ISSN : 0976-8491 (Online) | ISSN : 2229-4333 (Print)

Web Profitability Mart

[1]Nagaraj G Cholli, [2]I M Umesh, [3]Dr. Srinivasan G N
[1,2,3]Dept. of Information Science & Engg., R V College of Engg., Bangalore, Karnataka, India

Abstract
This paper tries to enhance web profitability by following the patters of clicks followed by users, that are click stream data, and then it tries to match user's behavior to existing patters by using Vectors. Here Vectors deploy existing three existing algorithms, which are Shingling, Min-hashing and Locality sensitive hashing to find the closest match. Based on this match and patterns, users are presented relevant advertisements in future and a pattern of their behaviors is deduced. These patterns can hence be used for future users as a genre of their choice.

Keywords
Web Profitability, Click Stream Data, Vectors

I. Introduction

In the current era of E-commerce online stores offer great levels of interactive marketing. However there is always a compromise between what users want and what they are led to. Currently Online stores target visitors using many type of information like their demographic areas, their interests, their behavior over web, their purchase history and even how they got to the website. Even web sites are collaborating with many social sites to understand the need of the market. Social sites serve as a great platform for easily available data and hence can lead to data mining and virtual surveys to understand the popularity of products in masses. Although many algorithms are available in market to understand user behavior by probabilistic studies or any other means, here in this paper we match the patterns using vector analysis supported by many existing algorithms. These ways when put together form a web profitability mart which can be greatly beneficial for Organizations dealing in E-commerce. We start observing the Click Stream data, which is basically how different users behave while purchasing on a website and the path which users take to reach the website. This study helps to differentiate between potential customers and naïve customers. Then we try to make a data mart using data available over associated social websites. Data marts are further refined to proper schemas in order to differentiate and associate data as per needs. A set of patterns are deduced from these data marts over which data mining is performed. Finally Vector analysis is performed over these results to give concrete results. Vector analysis is supported by Shingling, Min-hashing and Locality Sensitive hashing algorithms.

II. Understanding the Need

Most of the time while searching for a product over internet there is a compromise between what actually users want and what they are led to. These days most of the internet portals, search pages, social networks, e-commerce and other websites are not necessarily designed in order to maximize user convenience and benefits. The reason is same as to why some retail stores place the most popular items (e.g. bread, milk) in the farthest possible place from the entrance. Indeed most of these intermediaries are in the business of matching consumers with products. Trouble is prior to visiting an intermediary; consumers are interested in only some products, which may not necessarily be the ones that yield the highest profits for the intermediary. If the latter was offering the perfect information service (i.e. One that enabled the consumers to find what they want most quickly and efficiently), it would be losing valuable potential revenues. Hence the incentive to attract users with products that they want a priori and then divert them towards products that they might be interested in ex-post (i.e. once there). Thus, consumers coming to the super market to buy daily staples (say, bread and milk) might be also induced to get expensive chocolate if they have to walk past the corresponding aisle anyway. In the same way, Google faces a subtle issue in designing its search result pages: consumers are mostly interested in the objective search results, but all revenues come from the sponsored search adds on the right hand side. The result is a compromise between what consumers want and what produces more revenue. So this infers that understanding users click stream path also becomes important.

III. Click Stream Data

Click stream data provides information about the sequence of pages or the path viewed by users as they navigate a web site. These results suggest that paths may reflect a user's goals, which could be helpful in predicting future movements at a web site. Path data may contain information about a user's goals, knowledge, and interests. The path brings a new facet to predicting consumer behaviour that analysts working with scanner data have not considered. Specifically, the path encodes the sequence of events leading up to a purchase, as opposed to looking at the purchase occasion alone. To illustrate this point consider a user who visits some random web site, www.flipkart.com. Suppose the user starts at the home page and executes a search for "Harry potter and chamber of secrets", selects the first item in the search list which takes them to a product page with detailed information about the book. Harry potter and chamber of secrets by J.K. Rowling (2003). Alternatively, another user arrives at the home page, goes to the Books category, surfs through a score of book descriptions, repeatedly backing up and reviewing pages, until finally viewing the same. Harry potter and chamber of secrets product page. Which user is more likely to purchase a book: the first or second? Intuition would suggest that the directed search and the lack of information review (e.g., selecting the back button) by the first user indicates an experienced user with a distinct purchase goal. The meandering path of the second user suggests a user who had no specific goal and is unlikely to peculiarities. These behaviors can be recorded using sessions and specialized software over the server. These patterns can serve to determine the behaviors of customers and hence can lead to web profitability and making patterns.

IV. Vectors for Similarities

In order to understand how vectors can be used to map similarities first we need to understand what vector is and what does dot product of a vector mean. Thus a vector is a geometric entity characterized by a magnitude and a direction. A vector is defined as a directed line segment, or arrow, in a Euclidean space. The dot product of two vectors a and b (sometimes called the inner product, or, since its result is a scalar, the scalar product) is denoted by a · b and is defined as:

$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \, \|\mathbf{b}\| \cos\theta \qquad (1)$$

Where, in equation (1) θ is the measure of the angle between a and b,

this means that a and b are drawn with a common start point and then the length of a is multiplied with the length of that component of b that points in the same direction as a. Vector Politics attempts to deal with the inherent complexity of politics by adding additional dimensions to describe a person's viewpoints. Vectors are used to derive the interest and preferences of a particular user. We use the equation above to perform calculations on patterns observed. In equation (1):

- A is the choice intersect of the user.
- B are the intersects of choices already present is server.
- The smaller the dot product, more closely is the interests of users related.

Thus, it helps in understanding the user and providing the best filtered results, and is beneficial for E-commerce. This concept of vectors is applied to the signatures derived by three algorithms mentioned below which are first applied on data step wise.

A. Shingling

Shingling means converting the given data into sets. A k -shingle (or k -gram) for a document is a sequence of k characters that appears in the document. Example: k=2; doc = abcab. Set of 2- shingles = {ab, bc, ca}. Thus shingles can be considered as bags. The value K must be optimum to the data. To compress long shingles, we can hash them to 4 bytes. Represent a data by the set of hash values of its k-shingles.

B. Minhashing

In Minhashing, Shingles are converted into signatures which can be dealt with easily. Signatures are short integer vectors that represent the sets, and reflect their similarity. They are easier to deal with.

C. Locality Sensitive Hashing

In this method, candidate pairs are those pairs of signatures that we need to test for similarity. These pairs can further be tested By Vector analysis as mentioned in section IV. Thus these algorithms can be used to polish the results deduced by vector analysis. They make the results more trustable. Results approximately close to zero are considered perfect to the existing patterns. More the results deviate from zero less sparse becomes the proximity of the consumer to its target.

V. Construction of Data Mart

A data mart is the access layer of the data warehouse environment that is used to get data out to the users. The data mart is a subset of the data warehouse which is usually oriented to a specific business line or team. Data marts are usually constructed for E-commerce using various sources like:

1. Social web sites, clicks, comments, like and dislike of users are recorded which can act as useful data later. This strategy is very useful for studying and strategizing for current market.
2. Online surveys, made available by various websites.
3. Offline surveys, through magazines or in person.
4. Blogs and reaction of people to various blogs.
5. Old trends of data available.
6. Click stream data that's behavior of users over websites.
7. Data sources of competitors can be analyzed.

Once the Data mart is ready, next step is Data mining. Data mining is extraction of useful information from data. This useful information can be highly useful for organizations and can behave as signatures or trends. One very important step involved in this process is data cleaning. Data available may be highly unorganized or there may be lots of missing or null values. There can also be data which is in consistent, so all this data need to be cleaned using various data cleansing tools. Next step is formation of data cube. That's organizing data in an ordered schema which can be a star schema or snow flake schema. Once data cube is generated, we have a perfect architecture to deploy data mining activities on data. The useful information hence derived act as trend setters or signatures. These signatures can then be compared with new ones using Vector Analysis Techniques.

VI. Results

Now we can proceed to the interesting part of our data warehouse: retrieving information. The average number of minutes from login to order

```
SQL>
SQL> select round(avg(minutes_login_to_order),2 )
     from clickstream_fact where minutes_login_to_or
     <30;
```

The average number of days from first being invited to the site by email to the first order.

```
SQL> select round(avg(days_first_invite_to_order),2)

     from clickstram_fact where days_first_invite_to

     _order <10
```

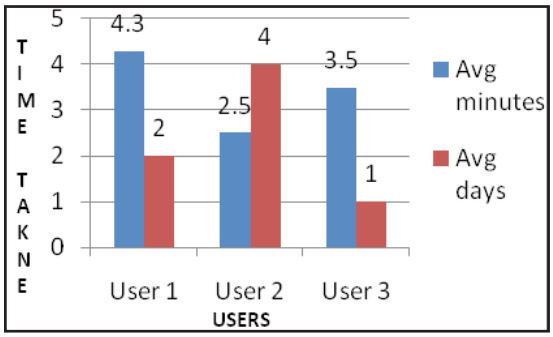

Fig. 1: Graphical Representation of the Result

The count of customers purchased similar set of items

```
SQL> /
select item_name,qty, customer_id from clickstream_fact order by customer_id;
```

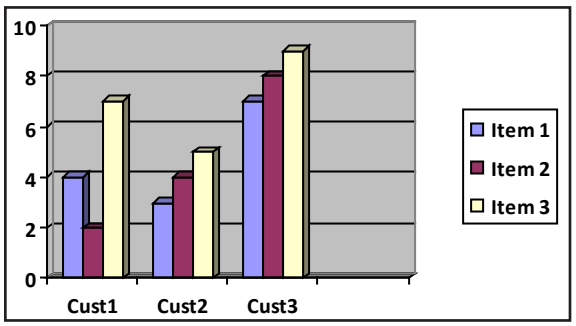

Fig. 2: Graphical representation of the result

The Concept of vectors is very much used in Matrimonial sites, where:

- People register themselves by entering certain attributes like their hobbies, interests, like and dislikes.

- Every member registered is plotted as a line on graph depending upon their attributes.
- The gradient of line is decided by the attributes they have mentioned.
- While forming pairs, different priority attributes are considered, along with, a Vector dot product of existing plots are computed in all combinations.
- The Dot products which lie closest form the best matching pair. Further refinement is done using priority attributes.

Very same approach can be implemented to enhance web profitability. Steps are:

- A data mart is made by collecting useful data from different sources.
- Data mining is done to refine and cleanse the collected data regarding current trends.
- Define the idea or product which an organization wants to launch in market.
- Use Shingling, Min-hashing and Locality Sensitive hashing algorithms to get the intensity of matches.
- Consider and note down important attributes in Current trends.
- Follow the same for new idea of product.
- Plot both on a graph as per the attributes, already decided, as straight line.
- The gradient of these straight lines is decided by the attributes.
- Perform a Vector Dot product between the Idea or product (line representing it), and various lines representing different market trends.
- Hence a Numerical value is derived for each case.
- Now, the smallest numerical value(i.e. the value closest to zero), marks the best match with the trend.
- The results can further be refined by using results derived by applying 3 algorithms already mentioned.
- Hence a very accurate result is obtained, and an organization can start to promote that idea or product will full enthusiasm and investments.

VII. Conclusion

Thus in this paper various methodologies to enhance web profitability has been mentioned. We noticed how actually web sites are framed and constructed. Most of the times there is a compromise between what users actually want and what they are led to while designing pages. Thus just as bread and milk are often found at far-away ends of super market. Web sites that match consumers with certain products have an incentive to steer users to products that yield the highest margins. Next we saw how click stream data can be highly valuable to judge potential customers and naive customers. We analyze the path which users take for the website and while surfing the website. This information is recorded and can act as virtual surveys for organizations. A data mart is then formed, in which various data cleaning activities are deployed. Then Data is arranged in form of relevant schema and finally Data mining is done over data to seek useful business information. This acts as trends or signatures for the latter process. Then we used Vectors to find the extent of deviation of what actually users want and what we have in store for them. Accordingly the users can be presented with potential advertisements and can be induced to buy products and increase organizations incentives. Three algorithms named Shingling, Min-hashing and Locality Sensitive hashing algorithms are used to divide data in sets, form signatures and do comparisons between signatures. The outputs hence deduced are highly accurate and thus potential customers can be intensively targeted. Thus these methodologies can greatly benefit an organization to form a rich web profitability mart and help them to increase their incentives, sales and profits.

References

[1] Alan L. Montgomery, Shibo Li, Kannan Srinivasan, John C. Liechty,"Modeling Online Browsing and Path Analysis Using Clickstream Data", November 2002 First Revision, September 2003 Second Revision, February 2004 Third Revision, February 2004, Associate Professor at Graduate School of Industrial Administration, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh; Assistant Professor of Marketing at Rutgers University, 228 Janice Levin Building, 94 Rockafeller Road, Piscataway, NJ; Professor of Management, Marketing, and Information Systems and Director of the Center for E-Business Innovation at the Graduate School of Industrial Administration, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA ; Assistant Professor of Marketing and Statistics at the Pennsylvania State University, 710 M Business Administration Building, University Park, PA.

[2] Prof Angrei Hagiu,"Why are Web Sites So Confusing?", Associate Professor, harvard Business School. ; Technology next, Vol. 1, Issue 3, October 2011.

[3] Jiawei Han, Micheline Kamber, Jian Pei,"Data Mining Concept and Techniques", 2011.

[4] Mena, Jesus,"WebMining for Profit: E-Business Optimisation", Butterworth-Heinemann, 2001.

[5] Sismeiro, Catarina, Randolph E. Bucklin,"Modeling Purchase Behavior at an Ecommerce Website: A Conditional Probability Approach", Anderson School at UCLA,Working Paper, 2003.

[6] Blattberg, R., J. Deighton,"Interactive Marketing: Exploiting the Age of Addressability", Sloan Management Review, Fall, 1991

[7] Moe, Wendy, W.,"Buying, Searching, or Browsing: Differentiating between Online Shoppers Using In-Store Navigational Clickstream", Journal of Consumer Psychology, 13 (1&2), pp. 29-40, 2003.

[8] Moe, Wendy, W., Peter S. Fader,"Dynamic Conversion Behavior at e-Commerce Sites", Management Science, forthcoming, 2004.

[9] Moe, Wendy W., Hugh Chipman, Edward I. George, Robert E. McCulloch,"A Bayesian Treed Model of Online Purchasing Behavior Using In-Store Navigational Clickstream", Working Paper, 2002.

[10] Montgomery, Alan L.,"Applying Quantitative Marketing Techniques to the Internet", Interfaces, 30, 2, pp. 90-108, 2001.

[11] New York Times,"Easier-To-Use Sites Would Help E-tailers Close More Sales", Bob Tedeschi, June 12, 2000.

[12] Park, Young-Hoon, Peter S. Fader,"Modeling Browsing Behaviour at Multiple Websites", Marketing Science, forthcoming. 2004.

[13] Redish, Janice, (2002),"Information-Rich Web Sites: Challenges and Opportunities", [Online] Available: http://www.redish.net/cmu.pdf. Last accessed October 2002.

[14] Underhill, Paco,"Why We Buy, the Science of Shopping, Touchstone Books", 1998.

[15] David Loshin,"Business, Intelligence: The Savvy Manager's Guide", Publisher: Morgan Kaufmann, Jul 2003.
[16] Mark W Humphries,"Data warehousing", Architecture and Implementation, Pearson Education.
[17] [Online] Available: http://en.wikipedia.org/wiki/Euclidean_vector
[18] [Online] Available: http://www.physics.uoguelph.ca/tutorials/vectors/vectors.html