http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (20): 5531-5535, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.5531.5535

Modeling and Simulation on Collective Intelligence in Future Internet-A Study of Wikipedia

Shiyu Du and Jiayin Qi
School of Economics and Management, Beijing University of Posts and Telecommunications,
Beijing, 100876, China

Abstract: Under the background of Web 2.0, network's socialization generates collective intelligence which can enrich human beings wisdom. However, what is the main factor that influences the performance of this behavior is still in research. In this study, the effect of number of Internet users that is represented by quantity, quality and variety of User-generated Content (UGC) is brought forward. Regarding Wikipedia as a study case, this study uses Agent-based modeling methodology and real data of Wikipedia for about 10 years to establish and simulate the model. The results verify that the size of group is indeed a necessary condition to generate collective intelligence. When the number of participants in Wikipedia reaches about 400000, the quantity of UGC increases exponentially, the quality of UGC reaches a satisfactory level and the variety of UGC can be guaranteed. This insight gives significance to show when mass collaboration will lead to collective intelligence which is an innovation than before.

Key words: Collective intelligence, agent-based modeling, swarm simulation, Wikipedia

INTRODUCTION

With the development of society's networking and network's socialization, the Internet is not only a public infrastructure, but also self-organized by human beings. New technology has brought a wave of User Generated Content (UGC) which causes nothing is too insignificant to observe online. Collective intelligence is a shared or group intelligence which is emerged from large numbers of individuals' cooperation and competition (Tapscott and Williams, 2008). It is a process of sharing knowledge and establishing consensus (Simpson and Weiner, 1989). Under specific conditions, self-organized users create and share resources for the same goal which urges the coordination of Internet resources (Boschetti, 2007). Meanwhile, it returns a controversy for whether our human knowledge will be enriched or threatened by this kind of collaboration online. Some believe that the mass collaboration online just generate seeming correct but not professional knowledge (Keen, 2007); While, another view is that the knowledge will be enriched with the increase of participation of mass and collective intelligence will dispense with the intervention of experts instead (Malone and Klein, 2007; Efthimios *et al*., 2012). They believe self-organized users' creating and sharing resources could urge the coordination of Internet resources better. However, there is no previous research on which size of group in mass collaboration could lead to collective intelligence.

Based on these, this study uses Agent-based modeling methodology and chooses Wikipedia which is a typical online mass collaboration application as a study case to explain: 1. Whether collective intelligence will appear due to the increasing size of the group. 2. At what level of size, collective intelligence will be achieved. 3. What is the trend of quantity, quality and variety of UGC in the process to generate collective intelligence.

THEORETICAL FOUNDATIONS

Till now, there are many relative researches focusing on the mechanism to generate collective intelligence. James Surowiecki (2005) stated that when the group is independent and diversified, the wisdom of crowd is amazing and even overmatches some individuals who have great intelligence. Williams and Tapscott (2006) found out three principles to generate great performance of mass collaboration: First is a collaboration goal; Second is attracting individuals to make contribution independently; Third is integrating small contributions together which will become an accumulation of great energy. Hayes and Malone (2010) illustrated that if a group has high degree homogeneity, then the production of them would be lack of variety. These researches gave a clear indication of the mechanism of collective intelligence at some degree. However, they are unilateral and lack of quantitative study for the factors.

Corresponding Author: Jiayin Qi, School of Economics and Management,
Beijing University of Posts and Telecommunications, Beijing, 100876, China

In this study, the mechanisms to generate collective intelligence in the Internet are defined as below:

First, the Internet users are independent and various; Second, a collaboration goal must exist and guide the individuals to generate useful UGC; Third, the dispersive contributions should be integrated together to form collective intelligence.

For the characteristics of collective intelligence, Lykourentzou *et al.* (2010) indicated that the quality and quantity of collaboration contents are two reasons to attract users to collaborate. And the difference between quantity and quality of the contents motivates different degrees of mass collaboration. At the same time, the increase of these two factors will promote the collective intelligence level. The study of Yaari *et al.* (2011) chooses the quality of UGC as one characteristic to indicate the success of Wikipedia. Moreover, the variety of UGC can also reflect the level of collective intelligence that mainly in the type of mass collaboration such as Wiki.

Therefore, this study is going to refer these ideas and use quantity, quality and variety of Wikipedia's UGC to reflect the performance of collective intelligence.

MODELING METHODOLOGY AND DATA COLLECTION

Agent-based modeling is a bottom-up modeling methodology which mainly contains three layers (Zhen and Cheng, 2009). The bottom is Agent layer. The middle is Individual-agent characteristic modeling layer. The top is Multi-agent System (MAS) layer.

Agent layer: It contains all the agents which reflect problem regions and system's responsibility. The processing method is to divide heterogeneous agents as corresponding agent classes and regards all homogenous agents as one agent class. In this model, each Internet user is supposed to be an agent.

Middle layer: Generally, this layer adopts a general model which is made up of 4 components: inner states, sensor, effector and environment. Inner states mainly define the attributes of agents. Each agent has a sensor to sense the environment in order to change its states according to the environment, these are called perceptions. Each agent also has an effector to effect the environment in order to change the states of the environment, these are called behaviors. Environment contains output variables of the system which are used to evaluate the performance of collective intelligence.

MAS layer defines interaction rules. This layer mainly solves 5 critical problems. First is the number of agents and agents' density in the Internet. Second is the communication channel between agents. The accessing method between agents is a combination of broadcast and unicast. Third is the communication protocol between agents. Here we design the communication protocol to be a combination of blackboard mechanism (global storage) and information delivery. Fourth is the structure between agents which designs the move and interaction principles. Fifth is the coordination between agents. Since agents are autonomous, when interaction happens, each agent decides its next step according to the previous defined rules.

For the experiment data, on one side, this research selects real data from January, 2001 to December, 2011 in Wikipedia's open platform. It includes changing data about the number of participants per month, increase quantity of entries per month, edit times per entry etc. On the other side, entries' marking system is also used to reflect the quality of an entry. In Wikipedia, it is made up of 4 parts: creditability, objectivity, completeness and readability. Full mark of each part is 5 and that is to say, total mark of an entry is 20 which reflect the highest quality of an entry. In this research, 105 randomly sampled Chinese entries in Wikipedia are used to reflect the relationship between the quality of an entry and edit times. Moreover, there are 11 different varieties of entries in Wikipedia platform. While at the same time, the educational level of participants has no obvious characteristics to differ.

Then we use software swarm for Java to simulate out the relationship between quantity, quality and variety of UGC with the number of agents respectively.

COLLECTIVE INTELLIGENCE MODEL

According to previous discussions, the assumptions and proposition for this model are defined as below:

- **Assumption 1:** Edits and modifications to an entry by Internet users are based on previous existed knowledge and quality of the entry. That is to say, for one entry, the more the edits times, the higher its quality is
- **Assumption 2:** The varieties of entries in Wiki have direct relationship with Internet users' varieties. The more the varieties of Internet users, the more varieties of entries will be
- **Hypothesis:** If the assumptions hold, then as the size of group increases, the higher the quantity, quality and variety of UGC will be. Moreover, the performance of collective intelligence would be better due to the increase of size which will attract more Internet users to participant into mass collaboration at some degree

Table 1: Fitting functions for number of agents

Time	Fit function	Coefficient	Degree of fitting
2001.1-2005.11	No. of agents = $\alpha \times \exp(\beta \times t)$	$\alpha = 0.3023$, $\beta = 0.2016$	0.9981
2006.1-2007.1	No. of agents = $\alpha \times t + \beta$	$\alpha = 39.82$, $\beta = -1082$	0.9974
2007.3-2011.11	No. of agents = $(\alpha - 0 \bullet 3 \bullet t) \bullet t + \beta$	$\alpha = 67.94$, $\beta = -1708$	0.9996

Table 2: Fitting functions for new UGC

Time	Fit function	Coefficient	Degree of fitting
2001.1-2007.1	Increase quantity of entries= $\alpha \times$ (no. of par.^β)+γ	$\alpha = 26990$, $\beta = 0.417$, $\gamma = -16330$	0.9493
2007.3-2011.11	No definite fitting function but except for some specific months, the values are drifting between 7000 and 10000		

Table 3: Fitting functions for edit times

Time	Fit function	Coefficient	Degree of fitting
2001.1-2005.1	Edit times = $\alpha \times t^{\wedge}\beta + \gamma$	$\alpha = 2.118$, $\beta = 0.4201$, $\gamma = 2.889$	0.9288
2005.3-2011.11	Edit times = $\alpha \times t^{\wedge}\beta + \gamma$	$\alpha = 1.139$, $\beta = 0.4653$, $\gamma = 7.39$	0.9993

Table 4: Fitting functions for UGC quality

Fit Function	Coefficient	Degree of Fitting
UGC quality = $(\alpha \times x^{\wedge}3 + \beta \times x^{\wedge}2 + \gamma \times x + \lambda) / (x + \mu)$	$\alpha = -4.96e\text{-}008$, $\beta = 0.0008735$, $\gamma = 14.02$, $\lambda = 245.9$, $\mu = 29.8$	with 95% confidence bounds

The model is designed as below:

- **Agent layer:** Since Wikis are self-organized and participating equally in generating and editing entries, due to the homogeneous principle, in this layer, the agents are designed just as one class: InternetUser
- **Individual-agent characteristic modeling layer**
  Inner states: Each agent has its own features which include the variety of generated UGC, individual's position in the Internet environment etc.,
- **Sensor:** In this model is viewing entries: viewUGC()
- **Effector:** In this model is generating entries and editing entries: generateUGC( ), editUGC( )
- **Environment:** In this model, it includes the three output variables: quality of entries (UGCquality), quantity of entries (generateUGCacc) and variety of entries (calculateVar)
- **MAS layer:** Using the real data in Wikipedia, we obtain following rules for interaction

First is to define the fitting functions for number of agents vs. time in the Internet as Table 1 shows.

Second is to define the fitting functions for increase quantity of entries vs. participants per month as Table 2 shows.

Third is to define the fitting functions for edit times vs. participants per entry as Table 3 shows.

Forth is to define the fitting functions for quality of an entry vs. edit times as Table 4 shows.

SIMULATION RESULTS

Through simulation, we can find the relationship between the size of group and quantity, quality and variety of entries.

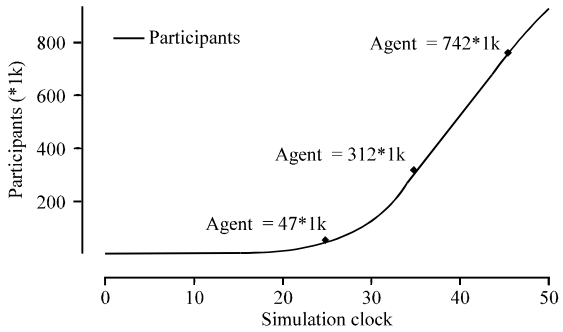

Fig. 1: Changing rate for number of agents

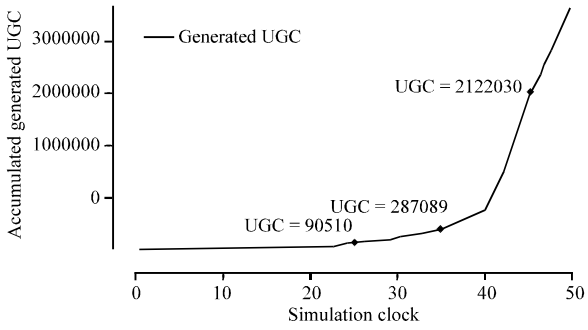

Fig. 2: Changing rate for accumulated number of UGC

Since the changes of Internet users are divided into three periods according to Table 1, here 3 simulation clock points (25s, 35s and 45s) are chosen to reflect the results which belong to each period respectively.

Fig. 1, 2 and 3 exhibit that, when simulation clock is 25 seconds, agent number is 47000, accumulated generated UGC is 90510 and quality of overall UGC is 14.89. When simulation clock is 35 seconds, agent number is 312000, accumulated generated UGC is 287089 and quality of overall UGC is 15.66. When simulation clock is

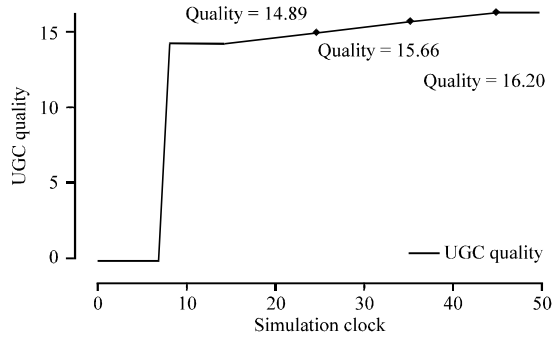

Fig. 3: Changing rate for quality of UGC

45 seconds, agent number is 742000, accumulated generated UGC is 2122030 and quality of overall UGC is 16.20. Moreover, from Swarm's raster and command lines in Eclipse, the results of variety of UGC can be obtained. We can find that, when simulation clock is 15s, agent = 6219, variety 2, 3, 7, 8, 10, 11 don't exist. When simulation clock is 25s, variety 10, 11 don't exist.

In conclusion, it is obvious to find that the quantity, quality and variety of UGC are increasing with the number of agents. When simulation clock is smaller than 20, the size of group is quite small. The quantity of UGC is also very low and increasing slowly. The number of participants is not enough to generate mass collaboration at this time. When the number of agents reaches about 400000, the quantity of accumulated generated UGC begins to increase exponentially. Collective intelligence starts to take into shape. So as for quality of UGC, when group size is quite small, comparatively the quality of UGC stays at about 14 which represent the intelligence of an individual. When the group size exceeds 1400000, the increasing rate is relatively less which shows the quality of UGC is approximately saturated. For the variety of UGC, when group size is quite small, for example, less and equal than 47000, the variety of UGC cannot be guaranteed which means collective intelligence has not been formed. With the increase of time and agents' number, we can find that all kinds of varieties of UGC exist.

If we suppose collective intelligence happens when the quality exceeds 16 (with 20 total marks, or 80% of all in another word), the group size is nearly 400000. The size is the same as the number of agents for quantity of UGC begins to increase exponentially and all varieties have existed. Therefore, it is confirmed that when number of Internet users exceeds 400000, mass collaboration has led to collective intelligence.

CONCLUSIONS

This model chooses mass collaboration of Web 2.0 Internet users as study point. For collective intelligence, the size of Internet users is an important factor influences this behavior. Using the study of Wikipedia, according to the model and simulation results, the performance of collective intelligence is definitely increasing with the increase of group size. When the number reaches about 400000, collective intelligence begins to take shape. At this time, the quantity of UGC increases exponentially. The quality of UGC begins to exceed 16 and all varieties of UGC have existed. Mass collaboration has led to collective intelligence. This result tells us, if we want to generate collective intelligence in order to manage the network resources better, administrators should try their best to attract enough participants to the Internet service. What's more, it shows that even though there is an existence of collective intelligence, it is still very hard to achieve 20 marks because this needs billions of users' participation and satisfies all kinds of people's tastes. When will best collective intelligence situation happen still needs further study.

ACKNOWLEDGEMENT

The study is supported by 973 Program (No. 2012CB315805, 2013CB329604) and National Natural Science Foundation of China (No. 71231002).

REFERENCES

Boschetti, F., 2007. Improving resource exploitation via collective intelligence by assessing agents' impact on the community outcome. Ecol. Econ., 63: 553-562.

Efthimios, B., A. Dimitris and M. Gregoris, 2012. Collective intelligence with web-based information aggregation markets: The role of market facilitation in idea management. Exp. Syst. Applic., 39: 1333-1345.

Hayes, T. and M.S. Malone, 2010. No Size Fits All: From Mass Marketing to Mass Handselling. China Machine Press, China.

Keen, A., 2007. The Cult of the Amateur: How Today's Internet is Destroying our Culture. Doubleday, New York, pp: 214-218.

Lykourentzou, I., K. Papadaki, D.J. Vergados, D. Polemi and V. Loumos, 2010. CorpWiki: A self-regulating wiki to promote corporate collective intelligence through expert peer matching. Inform. Sci., 180: 18-38.

Malone, T.W. and M. Klein, 2007. Harnessing collective intelligence to address global climate change. Innovations, 2: 15-26.

Simpson, J.A. and E.S.C. Weiner, 1989. The Oxford English Dictionary. Vol. 2, 2nd Edn., Clarendon Press, Oxford.

Surowiecki, J., 2005. The Wisdom of Crowds. Knopf Doubleday Publishing Group, USA., ISBN: 9780307275059, Pages: 296.

Tapscott, D. and A.D. Williams, 2008. Wikinomia: The Global Co-Operation which Changes Everything. Wydawnictwa Akademickie i Profesjonalne, Warszawa, ISBN: 9788360807231, Pages: 416.

Williams, A.D. and D. Tapscott, 2006. Wikinomics: How Mass Collaboration Changes Everything. Portfolio, New York, pp: 124-150.

Yaari, E., S. Baruchson-Arbib and J. Bar-Ilan, 2011. Information quality assessment of community generated content: A user study of Wikipedia. J. Inform. Sci., 37: 487-498.

Zhen, L. and Y.J. Cheng, 2009. Swarm for Java Simulation and Programming Implementation. China Machine Press, China.