Coverage Patterns For Efficient Banner Advertisement Placement

S Bhargav
International Institute of
Information Technology
Hyderabad, India
bhargav.life@gmail.com

P Krishna Reddy
International Institute of
Information Technology
Hyderabad, India
pkreddy@iiit.ac.in

R Uday Kiran
International Institute of
Information Technology
Hyderabad, India
uday.rage@gmail.com

ABSTRACT

In an online banner advertising scenario, an advertiser expects that the banner advertisement should be displayed to certain percentage of web site visitors. In this context, to generate more revenue for a given web site, the publisher has to meet the demands of several advertisers by providing appropriate sets of web pages. To help the publishers and advertisers, in this paper, we propose a model of coverage patterns and a methodology to extract potential coverage patterns by analyzing click stream data. Given web pages of a site, a coverage pattern is a set of web pages visited by a certain percentage of visitors. The proposed approach has the potential to enable the publisher in meeting the demands of several advertisers. The efficiency and advantages of the proposed approach is shown by conducting experiments on real world data sets.

Categories and Subject Descriptors

H.4 [**Information Storage and Retrieval**]: Online Information Services - commercial services, web-based services

General Terms

Algorithms, Design, Experimentation

Keywords

Click stream mining, online advertising, internet monetization, computational advertising

1. INTRODUCTION

Online advertising is an important factor for the growth of internet economy. Banner advertising is one of the dominant mode of online advertising in addition to the contextual and sponsored search advertising. Most of the research work on online banner advertisement domain has been focused on bidding [3] and scheduling of advertisements [5]. In case of banner advertisement, a publisher receives requests from several advertisers for a set of web pages to place the advertisement covering certain percentage of web site visitors. In this situation, for a given web site, the publisher has a problem of providing appropriate set of web pages for each advertiser satisfying the respective coverage. In this paper, a model of coverage pattern is proposed to identify multiple sets of web pages where each set covers a given portion of the web site visitors. Given a set of click stream transactions, an algorithm is proposed to identify the coverage patterns. By extracting several coverage patterns, the proposed approach provides flexibility to the publisher in providing multiple advertising options to the advertisers.

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

2. PROPOSED MODEL AND ALGORITHM

2.1 Model of Coverage Patterns

To extract the coverage patterns, we consider the transactions generated from click stream data. Each transaction contains a set of web pages. Let $W = \{w_1, w_2, \cdots, w_n\}$ be a set of identifiers of web pages of a given web site and D be a set of click stream transactions, where each transaction T is a set of web pages such that $T \subseteq W$. Let T^{w_i}, $w_i \in W$, be the set of all $TIDs$ in D that contain the web page w_i. A set of web pages in W i.e., $X = \{w_p, \cdots, w_q\}$, $1 \leq p \leq q \leq n$, is called a pattern. A pattern containing k number of web pages is called a k-pattern.

DEFINITION 1. *(Relative frequency of a web page $w_i \in W$.) Normalizing the frequency of a web page $w_i \in W$ by transactional database size $|D|$ gives the relative frequency of w_i and is denoted as $RF(w_i)$. That is, $RF(w_i) = \frac{|T^{w_i}|}{|D|}$.*

DEFINITION 2. *(Coverage-support of a pattern X.) The ratio of number of transactions containing at least one item in X to the transactional database size is called the coverage-support of X and denoted as $CS(X)$. That is, $CS(X) = \frac{|(T^{w_p} \cup \cdots \cup T^{w_q})|}{|D|}$.*

For a pattern X, $CS(X) \in [0, 1]$. If $CS(X) = 0$, no single web page of X has appeared in the entire transactional database. If $CS(X) = 1$, every transaction in T contains at least one web page $w_j \in X$. It can be noted that once a pattern X satisfies the user-specified threshold value, say minimum coverage-support ($minCS$), then adding other items (in particular items co-occurring with X) will also satisfy the $minCS$. However, such patterns are uninteresting in the context of banner advertisement because they do not increase the coverage-support of a pattern significantly. To capture this aspect, we introduce another parameter known as the *overlap ratio*.

DEFINITION 3. *(Overlap ratio of a pattern.) Overlap ratio of a pattern $X = \{w_p, \cdots, w_q, w_r\}$, where $1 \leq p \leq q \leq r \leq n$ and $|T^{w_p}| \geq \cdots \geq |T^{w_q}| \geq |T^{w_r}|$, is the ratio of the number of transactions common in $X - \{w_r\}$ and $\{w_r\}$ to the number of transactions in w_r (i.e., minimum number of transactions in either $X - \{w_r\}$ or $\{w_r\}$). It is denoted as $OR(X)$ and is measured as follows.*

$$OR(X) \quad = \quad \frac{|(T^{w_p} \cup T^{w_{p+1}} \cup \cdots \cup T^{w_q}) \cap (T^{w_r})|}{|T^{w_r}|}$$

For a pattern X, $OR(X) \in [0, 1]$. If $OR(X) = 0$, there exists no common transactions between $X - \{w_r\}$ and $\{w_r\}$. If $OR(X) = 1$, w_r has occurred in all the transactions where at least one web page $w_j \in (X - \{w_r\})$ has occurred. The parameter *overlap ratio* also does not satisfy the *downward closure property* if a pattern is considered as an unordered set of web pages. However, it was observed that this measure satisfies the *downward closure property* if a pattern is an ordered set in which web pages are sorted in descending order of their frequencies. This property is known as the *sorted closure property* [4].

Now, we define the coverage pattern as follows.

DEFINITION 4. *(Coverage pattern X.) A pattern* $X = \{w_p, \ldots, w_q, w_r\}$*, where* $1 \leq p \leq q \leq r \leq n$*, is said to be a coverage pattern if* $CS(X) \geq minCS$ *and* $OR(X) \leq maxOR$ *and* $RF(w_i) \geq minRF$ $\forall w_i \in X$*. The variables* $minCS$*,* $maxOR$ *and* $minRF$ *represent the user-specified minimum coverage support, maximum overlap ratio and minimum relative frequency respectively.*

Problem statement: Given a transactional database T, W and user-specified $minRF$, $minCS$ and $maxOR$, identify the set of coverage patterns such that

i. If X is a coverage 1-pattern (i.e., $k = 1$), then $RF(w_i) \geq minRF$ and $RF(w_i) \geq minCS$, $w_i \in X$.

ii. Otherwise (i.e., when $k > 1$), each coverage pattern X must have $CS(X) \geq minCS$, $OR(X) \leq maxOR$ and $RF(w_i) \geq minRF$, where $w_i \in X$.

2.2 Algorithm

In this section, we explain the coverage patterns mining (CPMine) algorithm. First, we define the notion of non-overlap pattern X.

DEFINITION 5. *(Non-overlap pattern X.) A pattern X is said to be* non-overlap *if* $OR(X) \leq maxOR$ *and* $RF(w_i) \geq minRF$ $\forall w_i \in X$.

The CPMine algorithm considers $minRF$, $minCS$ and $maxOR$ parameters for mining coverage patterns. CPMine [1] employs *level-wise* search to discover the complete set of coverage patterns. In *level-wise* search, k-patterns are used to explore $(k + 1)$ patterns. In the first scan of the database, CPMine discovers the set of all frequent items whose relative frequency is greater than $minRF$ (denoted as NO_1) and coverage 1-patterns (denoted as L_1). Then, the items in NO_1 are sorted in the descending order of their frequencies. This is the important exception that has to be carried out in the CPMine algorithm to efficiently mine coverage patterns. Each item $i \,\epsilon\, NO_1$ is of the form $< i, T^i >$ where T^i denotes a set of transaction ids' which contain an item i. Using NO_1 as a *seed set*, candidate patterns C_2 are generated by $NO_1 \bowtie NO_1$. From C_2, the patterns that satisfy $minCS$ and $maxOR$ are generated as coverage 2-patterns, L_2. Simultaneously, all candidate 2-patterns that satisfy $maxOR$ are generated as non-overlap 2-patterns NO_2. Since overlap patterns satisfy *sorted closure property*, C_3 is generated by combining $NO_2 \bowtie NO_2$. From C_3, L_3 and NO_3 are discovered. The above described process is repeated until no new non-overlap patterns are found, or no new candidate pattern can be generated. The set L = $\{L_1 \cup L_2 \cdots \cup L_k\}$ represents the set of discovered coverage patterns.

3. EXPERIMENTS

Figure 1 shows how the number of patterns extracted with CPMine algorithm vary with $minCS$ and $maxOR$. The experiments are conducted on BMS-POS data set [6]. At the fixed $maxOR$ value, it can be observed that the number of coverage patterns decreases as $minCS$ value increases. Normally, as $minCS$ value increases the number of pages in the set increases. Several small sets fail to qualify the $minCS$ threshold value. At the fixed $minCS$ value, it can also be observed that the number of patterns increases as $maxOR$ value increases. It is due to the fact that as $maxOR$ value is increased more number of web pages qualify $minCS$ value. As a result, the number of coverage patterns also increases.

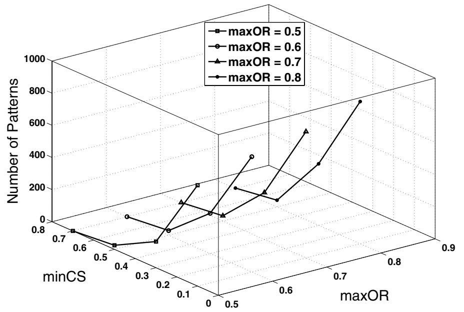

Figure 1: Performance of CPMine algorithm.

Table 1 shows the sample of two coverage patterns with the corresponding coverage supports extracted from the click stream transaction data set of an online news portal MSNBC [2] using proposed approach. It can be observed that there is no intersection between the first and second coverage patterns, even though they differ little in terms of coverage support value. The publisher can provide either of the option to the advertiser.

Table 1: Two patterns extracted from MSNBC dataset [2]

Coverage pattern	Coverage support
local, on-air, health, tech	0.47
news, weather, sports, bulletin-board service	0.44

4. CONCLUSION

In this paper, we have proposed an efficient approach to help the publisher of the web site to manage the requests of several advertisers for banner advertisement placement. We have proposed a model of coverage patterns and an iterative approach to extract the same. The efficiency and advantages of proposed approach are shown by conducting experiments on real world data sets. As a part of future work, in addition to building a tool for banner advertisement placement, we are planning to extend the notion of coverage patterns for extracting potential knowledge patterns in other domains.

5. REFERENCES

[1] S. Bhargav and P. K. Reddy. Data analysis approaches for improved online banner advertisement placement and identifying suitable crop cultivation period. Technical report, IIIT, Hyderabad, India, 2011.

[2] A. Frank and A. Asuncion. UCI ML repository, 2010.

[3] A. Ghosh, B. I. Rubinstein, S. Vassilvitskii, and M. Zinkevich. Adaptive bidding for display advertising. In *In Proc. of 18th World wide Web Conference(WWW 09)*, 2009.

[4] B. Liu, W. Hsu, and Y. Ma. Mining association rules with multiple minimum supports. In *In Proc. of the fifth ACM SIGKDD (KDD 99)*, 1999.

[5] M. Mahdian, H. Nazerzadeh, and A. Saberi. Allocating online advertisement space with unreliable estimates. In *Proceedings of the 8th ACM conference on Electronic commerce*, 2007.

[6] Z. Zheng, R. Kohavi, and L. Mason. Real world performance of association rule algorithms. In *In Proc. of the seventh ACM SIGKDD(KDD 01)*, 2001.