

ANALYTICS ON BEHAVIOR OF USERS ON LARGE DATASETS

MONA GUPTA



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY DELHI
APRIL 2017

ANALYTICS ON BEHAVIOR OF USERS ON LARGE DATASETS

by

MONA GUPTA

Department of Computer Science and Engineering

Submitted

in fulfillment of the requirements of the degree of Doctor of Philosophy

to the



Indian Institute of Technology Delhi

APRIL 2017

Certificate

This is to certify that the thesis titled **Analytics on Behavior of Users on Large Datasets** being submitted by **Mona Gupta** for the award of **Doctor of Philosophy in Computer Science & Engg.** is a record of bona fide work carried out by her under my guidance and supervision at the Department of Computer Science & Engineering, Indian Institute of Technology Delhi. The work presented in this thesis has not been submitted elsewhere, either in part or full, for the award of any other degree or diploma.

Amitabha Bagchi
Associate Professor
Dept. of Computer Science & Engg.
Indian Institute of Technology Delhi.

Acknowledgements

It gives me immense pleasure to acknowledge several people who have contributed towards my dissertation.

I really thank my supervisor **Dr. Amitabha Bagchi** for giving me an opportunity to work under him and guiding me through out my research work. Without his motivation, support and guidance this work could not have reached this extent.

I also thank **Dr. Parag Singla** who gave me a chance to work under his supervision and provided continous guidance and support for doing my research.

I would also like to thank Rudra Mohan Tripathy, Happy Mittal, Mayank Srivastava, Tushar Dogra, and Akhil Jain with whom I learned a lot while working with them on various projects.

I would like to thank all my fellow students – Kanika Modi, Yashoteja Prabhu, Nisha Jain, Manoj Gupta for all the help, discussions and great company. I thank the staff of the Department of Computer Science and Engineering for providing access to the required computer resources in the lab.

I am grateful to my husband for all the support.

Mona Gupta

Abstract

In recent times, there has been a tremendous amount of increase in the number of users participating in social networks, online comparison shopping, using social media to retrieve/share information. So, it really becomes necessary to analyze the network characteristics, understanding the temporal behavior of users involved in them, and characterizing this behavior in various online platforms such that the output of these various platforms increases. The work in this thesis helps in magnifying the output of various platforms by employing the analytics on the behavior of the users participating in these platforms. We have analyzed four different large datasets and present analytics on the behavior of users and characterize them.

The first of our work analyzes the game of cricket using a complex network model. Our study examines Test cricket, One Day Internationals (ODIs) and T-20 cricket matches of the Indian Premier League (IPL). Each team is represented by a weighted network model. Different social network metrics are then applied on the created networks to get insight into the nature of social relationships between players of a team and the effect of these relationships over the performance of a team. We also looked at evolution of teams performances with respect to their network properties for every decade. Our findings suggest that, the average degree and average clustering coefficient follow similar trends as the performance of the team in Test cricket but not in One day cricket and T20. So partnership actually plays a more significant role in team performance in Test cricket as compared to One day cricket and T20 cricket.

In the next of our work, we delved into the temporal aspects of user behavior and relate it to the evolution of particular topics being discussed on the Twitter OSN. The hypothesis behind the work being that users deviate from their usual tweeting hours when a major event occurs. A metric called 'tweet strength' is introduced which gives more weight to tweets by users who in general tweet less or those who do not usually tweet at a given time. We also studied the evolution of a set of topics through the lens of tweet strength to identify the class of users driving the popularity of a topic at different times. We also study word-of-mouth diffusion mechanism through the network by defining a copying behavior. When a follower of a user tweets on the same topic as the user, the follower is said to have copied. We further make the definition time-dependent by imposing temporal thresholds and are able to capture the recency effect in the word of mouth mechanism of diffusion.

Since the number of users participating in online comparison shopping is increasing manifold, we wanted to study the behavior of users on an online comparison shopping website. For this, we characterized the session-level behavior of users on an Indian mobile phone comparison shopping website collected over a period of one year. There are three aspects to our study: data analysis, correlation between news sources of product information and popularity of a handset, and behavior

prediction. We have used KL divergence to show that a time-homogeneous Markov chain is observed when the number of clicks varies from 5 to 30. Our results depict that Markov Chain model does not hold in entirety for comparison shopping setting but tells us how far the Markov Chain model holds for this setting. Our analysis corroborates intuition that increasing price leads to decrease in popularity. After the strong correlation between various variables and user behavior was found, we predict the users macro (the overall sales of handset) and micro behavior (whether a user will convert or exit the site) using Markov Logic Networks. Our predictive model validates the intuition that past browsing behavior is an important predictor for future behavior. Methodology of combining data analysis with machine learning is, in our opinion, a new approach to the empirical study of such data sets.

Information spreads widely through conversations on an original post through retweets and mentions on Twitter. If a conversation is large then a user either has to go through all the tweets or while going through more recent tweets, one may miss important tweets to become a part of conversation. Thus, to increase the user engagement in a conversation, we want to show users the tweets which are more relevant in addition to being recent. For this, we are working on a Twitter dataset. We have analyzed the basic characteristics of Twitter conversations and have seen how they evolve with the size of the cascade. We are in the process of developing a function which can score different tweets in a cascade in order to be able to strengthen our heuristic that features such as branching may make a particular node more interesting than others.

सामान्य सारांश

हालही में, सामाजिक नेटवर्क, इंटरनेटपर तुलनात्मक खरीदारी, जानकारी देने के लिए और हासिल करने के लिए सामाजिक साधनों का प्रयोग, वगैरा में सहभागी होनेवाले सदस्यों की संख्या में बहुत ही जादा बढ़ोतरी हुई है। तो यह जरूरी हो गया है कि हम नेटवर्क की विशेषताओं का विश्लेषण करें, उनमें जुड़े वापरकर्ताओं के अस्थायी व्यवहार को समझे, इस व्यवहार की रूपरेखा को विविध प्लॅटफॉर्म (मंच) की श्रेणी में ढालें तांकि इन विभिन्न मंचों की उपज में बढ़ोतरी हो। इस शोध प्रबंध में किए गए काम से, इन मंचों में सहभागी वापरकर्ताओं के, व्यवहार की विश्लेषिकी पर काम करें, तांकि इन विभिन्न मंचों की उपज को और भी जादा बढ़ाया जा सके। हमने चार बड़े विवरणसेट का विश्लेषण किया है और वापरकर्ताओं के व्यवहार की विश्लेषिकी श्रेणीबद्ध करके सादर करते हैं।

हमारे पहला विश्लेषण का काम है, क्रिकेट का खेल जिसमें पेचिदा नेटवर्क प्रतिमान का प्रयोग किया गया है। हमारा अध्ययन पांच दिनों की क्रिकेट प्रतियोगिता, एक दिन की आंतरराष्ट्रीय क्रिकेट प्रतियोगिता (ओडीआय) और इंडियन प्रिमियर लीग (आयपीएल) के टी-२० क्रिकेट प्रतियोगिता का परीक्षण करता है। हरेक गिरोह का प्रतिनिधित्व प्रभावी नेटवर्क प्रतिमान ने किया है। फिर गिरोह के खिलाड़ियों के बीच के, सामाजिक रिश्ते को, अंदर से समझने के लिए और इस रिश्ते का परिणाम, गिरोह के प्रदर्शन में क्या होता है, यह देखने के लिए विभिन्न सामाजिक नेटवर्क साँचें, तैयार किए गए नेटवर्कपर लागू किए जातें हैं। हमने हर दशक के गिरोह प्रदर्शन के विकास को, उनके नेटवर्क लक्षणों के संदर्भ में भी देखा है। हमारे निष्कर्ष सूचित करते हैं कि, सामान्य स्तर और सामान्य समुह के साथ का कौशल्य, इसकी प्रवृत्ति ठीक, गिरोह के पांच दिनों की प्रतियोगिता के प्रदर्शन जैसी ही है लेकिन एक दिन की क्रिकेट प्रतियोगिता या टी-२० क्रिकेट प्रतियोगिता जैसी नहीं है।

हमारे अगले काम में, हमने वापरकर्ताओं के व्यवहार के अस्थायी पहलुओं को ट्वीटर ओएसएनपर चर्चा किए जानेवाले खास विषयों के विकास से जोडकर देखा। इस काम के पीछे प्रकल्पना है कि, जब कोई बड़ी घटना होती है तो वापरकर्ता उनके सामान्य ट्वीटींग घंटों में बदलाव लाते रहतें हैं। एक परिमाण है 'ट्वीट स्ट्रेंथ' जिसकी शुरुआत हुई है, जो उन ट्वीट्स को जादा मायने देता है, जिन्हे वापरकर्ता, जो साधारण तौर पर कम ट्वीट करतें हैं या वो जो एक तय समय में ट्वीट नहीं करतें हैं। हमने संग्रहित विषयों के विकास का भी ट्वीट स्ट्रेंथ को लागू कर, अध्ययन किया, यह पहचानने के लिए कि, विभिन्न समयपर, कौनसी वापरकर्ताओं की श्रेणी, विषय की लोकप्रियता

का नियमन करती है। हमने नकल उतारने के व्यवहार को परिभाषित करके नेटवर्क पर मौखिक प्रसार प्रणाली का भी अध्ययन किया। जब कोई वापरकर्ता का अनुगामी, वापरकर्ता के ही विषयपर ट्वीट करता है, तो अनुगामी ने नकल उतारी है ऐसा कहा जाता है। हमने आगे चलकर प्रभावशाली अस्थायी चौखट बिठाकर समय-निर्भरता की भी परिभाषा की है और हाल ही के मौखिक प्रसार प्रणाली में उसका परिणाम पकड़ में आ रहा है।

हालांकि इंटरनेट वापरकर्ताओं का सहभाग तुलनात्मक खरीदारी में काफी बढ़ गया है, तो हम चाहते थे कि इंटरनेट तुलनात्मक खरीदारी वेबसाइट के वापरकर्ताओं के व्यवहार का भी अध्ययन किया जाए। इसीलिए, हमने एक साल की, भारतीय मोबाइल फोन तुलनात्मक खरीदारी की वेबसाइट से, वापरकर्ताओं के सत्र-स्तर के व्यवहार की, जानकारी हासिल की है। हमारे अध्ययन के तीन पहलू हैं: विवरण का विश्लेषण, उत्पाद जानकारी के नए स्रोत का आपसी संबंध और हँडसेट की लोकप्रियता, और व्यवहार का पूर्वकथन। हमने केएल डायवर्जन्स का प्रयोग किया है यह दिखाने के लिए कि, समय-सदृश मार्कोव् चैन को तब देखा गया है जब क्लिक करने की संख्या ५ से ३० तक बदलती है। हमारे निष्कर्ष यह बताते हैं कि मार्कोव् चैन प्रतिमान, तुलनात्मक खरीदारी में पूरी तरह से खरा नहीं उतरता है लेकिन हमें यह जरूर बताता है कि इस व्यवस्था में मार्कोव् चैन प्रतिमान कहां तक खरा उतरता है। हमारा निष्कर्ष इस अंतः प्रज्ञा की पुष्टि करता है कि जब कीमतों में बढ़त होती है तो लोकप्रियता कम होती है। विभिन्न परिवर्तनीय और वापरकर्ता के व्यवहार के मजबूत सहसंबंध पता चलने के बाद, हम मार्कोव् लॉजिक नेटवर्क का इस्तेमाल करके वापरकर्ता के व्यवहार का अनुमान करते हैं कि मोटे (हँडसेट की कुल बिक्री) तौर पे और सूक्ष्म तौर पे (क्या वापरकर्ता खरीदेगा या साइट बंद कर देगा)। हमारा अनुमान प्रतिमान अंतः प्रज्ञा की पुष्टि करता है कि भूतकालीन ब्राउसिंग व्यवहार, भविष्यकालीन व्यवहार का अनुमान लगाने में महत्वपूर्ण है। विवरण विश्लेषण के साथ मशीनों के सीखने की जोड़ देने की कार्यप्रणाली, हमारी राय में, ऐसे विवरण समुह के आनुभविक अध्ययन का नया तरीका है।

ट्वीटरपर मूल पोस्ट के पुनःट्वीट्स और पुनरुच्चार की बातचीत से जानकारी विस्तार से फैल जाती है। अगर बातचीत लंबी है तो वापरकर्ता को सभी ट्वीट्स को पढ़ना पड़ता है या बिल्कुल हाल ही के ट्वीट्स को ही पढ़ने से, हो सकता है की कोई महत्वपूर्ण ट्वीट बातचीत से छुट जाए। तो वापरकर्ता को बातचीत में रखने के लिए, हम वापरकर्ता को दिखाना चाहते हैं वे ट्वीट्स जो जादा संबंधित है और हाल ही के है। इसलिए, हम ट्वीटर विवरण समुह पर काम कर रहे हैं। हमने ट्वीटर बातचीत के मूलभूत पहलूओं का विश्लेषण किया है और देखा है कि कैसे वे प्रपात में विकसित होती है। हम अब ऐसे फंक्शन का विकसन करने में लगे हुए हैं, जो प्रपात में विभिन्न ट्वीट्स की गणना करें ताकि हमारे अनुमान की पुष्टि करें जैसे कि फैलाने में एखाद खास बिंदू अन्य से जादा दिलचस्प बनें।

Contents

1	Introduction	1
2	Complex Network Characteristics and Team Performance in the Game of Cricket	5
2.1	Related Work	6
2.2	Dataset Description:	7
2.3	Introduction to Cricket	7
2.3.1	Motivation	8
2.4	Cricket : A Complex Network	9
2.5	Observation from Cricket Network	15

2.5.1	Performance Vs Network Characteristic	15
2.5.2	Formats of the Game Vs Network Characteristics	16
2.6	Contributions of the work	17
3	Temporal Analysis of User Behavior and Topic Evolution on Twitter	21
3.1	Related Work	23
3.2	Methodology	24
3.2.1	Dataset Description	24
3.2.2	Timezone Shifts	25
3.2.3	Tweet Strength	25
3.3	Topic evolution and user behavior	27
3.3.1	Evolution of Tweet Strength	27
3.3.2	Role of frequent users	29
3.4	Diffusion of a Topic	31

3.5	Contributions of the Work	34
-----	-------------------------------------	----

4 Analysis and Characterization of Comparison Shopping Behavior in the Mobile

Handset Domain		35
-----------------------	--	-----------

4.1	Basic Characterization	37
-----	----------------------------------	----

4.1.1	Dataset Description	37
-------	-------------------------------	----

4.1.2	Search Query Analysis	38
-------	---------------------------------	----

4.1.3	Time-based Characterization	39
-------	---------------------------------------	----

4.1.4	Location	40
-------	--------------------	----

4.1.5	Click to Buy (Conversion)	41
-------	-------------------------------------	----

4.1.6	Repeat Users	43
-------	------------------------	----

4.2	Price and Brands	45
-----	----------------------------	----

4.2.1	Analysis Based on Brands	45
-------	------------------------------------	----

4.2.2	Analysis Based on Price	47
-------	-----------------------------------	----

4.3	News Feeds	50
4.3.1	Characterization using KL Divergence	54
4.3.2	Learning State Transition Probabilities	56
4.4	Predicting Future Behavior	58
4.4.1	Choosing the Learning Model	59
4.4.2	Methodology for Predicting Micro Behavior	60
4.4.3	Methodology for Predicting Macro Behavior	60
4.4.4	Experiments	61
4.4.5	Contributions of the Work	64

5

Analyzing and Summarizing the Conversations on Twitter

67

5.1	Related Work	68
5.2	Dataset	69
5.3	Extraction of Conversations	70

5.4	Basic Characterization	71
5.5	Methodology for Future Work	73
6	Conclusions	77
	Publications	79
	References	81
	Biography	89

List of Figures

2.1	Performance of teams	8
2.2	Degree Distribution	10
2.3	Average weight as a function of end-point degrees for AUSTRALIA	11
2.4	Strength Distribution for Test Cricket	12
2.5	Average weighted clustering coefficient as a function of k	13
2.6	Average weighted nearest-neighbor degree as a function of k.	14
2.7	Strength vs. Degree	16
2.8	Comparison between real data and randomized weights for Australia	17
2.9	Time Evolution of Australia and West Indies in Test.	18
2.10	Time Evolution of Australia, West Indies and South Africa in One day	19
2.11	Time Evolution in IPL	20
3.1	Evolution of Median of the tweet strength distribution	28
3.2	Evolution of the mean of top 10% of tweet strengths	29
3.3	Evolution of the ratio of the largest strongly connected component to the second largest	30
3.4	Evolution of fraction of copy tweets	31
3.5	Rank ordered distribution of number of tweets on the topic per user	33
4.1	Month vs. No. of Sessions	40

4.2	Date vs. No. of Sessions	40
4.3	Day vs. No. of Sessions	41
4.4	Hour vs. No. of Sessions	41
4.5	Time Spent vs Hour	42
4.6	No. of Sessions vs Time Spent	42
4.7	No. of Sessions vs Time Spent	43
4.8	Day vs. No. of Distinct Converts	43
4.9	Date vs. No. of Distinct Converts	44
4.10	Hour vs. No. of Distinct Converts	44
4.11	Percentage of converts with varying time spent	45
4.12	Number of visits by the number of users	46
4.13	Session length with repeat number	47
4.14	Session length of convert users with repeat number	48
4.15	Percentage share in visits of each brand	49
4.16	Percentage share in converts of each brand	50
4.17	Convert percentage of various brands	51
4.18	Month-wise converts of various brands	52
4.19	Number of compares between various brands	53
4.20	Price offered by various brands	54
4.21	Price-wise converts of various brands	55
4.22	Variability in Price while comparing	56
4.23	Change in price and no. of converts for Sony Xperia	57
4.24	Change in price and no. of converts for 3 different phones	58
4.25	Micromax Power A96	59
4.26	Nokia X	60

4.27	Galaxy S5	61
4.28	Galaxy S5 derivative on Smartprix	62
4.29	KL divergence vs time step	62
5.1	Cascade Size Distribution	71
5.2	Cascade Depth Distribution	72
5.3	User Distribution	73
5.4	Cascade Depth Distribution	74
5.5	Hashtag Distribution	74
5.6	Ratio of lifetime of a Cascade	75

List of Tables

2.1	Details of Test Cricket Network	9
2.2	Details of One day Cricket Network	9
2.3	Details of IPL Cricket Network	9
3.1	Data set summary.	25
4.1	User and Session Statistics for Smartprix Dataset	38
4.2	No. of Price Decreases in Various Categories	51
4.3	Average increase in converts after price decrease	52
4.4	No. of Price Increases in Various Categories	53
4.5	Average increase in converts after price increase	54
4.6	User and Session Statistics for Smartprix Dataset	55
4.7	Markov Chain Probabilities	57
4.8	Task 1: User will convert in this session	63
4.9	Task 2: Exit within 3 clicks	63
4.10	Performance of different algorithms on two tasks	64
4.11	Task 3: Next day visits	64
5.1	Characteristics of Twitter Dataset	69