Research Article

Distilling Big Data: Refining Quality Information in the Era of Yottabytes

Sivaraman Ramachandramurthy,¹ Srinivasan Subramaniam,² and Chandrasekeran Ramasamy³

¹Anna University, Regional Office, Madurai 625007, India
²Department of Computer Science and Engineering, Anna University, Regional Office, Madurai 625007, India
³Department of Computer Science and Engineering, Annamalai University, Chidambaram 608002, India

Correspondence should be addressed to Sivaraman Ramachandramurthy; rsiva.raman@yahoo.com

Received 15 June 2015; Revised 1 August 2015; Accepted 9 August 2015

Academic Editor: Venkatesh Jaganathan

Copyright © 2015 Sivaraman Ramachandramurthy et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Big Data is the buzzword of the modern century. With the invasion of pervasive computing, we live in a data-centric environment, where we always leave a track of data related to our day-to-day activities. Be it a visit to a shopping mall or hospital or surfing Internet, we create voluminous data related to credit card transactions, user details, location information, and so on. These trails of data simply define an individual and form the backbone for user-profiling. With the mobile phones and their easy access to online social networks on the go, sensor data such as geo-taggings and events and sentiments around them contribute to the already overwhelming data containers. With reductions in the cost of storage and computational devices and with increasing proliferation of Cloud, we never felt any constraints in storing or processing such data. Eventually we end up having several exabytes of data and analysing them for their usefulness has introduced new frontiers of research. Effective distillation of these data is the need of the hour to improve the veracity of the Big Data. This research targets the utilization of the Fuzzy Bayesian process model to improve the quality of information in Big Data.

1. Introduction

Big Data systems are the eventual outcomes of today’s data centric world. In this 21st century, we cannot imagine passing a single day without the touch of Internet. In fact the World Wide Web, online social networks, and extensive computerization of our everyday activities have put forth new phenomena in managing the data thrown out by these processes. By virtue of its multidimensional nature, comprising volume, velocity, variety, variability, and veracity (5Vs), Big Data poses serious research issues and its complexity is always in the increasing trend. The beauty of Big Data is the fact that it incorporates all types of data, not just pictures, texts, sounds, emails, and so forth, but also the inputs and outputs from sensors of any kind. Also these data can be either unstructured type arising out of social media transactions or even multistructural type which is the outcome of interactions between people and machines. The architecture of Big Data is different from the data warehousing solutions, as this architecture should cater to the 5Vs and should produce meaningful results [1].

According to a report by Forbes, the seed for the term Big Data was first sowed on the Communications of the ACM by Steve Bryson in the year 1999. This paper has highlighted the fact that advances in powerful computing threw out huge volumes of data and understanding from these data is a significant endeavour. Since then, the growth of Big Data occurred in a rapid manner with the increase in computing power and reduction in the cost of storage devices. Earlier, data capacity in the order of gigabytes (GB) looked big, but now terabytes (TBs) have become so common such that even desktop computers sport hard disk drives of 2TB. According
to a recent study [2], the digital universe is doubling in its size every two years and the amount of data we generate and copy annually will become 44 trillion gigabytes or 44 zettabytes by 2020. More than 2 billion people and several million enterprises are doing their works online, thereby generating mass volumes of data. A single celeb-selfie tweet by Ellen DeGeneres who hosted the Oscars of 2014 was viewed 26 million times across the world within a short span of 12 hours, which shows the power of a single tweet to generate enormous data. In 2013, only 22% of the entire information processed in the digital universe was found to be useful if it has been properly tagged and a meagre 5% contained quality data. By 2020, the useful percentage of data may grow up to 35%, due to the contribution of data from embedded devices.

Parallel developments in the arena of Internet of Things (IoT) will accelerate the growth of Big Data with the sensors in mobile phones and other handheld devices pumping their own contributions [3]. New data sources leading to novel methodologies for data distilling and applying the outcomes to business processes so as to reap profits have become the cliché of successful businesses across the world. In short, smart work has outpaced the concept of hard work and proper data distillation techniques are needed to extract quality valuable data from the data dumps.

The people of today’s Internet age have become more and more conscious about the markets. It has become increasingly tough to sell products to them and conventional marketing methodologies have become obsolete. In the financial sector, decision making has been revolutionized. Predictive analysis has entered into new dimensions about the functioning of stock markets and can provide a detailed insight about various investment portfolios. On the healthcare front, Big Data can help you arrive at tailor made solutions based on your past medical history. The new path breaking strategies in human genome mapping with Big Data tools, wherein genes are being mapped to the diseases as per the earlier medical records, can help doctors pinpoint the actual medicine to be administered.

2. Social Sensing

Social sensing is the new technology of human centric sensing, wherein the sensors are attached to the humans in a wearable form or use the mobile devices like cellphones which carry the sensors like GPS, camera, and so forth. As these mobile devices are capable of automatically connecting to the Internet, the data sensed by them can be accessed in no time across the connected applications and devices. Numerous applications like Cabsense, Bikenet, and Cenwits use the social sensing for their operations. In the field of healthcare, this concept has created wonders in continuously monitoring patient’s vital data, emergency response, prediction about diseases, lifestyle effect on health, and so forth. Social sensing can be either voluntary/intentional or involuntary/unintentional. The former involves the participation from the user to identify the symptoms and causes whereas the latter does not have any user interaction and can happen without the knowledge of the user. Examples of user assisted sensing include marking the routes with more traffic and spotting the litter. The unassisted social sensing utilizes the built-in GPS of the hand held smart devices to automatically geo-tag the user and can use accelerometer to detect motion, and so forth. Reporting of natural calamities and civil unrest, on Twitter, Facebook, and YouTube, provides a global forum for sharing data on real time basis. Several research attempts are being made to understand the functioning of these online social sensing applications in propagation of information, characteristics, and tipping points.

3. Literature Review

With rapid decrease in the storage costs and dramatic increase in the computational capabilities, the size of the data is no longer a heinous factor and the quality of the data under study has become the primary concern. In one study by IBM, it was found that US $3.1 trillion was lost every day due to poor data quality and the survey has shown that 27% of the respondents were not sure about their own data. Since majority of the data come from anonymous unverified and untagged sources, the quality of the data needs to be flagged before it is used in any task [4]. Until recently, the Big Data has been characterized only by 3Vs, namely, volume, velocity, and variety, and the content looks more like a soup with the content providers scooping out and dumping them into information systems [5]. But with the introduction of data quality as a key factor, the fourth dimension of Big Data, namely, veracity, came into the picture which focuses on Information Quality (IQ) [6]. However big the quantity of data is, it will not be of any use if it does not have any meaningful extractable information and this research aims at providing a solution to the problem of data quality. Uncertainty in the data leads to unreliable information. This issue of uncertainty may be broadly classified into three heads, namely, process uncertainty, data uncertainty, and model uncertainty. The examples for process uncertainty are uncertain travel times in a GPS based tracking environment, yield of semiconductors, and so forth, and the data uncertainty involves inconsistency, ambiguity, incompleteness, conflicting data, and so forth. Model uncertainty comes into picture in the event of prediction and forecasting services where all models are approximations only. The data quality can generally be grouped under several categories like accuracy, believability, reputation, objectivity, factuality, consistency, freedom from bias, correctness, and unambiguousness. But with the introduction of uncertainty, all these factors have to be revisited to evolve a perfect working solution [7–11].

Figure 1 shows the 4Vs of Big Data as proposed by IBM [11]. Meeker and Hong [12] have discussed the opportunities and challenges in the Big Data related to the warranty and insurance claims in a typical industrial environment. With the technological advancements, leading to the implantation of sensors within the products, we can easily measure various parameters like use rate, system load, and environmental factors from these embedded sensors. The concepts represented in this research refer to the Internet of Things and the data are referred to as the SOE (System Operation/Environmental)
Data and they can be well used to analyse the reliability issues. Locomotive engines, aircraft engines, power distribution transformers, CT scan devices, wind turbines, and solar energy power invertors are some examples where such sensor data are available to predict the reliability of the system.

Uddin et al. [13] extensively analysed the social sensing and have developed algorithms for improving the source selection in social networks like Twitter, Flicker, and so forth. These social networks provide unprecedented opportunities to develop sensor applications where humans either act as sensors or play as sensor operators by posting the observations or measurements in an online medium. As the observations posted by one user are available to other users, any wrong or faulty information by one user may pass on to others since there is no mechanism to verify its correctness. Baseless rumours that will mislead the facts may also propagate in these online networks thereby creating bias. Uddin et al. have found that social sensing brings in new type of inaccuracy called unknown dependence between the sources which affects the creditability of data. They have adopted the concept of diversifying the sources to extract quality information. However, their technique used the follow relationships in the Twitter as a method of measurement and does not focus on the geographic locations and community they belong to in the process of assessment.

Maximum likelihood estimation (MLE) method [14] was adopted to find the accuracy of the participants and this technique utilizes the Cramer-Rao bound for estimation purposes. This method provides an insight into participants reliability without knowing their reliability a priori. Fisher information was initially computed by using the asymptotic normality of MLE. Though satisfactory results were obtained, a tighter Cramer–Rao bound can be derived to measure the truthfulness of the variable.

Aggarwal and Abdelzaher [15] discussed in detail the social sensing applications in the web like Google latitude, Citysense, Macrosense, Wikitude, Green GPS, and so forth. With the development of miniaturized sensor technology, sensors can even be embedded in the cloth fabric and the user can wear the attire without any obstruction or protrusion. 3G/4G mobile wireless networks have virtually bridged the bandwidth performance between wired and wireless networks. Online real time stream processing systems have also been developed to fetch and analyse the sensor data on the fly. But these real time systems bring in several issues like data compression, challenges in handling multiple data streams, data collection in mobile devices, and so forth. Also privacy and trust factors always collide with each other as one demands data fidelity and the other denies it. With the inroads created by Big Data, certain new dimension has also been put forth by Mishra and Sharma [16] by introducing the concept of value and an in-depth study has been made in the significant developments in the area of national development, industrial upgrades, interdisciplinary research, and better perception of the present scenarios of any industry. An interesting work carried out by Najafabadi et al. [17] has focussed on the importance of applying deep learning methods and evolving strategies to yield better results in the analysis of Big Data. They have introduced the concept of feeding a part of the available data corpus into deep learning algorithms and utilizing the remaining input data corpus for pattern recognition to extract data abstractions and representations. Thus the variety characteristic of Big Data, which focuses on the varied input data types for training the data representations and the target data source for generalizing the representations, becomes a problem of domain adaptation in deep learning. Their research has also explored the selection of criteria to provide useful semantic meaning.

The veracity of a Big Data environment depends on various factors, which include trustworthiness and expertise. These two factors are responsible for providing credible judgments about the quality of data. Objectivity, truthfulness, and credibility have been designated as the three main theoretical dimensions of veracity [18]. Various Big Data problems can be characterized by these dimensions based on quality.
4. System Design

One of the critical questions in the research of social sensing is about the reliability of the data collected by the humans. The process of contributing information is open and we never had any filtering mechanism to identify the reliability of the humans who act as sensors. Trust analysis of social sensing data is still an underdeveloped area that needs new techniques to improve reliability and enhance quality of information. This paper aims at creating a suitable framework to measure the reliability of data sensing sources and also to study the exactness of their responses. This approach is based on the Fuzzy Bayesian Interpretation model [19–23] where the data is processed on the server side using software only and no specialized hardware is required. Though this paper has addressed the explosion of data and has covered in depth the size of zettabytes and yottabytes, this research has utilized the simulation of an information network comprising “n” number of sources and assertions. A network model based approach can be visualised for solving the reliability issues, with the nodes representing the unreliable information and the links representing the abstract relations. Let there be sources $S_1, S_2, S_3, \ldots, S_s$, who collectively affirm information $C_1, C_2, C_3, \ldots, C_c$. If these sources and assertions are made to represent the nodes, then the links between the corresponding source nodes and the assertion nodes are represented by $C_{ij}$. Then the creditability of the sources is given by $C_{\text{cred}}(S_i)$ and that of the assertion given by $C_{\text{cred}}(C_j)$.

5. Methodology

5.1. Bayes Interpretation. We propose to use the Bayesian interpretation to rank the various sources providing information based on the creditability values. The statement of Bayes’ theorem is

$$p(h \mid e) \propto p(e \mid h) p(h),$$

where $p(h \mid e)$ and $p(e \mid h)$ are conditional probabilities. For example, if $p(e \mid h)$ is the probability of $e$ based on the condition of $h$, according to Bayes rule, whenever an evidence of $e$ is received, $p(h)$ has to be updated by $p(h \mid e)$. We have to multiply the prior probability $p(h)$ by the prior likelihood $p(e \mid h)$ followed by the normalization such that all probabilities sum to 1. The resulting posterior probability $p(h \mid e)$ is the revised probability assignment for $h$. Therefore, the new probability of $h$ is proportional to its original probability, multiplied by the likelihood of evidence $e$ given $h$ [24].

A Bayesian model can well be explained by using the following template and it forecasts an outcome through four factors, namely, prior probabilities, prior likelihoods, sensory input, and the utility function. This shows that the proposed model is deterministic and it has been depicted in Figure 2. However certain sensory malfunctions, internal noise leading to corruption can be the basis for exceptions. Certain models also replace the expected utility maximisation with probability matching leading to a nondeterministic process.

A Bayesian interpretation proposed by Dong Wang et al. [25] involves the evaluation using multiple sources $S_1, S_2, \ldots, S_j$ with assertions $C_j$. Then according to the Bayes theorem,

$$P(C_j \mid S_1C_j, S_2C_j, \ldots, S_kC_j)$$

$$= \frac{P(S_jC_j, S_2C_j, \ldots, S_kC_j \mid C_j)}{P(S_1C_j, S_2C_j, \ldots, S_kC_j)} P(C_j).$$

Here an implicit assumption is made that the probability of a source making any assertion in low and no change in posterior probability can be made from the nonavailability of claim. This shows that only existing claims are considered. Assuming, for any given assertion, the probability that any two sources claiming the same status are independent, then the equation can be rewritten as follows:

$$P(C_j \mid S_1C_j, S_2C_j, \ldots, S_kC_j)$$

$$= \frac{P(S_jC_j \mid C_j)}{P(S_1C_j, S_2C_j, \ldots, S_kC_j)} P(C_j).$$

If the change in posterior probability is found to be small and this case arises when evidence is collected from several unreliable sources,

$$\frac{P(S_kC_j \mid C_j)}{P(S_kC_j)} = 1 + \delta^f_{ikj},$$

Similarly,

$$\frac{P(S_kC_j \mid C_j)}{P(S_kC_j)} = 1 + \delta^f_{ikj},$$

where

$$\left|1 + \delta^f_{ikj}\right| \ll 1,$$

$$\left|1 + \delta^f_{ikj}\right| \ll 1.$$

Hence the equation can be rewritten as

$$P(S_1C_j, S_2C_j, \ldots, S_kC_j) \approx \prod_{k=1}^{K_j} P(S_kC_j).$$
Substituting the same in (2) we get
\[
P(C_j^t | S_n C_j, S_{n+1} C_j, \ldots, S_k C_j) = \frac{P(S_n C_j | C_j^t) \ldots P(S_k C_j | C_j^t)}{P(S_n C_j) \ldots P(S_k C_j)} P(C_j^t),
\]
and rewriting it gives the following:
\[
P(C_j^t | S_1 C_j, S_2 C_j, \ldots, S_k C_j) = \frac{P(S_1 C_j | C_j^t) \ldots P(S_k C_j | C_j^t)}{P(S_1 C_j) \ldots P(S_k C_j)} \times P(C_j^t).
\]
Substituting from (7),
\[
P(C_j^t | S_1 C_j, S_2 C_j, \ldots, S_k C_j) = P(C_j^t) \left(1 + \sum_{k=1}^{K_j} \delta_{kj}^t \right).
\]
As higher orders of \(\delta_{kj}^t\) can be neglected,
\[
\delta_{kj}^t = \frac{P(S_k C_j | C_j^t) - P(S_k C_j)}{P(S_k C_j)}.
\]

5.2. Creditability of Sources. Let \(S_i\) be the source of information making claims \(\text{Claims}_i\), and \(J_k\) denotes the \(k\)th claim of \(\text{Claims}_i\) and \(|\text{Claims}_i| = L_i\). By the Bayes theorem,
\[
P(S_i^t | S_i C_{j1}, S_i C_{j2}, \ldots, S_i C_{jL}) = \frac{P(S_i C_{j1}, S_i C_{j2}, \ldots, S_i C_{jL} | S_i^t)}{P(S_i C_{j1}, S_i C_{j2}, \ldots, S_i C_{jL})} P(S_i^t).
\]
As earlier assuming the conditional independence,
\[
P(S_i^t | S_i C_{j1}, S_i C_{j2}, \ldots, S_i C_{jL}) = \frac{P(S_i C_{j1} | S_i^t) \ldots P(S_i C_{jL} | S_i^t)}{P(S_i C_{j1}, S_i C_{j2}, \ldots, S_i C_{jL})} P(S_i^t).
\]
Since the change in posterior probability arising out of any single claim is negligible,
\[
\frac{P(S_i C_{jk} | S_i^t)}{P(S_i C_{jk})} = 1 + \eta_{ijk}^t,
\]
where \(|\eta_{ijk}^t| \ll 1\).

Hence,
\[
P(S_i^t | S_i C_{j1}, S_i C_{j2}, \ldots, S_i C_{jL}) - P(S_i^t) = \sum_{k=1}^{Li} \eta_{ijk}^t,
\]
\[
\eta_{ijk}^t = \frac{P(S_i C_{jk} | S_i^t) - P(S_i C_{jk})}{P(S_i C_{jk})}.
\]

5.3. Iterative Algorithm. In the preceding sections, posterior probability was derived based on the assumption that either an assertion is true or the source is truthful and they were derived from \(\delta_{kj}^t\) and \(\eta_{ijk}^t\):
\[
P(S_i C_{j} | C_j^t) = \frac{P(S_i C_{j} | C_j^t)}{P(C_j^t)},
\]
where
\[
P(S_i C_{j} | C_j^t) = P(S_i \text{ speaks}) \cdot P(S_i \text{ Claims } C_j | S_i \text{ speaks}) \cdot P(C_j^t \text{ speaks}, S_i \text{ Claims } C_j).
\]
Since the ground truth is unknown, the probability is estimated by considering the best information we have which is \(P(S_i^t | S_i C_{j1}, S_i C_{j2}, \ldots, S_i C_{jL})\). Thus,
\[
P(S_i C_{j}, C_j^t) = P(S_i \text{ speaks}) \cdot P(S_i^t | S_i C_{j1}, S_i C_{j2}, \ldots, S_i C_{jL}) \cdot C_{\text{true}}
\]
Substituting this into the above equations we get
\[
P(S_i C_{j}, C_j^t) = \frac{P(S_i \text{ speaks}) P(S_i^t | S_i C_{j1}, S_i C_{j2}, \ldots, S_i C_{jL})}{C_{\text{true}}}
\]
Similarly,
\[
P(S_i C_{j}) = \frac{P(S_i \text{ speaks})}{C_{\text{true}}}
\]
Therefore by rearranging we get
\[
\delta_{kj}^t = \frac{P(S_k C_{j} | C_j^t) - P(S_k C_{j})}{P(S_k C_{j})},
\]
\[
\delta_{kj}^t = \frac{P(S_i C_{jk} | S_i C_{j1}, S_i C_{j2}, \ldots, S_i C_{jL}) - 1}{C_{\text{true}}/c}
\]
If we consider the fractions of all true assertions to the true assertions total as prior probability, for a source to be truthful, \(P(S_i^t)\), then the above equation can be rewritten as
\[
\delta_{kj}^t = \frac{P(S_i^t | S_i C_{j1}, S_i C_{j2}, \ldots, S_i C_{jL}) - 1}{P(S_i^t)}.
\]
Substituting $\delta_{ij}^k$ we get
\[
P\left( C_j | S_{ij}C_j, S_{i2}C_j, \ldots, S_{ik}C_j \right) - P\left( C_j \right)
\]
\[
\frac{P\left( C_j \right)}{P\left( C_j \right)}
\]
\[
= \sum_{i=1}^{K_i} P\left( S_i | S_jC_{i1}, S_jC_{i2}, \ldots, S_jC_{iL} \right) - P\left( S_i \right)
\]
Similarly it can be proved that
\[
\eta_{ijk} = \frac{P\left( S_i | C_{ij}^1, S_jC_{i2}, \ldots, S_jC_{iL} \right) - P\left( S_i \right)}{P\left( C_j \right)}
\]
Comparing the above equations we derive the creditability rank scores $\text{Rank}(S_i)$ and creditability rank assertions $\text{Rank}(C_j)$ using the iterative fact finding method:
\[
\text{Rank}C_j = \frac{P\left( C_j | S_{ij}C_j, S_{i2}C_j, \ldots, S_{ik}C_j \right) - P\left( C_j \right)}{P\left( C_j \right)}
\]
\[
= \sum_{k \in \text{Sources}} \text{Rank}S_k,
\]
\[
\text{Rank}S_i = \frac{P\left( S_i | S_jC_{i1}, S_jC_{i2}, \ldots, S_jC_{iL} \right) - P\left( S_i \right)}{P\left( S_i \right)}
\]
\[
= \sum_{k \in \text{Claims}} \text{Rank}C_k.
\]
When the creditability ranks are found such that they satisfy these equations, along with the assumption that with the prior probability assertion is true being initialized to $p(tru/a) = C_{true/c}$, it gives us the
\[
P\left( C_j | \text{network} \right) = p_a^t \left( \text{Rank} \left( C_j \right) + 1 \right).
\]
Similarly if $p_a^t$ is the prior probability of a chosen source telling truth then
\[
P\left( S_i | \text{network} \right) = p_a^t \left( \text{Rank} \left( S_i \right) + 1 \right).
\]
Hence the abovementioned Bayesian analysis provides the platform for finding the probability for each individual source $S_i$ and assertions $C_j$ to be true.

5.4. Fuzzy Bayesian Inference. A fuzzy posterior approximator has been applied to yield
\[
F(\theta | x) = \sum_{j=1}^{m} p_j(\theta) C_j(x | \theta),
\]
where
\[
C_j(x | \theta) = \frac{C_{th}G(x | \theta)}{\sum_{i=1}^{m} \int_{d} G(x | u) p_i(u) C_{th}i du}.
\]
Adding fuzzy rules and weights we get
\[
F(\theta | x) = \sum_{i=1}^{m} w_i a_F(\theta) V_F iC_F, i
\]
where the weights $w_F$ and then part set volumes $V_F$ are given by
\[
w_F = w_p, w_h, k,
\]
\[
V_F = V_p, V_h, k.
\]

6. Results and Discussions
We carried out the experiments using MATLAB 2013, simulating the source and assertion information. We evaluated the results by using extensive simulations by generating up to 100 sources and 1000 assertions. Experiments were carried out to verify the accuracy and correctness of the probability that a given source is truthful or the given assertion is found truthful by the proposed Fuzzy Bayesian Scheme of interpretations. This technique is then compared with the existing methods like Page Rank and the results are tabulated. To test the results we randomly generated the source and assertions and partitioned these assertions into true and false. $p_i$ is the random probability assigned to each source $S_i$ generating $L_i$ claims. Every claim has $p_i$ probability of being true and $1 - p_i$ probability of being false. We compared the proposed Fuzzy Bayesian Interpretation scheme with that of the Page Rank algorithm. For every source correctness probability distribution, the results were averaged over 100 datasets in a time series. The number of true and false assertions was fixed as 1000 and the claims per source are datasets in a time series. The number of true and false assertions was changed from 0.2 to 0.8. It has been observed that the error of source correctness probability prediction decreases with the increase in the number of sources. Relatively small error percentage of 5% to 8% was encountered by the source correctness probability. This probability is very accurate in the order of 0.1, when the large number of sources is considered. Figure 3 shows the graph between the prediction accuracy and the varying number of sources for prediction error of source probabilities, false positive of assertions, and false negative of assertions. We varied the count of true assertions to the overall total of assertions in the network and fixed the total number of assertions as 3000 and the number of sources as 30. The ratio of true to total assertions was changed from 0.2 to 0.8. It has been observed that the error of source correctness probability prediction decreases with the increase in the ratio of true assertions. Alternatively, the numbers of false negatives increase, as more and more true assertions are misclassified as false. Figure 4 shows the graph between the prediction accuracy and the varying number of assertions for prediction error of source.
Figure 3: Prediction accuracy versus varying number of sources.

Figure 4: Prediction accuracy versus varying number of assertions.
probabilities, false positive of assertions, and false negative of assertions. At last, these results are compared with the Page Rank algorithm and the results are displayed as graph. The results show that Fuzzy Bayesian Interpretation outperforms the Page Rank algorithm for the given number of sources and the same has been shown in Figure 5. The observed results have confirmed the advantages of Fuzzy Bayesian approach over the conventional Bayesian analysis [24] in comparison with Page Rank.

7. Conclusion

In this research, an analysis of information networks was presented using the Fuzzy Bayesian approach and it has been successfully verified to assess the creditability of sources and assertions. Though several rank based algorithms are present, this paper converted the rank to a fuzzy probability method to find the truthfulness of a source and its assertion. With the ever increasing end in the smart devices under the category “Internet of Things,” today’s businesses have understood the importance of Big Data as they have opened up new frontiers in the field of business intelligence. Traditionally business intelligence adopted a revolutionary approach by bringing in the Big Data in Hadoop based environments and deriving the reporting, modelling, and prediction systems. Now, the businesses are moving towards evolutionary approaches or even hybrid approaches that use the combination of techniques of Big Data and business intelligence. With the Big Data becoming the primary input for business intelligence tools, social sentiment analysis and usage patterns form the backbone to market any product across the world leading to the concept of social sensing. This is a rapidly developing technology and, with the ever increasing mobile devices proliferation, this technique gains momentum. With the miniaturization of the sensors and the ease with which they are embedded into mobile devices, social sensing will definitely have a widespread acceptance and usage. Though the mobile device and the associated sensors possess high degree of reliability, when humans act as the operators of these sensors, the reliability becomes a question mark. Hence several researches similar to the one presented in this paper are being conducted across the world to improve the veracity of the Big Data. With the explosion of Cloud Computing, the concept of Cloud supported analytics has already evolved and is becoming multidimensional [26]. This new paradigm enables the infrastructure to scale up or down dynamically, adapting the system to the actual demand. Analytics as a Service (AaaS) or Big Data as a Service (BDaaS) is another budding area which is quite interesting and can take us towards ultimatum. Though the present study has incorporated only the Fuzzy Bayesian and Page Rank algorithms, future studies can be done using several advanced techniques like Maximum Likelihood Estimation and Expectation Maximisation to bring in more reliability in an undependable system.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] M. R. Trifu and M. L. Ivan, “Big data: present and future,” Database Systems Journal, vol. 5, no. 1, 2014.
[2] http://www.emc.com/leadership/digital-universe/2014view/executive-summary.htm.
[3] http://www.emc.com/leadership/digital-universe/2014view/high-value-data.htm.
[4] A. Dasgupta, Big Data: The Future Is in Analytics, Geospatial World, 2013.
[5] M. J. Bates, “After the dot-bomb: getting web information retrieval right this time,” First Monday, vol. 7, no. 7, 2002.
[6] M. Schroock, R. Shockley, J. Smart, D. Romero-Morales, and P. Tufano, “Analytics: the real-world use of big data. How innovative enterprises extract value from uncertain data,” Tech. Rep. GBE03519-USEN-00, IBM Institute for Business Value in Collaboration with Saïd Business School at the University of Oxford, 2012.
[7] R. Y. Wang and D. M. Strong, “Beyond accuracy: what data quality means to data consumers,” Journal of Management Information Systems, vol. 12, no. 4, p. 5, 1996.
[8] R. W. Zmud, “An empirical investigation of the dimensionality of the concept of information,” Decision Sciences, vol. 9, no. 2, pp. 187–195, 1978.
[9] Y. Wand and R. Y. Wang, “Anchoring data quality dimensions in ontological foundations,” Communications of the ACM, vol. 39, no. 11, pp. 86–95, 1996.
[10] W. H. DeLone and E. R. McLean, “Information systems success: the quest for the dependent variable,” Information Systems Research, vol. 3, no. 1, pp. 60–95, 1992.
[11] I. Claverie-Berge, “Solutions Big Data IBM,” http://www-05.ibm.com/fr/events/netezzaDM_2012/Solutions_Big_Data.pdf.
[12] W. Q. Meeker and Y. Hong, “Reliability meets big data: opportunities and challenges,” Quality Engineering, vol. 26, no. 1, pp. 102–116, 2014.
[13] Y. S. Uddin, T. Al Amin, H. Le, T. Abdelzaher, B. Szymanski, and T. Nguyen, “On diversifying source selection in social sensing.”

![Figure 5: Prediction comparison.](image-url)
in Proceedings of the 9th International Conference on Networked Sensing Systems (INSS '12), pp. 1–8, Antwerp, Belgium, June 2012.

[14] D. Wang, T. Abdelzaher, L. Kaplan, and C. C. Aggarwal, “On quantifying the accuracy of maximum likelihood, estimation of participant reliability in social sensing,” in Proceedings of the 8th International Workshop on Data Management for Sensor Networks (DMSN '11), 2011.

[15] C. C. Aggarwal and T. Abdelzaher, “Social sensing,” in Managing and Mining Sensor Data, pp. 237–297, Springer, New York, NY, USA, 2013.

[16] R. Mishra and R. Sharma, “Big data: opportunities and challenges,” International Journal of Computer Science and Mobile Computing, vol. 4, no. 6, pp. 27–35, 2015.

[17] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, “Deep learning applications and challenges in big data analytics,” Journal of Big Data, vol. 2, article 1, 2015.

[18] T. Lukoianova and V. Rubin, “Veracity roadmap: is big data objective, truthful and credible?” Classification Research Online, vol. 24, no. 1, 2014.

[19] R. Viertl, “Foundations of fuzzy Bayesian inference,” Journal of Uncertain Systems, vol. 2, no. 3, pp. 187–191, 2008.

[20] O. Osoba, S. Mitaim, and B. Kosko, “Bayesian inference with adaptive fuzzy priors and likelihoods,” IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 41, no. 5, pp. 1183–1197, 2011.

[21] M. A. Gil, “Bayesian decision making with previous probabilistic uncertainty and actual fuzzy imprecision,” Kybernetes, vol. 17, no. 3, pp. 52–66, 1988.

[22] P. V. Golubtsov and S. S. Moskaliuk, “Fuzzy logic, informativeness and bayesian decision-making problems,” Hadronic Journal, vol. 26, no. 5, pp. 589–630, 2003.

[23] O. V. Verevka and I. N. Parasyuk, “Mathematical fundamentals of constructing fuzzy Bayesian inference techniques,” Cybernetics and System Analysis, vol. 38, no. 1, pp. 89–99, 2002.

[24] Bayesian Perceptual Psychology, http://www.philosophy.ucsb.edu/docs/faculty/papers/bayesian.pdf.

[25] D. Wang, T. Abdelzaher, H. Ahmadi et al., “On Bayesian interpretation of fact-finding in information networks,” in Proceedings of the 14th International Conference on Information Fusion (Fusion '11), Chicago, Ill, USA, July 2011.

[26] M. D. Assunção, R. N. Calheiros, S. Bianchi, M. A. Netto, and R. Buyya, “Big data computing and clouds: trends and future directions,” Journal of Parallel and Distributed Computing, vol. 79–80, pp. 3–15, 2015.