Optimal Group Composition for Crowdsourced Knowledge Building

Utkarsh Katiyar (2016eeb1103@iitrpr.ac.in)

Arunaksha Talukdar(2016csb1032@iitrpr.ac.in)

Abhishek Singh (2016csb1028@iitrpr.ac.in)

1. Abstract

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.

2. Introduction

The advancements in the Internet technologies have facilitated collaboration at an unprecedented scale by providing several crowdsourced portals. These portals are successfully accomplishing the job of collective problem solving, collaboratively accumulating facts and building knowledge on top of the existing information. Some of the examples are the online openly editable encyclopedia Wikipedia and Q&A portals such as Stack Exchange and Quora etc. We conjecture that these portals emulate various aspects of the universal knowledge building process.

For our project, we have taken datasets of Q&A portals under consideration. Q&A portals can help us understand other aspects of the phenomenon. Although these portals differ in their features, they all aim at building a repository of the world's knowledge by a collective effort of the crowd. Further, the diversity in the approaches employed by these crowdsourced mediums is advantageous in understanding different characteristics of the process. For instance, while Wikipedia may help in uncovering the aspects of co-ordination and conflicts while building repositories of knowledge, Q&A portals may assist in deciphering the importance of discussions and incentivization etc. Additionally, these portals are able to store each and every footprint of the process digitally, which is otherwise not possible to acquire.

With so many platforms enabling an unprecedented collaboration among people from every nook and cranny of the universe, we may get a huge amount of data for exploration. We, therefore, envision that proper analyses conducted on the big data taken from these portals may shed light on deciphering how new knowledge gets created. These analyses may further help in uncovering several other characteristics of collaborative knowledge creation facilitating unprecedented improvements in the domain. Therefore, we suggest use of these portals for understanding how knowledge building happens in general. The forthcoming sections provide the details of knowledge building domain along with the challenges associated with the realization of the proposed vision.

3. Sources of Knowledge Units in a Knowledge Building System :

The knowledge generated in the system can be classified on the basis of whether it is an outcome of group dynamics or not. The total knowledge can be divided into two categories:

(I) Internal Knowledge and

(II) Triggered Knowledge.

- Internal Knowledge: Internal knowledge is a subset of the user's knowledge which is added to the system independent of the effect of group dynamics. This is precisely the knowledge that the users would have added to the system if they had been participating in the knowledge building process individually (and not in a group).

- Triggered Knowledge: This is the kind of knowledge that gets added to the system as a result of the group dynamics. When people participate in the knowledge building

process as a group, they get triggered on seeing each others' ideas and hence, generate more knowledge.

3.1. Assumptions :

A knowledge building system is a complex system with a number of parameters. The current model primarily focuses on triggering among the knowledge units. Therefore, to keep the model simplistic as well as to highlight the phenomenon of triggering, it makes a few assumptions:

(1) All the users of the system are available at time t = 0.

(2) Closed System: The model assumes that the users of the system gain no knowledge from the outside environment. Therefore, the internal knowledge gets added to the system at time t = 0 only.

(3) Uni-specialist Assumption: Uni-specialist is a user who contributes in only one activity. On the other hand, a multi-specialist contributes in multiple activities. Given the ecosystem observation, the current model assumes every user to be a uni-specialist, although there may be a very small number of multi-specialists too.

(4) Constant Triggering: Triggering among users may depend on multiple factors, leading to varying triggering values over time. For simplicity, we take a fixed matrix capturing the triggering factors averaged over the entire duration of the process.

(5) Knowledge produced at time t is triggered by the knowledge produced at time t − 1 only. This is as per the triggering phenomenon. As new knowledge units come, they trigger other related units.

It should be noted that due to the complex nature of a knowledge building system, we explicitly control for a few parameters in the form of assumptions so that we can focus on explaining the dynamics of the triggering phenomenon through the model.

4. Modelling Our Approach:

We have used AADE architecture to go about solving the problem our project presented to us. AADE architecture consists of four phases :

(1). Acquisition(A) : where we acquire different kind of datasets from various knowledge building portals.

(2). Analyze(A) : where we analyze the acquired data through various means to look for patterns and various parameters affecting knowledge building portals.

(3). Dynamics(D) : Based on the annalysis, we try to understand the dynamics of various KB portals.

(4). Extrapolate(E) : In this phase, we extrapolate the insights obtained through analysis for their general applicability.

So, first off, we acquire datasets from various sites(Q&A portals mainly in our case) and using K-means algorithm we find out the clusters of questioners (people who mainly ask questions on the given site) and answerers (people who are more inclined to answering questions than asking questions). Using this data, we create a diagonal matrix N of dimension 2x2 where N_{11} represents the no. of questioners while N_{22} represents the no. of answerers.

Then, we create a column matrix K of order 2x1. K matrix signifies the current knowledge building happening in a site. That is, for a Q&A portal, K_{11} will represent the number of questions asked while K_{22} will represent the numbers of answers given.

To take internal knowledge into account, we create one another column matrix R of dimension 2x2 whose elements represent average internal knowledge of each category per user.

Now, consider a knowledge building system which allows *n* users to participate. Further, consider that there are *m* activities in the system that these users may participate in. We divide these *n* users into *m* categories, where the users in a category trigger the users in other categories with varying degrees given by a Triggering matrix (T) :

$$T = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}$$

in which diagonal elements represent self triggering i.e how much a questioner triggers a questioner while off diagonal elements represent cross triggering i.e, how much a questioner triggers a person who is more inclined towards answering.

Note: We consider self triggering to be 10^{-2} taking into account the fact that questioners will not trigger questioners as much as they will trigger answerers and vice versa. Now there can be two cases depending on the value of the spectral radius of NT , i.e. ρ(NT) which is equal to max{|λ1 |,. . . , |λe |} where λi 's are the eigenvalues of the matrix NT . If ρ(NT)≥ 1, the knowledge keeps on increasing exponentially with time and reaches infinity. However, if ρ(NT) ≤ 1, then initially the knowledge production is high, which keeps decreasing and ultimately converges. That's why, we get the following closed form for the total knowledge produced in the system.

$$Kc\ (\infty) = (I - NT)^{-1} N\ R$$

From datasets provided, we can easily calculate N and K matrix. Then, using the information provided to us by our mentor in her paper(cited paper mentioned in references), we formulate a set of mathematical equations as shown below :

$$K(t)=NTK(t-1) \quad \ldots(1)$$

$$\Rightarrow K(t)=NT^{T}K(0) \quad \ldots(2)$$

$$\Rightarrow K(t)=NT^{T}(NR) \quad \ldots(3)$$

For a large t, i.e t tending to ∞

$$K_i=(I-NT)^{-1}(NR) \quad \ldots(4)$$

We calculate T by putting values of N, R and K_i in the following equation:

$$N^{-1}(K_i)-R=T(K_i) \quad \ldots(5)$$

Once we have found T matrix, we get a fair idea of the kind of triggering that's going on in a website and to what extent one category of users is influencing other category.

Now that we have all the required variables in our hand, we can move on to finding the optimum distribution of users across various categories on the given site for the required knowledge creation.

Lets say you own a website which is basically a Q&A portal. Knowledge that has been build over a fairly long period of time on your site is given by matrix K_i. You know the numbers of users of different categories on your site i.e, N matrix and also know how users of one category trigger users of other category i.e, you know triggering matrix T as well. Now, you

aim to create more knowledge on your website given by column matrix K_f of dimension 2x1 where elements of K_f are greater than or equal to corresponding elements of K_i. A new variable square matrix N_f is taken into account which will store the number of users required across different categories on a site to create the knowledge envisioned or required by site owner. To find N_f, we feed the known values in equation (5) which becomes

$$K_f = N_f(T*K_f + R) \qquad \ldots(6)$$

Using equation (6), we calculate N_f i.e, the new number of users that should be present in different categories for the required knowledge creation. There's a possibility that total of number of users in N_f matrix come out to be greater than the total number of users existing currently on a website given by summation of elements in N matrix. So, as an owner, you now have two choices to go on with:

(1). You may try to increase users on your site in accordance with the newly obtained N_f matrix following approaches similar to targeted advertisements; or

(2). You may scale down the new number of users obtained to existing number and then distribute them across different categories as needed. For instance,

$$N = \begin{bmatrix} 40 & 0 \\ 0 & 60 \end{bmatrix}$$

$$N_f = \begin{bmatrix} 80 & 0 \\ 0 & 120 \end{bmatrix}$$

Then, scaled down matrix (N_s) :

$$N_s = \begin{bmatrix} 40 & 0 \\ 0 & 60 \end{bmatrix}$$

Thus, as an owner, you can try and distribute 40 percent users in questioners category and 60 percent of users in answerers category(for this particular case). Also, you get an idea of the scale by which you need to increase number of users across different categories in the long run for your desired goal.

4.1. Discussing Catalyst and Impediments:

The current scenario is not suitable enough to efficiently perform the objectives of the project at hand.

In acquisition phase, some sites usually do not release their underlying datasets. Also, the ones that do are usually too big in size and scattered over various files making it very hard to work with them and to analyze them properly. So, there's a need for organized datasets of various crowdsourced sites.

For analysis phase, sites may keep a track record of various users and their interactions to ensure better analysis. Privacy maybe a concern here but sites like Quora are doing it openly. They notify users like "X upvoted this answer after seeing your upvote". Such data, if made available for various processes, can help us generalize Triggering matrix to a larger extent and hence can lead us to better results. Similarly, in Dynamics and Extrapolation phase, we have to make many assumptions otherwise various real world non-quantifiable parameters come into picture making the problem at hand much more complicated.

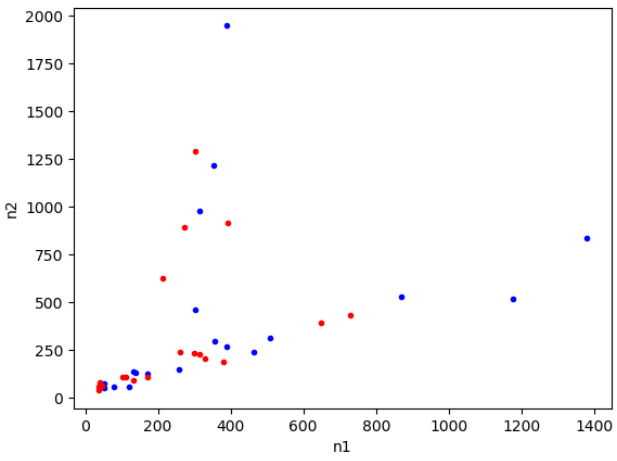
5. Verification of Model:

To verify our model, we calculated Triggering matrix T at two points of time t (say 90% and 100%). Time percentage represents the fraction of time we are applying our model to calculate T matrix and hence the final optimum distribution. Putting it simply, we first extract the 90% data from a given dataset i.e, we are considering 32.4 months of data if data has been recorded for a span of 36 months. Right now we have K_{90} and N_{90} using which we calculate T_{90} .

Later, we consider 100% of data i.e, we now have K_{100}, T_{90} and using these we calculate N_{100}.

We observed that calculated value of N_{100} comes out to be very close to actual value of N_{100} found through simple counting. Thus, verifying our prediction and calculation of optimum numbers of users required to create desired amount of knowledge.

Image below corroborates the same closeness between values of $N_{100}^{predicted}$ *(red dots) and* N_{100}^{actual} *(blue dots).*



6. Observations And Conclusions:

For our analysis, we took under consideration datasets of 20 websites having Q&A type knowledge building environment. For each dataset, we calculated the number of users and their inclinations towards various available categories using K-means algorithm. Then using the knowledge that had been built on that particular website for a long enough period of time, we calculated its Triggering matrix T using equations (3)-(6).

On calculating the Triggering matrix followed by prediction of ideal distribution of users in a knowledge building environment, we observed that our model is thoroughly followed by a site at different time stamps and for the sites with approximately same popularity and similar knowledge building period we get similar T matrix.

Moreover, it is observed that greater knowledge building in one of the categories usually triggers more knowledge creation in other categories as well thus implying the existence of a phenomenon similar to peer pressure in virtual world as well.

Another striking observation that came up through results is that even if there are more users in a particular category, it does not necessarily mean that major knowldege building will happen in that category only. (These observations can be corroborated by our findings uploaded on Github repository of our project under the folder Final_Results).

In this project, we were able to calculate 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. Our model shows how distribution of different kind of users on a website affects the process of knowledge building process and that if our model is followed iteratively on proper intervals and enforced as required by the administrators, then it'll definitely help in pacing up the process of knowledge building. This model can also be used to determine and target the kind of users needed by a website to catalyse knowledge building. It will also find its uses in deciding the incentivization schemes for a website.

We conclude the project with our findings and the thought that it may be a very complicated area considering all the real-time variables involved in the knowledge building processes but yet it offers scope for a lot of explorations when equipped with the right format of datasets and all the possible quantifiable parameters.

References:

1. Abstract and Introduction part inspired by the research paper on “Group Composition for Crowdsourced Knowledge Building: The Effect of User Interaction”, provided and written by Anamika Ma’am and S.R.S Iyenger Sir.
2. Datsets of various Q&A portals like Quora, Stack Exchange etc. provided by mentor.