Optimal Group Composition for Crowdsourced Knowledge Building

Team Members

Utkarsh Katiyar	2016eeb1103
Arunaksha Talukdar	2016csb1032
Abhishek Singh	2016csb1028

Under the Guidance of

Sudarshan Iyengar

Anamika Chhabra

Introduction

We all know that the amount of knowledge that humans possess is gradually increasing as a whole but we don't quite understand yet the procedure and conditions that lead to the creation of new knowledge. A know-how of the same can exponentially increase the existing pace of building knowledge. Our biggest difficulty, while analyzing various aspects of process of knowledge building, is our inability to acquire the underlying data of this complex process. However, current time shows great promise of improvements in the knowledge building domain due to the availability of several online knowledge building portals. In this report, we emphasize that these portals act as prototypes for universal knowledge building process. The analysis of big data availed from these portals may equip the knowledge building researchers with the much needed meta-knowledge. Since getting real world data is difficult, we have taken virtual data from different online knowledge building environments like Wikipedia, Quora, Stack Exchange etc.

Project Objective

01 Our first objective is to acquire datasets from various knowledge building environments, say Q&A portals like Quora, StackExchange etc.

02 Then, we aim to analyze the acquired data through various perspectives to determine the conditions and parameters affecting knowledge creation on a portal.

03 Depending on our analysis, we aim to calculate by taking few necessary assumptions, the way users are triggering each other in a given environment.

04 Lastly, we use our analysis and knowledge of dynamics of a given environment to gradually generalize the applicability of our concept on a universal level.

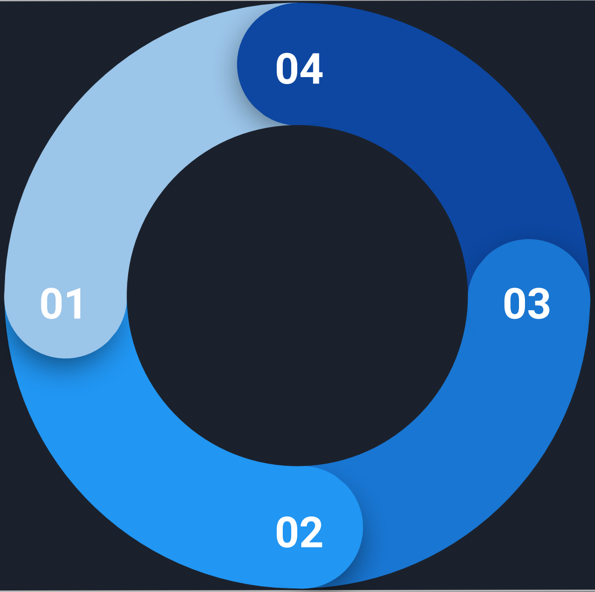
Assumptions

Closed System

Internal knowledge added at t=0 only.

Self Triggering

We assume the triggering within same categories is low.



Constant Triggering

Fixed triggering matrix average over a time

Continuity

Knowledge produced at time t is triggered by the knowledge produced at time t − 1

Modelling Our Approach

1. Calculate a diagonal N matrix with its elements being users from different categories using K-means algorithm.
2. Create a column matrix K of dimension 2x1 with its elements being no. of ques. asked and no. of ans. given respectively on a KB portal. K=[Q,A]
3. To take into account initial internal knowledge we use a column matrix R of dimension 2x1 with constant values of R= [0.01,0.01]
4. Using these values of N, R, K we calculate T matrix with the equation:

   N^-1 *(K)-R=T(K)
5. Once we know the amount of Triggering going on a website, we use this T matrix , with R and the required K matrix to calculate the new N matrix which will determine the required distribution on a website of various user across different categories.

Challenges

1. The current scenario is not suitable enough to efficiently perform the objectives of project at hand.
2. Some sites usually do not release their underlying datasets. And those that do, are usually too big in size and scattered over a number of files making it very cumbersome to be used properly.
3. The process may further get easier if sites start keeping a track of the ways in which one user triggers another user and so on. As this would help in better recognition of the dynamics being followed on a site and hence making the extrapolation easier.

Verification Of Model

1. Firstly, calculate Triggering matrix T from Result’s at time fractions of 90% and 100% of the span of total time for which the data has been calculated. (For instance, considering 32.4 months out of 36 months.).
2. Later we consider 100% of the data and using K_{100} and T_{90} We calculate $N_{1\,00}$ which comes out to very close to Actual Data.

Observation

1. The Category which Generates More Knowledge Trigger's other Categories more (Most Cases). Hence showing the Social Phenomena of Peer Pressure.
2. Even if there are more users in a particular category, it does not necessarily mean that major knowledge building will happen in that category only (For our case: Websites having Question asker's more still had total Knowledge generated coming from Answerer's category).
3. It's was observed that the Predicted value of Distribution of User's, calculated from data at 85% of final time came out to be almost same to the actual Distribution of User's at Final time.

Conclusion

1. In this project, we calculated the Triggering Matrix for a website and then used it to predict and verify the optimum distribution required to create the desired knowledge just as it was asked of us through the project.
2. With the help of our model, we can pace up the process of knowledge creation in an environment. We can also determine the kind of users needed by a particular KB portal or the kind of users that a KB portal should incentivize to maximize knowledge creation on their portal.
3. Our concluding remarks on the project are that this area offers a very good scope for explorations and advances should be made to better equip the knowledge building communities with more organized datasets.

Thank you !!