Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 85 (2016) 550 – 557

Procedia Computer Science

International Conference on Computational Modeling and Security (CMS 2016)

Hadoop Framework For Entity Resolution Within High Velocity Streams

S.Prabhakar Benny[1], Dr S.Vasavi[2] P.Anupriya[3]

[1]Research Scholar, JNTU Hyderabad,University College of Engineering for Women, Kakatiya University, Warangal,Telangana, India
Prab_ku@yahoo.co.in

[2]Department of Computer Science & Engineering,VR Siddhartha Engineering College, Vijayawada,vasavi.movva@gmail.com

[3]M.Tech student, [2]Department of Computer Science & Engineering,VR Siddhartha Engineering College, padilamanupriya99@gmail.com

Abstract

Large amount of data is being generated from sensors, satellites, social media etc. This big data (velocity, variety, veracity, value and veracity) can be processed so as to make timely decisions by the decision makers. This paper presents results of the proposed Hadoop framework that performs entity resolution in Map and reduce phase. MapReduce phase matches two real world objects and generates rules. The similarity score of these rules are used for matching stream data during testing phase. Similarity is calculated using 13 different semantic measures such as token-based similarity, edit-based similarity, hybrid similarity, phonetic similarity as well as domain dependent Natural language processing measures. Semantic measures are implemented using Hive programming. The proposed system is tested using e-catalogues of Amazon and Google.

Keywords: Big data; Entity Resolution;, Hadoop Framework; Hive; Stream Processing

1. Introduction

Stream data means real-time data that are communicated as tweets, posts, messages, e-catalogues etc., Stream data can be of different types such as structured data, unstructured data and semi sstructured data. The structured data depends on a data model such as relational model for storing and future access. Unstructured data such as text documents, news articles cannot be stored as a record into a file. Semi-structured data such as web data do not adhere to a strict data model structure. This paper considers semi structured data for Entity Resolution.

Entity Resolution(ER) arises in applications such as data integration, de-duplication. The problem is to identify which entity of one data set is same as which entity of another data set. ER applications include data cleaning, price comparison, biomedical research, outlier detection. For example consider online shopping product catalogs from various vendors. Identifying the entities that represent the same product from these catalogs helps not only price comparison but also for other features comparison. Big data technologies such as Hadoop can only process data batch by batch[8]. Due to this when one batch is finished, data is already aged by, at least, the time required by the batch [12]. In the proposed system Map phase reflects performing comparison (using similarity measures) and the reduce phase reflects entity resolution (check for matches). This paper summarizes the proposed system [19] for Entity Resolution process and present results of the proposed system. The paper is organized as follows: Section 2

doi:10.1016/j.procs.2016.05.218

presents literature survey on existing methods for ER. Detailed study on these approaches can be found in [19]. Section 3 presents results of the proposed system. Section 4 presents conclusion and future work.

2. Literature Survey

Entity resolution is the problem of identifying similar entities across multiple data sources that satisfies a given match function.

The Entity is represented as a set of attribute-value pairs. According to [15], An entity $e_i\text{€}E$ is defined as given in Eq. (1):

$$e_i=\{(a_{ij},v_{ij})|a_{ij}\text{€}N,v_{ij\text{€}}V\} \quad (1)$$

a: attribute names, v: values and E: entities.

$E = \{e_1,e_2,\ldots\ldots, e_m\}$ is set of entities

$M: E \times E \rightarrow \{\text{true, false}\}$ is a match function

The match function maps each pair of entities (e_i ,e_j) to true or false as follows:

$$M(e_i , e_j) \begin{cases} = \text{true if } e_i,e_j \text{ are having high similarity score} \\ = \text{false if } e_i,e_j \text{ are having less similarity score} \end{cases}$$

The similarity measure counts how close the two sets entities are and we can set similarity threshold‘t’.

$$M(e_i , e_j) \begin{cases} = \text{true if } e_i,e_j \text{ are having high similarity score} \geq t \\ = \text{false if } e_i,e_j \text{ are having less similarity score } < t \end{cases}$$

Blocking-based clean-clean ER framework over highly heterogeneous information spaces (HHIS) is proposed in [1]. This framework contains two layers that groups entity profiles into blocks and a set of blocking schemes that build blocks of high robustness in the context of HHIS. Meta-blocking approach is proposed in [2] that extracts the most similar pairs of entities. Also usage of Graph pruning eliminates the unwanted nodes to reduce the redundant comparisons. MapReduce based work is given in [3], where input data is partitioned and then sent to different nodes (mappers) in the cluster. Hadoop is used to perform the MapReduce tasks. A family of techniques for constructing hints is proposed in [5]. These hints are used to maximize the number of matching records. Rules are proposed in [6,7] to describe the complex matching conditions between records and entities.

Apache Flume is a distributed system for collecting, aggregating, and moving large amounts of data from multiple sources into HDFS or another central data store[8]. A*pache Sqoop* is a tool for transferring data between Hadoop and relational databases[8]. Oracle Warehouse Builder (OWB) provides ETL to support for designing a data warehouse and the data flow; these tasks are typically addressed by ETL tools such as OWB[9].

In [10], Informatica Power Center is an ETL tool, which is used for data integration. In [11], Pentaho Data Integration delivers comprehensive Extraction, Transformation and Loading (ETL) capabilities using a meta-data driven approach. Modified version of the Hadoop MapReduce framework that supports online aggregation is given in [13]. Current release of Apache Drill supports in-memory and beyond-memory execution are given in [14]. Supported file formats in the first beta drop of Cloudera Impala include text files and Sequence Files[16].

3. Proposed System

The main objectives of the proposed systemas shown in fig 1 are:

1. Tokenization of structured data
2. Perform the mapping of entities within stream data
3. Reducing the redundant mappings
4. Reducing the number of redundant comparisons
5. Rule generation

The training Phase starts with the removal of unwanted words. Word boundaries, stemming and stop word removal are done subsequently. Snowball (Porter2) algorithm is used for stemming [17]. After this the data is represented in tabular format i.e structured format. A total of 318 words list given by Cambridge University is used as a stop word dataset. Output from preprocessing is set of entities. This entity collection is used as input by token blocking step. Consider an example before preprocessing the entity e1 is having the title as “Resumemaker’s professional (v1.2),” then after applying the stemming algorithm and removing unwanted stuff we get the title as “Resumemaker professional” as shown in the Fig 2. This process is repeated for all the entities