Learning Social Influence from Past Data

Edbert Eddie Puspito

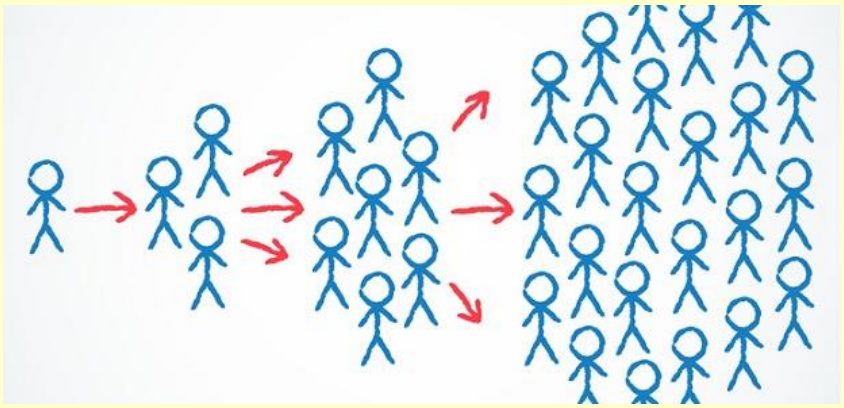
Advised by
Professor Raymond Chi-Wing Wong

Abstract

When a user performs an action in social networking websites, friends of this user may be influenced to perform the same or similar action. This phenomenon is called a social influence. Recent studies on data mining and machine learning try to model the phenomenon in social network websites and tried to answer the following question: if a group of Facebook users likes a photo, how many likes of this photo are there at the end?

This project proposes a new approach to calculate the probabilities of an influence of one social networking user over another user. Past research studies calculated the probabilities in the perspective of the user who performs the action. We shifted the paradigm and tried to calculate the probabilities in the perspective of the person who observed the actions.

We verified our ideas and techniques using the last.fm dataset consisting of a social graph with 83K nodes and 2M edges, together with an action log consisting of 110M actions. Experimental results showed that the new model performed better than previous models in correctly predicting the users that will perform the action.



Viral Marketing

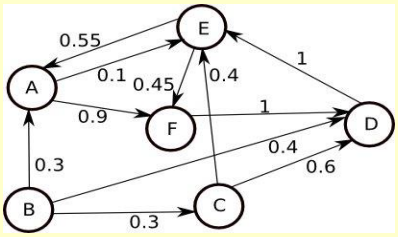
:-)

Idea

User	Action	Time
A	a1	5
A	a2	6
A	a3	9
B	a1	14

Data

Model



Network with influence probabilities

Aims

Find a more accurate model

Methodology

Observe

Learning Influence Probabilities In Social Networks

Amit Goyal
University of British Columbia
Vancouver, BC, Canada
goyal@cs.ubc.ca

Francesco Bonchi
Yahoo! Research
Barcelona, Spain
bonchi@yahoo-inc.com

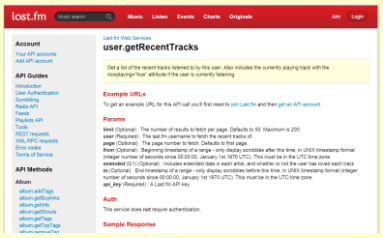
Laks V. S. Lakshmanan
University of British Columbia
Vancouver, BC, Canada
laks@cs.ubc.ca

ABSTRACT

Papers, papers and papers

Crawl

Last.fm Website



Last.fm API

We performed a literature surveys of existing influence probabilities models. We observed and tried to find flaws and weaknesses.

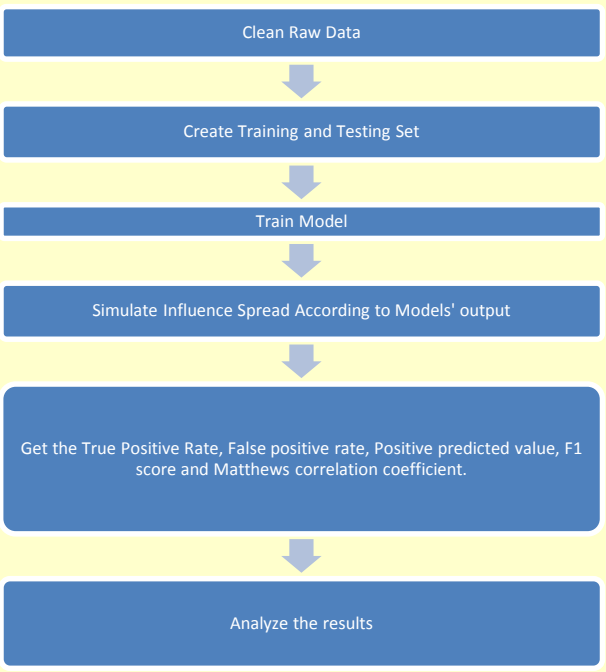
We chose Last.fm as the social networking website to work with. We observed the website to see how users interact with each other's and crawled the dataset using the provided API.

Design

Name	Formula (The probability of user a influencing user b on any action)
Existing models	
KISS	$P = \frac{\text{The number of times user } a \text{ successfully influenced user } b}{\text{The number of actions taken by user } a.}$
GOYAL	$P = \frac{\text{The total credit given to user } a \text{ by user } b \text{ for influencing him}}{\text{The number of actions taken by user } a.}$
Proposed models	
NEW	$P = \frac{\text{Total number of unique actions in which user } a \text{ influenced user } b}{\text{The number of unique actions done by } b}$
NEW_MOD	$P = \frac{\text{Total number of unique actions in which user } a \text{ influenced user } b}{\text{The number of unique actions that } a \text{ successfully influenced } b \; + \; \text{The number of unique actions done by } b, \text{but is not influenced from } a, \text{when user } a \text{ and user } b \text{ is active in the social network.}}$
NEW_MULTI	$P = \frac{\text{Total number of actions in which user } a \text{ influenced user } b}{\text{The number of actions done by } b}$
NEW_MOD_MULTI	$P = \frac{\text{Total number of actions in which user } a \text{ influenced user } b}{\text{The number of actions in which } a \text{ successfully influenced } b \; + \; \text{the number of actions done by } b, \text{but is not influenced from } a, \text{when user } a \text{ and user } b \text{ is active in the social network.}}$
NEW_WEIGHTED	$P = \frac{\text{Sum of total weighted actions in which user } a \text{ influenced user } b}{\text{Sum of total weighted actions in which } a \text{ successfully influenced } b \; + \; \text{The number of actions done by } b, \text{but is not influenced from } a, \text{when user } a \text{ and user } b \text{ is active in the social network.}}$
NEW_WEIGHTED_FULL	$P = \frac{\text{Sum of total weighted actions in which user } a \text{ influenced user } b}{\text{Sum of total weighted actions in which } a \text{ successfully influenced } b \; + \; \text{Sum of total weighted actions done by } b, \text{but is not influenced from } a, \text{when user } a \text{ and user } b \text{ is active in the social network.}}$

Models, models and models

Experiment



The flowchart

Based on the weaknesses we shift the calculation perspective. Existing models attempted to calculate the probabilities of influence from the perspective of the user who influences. But our new model attempts to calculate the probabilities of influence from the perspective of the user who is being influenced.

We created algorithms to train the models, simulating influence spread and evaluations.

Then we gathered and analyzed the result.

Result

Raw Data Statistics

	Raw Data	Training Set 1& 3	Training Set 2&4
Number of records	110,266,356	20,132,437	18,535,059
Number of users	83,526	1,670	7,947
Number of edges	2,151,749	7,884	81,404

Some of the experiment result

Set 1

Activation Threshold	0.1			0.25			0.5			0.75		
	PPV	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR
GOYAL	0.6401	0.0122	0.0001	1.0000	0.0000	0.0000	N/A	0.0000	0.0000	N/A	0.0000	0.0000
KISS	0.2145	0.1640	0.0073	0.3927	0.0475	0.0009	0.5992	0.0088	0.0001	0.7292	0.0000	0.0000
NEW	0.0673	0.3372	0.0570	0.1622	0.1404	0.0088	0.3184	0.0518	0.0014	0.4730	0.0001	0.0002
NEW_MOD	0.0626	0.3616	0.0661	0.1384	0.1568	0.0119	0.2356	0.0598	0.0024	0.3722	0.0002	0.0003
NEW_MULTI	0.0422	0.4717	0.1308	0.0781	0.2483	0.0358	0.1513	0.1230	0.0084	0.2415	0.0006	0.0020
NEW_MOD_MULTI	0.0267	0.4938	0.1422	0.0358	0.2745	0.0438	0.0543	0.1358	0.0112	0.1952	0.0008	0.0031
NEW_WEIGHTED	0.0142	0.9872	0.8348	0.0155	0.9627	0.7451	0.0174	0.8986	0.6198	0.0202	0.0176	0.4647
NEW_WEIGHTED_FULL	0.0267	0.7274	0.3238	0.0358	0.4949	0.1626	0.0543	0.3078	0.0654	0.0100	0.1882	0.0217

True Positive Rate, False positive rate, Positive predicted value of the models

Experiment set	1	2	3	4	Average
Models					
GOYAL	N/A	N/A	N/A	N/A	N/A
KISS	0.017	0.070	0.012	0.059	0.040
NEW	0.089	0.136	0.079	0.120	0.106
NEW_MOD	0.095	0.108	0.089	0.115	0.102
NEW_MULTI	0.136	0.144	0.129	0.124	0.133
NEW_MOD_MULTI	0.078	0.132	0.125	0.110	0.111
NEW_WEIGHTED	0.034	0.029	0.020	0.018	0.025
NEW_WEIGHTED_FULL	0.092	0.040	0.080	0.029	0.060

F1 Score of the models on activation rate of 0.5

Experiment set	1	2	3	4	Average
Models					
GOYAL	N/A	N/A	N/A	N/A	N/A
KISS	0.072	0.123	0.058	0.107	0.090
NEW	0.124	0.126	0.112	0.112	0.119
NEW_MOD	0.113	0.110	0.101	0.108	0.108
NEW_MULTI	0.127	0.139	0.122	0.130	0.130
NEW_MOD_MULTI	0.121	0.134	0.120	0.124	0.125
NEW_WEIGHTED	0.063	0.015	0.045	0.012	0.034
NEW_WEIGHTED_FULL	0.105	0.058	0.111	0.058	0.083

Matthews correlation coefficient of the models on activation rate of 0.5

Conclusion

In our project, we have proposed 6 new model variants for calculating probabilities between users and another connected users in social network and compared their performance with existing models.

Out of those 6 models, we found out that NEW, NEW_MOD, NEW_MULTI and NEW_MOD_MULTI model have better performance than existing models. The NEW_MULTI model also have the best performance among all models.

Some interesting points were raised during the project that may serve as the direction for future research.

1. The NEW_MULTI model have better performance than NEW_MOD_MULTI model. However, in the NEW_MOD_MULTI model, we does not count actions that happened before both user are active on the social network. In theory, it should lead to a more precise and accurate model, but experiment result showed otherwise.

2. From the experiments, it would not make sense if the influence from an action can last for more than 180 days in Last.fm. It seems that influence decay is affected by the nature of action and the model. It is probable that influence persist for more than 180 days.

3. When progressing with the project, we considered that the nature of actions may affected the influence. We used cosine similarity to find the similarity scores between users and influence probabilities. We found that there is a weak positive correlation. However, we are not able to find a method to integrate the users' similarity to the models.

4. We used various evaluation methods and felt that those methods may not be suitable in determining influence propagation models performance. As the F1 scores and MCC of the models were very low, at around 0.1 while the maximum score was 1. Also ROC curve did not account that GOYAL model have 0 true positive case on activation threshold of 0.5.