Review of crowd computing as a sourcing Tool and an innovative field of data mining

Reena Hooda

Assistance Professor, CSE, IGU Meerpur Rewari (Haryana)-India

Abstract: Human population is the biggest source that is all time available and easily reachable at low costs. The major concern of data mining technologies is to find the interesting patterns and new scopes of applicability. Crowd computing involves elements like machine learning that is current application of artificial intelligence, electronic systems crowd sourcing and connectivity. By owning a system, internet facility, one can establish a full-fledged business using crowd a source to perform certain tasks without any requirements for further infrastructural facilities like buildings, machines etc. Crowd sourcing offers two notions, one is for performing tasks and other for providing data for knowledge discovery or a s training data set for computational systems. Keep in view all sides, present papers highlights the evolution of crowd computing, crowd sourcing, significance for data mining along with problems, and necessities of preprocessing methods.

Keywords: Crowd Sourcing, Training Data Set, Data Mining, Preprocessing.

I.   Introduction

Data Mining aims at exploring new facts by learning through training data sets; it is a knowledge discovery tool that comprises many statistical and advance learning methods to give precise solutions at minimum cost, less time while maintaining trustworthiness of the information. More availability of training data, more accurate will be the results. So crowd data as a training data enhance performance of data mining methodologies. As a source, crowd not only provides the training data however may also perform some basic computational tasks like data entries to minimize the data entry burden and speed up the operations. Or more precisely, crowd can perform innovative tasks based on their own imaginations. Thus it is advantageous to use this source form both perspectives. The term crowd computing sounds like cloud computing, initially introduced by Eric Brown in 2009. It is a biggest source of information, sharing of thoughts, overriding the professional hierarchies and usually done at all the times even in free times or for relaxing after a hard mental work. Crowd computing is countless knowledge exhibiting tool that combines crowd sourcing, automation, machine computing, human computing, and cloud computing plus personal communication skills. [2] It is that source which is more powerful than the Internet and is a new stream of research less known to seekers [2]. Crowd computing might help in developing new technologies, searching new spheres, getting new ideas from so called human power i.e. an extremely accessible source for problem solving [4]. It is the combination of human mind with artificial intelligence even realizing its power by the businesses for cutting cost, speedy performances with improved productivity. [1] Biggest example is the Wikipedia that took only 100 hrs. in its developing [5] without any need to hire people. Individuals are slow however, if their work is collaborated, can yield a huge amount of task and various solutions to a particular problem. In a simple way, in crowd computing, a complex problem is partitioned into several micro tasks to be performed by individuals in a small duration of time otherwise may take days, months or years to solve. Learning process is circular, every person has some specific knowledge which he learned or gained form surroundings, collaboratively resulted in to new technology development, solving biggest time consuming problems. Machine learns from the human inputs, their changed behavior in various circumstances that drives the automation ultimately facilitates the persons to smoothen their life styles. [5] Even many companies like Amazon or the Facebook rely on the crowd for their success that cooperate this crowd all

over the world. The term crowd computing available in literature only and less used practically, however it has a great future.

II. Historical Background and Applications

As documented, in 18th century, British Royal Astronomers took the help of crowd to create maps of stars and seas through the spreadsheets provided to them by mail. Another example of crowd computing is the translation of ancient documents discovered during 1800s in Egypt In 2011, by means of launching a website to speed up the translation. [1] Scanned books are digitalized by general public by typing while sign-up for a particular website or filling a form or downloading some application. [5] First wave of the crowd computing include the data labeling for example Mechanical Turk of Amazon, second include the human computation for example Artificial based applications. [6] As the name crowd, it covers all human lives, every sphere of life from eating & playing habits, interests & hobbies, house interiors, color preferences, eco-friendly conditions, approachable living areas, markets, customers etc. [7] All kind of information can be gathered while sitting at a place. Local advantage may be taken by getting skilled work and can only be performed by those local workers like handicrafts designs & patterns, local histories and beliefs. It is even useful for the manufacturers of the home made products with local printings, ideologies and filmmaker to get a theme for movie from these sources. An animation, fictions created by this source with new ideas is another good application and job opportunity.

III. Characteristics and Elements of Crowd Computing

Following may be the key features of crowd computing:

- Partitioning of total work into smaller chunks so that can be easily handled and managed by the crowd, also called micro tasking. [1]
- Partitioning requires the management skills and technical expertise so that optimal partitioning can be done that seems feasible and technically correct.
- Automation or the robotics to perform certain tasks or activities as programmed by humans. So artificial intelligence to merge human experiences, experts, machine computing etc. are comprised in this area. [1]
- Monitoring System to note and store the identified changes in normal behaviors. [1]
- Availability of target crowd that may comprise the users, specialist, developers, and data handlers.
- Re-engineering so validate the overall task and minimizing the deficiencies.
- Networking facilities, capacity to transfer data.
- Data warehouse to store and extract large heterogeneous data.
- Crowd sourcing: it is an outsourcing task by which a problem is solved by informal group of people. [3]
- Big data analysis is required with heterogeneous and gigantic valued features like different age groups, different financial status, education, knowledge level, geography, government regulations, local beliefs etc.
- Crowd Sourcer who gives the problem to general public or the group of experts to get the solution, for which incentives may be either monetary or non-monitory like prizes or any kind of recognitions. [3]
- Data preprocessing tasks and integration tools to incorporate different profiles and expertise [10] and keep their accuracy.
- Tools for handling spammers, unauthorized accesses and quality control [10].

- Interaction between humans (Crowd) and machine is must [9]. As crowd from different areas with different levels of knowledge so flexible interface is required.

Figure 1: Shows Crowd Computing, interactions between crowd and systems via internet [11--24]

IV. Advantages of Crowd Computing

Following are major advantages of Crowd Computing:

- With crowd sourcing, data can be collected from where Internet can reach; data computation can be performed from remote sites.
- It can be helpful in enhancing word dictionary by finding new local vocabularies. [8]
- All time available source, easy to get new ideas as crowd is involved.
- Enhance the group activities, participation and team work. Participation as a game playing or as a challenge increase the response ratio and performance.
- Sometimes no need to pay for work as done by game playing or hobby or entail little payments like small gifts etc.
- It may involve continuous work for long projects so provide job opportunities while working from home. Thus beneficial to unemployed crowd and for economic growth in tern[9].
- Diversifications of workers are good as sourcer might get local information along with the task assigned. [10]
- Better publicity among crowd as crowd is also a consumer.

V. Problems in Crowd Computing

- Selecting a particular crowd for a specific application is biggest issue. Further due Internet, sourcer cannot bar the number and type of crowd participation.
- Require the understanding of local language, customs and natural language processing tools [8].
- Bounding the participative crowd for a limited area or limited number may decrease the importance of training data set resulted in lack of accuracy, and overcrowded are again difficult to integrate.

- Sophisticated data preprocessing tools are required because the micro task may be biased due to individual involvements or group's interests. Besides this key token used may be local and difficult to comprehend as machine is not intelligent as man and there is limitations of algorithms because they work merely on the bases of predefined assumptions and failed in uneven situations.

- Other problem of the filling of missing data because task is distributed in diversified crowd. And if the data is not missing it may be noisy. [9] So data preprocessing algorithms may require deep learning and adequate model building to make the ETL (extract, transform and load) process successful.

- Data integration methods must be carefully selected while getting from remote sites.

- Big data analysis required availability of sophisticated tools for storage, interpretation and representation the data may prove costly however, open source tool may be employed.

- Some standard must be developed to get uniformity between micro tasks.

- Sampling method selection also a big problem as if sample is too small then it will loss the advantage of crowd with poor training data; if too big, it will not be optimal; or a new question –can sampling be successful in crowd computing- is hard to answer.

- Background knowledge is must to know the features of crowd for making good clusters.

- Training to crowd may be required to help them entering correct data as they play important role as a training data or entering large texts in particular format only.

- Problems of spammers and false profiles may mislead the distribution of tasks.

- Crowd is diversified and expert in different areas however machines are not, so a balance between the source and machine is must.

VI.   Conclusions

Crowd computing is best suited to overpopulated countries like India to get work at low costs. As every mind have some idea therefore crowd can contributed in better imaginations, innovations in science & technology, graphics etc., from basic to sophisticated jobs. Crowd is all time available, it is an easily accessible source of huge data and for performing tasks which may be accomplished in long period of time. There may be some problem of infrastructure however, with increased nets of technology, internet facilities, cheaper hardwares & softwares tools and cloud computing, future of crowd computing is bright. Constraint is that spammers, hackers, false profiles etc. must be tackled along with management of diversified inputs.

References

[1] https://en.wikipedia.org/wiki/Crowd_computing

[2] https://dl.acm.org/citation.cfm?id=2513470

[3] https://www.techopedia.com/definition/27816/crowdsourcing

[4] https://www.microsoft.com/en-us/research/video/crowd-computing/

[5] https://www.slideshare.net/CrowdComputingSystems/what-is-crowd-computing

[6] https://www.slideshare.net/mattlease/the-rise-of-crowd-computing-december-2015?next_slideshow=1

[7] Jing Gao [2016], "Mining Reliable Information from Passively and Actively Crowdsourced Data". ACM ISBN 978-1-4503-4232-2/16/08. Retrieved from: http://hanj.cs.illinois.edu/pdf/kdd16_jgao.pdf

[8] Sai R. Gouravajhala [n.d.], "Towards Crowd-Assisted Data Mining". Retrieved from : https://web.eecs.umich.edu/~wlasecki/pubs/crowd-analytics_hcml2016.pdf.

[9] http://crowdsourcing-class.org/

[10] Omar Alonso [2011], "Perspectives on Infrastructure for Crowdsourcing ". Crowdsourcing for Search and Data Mining (CSDM 2011). Retrieved from: http://ir.ischool.utexas.edu/csdm2011/proceedings/csdm2011_proceedings.pdf.

[11] https://community.powerbi.com/t5/Desktop/Columned-chart-not-showing-data-from-all-the-months/td-p/346884

[12] https://hms.harvard.edu/about-hms/graduation

[13] https://pngtree.com/freepng/running-lady_3344773.html

[14] https://es.pngtree.com/freepng/a-reporter-for-an-interview-with-a-lady_3335891.html

[15] http://clipartmag.com/cartoon-girl-drawing

[16] http://www.alamy.com/stock-photo-doctor-with-mustache-in-uniform-professional-cartoon-character-vector-142382266.html

[17] https://365psd.com/vector/silhouettes-of-businessman-32568

[18] http://tomtunguz.com/how-to-use-ml/

[19] https://www.sparkcognition.com/webinar-machine-learning-utility/

[20] https://www.igcseict.info/theory/1/types/index.html

[21] http://www.standard.net/National/2018/03/25/What-Facebook-s-privacy-policy-allows-may-surprise-you

[22] https://community.powerbi.com/t5/Desktop/Columned-chart-not-showing-data-from-all-the-months/td-p/346884

[23] http://www.rhythmsystems.com/blog/10-tips-to-make-the-most-of-your-networking

[24] https://www.corestack.io/blog/yang-for-the-impatient/