Association for Information Systems
AIS Electronic Library (AISeL)

MWAIS 2013 Proceedings | Midwest (MWAIS)

5-24-2013

Identifying Interesting Knowledge Factors from Big Data for Effective E-Market Prediction

Santhosh Kumar Lakkaraju
sklakkaraju@pluto.dsu.edu

Identifying Interesting Knowledge Factors from Big Data for Effective E-Market Prediction

Santhosh Kumar Lakkaraju
Doctoral Student
sklakkaraju@pluto.dsu.edu

ABSTRACT

Knowledge management plays an important role in disseminating valuable information. Knowledge creation involves analyzing data and transforming information into knowledge. Knowledge management plays an important role in improving organizational decision-making. It is evident that data mining and predictive analytics contribute a major part in the creation of knowledge and forecast the future outcomes. The ability to predict the performance of the advertising campaigns can become an asset to the advertisers. Tools like Google analytics were able to capture user logs. Large amounts of information ranging from visitor location, visitor flow throughout the website to various actions the visitor performs after clicking an ad resides in those logs. This research approach is an effort to identify key knowledge factors in the marketing sector that can further be optimized for effective e-market prediction.

Keywords

Knowledge management, knowledge creation, predictive analytics, prediction, advertising, E-Marketing

INTRODUCTION

We are living in a knowledge-based economy, and survival in such situations depends on the ability to convert information to knowledge. We have been observing lots of innovations in our day-to-day life. Information improves decision-making and stimulates innovation. But there is a great need to capture such information and transform it into knowledge so that it paves a path for innovation (Hair, 2007). Knowledge creation, application and succession typically depend on the researcher's ability to understand, analyze and handle the information. Extensive research has been done on how to capture, analyze and transform information into knowledge (Awad et al., 2008; Bran et al., 2009; Duggan et al., 2008; Ranjan et al., 2011; Zhang et al., 2008; Zhang et al., 2011). Advertiser's ability to place ads with better quality and better productivity is constantly tested by the search engines and there is an increasing need for models that are capable enough to predict the performance of ads and return on investment (Adar et al. 2007; Apte et al., 2003; Apte et al. 2002; Ashkan et al. 2009; Guazzelli et al. 2009).

A substantial growth was observed in the field of predictive analytics over the past few years (Hair, 2007; Shmueli et al., 2011). Predictive models are highly dependent on the grounded data collected through the means of analytical tools like Google analytics. Such data when effectively mined and applied through the predictive models can predict future performances. There is a growing need for such predictive applications in the recent years (Oliner et al., 2012; Hair, 2007). Analytical tools were able to capture data of multiple instances from a single visitor and a single campaign is capable of capturing thousands of visitors to a website. All such visitor instances collectively become a huge volume of information which doesn't have any meaning without the data mining tools (Regelson et al., 2006; Richardson et al., 2007). Such information is typically used to predict visitor behavior and in turn contribute to knowledge discovery (Hair, 2007). Data mining and predictive analytics are two inter-dependent processes where data mining typically identifies patterns and underlying relationships within the collected data, and predictive analytics/models utilizes both structured and unstructured data examining the visitor patterns obtained through mining the visitor logs to the website and formulates the predictors which predict the future performances (Hair, 2007; Yang et al., 2001). Oliner et al., (2012) states that "*prediction/predictive models from a business point of view can predict market strategy, ad placement etc. by providing a range of predicted values significantly termed as confidence intervals within which the true value is likely to lie".* Shmueli et al., (2011) defines the success of the predictive model rely on "*the model's ability to generate accurate predictions of the new outcomes*".

Electronic marketing revolutionized the conventional marketing industry and became a primary source for the marketers to reach numerous people across the globe. Search engines like Google provide interactive interfaces like Ad-words and Analytics, through which the advertisers place text ads, video ads, ads on social networking sites like Facebook and Twitter to attract traffic to their respective websites. Different advertisers have different motives in approaching search engine marketing. There are advertisers selling products, services or content directly online through their websites, advertisers depending on leads, drive traffic to their websites, increase brand awareness, enhance reputation and some trying to improve customer service and aim at customer satisfaction. Search engine marketing systems are really tricky, driving traffic to the websites doesn't solve the actual problem, the directed traffic should be able to generate a conversion else the advertiser

would result in spending money with no return on investments. To overcome such situations advertisers are depending on data mining and predictive modeling tools (Hair, 2007; Oliner et al., 2012). These tools capture large amounts of structured (analytical, numerical information) and unstructured (information in the form of text/images) visitor information. By applying tacit knowledge to this structured/unstructured information we can create and document explicit knowledge (Hair, 2007, Zhang et al., 2011). It is evident that human intelligence is highly capable of analyzing unstructured information and machine intelligence is capable of handling structured information. Later sections will provide a brief overview on advertising and common terminology that we use in advertising, extensive literature review on existing research and extract different knowledge factors that were identified from the literature.

MOTIVATION

Search engine marketing industry is advancing rapidly in recent years. Predictive analytics and data mining techniques are being increasingly applied on marketing data for making better return on investments. Markov model, Support Vector Machine's, logistic regression, discriminant analysis and decision trees are some of the popular techniques that are applied to big data and extract knowledge (Awad et al. 2008; Regelson et al. 2006; Richardson et al. 2007; Thorleuchteret al. 2012). It is evident that data logs are highly used and efficient enough for making better business predictions. Predictive models are capable enough in guiding the advertisers to choose an optimal marketing strategy and effective ad placement (Oliner et al., 2012). Using historical data like user logs and visitor actions are common strategies applied by the researchers who intend to develop predictive models. This research provides a systematic literature review of the existing literature from a knowledge management perspective, identifying the key knowledge factors that were addressed in existing literature and present an overview on how to optimize them for making better market predictions.

OVERVIEW ON ADVERTISING

Popular search engines like Google, Yahoo and Bing are acting as possible interfaces between the advertisers and customers. The working mechanism of online advertising and search engine marketing is clearly explained in this section. The advertiser develops a web site with multiple pages containing information about an industry, educational programs or product sales are usually developed with a site map that helps the know bots and crawlers to easily navigate through the web site. Good sets of relevant keywords are used throughout the text embedded within the web pages. User clicks on a particular ad and gets redirected to the advertisers web page, which is formally denoted as a landing page (Punera et al. 2010). The landing page should be structured in an attractive way that it channels the visitors to generate conversions. Search engines act as a primary source for driving traffic to those web sites. Search engines provide thousands of results for every keyword and it is observed that users are less likely to browse beyond first few pages, which apparently imply that search engine positioning is highly important for the web sites to survive in the market. Richardson et al. (2007) implied that the click through rate of an ad is directly dependent on the ad position and ads appearing in the latter pages have less visual attention and are less likely to get clicked resulting in the loss of revenue. Basic terminology used in online advertising is explained in table1.

Terminology	Definition
Conversion	The actions made by the visitor on the web site those benefits the advertiser.
Impression	Every time an ad appears on a search result is counted as an impression.
Click	It occurs when a user clicks the ad leading him/her to the web site is a click.
Lead	A visit to a page that enters the conversion cycle and potentially lead to conversion.
Click Through Rate (CTR)	(Number of clicks)/ (number of impressions) i.e. clicks to impression ratio.
Bounce Rate	The percentage of visitors who enters the conversion cycle and exits the web page without making a conversion.

Table 1: Basic terminology used in e-marketing (Grappone et al. 2006; Introna et al. 2000; Jones, 2008; Moran et al. 2009)

LITERATURE REVIEW

A substantial growth was observed in the field of predictive analytics over the past few years (Hair, 2007; Shmueli et al., 2011). Predictive models are highly dependent on the grounded data collected through the means of analytical tools like Google analytics. Such data when effectively mined and applied through the predictive models can predict future performances. There is an increasing need for such predictive applications in the recent years (Oliner et al., 2012; Hair, 2007). Analytical tools were able to capture data of multiple instances from a single visitor and a single campaign is capable of capturing thousands of visitors to a website. All such visitor instances collectively become a huge volume of information, which doesn't have any meaning without the data mining tools. Such information is typically used to predict visitor behavior (Hair, 2007). Data mining and predictive analytics are two interdependent processes where data mining typically identifies patterns and underlying relationships within the collected data, and predictive analytics/models utilizes both structured and unstructured data examining the visitor patterns obtained through mining the visitor logs to the website and formulates the predictors which predict the future performances (Hair, 2007). Oliner et al., (2012) states that "*prediction/predictive models from a business point of view can predict marketing strategy, ad placement etc. by providing a range of predicted values significantly termed as confidence intervals within which the true value is likely to lie".* Shmueli et al., (2011) defines the success of the predictive model rely on "*the models ability to generate accurate predictions of the new outcomes*".

Regelson et al. (2006) propose the use of keyword clusters to predict the click through rates of ads. Keyword clustering is grouping similar search terms together as clusters and captured broad estimates of click through rates. (Regelson & Fain, 2006) hypothesize that different search terms have different likelihood of receiving click through rates. (Regelson & Fain, 2006) tried to capture the periodic click through rates of existing search terms and marked patterns within the obtained historical click through data and provided the future outcomes from those patterns. (Regelson & Fain, 2006) synthesized that keyword clusters are capable of producing more accurate estimates of click through rates for search terms without historical information.

Richardson et al., (2007) developed a predictive model to estimate the click-through rate for new ads. Richardson et al., (2007) chose each ad as a single individual complete system, applied logistic regression and measured the predictor variable. Richardson et al., (2007) collected historical information from a set of active ads in Microsoft and web search engines. The historical information comprises of URL, keywords, ad-content, URL, clicks and respective views. Richardson et al., (2007) initially estimated individual term's click through rate followed by the related term click through rates and further estimated the ad-quality. Richardson et al., (2007) further applied logistic regression to predict the query independent click through rates of an advertisement.

Bounce rates was a key issue bugging the advertisers for a long time. Sculley et al., (2009) defined bounce rate of an ad as a tendency of a visitor to click an ad and immediately moves on to some other task. Sculley et al., (2009) provided a quantitative and qualitative analysis to successfully depict bounce rate as a measure of visitor satisfaction and illustrated a predictive model to analyze different features of an ad that contributes to the increase or decrease of its bounce rates. Sculley et al., (2009) tried to address how an advertiser can quantify user satisfaction based on the institutional bounce rate, how the search queries typically involve in increase or decrease of bounce rates and how can such bounce rates be predicted in the absence of historical information. Sculley et al., (2009) applied logistic regression and support vector regression to predict the possible bounce rate and further analyze the likelihood of visitor satisfaction.

Wen et al. (2002) proposed a methodology to analyze various search queries obtained from visitor logs, group similar queries together, form query clusters and analyze them with respect to that of the keyword clusters. Wen et al. (2002) hypothesized that a combination clustering of search terms and queries will yield better results that keyword alone or queries alone. Wen et al. (2002) applied incremental DBSCAN algorithm for clustering queries based on the content-word similarity and provided empirical evidence how the similarity functions typically effect the clustering results. Through this study Wen et al. (2002) synthesize that query clustering strategy is more efficient than keyword clustering alone.

The below given table clearly elucidate different knowledge factors that were addressed in each of these existing literature.

KA – Knowledge Acquisition, KC – Knowledge Creation and KM – Knowledge Management

Author	Objective	Knowledge factors	Knowledge
(Adar et al., 2007), (Punera & Merugu, 2010) (Raj, Dey, & Gaonkar, 2011)	Analyze World Wide Web data and identified behavioral patterns to understand predictive power.	Visitor/ User Behavior	KA, KC
(Apte et al., 2003, 2002)	Probabilistic modeling for insurance risk management and text mining.	Scalability and Reliability	KA, KC
(Ashkan et al., 2009) (Regelson & Fain, 2006)	Predictive model to estimate the ad click through rate through query intent analysis	Click Through Rates	KA, KC
(Awad et al., 2008b)	Predict the WWW surfing by using multiple evidence combination	Latency (time delay), personalization	KA, KC
(Duggan & Payne, 2008)	A predictive model to predict user domain knowledge from search behaviors	Domain Knowledge	KA, KC
(Gruhl, Guha, Kumar, Novak, & Tomkins, 2005)	Correlated postings in blogs, media and web to draw conclusions on predictive power	Queries	KC, KA
(Huffman & Hochster, 2007)	A predictive approach to approximate the visitor satisfaction based on query relevance	Query relevance	KC, KA
(Oliner et al., 2012)	Use of log analysis for making better predictions	Visitor Log	KC, KA, KM
(Introna & Nissenbaum, 2000)	Explain the role of indexing in achieving search engine recognition	Indexing	KC, KA, KM
(Sculley et al., 2009)	Proposed a predictive model to predict the bounce rates in sponsored search	Bounce Rates	KC, KA
(Wen et al., 2002)	Applying the knowledge of query clustering using user logs for identifying better method	Query clustering, keyword clustering	KC, KA, KM
(White, Dumais, & Teevan, 2009), (X. Zhang, Cole, Street, & Belkin, 2011)	Characterize the influence of domain expertise on web search behavior	Domain knowledge and user behavior	KC, KA
(Ghose & Yang, 2008)	Analyzed the firms behavior for predicting the E-marketing performance	Behavior	KC, KA
(Zhu et al., 2009)	Analyze historical click through data to optimize the search engine revenue	Click Through Data	KC, KA
(Yang et al., 2001)	Applying the web log mining for higher precision of advertising	Patterns from historical data	KC, KA

Table 2: Different knowledge factors identified from existing literature

KNOWLEDGE FACTORS

From table 2 the following key knowledge factors were extracted from existing e-market and predictive analytics literature.

- Visitor/User Behavior,
- Scalability and reliability,
- Click Through Rates/Click Through Data,
- Latency (Time Delay), Personalization,
- Domain Knowledge,
- Query relevance, Query Clustering, Keyword Clustering,
- Indexing,
- Visitor Log, Patterns from user log,
- Bounce Rate.

These nine knowledge factors have their own importance and impact on the advertisers return on investment. These nine knowledge factors have their own importance in predicting specific visitor actions on an advertiser's web page.

CONCLUSION

Predictive analytics in E-marketing has increasing importance in recent years. The advancements in the field of data mining encouraged the researchers to identify different possibilities to tackle big data. New knowledge is applied, often created and shared. This research successfully analyzed existing literature on predictive analytics and E-marketing from a knowledge management perspective and tried to identify critical success factors from them. Through the application of knowledge management concepts nine critical success factors were identified which can typically bring a major impact on E-marketing and advertising sectors. Further this research can be extended in identifying different approaches to optimize these knowledge factors.

ACKNOWLEDGEMENTS

I thank Dr. Omar El-Gayar, Dr. Amit Deokar and Dr. Surendra Sarnikar for their suggestions and encouragement to conduct this research, and all the reviewers for their valuable comments.

REFERENCES

1. Adar, E., Weld, D. S., Bershad, B. N., & Gribble, S. D. (2007). Why We Search : Visualizing and Predicting User Behavior. *Proceedings of the 16th international conference on World Wide Web (WWW '07)*. Retrieved from http://doi.acm.org/10.1145/1242572.1242595
2. Apte, C. V., Hong, S. J., Natarajan, R., Pednault, E. P. D., Tipu, F. a., & Weiss, S. M. (2003). Data-intensive analytics for predictive modeling. *IBM Journal of Research and Development*, *47*(1), 17–23. doi:10.1147/rd.471.0017
3. Apte, C. V., Natarajan, R., Pednault, E. P. D., & Tipu, F. a. (2002). A probabilistic estimation framework for predictive modeling analytics. *IBM Systems Journal*, *41*(3), 438–448. doi:10.1147/sj.413.0438
4. Ashkan, A., Clarke, C. L. A., Agichetein, E., & Guo, Q. (2009). Estimating Ad Clickthrough Rate through Query Intent Analysis. *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 01 (WI-IAT '09)*, *1*. Retrieved from http://dx.doi.org/10.1109/WI-IAT.2009.39
5. Awad, M., Khan, L., & Thuraisingham, B. (2008a). Predicting WWW surfing using multiple evidence combination. *The VLDB Journal*, *17*(3), 401–417. doi:10.1007/s00778-006-0014-1
6. Awad, M., Khan, L., & Thuraisingham, B. (2008b). Predicting WWW surfing using multiple evidence combination. *The VLDB Journal*, *17*(3), 401–417. doi:10.1007/s00778-006-0014-1
7. Bran, C. A., Malone, T., Lewis, D., & Burton, J. (2009). The Knowledge Worker of the Future. *ACM*, 851–852.
8. Duggan, G. B., & Payne, S. J. (2008). Knowledge in the Head and on the Web : Using Topic Expertise to Aid Search. *ACM*, 39–48.

9. Ghose, A., & Yang, S. (2008). Analyzing Search Engine Advertising : Firm Behavior and Cross-Selling in Electronic Markets. *Proceedings of the 17th international conference on World Wide Web (WWW '08)*, 219–226. Retrieved from http://doi.acm.org/10.1145/1367497.1367528

10. Grappone, J., & Couzin, G. (2006). *Search Engine Optimization: An hour A day*. Wiley .

11. Gruhl, D., Guha, R., Kumar, R., Novak, J., & Tomkins, A. (2005). The predictive power of online chatter. *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (KDD '05)* (pp. 78–87). ACM Press. doi:10.1145/1081870.1081883

12. Guazzelli, A., Stathatos, K., & Zeller, M. (2009). Efficient deployment of predictive analytics through open standards and cloud computing. *ACM SIGKDD Explorations Newsletter*, *11*(1), 32. doi:10.1145/1656274.1656281

13. Hair, J. F. J. (2007). Knowledge creation in marketing: the role of predictive analytics. *European Business Review*, *19*(4), 303–315. doi:10.1108/09555340710760134

14. Huffman, S. B., & Hochster, M. (2007). How Well does Result Relevance Predict Session Satisfaction ? *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '07)*, 567–573. Retrieved from http://doi.acm.org/10.1145/1277741.1277839

15. Introna, L., & Nissenbaum, H. (2000). Defining the Web: the politics of search engines. *Computer*, *33*(1), 54–62. doi:10.1109/2.816269

16. Jones, K. B. (2008). *Search Engine Optimization* (Vol. Wiley Publ).

17. Moran, M., & Hunt, B. (2009). *Search Engine Marketing, Inc. Driving search traffic to your company's website* (Second edi.). IBM.

18. Nagar, Y., & Malone, T. W. (2011). Making business predictions by combining human and machine intelligence in prediction markets. *Proceedings of the 32nd International Conference of Information Systems (ICIS '11)* (pp. 1–16). Shanghai, China.

19. Oliner, A., Ganapathi, A., & Xu, W. (2012). Advances and challenges in log analysis. *Communications of the ACM 55*, *55*(2), 55–61. doi:10.1145/2076450.2076466

20. Punera, K., & Merugu, S. (2010). The Anatomy of a Click : Modeling User Behavior on Web Information Systems Categories and Subject Descriptors. *Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM '10)*, 989–998. Retrieved from http://doi.acm.org/10.1145/1871437.1871563

21. Raj, N., Dey, L., & Gaonkar, B. (2011). Expertise Prediction for Social Network Platforms to Encourage Knowledge Sharing. *IEEE/WIC/ACM*, 380–383. doi:10.1109/WI-IAT.2011.93

22. Ranjan, J., & Bhatnagar, V. (2011). Role of knowledge management and analytical CRM in business: data mining based framework. *The Learning Organization*, *18*(2), 131–148. doi:10.1108/09696471111103731

23. Regelson, M., & Fain, D. C. (2006). Predicting Click-Through Rate Using Keyword Clusters. *Proceedings of the Second Workshop on Sponsored Search Auctions*.

24. Richardson, M., Dominowska, E., & Ragno, R. (2007). Predicting Clicks : Estimating the Click-Through Rate for New Ads. *Proceedings of the 16th international conference on World Wide Web (WWW '07)*, 521–529.

25. Sculley, D., Malkin, R., Basu, S., & Bayardo, R. J. (2009). Predicting Bounce Rates in Sponsored Search Advertisements. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '09)*. Retrieved from http://doi.acm.org/10.1145/1557019.1557161

26. Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS Quarterly*, *35*(3), 553–572.

27. Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2012). Analyzing existing customers' websites to improve the customer acquisition process as well as the profitability prediction in B-to-B marketing. *Expert Systems with Applications*, *39*(3), 2597–2605. doi:10.1016/j.eswa.2011.08.115

28. Wen, J.-R., Nie, J.-Y., & Zhang, H.-J. (2002). Query clustering using user logs. *ACM Transactions on Information Systems*, *20*(1), 59–81. doi:10.1145/503104.503108

29. White, R. W., Dumais, S. T., & Teevan, J. (2009). Characterizing the influence of domain expertise on web search behavior. *ACM*, 132. doi:10.1145/1498759.1498819

30. Yang, Q., Va, C., & Zhang, H. H. (2001). Mining Web Logs for Prediction Models in WWW Caching and Prefetching. *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '01)*, 473–478. Retrieved from http://doi.acm.org/10.1145/502512.502584

31. Zhang, G., Zhou, F., Wang, F., & Luo, J. (2008). Knowledge Creation in Marketing Based on Data Mining. *International Conference on Intelligent Computation Technology and Automation (ICICTA)*, 782–786. doi:10.1109/ICICTA.2008.45

32. Zhang, X., Cole, M., Street, H., & Belkin, N. J. (2011a). Predicting Users ' Domain Knowledge from Search Behaviors. *ACM*, 2004–2005.

33. Zhang, X., Cole, M., Street, H., & Belkin, N. J. (2011b). Predicting Users ' Domain Knowledge from Search Behaviors. *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval (SIGIR '11)*, 2004–2005. Retrieved from http://doi.acm.org/10.1145/2009916.2010131

34. Zhu, Y., Wang, G., Yang, J., Wang, D., Yan, J., Hu, J., & Chen, Z. (2009). Optimizing search engine revenue in sponsored search. *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '09*, 588. doi:10.1145/1571941.1572042