Swarm Intelligence in Big Data Analytics

Shi Cheng[1,⋆], Yuhui Shi[2], Quande Qin[3], and Ruibin Bai[1]

[1] Division of Computer Science, University of Nottingham Ningbo, China
[2] Department of Electrical & Electronic Engineering,
Xi'an Jiaotong-Liverpool University, Suzhou, China
[3] College of Management, Shenzhen University, Shenzhen, China
shi.cheng@nottingham.edu.cn, yuhui.shi@xjtlu.edu.cn

Abstract. This paper analyses the difficulty of big data analytics problems and the potential of swarm intelligence solving big data analytics problems. Nowadays, the big data analytics has attracted more and more attentions, which is required to manage immense amounts of data quickly. However, current researches mainly focus on the amount of data. In this paper, the other three properties of big data analytics, which include the high dimensionality of data, the dynamical change of data, and the multi-objective of problems, are discussed. Swarm intelligence, which works with a population of individuals, is a collection of nature-inspired searching techniques. It has effectively solved many large-scale, dynamical, and multi-objective problems. Based on the combination of swarm intelligence and data mining techniques, we can have better understanding of the big data analytics problems, and designing more effective algorithms to solve real-world big data analytics problems.

1 Introduction

Nowadays, the big data analytics has attracted attentions from more and more researchers. The big data is defined as the dataset whose size is beyond the processing ability of typical database or computers. Four objects are emphasized in the definition, which are capture, store, management, and analysis [26]. However, the definition only focuses on the amount of data. There are three other properties which also need to be emphasized in the big data analytics research.

The dimensionality of data affects the performance of algorithms. Many methods suffer from the "curse of dimensionality", which implies that their performance deteriorates quickly as the dimension of the search space increases [12,18]. The big data analytics also suffers this problem. Handling large scale data with a good performance in limited time should be concerned in the big data analytics.

The content of the big data is increasing over time, and the target of big data analytics also changes with time. The algorithm should be able to handle the dynamical changing data, and to adjust the target of data analytic.

⋆ The authors' work is partially supported by National Natural Science Foundation of China under Grant Number 60975080, 61273367; and by Ningbo Science & Technology Bureau (Science and Technology Project Number 2012B10055).

H. Yin et al. (Eds.): IDEAL 2013, LNCS 8206, pp. 417–426, 2013.

The data came from different sources in the large dataset. Generally, more than one objective needs to be satisfied at the same time in these large dataset. The most traditional methods can only be applied to continuous and differentiable functions, and have to perform a series of separate runs to satisfy different objectives.

Swarm intelligence is a set of search and optimization techniques [14,15,22]. To search a problem domain, a swarm intelligence algorithm processes a *population* of individuals. Different from traditional single-point based algorithms such as hill-climbing algorithms, each swarm intelligence algorithm is a population-based algorithm, which consists of a set of points (population of individuals). Each individual represents a potential solution of the problem being optimized. The population of individuals is expected to have high tendency to move towards better and better solution areas iteration over iteration through cooperation and/or competition among themselves.

In this paper, the difficulty of big data analytics problem is analysed. Big data analytics are divided into four components: handling large amount of data, handling high dimensional data, handling dynamical data, and multi-objective optimization. Most real world big data problems can be modelled as a large scale, dynamical, and multi-objective problems. Swarm intelligence has shown significant achievement on these problems. With the swarm intelligence, more effective methods can be designed and utilized in the big data analytical problems.

This paper is organized as follows. Section 2 and 3 review the basic concepts of big data analytics and the swarm intelligence methods, respectively. Swarm intelligence utilized in the big data analytics problems is introduced in Section 4. Two applications of big data analytics, the intelligent transport system and the wireless sensors networks, are brief reviewed in Section 5, followed by conclusions in Section 6.

2 Big Data Analytics

The big data is defined as the dataset whose size is beyond the processing ability of typical database or computers. Four objects are emphasized in the definition, which are capture, store, management, and analysis [26]. The big data analytics is to automatically extract knowledge from large amounts of data. It can be seen as mining or processing of massive data, and "useful" information could be retrieved from large dataset [29]. The properties of big data analytics can be concentrated in three parts: large volume, variety of different sources, and fast increasing speed, i.e., velocity. The algorithms should be effective to solve large-scale, dynamic big data analytics problems.

The knowledge discovery in databases (KDD) is the process of converting raw data into useful information. Data mining (the analysis step of KDD), is the process that attempts to discover useful information (or patterns) in large data repositories [16]. The data mining field includes many subfields, such as classification analysis, clustering analysis, and association analysis, just to name a few.

The big data may contain many kinds of unstructured or semi-structured data; these data need to be transformed as structured data. A kind of data's attribute will be transformed as a feature of data, and thus an example of data is transformed as a vector which contains many features. The dimension of the feature space is equal to the number of different attributes that can appear in the data set. This indicates that the dimension of big data analytics problems is much higher than the traditional problems.

The data clustering methods also can be applied to the swarm intelligence [32]. In the brain storm optimization algorithm, every solution is spread in the search space. The distribution of solutions can be utilized to reveal the landscapes of a problem. From the clustering analysis of solutions, the search results can be obtained [7].

3 Swarm Intelligence

Many real-world applications can be represented as optimization problems of which algorithms are required to have the capability to search for optimum. Most traditional methods can only be applied to continuous and differentiable functions [32]. The meta-heuristic algorithms are proposed to solve the problems, which the traditional methods cannot solve or at least be difficult to solve. Recently, kind of meta-heuristic algorithms, termed as swarm intelligence, are attracting more and more attentions from researchers.

Swarm intelligence (SI), which is based on a population of individuals, is a collection of nature-inspired searching techniques [22]. To search a problem domain, a swarm intelligence algorithm processes a *population* of individuals. Each individual represents a potential solution of the problem being optimized. In swarm intelligence, an algorithm maintains and successively improves a population of potential solutions until some stopping condition is met. The solutions are initialized randomly in the search space, and are guided toward the better and better areas through the interaction among solutions over iterations [5, 32].

As a general principle, the expected fitness of a solution returned should improve as the search method is given more computational resources in time and/or space. More desirable, in any single run, the quality of the solution returned by the method over iterations should improve monotonically – that is, the quality of the solution at time $t + 1$ should be no worse than the quality at time t, i.e., $fitness(t + 1) \leq fitness(t)$ for minimum problems [17]. There exist many swarm intelligence algorithms, among them most common ones are the particle swarm optimization (PSO) algorithm [21], which was originally designed for solving continuous optimization problems, and the ant colony optimization (ACO) algorithm, which was originally designed for discrete optimization problems [13].

The most important factor affecting a swarm intelligence algorithm's performance may be its ability of exploration and exploitation [6]. Exploration means the ability of a search algorithm to explore different areas of the search space in order to have high probability to find good promising solutions. Exploitation, on the other hand, means the ability to concentrate the search around a promising

region in order to refine a candidate solution. A good optimization algorithm should optimally balance the two conflicted objectives.

4 Swarm Intelligence in Big Data Analytics

Data mining has been a popular academic topic in computer science and statistics for decades, swarm intelligence is a relatively new subfield of computational intelligence (CI) which studies the collective intelligence in a group of simple individuals. In the swarm intelligence, useful information can be obtained from the competition and cooperation of individuals.

Generally, there are two kinds of approaches that apply swarm intelligence as data mining techniques [27]. The first category consists of techniques where individuals of a swarm move through a solution space and search for solution(s) for the data mining task. This is a search approach; the swarm intelligence is applied to optimize the data mining technique, e.g., the parameter tuning. In the second category, swarms move data instances that are placed on a low-dimensional feature space in order to come to a suitable clustering or low-dimensional mapping solution of the data. This is a data organizing approach; the swarm intelligence is directly applied to the data samples, e.g., dimensionality reduction of the data.

Swarm intelligence, especially particle swarm optimization or ant colony optimization algorithms, is utilized in data mining to solve single objective [1] and multi-objective problems [9]. Based on the two characters of particle swarm, the self-cognitive and social learning, the particle swarm has been utilized in data clustering techniques [10], document clustering, variable weighting in clustering high-dimensional data [25], semi-supervised learning based text categorization, and the Web data mining [28].

In a swarm intelligence algorithm, there are several solutions exist at the same time. The premature convergence may happen due to the solution getting clustered together too fast. However, the solution clustering is not always harmful for optimization. In a brain storm optimization algorithm, the clustering analysis is utilized to reveal the landscapes of problems and to guide the individuals to move toward the better and better areas [32]. Every individual in the brain storm optimization algorithm is not only a solution to the problem to be optimized, but also a data point to reveal the landscapes of the problem [7]. The machine learning and data mining techniques can be combined to produce benefits above and/or beyond what either method could achieve alone [31].

The big data analytics is required to manage immense amounts of data quickly [29]. The amount of data are attracting more and more attentions, however, the dimension of data and the number of objective of problems also increase the "hardness" of problems. Three kinds of difficulties should be overwhelmed to solve big data problems:

4.1 Large Scale Optimization

The big data analytics requires a fast mining on the large scale dataset, i.e., the immense amounts of data should be processed in a limited time. The analytic

problem can be modelled as optimization problems. In general, optimization concerns with finding the "best available" solution(s) for a given problem within allowable time, and the problem may have several or numerous optimum solutions, of which many are local optimal solutions. Normally, the difficulty of problem will increase with the increasing of the number of variables and objectives. Specially, problems with large number of variables, e.g., more than thousands variables, are termed as large scale problems.

Many optimization methods suffer from the "curse of dimensionality", which implies that their performance deteriorates quickly as the dimension of the search space increases [2,11,18]. There are several reasons that cause this phenomenon.

First, the solution space of a problem often increases exponentially with the problem dimension and more efficient search strategies are required to explore all promising regions within a given time budget. The evolutionary computation or swarm intelligence is based on the interaction of a group of solutions. The promising regions or the landscape of problems are very difficult to reveal by small solution samples (compared with the number of all feasible solutions).

The "empty space phenomenon" gives an example of problems getting hard when the dimension increases [30]. The number of possible solutions is increased exponentially when the dimension increasing. The search performance of most algorithms is based on the previous search experience. Considered the limitation of computational resources, the percentages of data points have been retrieved will close to zero when the dimension increased to a large number. The performance of algorithms is affected by the increasing of problems' dimension.

Second, the characteristics of a problem may change with the scale. Problems will become more difficult and complex when the dimension increases. Rosenbrock's function, for instance, is unimodal for two dimensional problems but becomes multimodal for higher dimensional problems. Because of such a worsening of the features of an optimization problem resulting from an increase in scale, a previously successful search strategy may no longer be capable of finding an optimal solution.

Third, the direction of "good" solutions is difficult to determine. The swarm intelligence takes an "update on each dimension, evaluation on whole dimensions" strategy. An algorithm is very difficult to determine which one is better when two solutions both have some good parts and their fitness values are equally bad. The similar scenario also happens in multiobjective optimization. In Pareto domination measurement, nearly all solutions are Pareto non-dominated when the number of objects is larger than 10.

The last, the bias is accumulated. In the swarm intelligence, each solution is updated dimension by dimension, and the fitness value is calculated for the whole solution. The solution update depends on the combination of several vectors, i.e., the current value, the difference between current value and previous best value, the differential between current value and neighbor best value, or the difference between two random solutions, etc. In the low dimensional space, the direction of the vector combination has the high probability to point to the global optimum. However, the distance metric, which is utilized in low dimension space, is not

effective in high dimensional space. The search direction is far away from the global optimum due to the bias accumulation.

Many effective strategies are proposed for high dimensional optimization problems, such as problem decomposition and subcomponents cooperation, parameter adaptation, surrogate-based fitness evaluations [20]. Based on the swarm intelligence, an effective method could find good solutions for large scale problems, both on the time complexity and result accuracy.

4.2 Handling High Dimensional Data

The "curse of dimensionality" also happens on the high dimensional data mining problems [11, 12, 18]. Many algorithms' performance deteriorates quickly as the dimension of the data space increases. For example, the nearest neighbor approaches are very effective in categorization. However, for high dimensional data, it is very difficult to solve the similarity search problem due to the computational complexity, which was caused by the increase of dimensionality.

Many methods are proposed on the high dimension data mining problems. Transforming the high dimensional mining problems into low dimensional space via a "projection" operation is an effective way. The locality sensitive hashing algorithm is proposed to find nearest neighbors in the high dimensional space [33]. This algorithm is based on hashing functions with strong "local-sensitivity" in order to retrieve nearest neighbors in a Euclidean space with a complexity sublinear in the amount of data.

The data mining problem can be transformed as an optimization problem, because many researches have been taken on the large scale optimization problems. Swarm intelligence, especially particle swarm optimization or ant colony optimization algorithms, is utilized in data mining to solve single objective [1] and multiobjective problems [9].

4.3 Handling Dynamical Data

The big data, such as the web usage data of Internet, real time traffic information, rapidly changes over time. The analytical algorithms need to process these data swiftly. The dynamic problems, sometimes termed as non-stationary environments, or uncertain environments [19], dynamically change over time. Swarm intelligence has been widely applied to solve stationary and dynamical optimization problems.

Swarm intelligence often has to solve optimization problems in the presence of a wide range of uncertainties. Generally, uncertainties in optimized problems can be divided into the following categories.

1. The fitness function or the processed data is noisy.
2. The design variables and/or the environmental parameters may change after optimization, and the quality of the obtained optimal solution should be robust against environmental changes or deviations from the optimal point.

3. The fitness function is approximated [20], such as surrogate-based fitness evaluations, which means that the fitness function suffers from approximation errors.
4. The optimum in the problem space may change over time. The algorithm should be able to track the optimum continuously.
5. The target of optimization may change over time. The demand of optimization may adjust to the dynamical environment, for example, there should be a balance between the computing efficiency and the computational cost for different computing loads.

In all these cases, additional measures must be taken so that swarm intelligence algorithms are still able to solve satisfactorily dynamic problems [19].

4.4 Multi-objective Optimization

Different sources of data are integrated in the big data research, and in most of the big data analytics problems, more than one objective need to be satisfied at the same time. According to the number of objectives, optimization problems can be divided as single objective and multiobjective problems. For the multiobjective problems, the traditional mathematical programming techniques have to perform a series of separate runs to satisfy different objectives.

Multiobjective Optimization refers to optimization problems that involve two or more objectives, and a set of solutions is sought instead of one. A general *multiobjective optimization problem* can be described as a vector function $\mathbf{f}$ that maps a tuple of n parameters (decision variables) to a tuple of k objectives.

Unlike the single objective optimization, the multiobjective problems have many or infinite solutions [3]. The optimization goal of an MOP consists of three objectives:

1. The distance of the resulting nondominated solutions to the true optimal Pareto front should be minimized;
2. A good (in most cases uniform) distribution of the obtained solutions is desirable;
3. The spread of the obtained nondominated solutions should be maximized, i.e., for each objective a wide range of values should be covered by the nondominated solutions.

In a multiobjective optimization problem, we aim to find the set of optimal tradeoff solutions known as the Pareto optimal set. Pareto optimality is defined with respect to the concept of nondominated points in the objective space. Swarm intelligence methods can effectively solve the multiobjective problems. Several new techniques are combined in the swarm intelligence techniques to solve multiobjective problems with more than ten objectives, in which almost every solution is Pareto nondominated in the problems. These techniques include objective decomposition, objective reduction [4], and clustering in the objective space [32].

5 Applications

The big data is created in many fields in everyday life. With the big data analytics techniques and swarm intelligence methods, more effective applications or systems can be designed to solve real world problems. The intelligent transportation system and wireless sensor networks are two typical examples of big data analytics application.

5.1 Intelligent Transportation System

The traffic problems are arising in many cities now. The traffic and transportation system is affected by many factors, such as the number of vehicles, weather, accidents, etc., and the traffic information changes in real time. The purpose of intelligent transport is to build more rapid, safe, and more efficient traffic and transportation systems by constructing the intelligent vehicles and road environment [34]. There are more than one objectives which need to be satisfied at the same time in intelligent transport systems, for example, rapid transportation, environmental pollution, transportation scheduling; and many of these objectives are conflicted with each other.

5.2 Wireless Sensor Networks

Based on the wireless sensor networks, the physical world is turning to be a kind of information system [8]. Different sensors are connected to form a networks; information is transferred in this network by communication techniques. The physical world's information from sensor networks can be collected almost anywhere at any time. The sensor networks and communication techniques have constructed a new paradigm, which is called the internet of things [8].

The wireless sensor networks have been applied to many real-world problems, such as environmental surveillance, transportation monitoring, engineering surveying, and industrial control, just to name a few [24]. Massive data will be generated from the long term and/or large scale wireless sensor network system. The goal of data analysis is to make the fastest possible revelation toward the "useful" information. Swarm intelligence is an effective way to handle these data, and to obtain "useful" information [23].

6 Conclusions

The big data has attracted more and more attentions currently. Most of the big data researches focus on the huge amount of data, however, handling the high dimensional data and the multiple objectives are also important in solving big data problems.

In this paper, the difficulty of big data analytics problem is analysed. Big data analytics are divided into four components: handling large amount of data, handling high dimensional data, handling dynamical data, and multi-objective

optimization. Most real world big data problems can be modelled as a large scale, dynamical, and multi-objective problems.

This paper is not to survey what have been done in the past, but to suggest the potential of swarm intelligence in big data analytics. Big data involves high-dimensional problems and a large amount of data. Swarm intelligence studies the collective behaviours in a group of individuals. It has shown significant achievements on solving large scale, dynamical, and multi-objective problems. With the application of the swarm intelligence, more rapid and effective methods can be designed to solve big data analytics problems.

References

1. Abraham, A., Grosan, C., Ramos, V. (eds.): Swarm Intelligence in Data Mining. SCI, vol. 34. Springer, Heidelberg (2006)
2. Bellman, R.: Adaptive Control Processes: A guided Tour. Princeton University Press, Princeton (1961)
3. Bosman, P.A.N., Thierens, D.: The balance between proximity and diversity in multiobjective evolutionary algorithms. IEEE Transactions on Evolutionary Computation 7(2), 174–188 (2003)
4. Brockhoff, D., Zitzler, E.: Objective reduction in evolutionary multiobjective optimization: Theory and applications. Evolutionary Computation 17(2), 135–166 (2009)
5. Cheng, S.: Population Diversity in Particle Swarm Optimization: Definition, Observation, Control, and Application. Ph.D. thesis, Department of Electrical Engineering and Electronics, University of Liverpool (May 2013)
6. Cheng, S., Shi, Y., Qin, Q.: Population diversity of particle swarm optimizer solving single and multi-objective problems. International Journal of Swarm Intelligence Research (IJSIR) 3(4), 23–60 (2012)
7. Cheng, S., Shi, Y., Qin, Q., Gao, S.: Solution clustering analysis in brain storm optimization algorithm. In: Proceedings of The 2013 IEEE Symposium on Swarm Intelligence (SIS 2013), pp. 111–118. IEEE, Singapore (2013)
8. Chui, M., Löffler, M., Roberts, R.: The internet of things. McKinsey Quarterly 2, 1–9 (2010)
9. Coello, C.A.C., Dehuri, S., Ghosh, S. (eds.): Swarm Intelligence for Multi-objective Problems in Data Mining. SCI, vol. 242. Springer, Heidelberg (2009)
10. Cohen, S.C.M., de Castro, L.N.: Data clustering with particle swarms. In: Proceedings of the 2006 IEEE Congress on Evolutionary Computations (CEC 2006), pp. 1792–1798 (July 2006)
11. Domingos, P.: A few useful things to know about machine learning. Communications of the ACM 55(10), 78–87 (2012)
12. Donoho, D.L.: Aide-Memoire. High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality. Tech. rep., Stanford University (August 2000)
13. Dorigo, M., Maniezzo, V., Colorni, A.: Ant system: optimization by a colony of cooperating agents. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 26(1), 29–41 (1996)
14. Dorigo, M., Stützle, T.: Ant Colony Optimization. MIT Press (June 2004)
15. Eberhart, R., Shi, Y.: Computational Intelligence: Concepts to Implementations, 1st edn. Morgan Kaufmann Publisher (2007)

16. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI Magazine 17(3), 37–54 (1996)
17. Ficici, S.G.: Monotonic solution concepts in coevolution. In: Genetic and Evolutionary Computation Conference (GECCO 2005), pp. 499–506 (June 2005)
18. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer Series in Statistics. Springer (February 2009)
19. Jin, Y., Branke, J.: Evolutionary Optimization in Uncertain Environments – A Survey. IEEE Transactions on Evolutionary Computation 9(3), 303–317 (2005)
20. Jin, Y., Sendhoff, B.: A systems approach to evolutionary multiobjective structural optimization and beyond. IEEE Computational Intelligence Magazine 4(3), 62–76 (2009)
21. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of IEEE International Conference on Neural Networks (ICNN), pp. 1942–1948 (1995)
22. Kennedy, J., Eberhart, R., Shi, Y.: Swarm Intelligence. Morgan Kaufmann Publisher (2001)
23. Kulkarni, R.V., Venayagamoorthy, G.K.: Particle swarm optimization in wireless-sensor networks: A brief survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 41(2), 262–267 (2011)
24. Liu, Y., Zhou, G., Zhao, J., Dai, G., Li, X.Y., Gu, M., Ma, H., Mo, L., He, Y., Wang, J., Li, M., Liu, K., Dong, W., Xi, W.: Long-term large-scale sensing in the forest: recent advances and future directions of greenorbs. Frontiers of Computer Science in China 4(3), 334–338 (2010)
25. Lu, Y., Wang, S., Li, S., Zhou, C.: Particle swarm optimizer for variable weighting in clustering high-dimensional data. Machine Learning 82(1), 43–70 (2011)
26. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H.: Big data: The next frontier for innovation, competition, and productivity. Tech. rep., McKinsey Global Institute (May 2011)
27. Martens, D., Baesens, B., Fawcett, T.: Editorial survey: swarm intelligence for data mining. Machine Learning 82(1), 1–42 (2011)
28. Pal, S.K., Talwar, V., Mitra, P.: Web mining in soft computing framework: Relevance, state of the art and future directions. IEEE Transactions on Neural Networks 13(5), 1163–1177 (2002)
29. Rajaraman, A., Leskovec, J., Ullman, J.D.: Mining of Massive Datasets. Cambridge University Press (July 2012)
30. Scott, D.W., Thompson, J.R.: Probability density estimation in higher dimensions. In: Gentle, J.E. (ed.) Computer Science and Statistics: Proceedings of the Fifteenth Symposium on the Interface, pp. 173–179 (1983)
31. Sheppard, J.W., Salzberg, S.L.: A Teaching Strategy for Memory-Based Control. Artificial Intelligence Review 11(1-5), 343–370 (1997)
32. Shi, Y.: An optimization algorithm based on brainstorming process. International Journal of Swarm Intelligence Research (IJSIR) 2(4), 35–62 (2011)
33. Slaney, M., Casey, M.: Locality-sensitive hashing for finding nearest neighbors. IEEE Signal Processing Magazine 25(2), 128–131 (2008)
34. Teodorović, D.: Transport modeling by multi-agent systems: A swarm intelligence approach. Transportation Planning and Technology 26(4), 289–312 (2003)