RESEARCH ARTICLE OPEN ACCESS

An Assessment Model to Evaluate Quality Attributes In Big Data Quality

Supriya Haribhau Pawar [1], Prof. Dr. Devendrasingh Thakore [2]
PG Scholar [1]
Department of Computer Engineering [1] & [2]
Bharati Vidyapeeth Deemed University College of engineering
Pune - India

ABSTRACT

Big data describe volume amount of structured and unstructured data. When unstructured data collected from different sources, maintain the quality of data is also important. The proposed system improves the quality of business data which are collected for business decision making. Data quality Assessment is a way for practitioners to understand the scope of how poor data quality effects on business process and develop a business case for data quality management. This paper contributes to providing a solution by introducing new assessment model to evaluate and manage the quality of social media data. Sentiment analysis is used for monitoring real-time data. Generate the new rules and attributes to assess the quality of data. Apply quality attributes on input data and assess only those data which are fit into the quality attribute dimensions. The proposed system improve solution by provide real time and validate data for the user.

Keywords : — Data quality, social media data, Sentiment Analysis, Quality attributes, evaluating quality attributes.

I. INTRODUCTION

Social media data is one of big data's most influential origins. The big data dimensions: Volume, variety, velocity, and veracity, produce some challenges not only to data analytics but also to the data system that manage the data. Social media platform offers marketers significant amount of data which can be used to make a marketing decision in the future [1]. Goals of social media data are an improvement of customer service, instant feedback on products and services through sentiment analysis. Use different metrics to reach different goals. When Social media data collected from various sources (Facebook, twitter), common challenges are arises in data like missing or incomplete data, unavailable of data streams, old data. The user wants to ensure the reliability of the data while collecting. When data is proceeding for analyzing some data, the user intends to make sure that the relevancy and quality of data are appropriate the particular solution. Reliable and valuable data enhance decision making of business. The evaluation of data quality happens in data processing phase in big data architecture, data extraction, data processing, and decision making. Quality evaluation of big data considers while data goes through the pipeline of big data system [2].

When unstructured data collected from different sources, maintain the quality of data is also important. Unstructured data (Input data) are important for marketer for taking a right decision and gain valuable customer insights, reduce marketing expenses and improve sales. The proposed system ensures the quality and trustworthiness of social media data for business and marketers decision making [3]. Data quality management is one where evaluate data quality and improve business decision making. Data quality management is continuous analysis, observation and improvement overall quality of the organization.

The purpose of this paper is to how to evaluate the quality of data and provide trustworthiness data to business for decision making. Introduce solution for data evaluation, in which data customer can select relevant quality attributes and metrics and evaluate quality attributes with evaluation metrics.

II. RELATED WORK

Big data is relevant to many components like government, healthcare, business management, social media, education, life science. Using big data these components improve their decision making, transparency, quality by providing continuous monitoring. The challenges are arises when the volume of structured and unstructured data coming from different sources. The reason for generating a large amount of data, big data-based application introduced new challenges and issues for quality assurance engineers [4]. These challenges are not only limited to data analysis but also to a big data system that manages all the information [1].

Recently social media data such as Twitter, Facebook increase business by providing insights into customer opinions, thoughts, and preferences. Design the platform that supports to monitoring and analysing customer feedback in social media network and identifies issues which are faced by customers. Internet users communicate and express their thoughts with thousands of other people. People use a social media platform to share their ideas and experiences with different customer products and services [2].

When data are coming from different sources maintaining or evaluating the quality of data is also important. Quality metrics are components of an effective quality management plan and measure properties of quality attributes. To assess and analyse the quality in any system, first need to characterize any quality attributes which are relevant to that system. Quality attributes such as Accuracy, performance, consistency, timeliness, completeness, relevancy [4] [3] are used to evaluate quality in social media data. Research aim [4] is providing trustworthiness metrics for information provenance and quality evaluation.

Quality metrics measure the performance of product and processes. Each metrics has following properties [2].

Description: Metric description.
Purpose: The purpose of metric.
Target: Where metric are used.
Formula: How the value of metric is achieved.
Range value: Value of range for the metric evaluation
Acceptable Value: Minimum value for accepted quality attributes.
Rules: The set of measurement value range and a set of constraints which define a set of target measurement.

The above observations and literature studies [1-6] indicate that quality evaluation is limited to only a few quality attributes, the purpose is to increase the quality assessment to introduce more quality attributes. Data quality management is one where evaluate data quality and improve business decision making. Data quality management is continuous analysis, observation and improvement overall quality of the organization. Unstructured sources like social media need data quality management for improving their data quality. Data quality management is one where evaluate data quality and improve business decision making. Data quality management is continuous analysis, observation and improvement overall quality of the organization. Ensuring data quality involves following steps:

A. *Quality assessment*- In quality evaluation phase, decide data source type to extract the data. It is a way for practitioners to understand the scope of how poor quality data affects on business process and develop a business case for data quality management.

B. *Quality design*- In the quality design phase, design and analysis the quality process and concentrate on the data elements that are consider based on the selected business user needs.

C. *Quality Transformation*- In quality transformation phase, define business related quality rules and perform measurement using metrics.

D. *Quality monitoring*- In quality monitoring phase, review expectations and refine rules and monitor data quality versus target.

III. PROPOSED SYSTEM

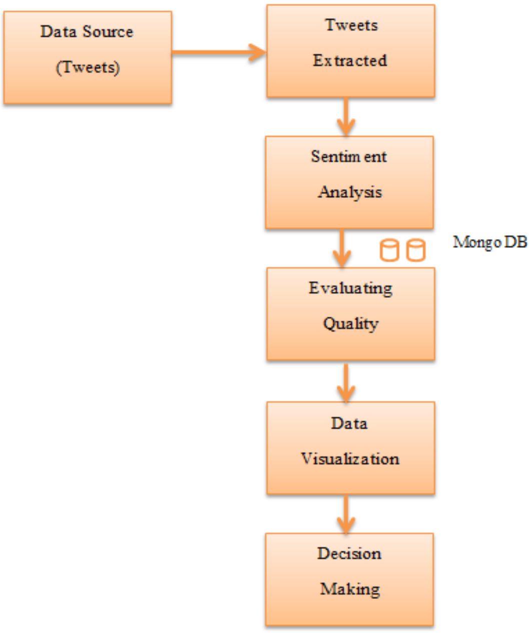

Fig. 1 Architecture of proposed system

The main purpose of the system is to collect social media data (e.g. tweets) to achieve customer insight that can be used in business decision making. The decision-making policies have a great importance in quality evaluation.

Social media source like tweeter where peoples are connecting to share their opinions and find out what happening in the world right now (Input data). Input data (unstructured data) are useful to business for improve data quality and decision making. End user specifies keywords related to interesting products. Data extractor extracts tweets via Twitter API or document based on the specified keywords.

Apply quality attributes on input data and assess only those data which are fit into the quality attribute dimensions. Sentiment analysis is applied on assess data. Sentiment analysis is the process of recognition whether extracted data is positive, negative or neutral. Input data is scanned for positive or negative words like sad, happy, great, and terrible. Algorithms are used to score the document to decide whether they indicate positive or negative sentiment.

Tweets are created based on metadata, and related quality attributes. Metadata are a help to users to validate the quality and value of data for business usage. Manage quality metadata and attributes; rules are needed to define in quality, i.e. which quality attributes can be used and where. The system applies below Quality attributes [5].

Quality Attribute	Description
Accuracy	Ensure that the input data is error free
Completeness	Check the extracted data is not missing.
Consistency	Implies that not two or more values conflict with each other.
Relevancy	The extracted information is helpful for the task. Non relevant data should not be considered.
Validity	Input data is valid in its purposed used.
Timeliness	The extracted input data is not old data. The timestamp is necessary when retrieving the data.
Believability	The extracted data is valid and credible.

Table 1: Quality attributes and it's Definition

The relevant data is visualized to the end user, and decision-making policy defines the valuable data for decision making by selecting only the data sets with the correct quality attribute value.

MongoDB stores data as a document in binary representation called Binary JSON (BSON). Documents that share a similar structure represented as collections. Reports are similar to rows and fields are similar to columns. Developers can start writing code and store the object as they are created. MongoDB provides document validation within the database. Queries executed and the result of the query based on text argument using Boolean operators. [10][11]

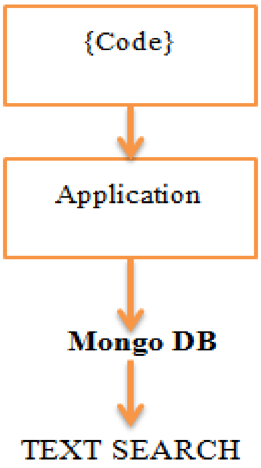

Fig. 2: MongoDB working

IV. PROPOSED RESULT

The proposed system provides real time and quality data to business for decision making. Identify data items or input data need to be assessed for business data quality. Apply quality attributes on input data and evaluate only those data which are fit into the quality attribute dimensions. For each quality attribute defines the range to reprinting good or bad quality data. Sentiment analysis determines positive and negative sentiment from text. Sentiment analysis API (or document) provides very accurate analysis of the emotion of the text from sources. The analysis of text presented in range (e.g. range between 0 to 5). The result of scores closer to 5 considered to be positive sentiment and scores closer to 0 will be of negative sentiment. Review the result and determine data quality is accepted or not.

V. CONCLUSION

The proposed system provides a solution for evaluating the quality of data for business decision making. The quality of data is assessed in each of data processing phase. The system analyses various data across the social web and helps to convert data into actionable insights. Sentiment analysis is used to review the social media data. The proposed system provides the visualized data and generates a report based on sentiment analysis.

ACKNOWLEDGEMENT

To prepare proposed methodology paper on "An assessment model to evaluate quality attributes in big data quality" has been prepared by Supriya Pawar and Prof. Dr. Devendrasingh Thakore.

Author would like to thank my faculty as well as my whole department, parents, friends for their support. Author has obtained a lot of knowledge during the preparation of this document.

REFERENCES

[1] A. Immonen, P. Pääkkönen, and E. Ovaska, "Evaluating the Quality of Social Media Data in Big Data Architecture," vol. 3536, no. c, 2015.

[2] A. Fabijan, H. H. Olsson, and J. Bosch, "Customer Feedback and Data Collection Techniques in Software R & D : A Literature Review," vol. 1, pp. 139–153, 2015.

[3] A. Immonen, M. Palviainen, and E. Ovaska, "Requirements of an Open Data Based Business Ecosystem," vol. 2, 2014.

[4] S. Bhatia, J. Li, W. Peng, and T. Sun, "Monitoring and Analyzing Customer Feedback Through Social Media Platforms for Identifying and Remedying Customer Problems," pp. 1147–1154, 2013.

[5] E. Bertino, "Big data - Opportunities and challenges: Panel position paper," Proc. - Int. Comput. Softw. Appl. Conf., pp. 479–480, 2013.

[6] J. R. C. Nurse, S. S. Rahman, S. Creese, M. Goldsmith, and K. Lamberts, "Information Quality and Trustworthiness : A Topical State-of-the-Art Review," no. Iccans, pp. 492–500, 2011.

[7] G. Cong et al., "Improving data quality: consistency and accuracy," Proc. 33rd Int. Conf. Very large data bases, vol. Vienna, Au, pp. 315–326, 2007.

[8] C. Tao and J. Gao, "Quality Assurance for Big Data Application – Issues , Challenges , and Needs."

[9] [M. Bobrowski, "Measuring data quality," Univ. Buenos …, no. 1428, pp. 99–002, 1999.

[10] S. Poe and S. V Vrbsky, "Comparing nosql mongodb to an sql db Comparing NoSQL MongoDB to an SQL DB," no. April 2013, 2015.

[11] L. Kumar, S. Rajawat, and K. Joshi, "Comparative analysis of NoSQL (MongoDB) with MySQL Database International Journal of Modern Trends in Engineering and Research (IJMTER)," pp. 120–128.

[12] S. H. Pawar, "A Study on Big Data Security and Data Storage Infrastructure," vol. 6, no. 7, pp. 539–542, 2016.

[13] P. T. Patil, "A Study on Evolution of Storage Infrastructure," vol. 6, no. 7, pp. 501–506, 2016.

[14] P. Malviya, "A Study Paper on Storage Area Network Problem-Solving Issues," vol. 4, no. 4, pp. 151–156, 2016.

[15] D. Thakore and A. R. Upadhyay, "A Framework to Analyze Object-Oriented Software and Quality Assurance," no. 5, pp. 254–258, 2013.

[16] DevendrasinghThakore, Dr. Akhilesh R. Upadhyay, "Analysing and Improving Quality Assurance in Document Search- Engine incorporating a Document-Ranking Algorithm for Text-Mining", International Journal of Latest Technology in Engineering, Management & Applied Science (IJLTEMAS), Feb 2013.

[17] Prof. D M Thakore, S U Kadam, "Increasing Scalability of Data Mining Algorithms for High Dimensional Data", International Journal of Advances in Computing and Information Research (IJACIR), May 2012.

[18] Kabugade, Rohan R., S. S. Dhotre, and S. H. Patil. "A Study of Modified O (1) Algorithm for Real Time Task in Operating System." Sinhgad Institute of Management and Computer Application NCI2TM (2014).

[19] Karande, Poonam, S. P. Dhotre, and Suhas Patil. "Task management for heterogeneous multi-core scheduling." Int. J. Comput. Sci. Inf. Technol 5.1 (2014): 636-639.