$) = 5+8+11 = 24. If we set the keyword-set-sensitivity threshold E$_1$ at 20 and the minimum support (frequency) at 4, keyword set {D, E}

Discussion

When we take a webpage it contains many words or phrases which are present in the black-list. So we need to remove those phrases. So we make use of the phrase finder algorithm which searches the webpage with a given URL for a particular phrase. If the phrase is found then it displays that the phrase is available in the webpage. If not is shows that the phrase is not available.

D$_1$	0	0	18	0	1
D$_2$	2	6	0	1	1
D$_3$	2	0	1	0	1
D$_4$	1	0	0	2	1
D$_5$	0	0	4	0	2
D$_6$	1	1	0	0	0
D$_7$	2	2	0	3	1

(b) Sensitivity table. The right column displays the sensitivity level.

KEYWORD	SENSITIVITY LEVEL
A	1
B	4
C	1
D	3
E	2

will not be an interesting keyword set by frequency-based Web monitoring systems (occurring in 3 pages), but will be a high sensitivity keyword set. Since {C, E} and {C} are high in sensitivity but {C} is covered by {C, E}, the Web content sensitivity, *WS*(D$_3$) = *S*({C, E}, D$_3$) = 3. Assume that the page-sensitivity threshold ⁚$_2$ is 15, D$_3$ is a low sensitivity page.

Table 1. A Web database and sensitivity table

(a) Web database. The rows represent the page vectors. The columns represent the frequencies in a particular page.

Web content sensitivity enables us to sort the Web pages in the order of sensitivity and take appropriate monitoring strategies.

5. Experimental results

5.1 Data sets

We evaluate our utility-based sensitivity mining approach on a real-world dataset, *news*. *news* contains 12007 news articles between January 2008 and July 2008 from *www.sina.com*, covering various topics. We obtained a keyword blacklist from a collaborator at Department of Security. The blacklist contains 1449 words that can be roughly categorized into 4 groups: (1) terrorism, violence, national security; (2) obscenity pornography; (3) fraud, threat, crimes, disaster; (4) others.

KEYWORD ID	A	B	C	D	E

2	111, 1030	383, 846
3	111, 613	580, 846
4	111, 613, 1030, 1334	383, 993
5	111, 638	580, 993
6	65	613, 638
7	108	757, 955
8	67	757, 1010
9	106	955, 1010
10	99, 212, 353, 1169	955, 1273

Utility mining model successfully discovered the national security related keyword sets, #1, #6, #7, #8 and #9. However, Apriori, the most widely used frequent itemsets mining algorithm, found only one set, {613, 638}, regarding to this issue. The rest keyword sets by utility model are all related to earthquake loss, casualties, consequences, etc. Since an 8.0 magnitude earthquake happened in Sichuan in May 2008, it is reasonable that a large number of messages were on this event. The outputs of Apriori are most related to a bribery case and a Los Angeles earthquake, which are actually not sensitive from the view of Department of Security, although they were reprinted by many Medias.

Table 3 presents the top 10 high sensitivity keyword sets when 4, 8, 1, 2 are set as the sensitivity values for group (1), (2), (3), (4), respectively. We observe that the top 5 keyword sets are still for Sichuan Earthquake, but #6, #7 and #9 are new ones. Actually, they are obscene phrases since we enhanced the sensitivity level of obscenity and pornography words. This result shows that our model is able to find different sensitive keyword sets by varying the sensitivity levels in the sensitivity table.

Table 3. Top 10 high sensitivity keyword sets

5.2 Results

During January 2008 and July 2008, China spent great efforts on security issues in order to protect the athletes and maintain the order at the stadiums for the Beijing 2008 Olympic Games. So, we assigned the highest sensitivity value to security related words. The sensitivity values of the keywords are set at 8, 4, 2, 1 from group (1) to (4). Table 2 presents the top 10 high sensitivity keyword sets by Two-Phase algorithm (keyword-set-sensitivity threshold is set at 2.5%) and the top 10 frequent keyword sets by Apriori (support threshold is set at 0.5%), respectively. (Single frequent words are excluded in Table 2.)

Table 2. Top 10 high sensitivity keyword sets and frequent keyword sets

Rank	High sensitivity	Frequent
1	86	383, 580

Rank	High sensitivity
1	86
2	111, 1030
3	111, 1334
4	111, 613, 1030
5	111, 1150, 1334
6	171
7	171, 600
8	65
9	171, 1408
10	67

Web content sensitivity of a page is the sum of the sensitivity of all the maximal high sensitivity keyword sets it contains. 20 new pages come to the top 100 sensitivity pages when we varied the sensitivity table, which report the recently happened notorious sex scandal in Hong Kong. (Due to the page limitation, we do not provide the URLs of those pages here). It is resulted from the enhanced value of group (2) words.

Again, it shows that our model is able to find different Web pages by varying the sensitivity table.

6. Conclusions

In this paper, we proposed a utility-based Web content sensitivity mining approach, which introduces the concept of utility into the Web content sensitivity problem. Two new terms, keyword set sensitivity and Web content sensitivity, were proposed to measure the sensitivity of a keyword set and a Web page, respectively. Two-Phase algorithm was used to discover high sensitivity keyword sets. We applied our approach on a real-world Web dataset. The results showed that our model can capture more sensitive pages than frequency-based methods. When varying the sensitivity table, different high sensitivity keyword sets were obtained as well as high sensitivity pages. Our approach enables the Internet regulators to take different strategies according to different sensitivity values.

This is a preliminary work and a number of problems need to be discussed in our future work:

1) Improve our approach by applying sentiment classification on the high sensitivity pages.
2) Web page layout actually delivers rich information. How to obtain hints of sensitive messages from the layout structure deserves exploration.
3) The number of hyperlinks, in-degree and out-degree, the number of clicks and duration, are good indicators of the popularity of a Web page. How to involve such information into our approach deserves investigation.

References

[1] Ying Liu, Wei-keng Liao, Alok Choudhary, “A Two- Phase Algorithm for Fast Discovery of High Utility Itemsets”, PAKDD, Hanoi, Vietnam, 2005.

[2] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, “Thumbs Up? Sentiment Classification using Machine Learning Techniques”, ACL-02 Conference on Empirical Methods in Natural Language Processing, pp.79-86, 2002.

[3] Peter Turney, Michael Littman, “Unsupervised Learning Semantic Orientation from a Hundred-billion-word Corpus”, Technical report EGB-1094, National Research Council Canada, 2002.

[4] A Solution to Web Public Opinion Monitoring and Analyzing, http://www.chinaeg.gov.cn/2005/77/749.aspx I3S Web Search System, http://www.i3s.ac.cn/solutions/websearch.html.

[5] Mining “hidden phrase” definitions from the web , Hung.v.Nyugen , P.Velamuru , D.Kolippakkam, H.Davulcu , H.Liu