A PROJECT REPORT

On

"ENTITY RESOLUTION"

Submitted to KIIT UNIVERSITY

In Partial Fulfillment of the Requirement for the degree of

B.Tech IN INFORMATION TECHNOLOGY

BY

SAKSHI RAI 1806059

SAMDARSHI KUMAR 1806060

SHASHIKANT SINGH 1806068

ANANYA 1806458

MANISHA GODARA 1806486

UNDER THE GUIDANCE OF

PROF. BHASWATI SAHOO

SCHOOL OF COMPUTER ENGINEERING

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

BHUBANESWAR, ODISHA -751024

MAY 2021

CERTIFICATE

This is to certify that the project entitled
"ENTITY RESOLUTION"
Submitted by

SAKSHI RAI 1806059

SAMDARSHI KUMAR 1806060
SHASHIKANT SINGH 1806068
ANANYA 1806458
MANISHA GODARA 1806486

in partial fulfillment of the requirements for the award of the **Degree of Bachelor of Technology** in **INFORMATION TECHNOLOGY** is a bonafide record of the work carried out under our guidance and supervision at School of COMPUTER ENGINEERING, Kalinga Institute of Industrial Technology, Deemed to be University.

Signature of Supervisor 2 (if applicable)	Signature of Supervisor 1
NAME OF THE SUPERVISOR 2	NAME OF THE SUPERVISOR 1
Academic affiliation	Academic affiliation
Organization	Organization

The Project was evaluated by us on 14/05/2021

EXAMINER 1	EXAMINER 2
EXAMINER 3	EXAMINER 4

Acknowledgement

We are profoundly grateful to **Prof. Bhaswati sahoo** for her expert guidance and continuous encouragement throughout to see that this project rights its target from its commencement to its completion.

SAKSHI RAI 1806059

SAMDARSHI KUMAR 1806060
SHASHIKANT SINGH 1806068
ANANYA 1806458
MANISHA GODARA 1806486

ABSTRACT

Data Analysis has multiple techniques and approaches under its sleeves, encompasses diverse techniques with a variety of requirements, in different business models, science and social science domains.

Substance goals is made to utilize in signs of certifiable elements in different records or notices by connecting and gathering. For instance, there could be various methods for tending to a similar individual in content, various locations for organizations, or photographs of a specific item.

This obviously has numerous applications, especially in government and general wellbeing information, web search, examination shopping, law requirement, and the sky is the limit from there. Also, as the volume and speed of information develop, the derivation crosswise over systems and semantic connections between elements turns into a more prominent test.

This paper gives a profound knowledge about ER that can decrease the intricacy by proposing canonical references to specific elements and duplicating and connecting substances.

Keywords:

- Workflow
- Analysing Attributes
- Blocking
- Binning
- Data Preparation
- Pairwise Matching

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLE

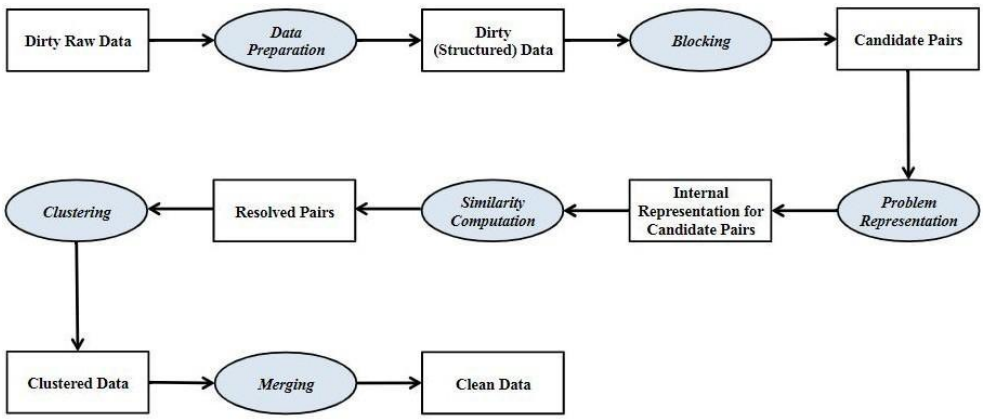
Chapter 1

1.1 Introduction To ER

Inside a solitary data framework or crosswise over various data frameworks, there may exist extraordinary portrayals for one genuine element. The distinctions may result from typographical blunders, shortenings, information designing, and so forth. Notwithstanding, these mistakes and irregularities can restrain the immaterialness of the information for value-based and scientific purposes, and, as needs be, limit the business estimation of the information. Hence, it is important to have the option to determine and explain such unique portrayals.

Entity Resolution is used for recognizing data that gives us insight into a similar genuine element. It is additionally known for a few different names. The general field of software engineering is too alluded to as information coordinating, record link, reference compromise, duplication, or object distinguishing proof. All the more explicitly, in the database area, ER is firmly identified with the strategies of closeness join. Before we portray our answers on empowering examination mindful information cleaning, in this section, we quickly survey the element goals issue (which is the focal point of this proposal) to provide context to our exploration.

fig 1.1



1.2 Problems In Entity Resolution

1.2.1 Traditional Challenges faced by Entity Resolution

- Name Ambiguity or Attribute Ambiguity

Suppose there are two records where one record contains Name of the user Priti and other data records contain the Name of the user Preeti.

- Poor Data entry

During manual entry, the record keeper enters some wrong information by mistake.

- Missing Values

When important values are not present in records.

- Data Formatting

When Same information is represented in a different way like the date format can be represented in a number of ways, The date can be in the format 1-MAY-2021 or in MAY-1-2021.

- Changing Attribute

When some document is created with some specific timestamp but we can manually change the attribute of the created
the document which can cause ambiguity.

- Data Truncation or Abbreviations

Short form of some word like Television can also be abbreviated as T.V.

1.2.2 Big Data ER Challenges

- Huge Data Records-

Maintaining Huge records is difficult so we need more efficient techniques to maintain data

Heterogeneity

Unclean Data Types, Unstructured quite a mixture of different types of data.

- More Connected

We need more good quality connection in order to get the golden record.

- Multi-Relational

Dealing with the structure of the entity suppose we there are two companies Big Bazar and Bazar these are not the same.

- Multi-Domain

Methods and tricks that can allow us to span across a larger domain.

1.3 Purpose of Entity Resolution

Our goal is to utilize in signs of certifiable elements across records or notices by connecting and gathering. For instance, there could be various methods for tending to a similar individual in content, various locations for organizations, or photographs of a specific item.

This obviously has numerous applications, especially in government and general wellbeing information, law requirement, web search, examination shopping, and the sky is the limit from there. Also, as the volume and speed of information develop, the derivation crosswise over systems and semantic connections between elements turns into a more prominent test.

What we need is to find the comparison matrix for getting the golden record which is made by comparison of the datasets.
The similarity score is simply a boolean data type it can either match or it is unmatched and also it can be the real values with distance functions.
The final step is after when we have constructed a vector of component-wise similarity for the given records we compute the probability for the pair if its a definite match or not.

Chapter 2

2.1 WORK FLOW

fig 2.1

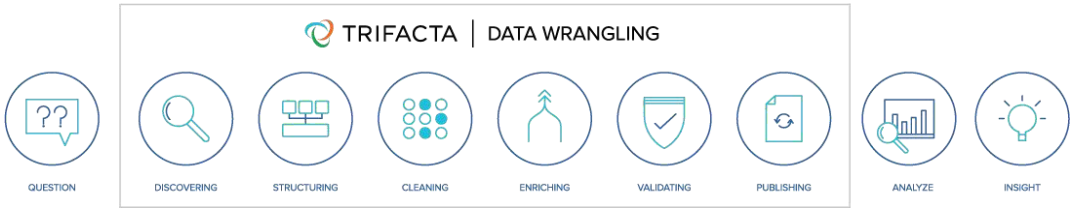
2.2 Data Preparation

The primary assignments of composition and information standardization required for planning for any Entity Resolution Algo. This errand, composition characteristics are coordinated (Example Telephone no versus Phone no), and compound qualities like locations are standardized. Information standardization includes changing overall string to upper or lower case and expelling spaces. Information cleaning, word reference queries are additionally significant. Starting information prep is a major piece of the work; brilliant standardization can go far!

The main objective is to build a system, for records, an "examination vector" of closeness scores of every segment trait. Comparability scores can just be a

Boolean match or unmatch)or they can be genuine qualities with separation capacities. Like for an instance, alter remove on printed qualities can deal with typographic blunders. Jaccard coefficients and other separation measurements can be utilized to think about sets. Indeed, even phonetic similitude can be utilized.

fig 2.2



2.3 Analyzing Attributes

Attribute analysis is one of the most important aspect in Entity Resolution.

It determines the type of columns which are present in any database. This gives us a clear view to perform any type of analytical operation on the dataset.

It is necessary to analyse the attributes as to know which operations needs to be performed on a certain type of data.

For Example - Binning can be performed only on Numeric Data. In this

module we are determining the following data types :-Constant

- Boolean
- Numeric With Length=1
- Numeric Continuous
- Uniformly Spaced
- Non-Uniformly Spaced

- Only Text
- Text with Number & Special Character
- Uniformly Distribute
- Date
-

-

Example:

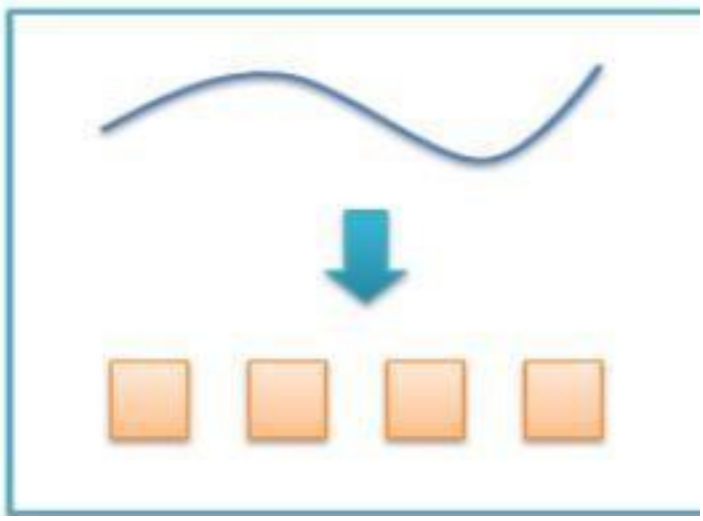
	Name	Type_col
0	PassengerId	Numeric Continuous
1	Pclass	Numeric With Length=1
2	Name	Text with Number & Special Character
3	Sex	Boolean
4	Age	Uniformly Distributed
5	SibSp	Numeric With Length=1
6	Parch	Numeric With Length=1
7	Ticket	Text with Number & Special Character
8	Fare	Numeric Continuous
9	Cabin	Text with Number & Special Character
10	Embarked	Only Text

2.4 Binning

It is a technique used in data processing and statistics. Numerical values are converted into its categorial counterparts using this process. This makes it more useful as the data is clustered together. When the possible data points are more than the observed data points, Binning as a process is more convinient to use. This reduces the noise and increases the identification process of the outliers and missing or invalid data.

There are two types of Binning Processes:-

1) Supervised, and
2) Unsupervised

fi 

Example:

	PassengerId	Pclass	Age	SibSp	Parch	Fare
0	0	2	2.0	0	0	0
1	0	0	3.0	0	0	0
2	0	2	2.0	0	0	0
3	0	0	3.0	0	0	0
4	0	2	3.0	0	0	0
5	0	2	NaN	0	0	0
6	0	0	5.0	0	0	0
7	0	2	0.0	2	0	0
8	0	2	2.0	0	1	0
9	0	1	1.0	0	0	0

2.5 Blocking

The main purpose of BLOCKING is to create Data Blocks. Data Blocks are small chunks of a large file which is the building blocks of a file system. Blocking is used to group similar entities into blocks. In these blocks, all comparisons are executed within the same block. This makes it easy to compare data as the size of data is reduced enormously and the complexity of data decreases gradually.

This intends to cluster similar entities so that the comparisons occur solely between co-occurring entities.

For Example - Suppose there are 1000 businesses in 1000 cities. It will take an ample amount of time to compare the data as such complex data for about as much as 1 trillion outcomes are possible. The execution, however, might take up to 12 days to compare each data. With the help of Blocking, we can co-relate the entities by grouping the business in each city. This reduces the comparison space and possible outcomes might reduce to only a billion possibilities. Thus, the time of execution gets reduced to approximately 16 minutes.

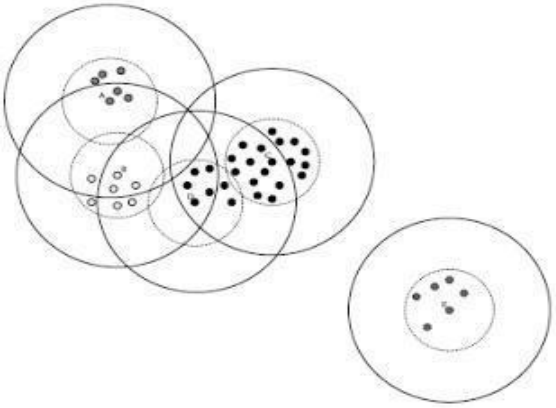
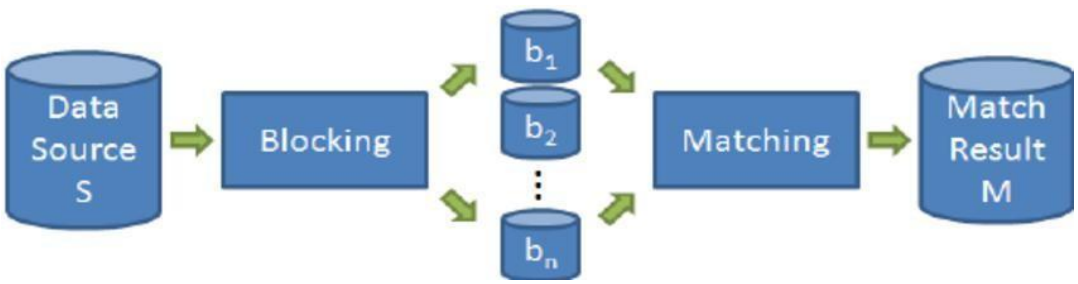


fig 2.3



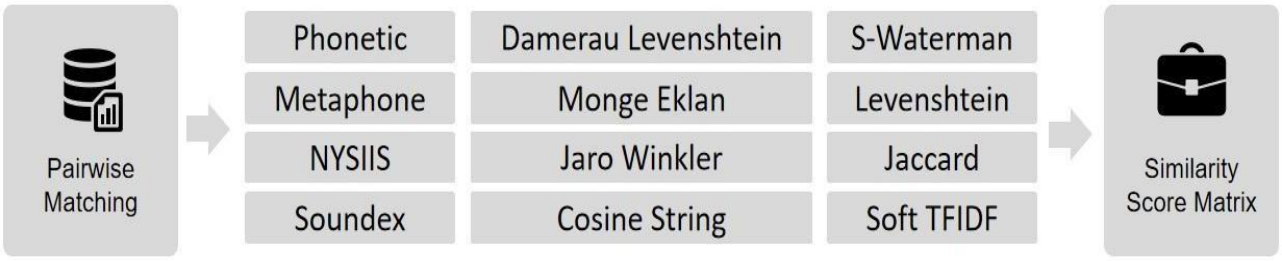
2.6 Pairwise Matching

It is an algorithm which determines the Similarity score of every data in a given dataframe. Similarity score is the score generated by some Similarity measuring functions like NYSIIS, Metaphone, Phonetic, Soundex, etc. These algorithms compares entities in pairs to check which of the two entities has higher priorities in terms of quantity, or if any two entities are similar.

There are a lot of proposals to generate similarity scores using this method. The two most common features are to use the mean of the components or to use Weighted Sum of the components, and by using Thresholds.

For example - Pair-Wise Mathching is most commonly used in voting systems and multiagent AI systems.

fig 2.4



Example:

j / i	0.0	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0
0.0	1.0	0.382571	0.532699	0.439488	0.681190	0.533750	0.563029	0.598446	0.413979	0.410752
1.0	NaN	1.000000	0.467336	0.664375	0.306781	0.241162	0.388742	0.228590	0.428832	0.660477
2.0	NaN	NaN	1.000000	0.551375	0.539465	0.396098	0.444757	0.421963	0.668599	0.480889
3.0	NaN	NaN	NaN	1.000000	0.418856	0.245248	0.504123	0.319967	0.510964	0.570572
4.0	NaN	NaN	NaN	NaN	1.000000	0.621446	0.655645	0.613013	0.481490	0.329167
5.0	NaN	NaN	NaN	NaN	NaN	1.000000	0.497234	0.444127	0.336801	0.268338
6.0	NaN	NaN	NaN	NaN	NaN	NaN	1.000000	0.497442	0.379787	0.332879
7.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.000000	0.442287	0.256362
8.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.000000	0.448167
9.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.000000

2.7 Important Algorithms

- Phonetic

Examines the method for describing the sounds of language and also illustrates the techniques of experimental phonetics, most of them require little more than a recorder, a camera, and a personal computer.

- Damerau-Levenshtein

The minimum distance between two strings it is also defined as the number of minimum edits required to find the result.

- S-Waterman

The smith-waterman Algo executes local sequence alignment; that is, for determining computational analysis.

- Monge-Elkan

The Monge-Elkan measure type of hybrid similarity measure that combines the benefits of sequence-based and set-based methods.

- Jaccard

Known for Intersection over Union and Jaccard similarity coefficient is a statistic used for gauging the similarity and diversity of sample sets.

- Soundex

The goal for homophones is to encode in the same representation so that they can be matched despite minor differences in representation.

- Cosine Similarity

We calculate the similarity between different pairs of the documents using the 'Cosine Similarity' algorithm.

- Jaro Winkler

A string metric system that determines the distance between two strings and helps in editing them.

- Monge Elkan

A general text string comparison method that uses an internal character-based measure of similarity grouped by a token level.

- NYSIIS

An algorithm developed by the New York State Identification and Intelligence System that converts any word into a phonetic code. It is mostly used on names pronounced in English

- Soft TF-IDF1,

It works on pre-existing fuzzy-matching functions on how frequently combinations of words appear in the data.

Chapter 3

Requirement Analysis

3.1 Hardware Requirement

For PYTHON implementation the sufficient requirements are RAM of 1.5 GB and Intel Core i3 – 2120 CPU @ 2.0 GHz.

3.2 Software Requirement

We need to install Anaconda, Python and Jupyter Notebook to run the python code on the jupyter notebook. We have to install dependencies in Python like numpy, pandas, matplotlib, scikit-learn, Stats-model, datetime.

Non-Functional Requirement

3.3 Security

Dataset should be loaded securely. Dataset should not be broken or uncleaned.

3.4 Usability

The scope of the framework is huge. All types of Businesses can easily use the system to enhance their Data Quality.

3.5 Portability

Since the project we work on is merely an executable file, it is portable.

CHAPTER 4
RESULT AND DISCUSSIONS

Entity resolution is becoming an increasingly important task as linked data grows, and the requirement for graph based reasoning extends beyond theoretical applications. With the advent of big data computations, this need has become even more prevalent. As data, noise, and knowledge grows greater the needs & opportunities for intelligent reasoning about entity resolution.there were five phases of project first we have collected the suitable dataset from Kaggle and then we performed data preprocessing that includes Cleaned the data, treat missing value out of range values, nulls, and whitespaces that obfuscate values, as well as outlier values that could skew analysis results then our second step is to analyze various attributes that were present in our document .The attributes that were present were as follows : constant , boolean , Numeric with length =1, Numeric Continuous, Uniformly spaced, Non- Uniformly spaced, Only text,Text with Number & special character,Uniformly Distributed and data then comes the binning in which Numerical values are converted into its categorial counterparts using this process. This makes it more useful as the data is clustered together .When the possible data points are more than the observed data points, Binning as a process is more convenient to use then comes blocking we clustered similar entities into blocks so that pairwise comparisons are executed only between co-occurring entities. In these blocks, all comparisons are executed within the same block. This makes it easy to compare data as the size of data is reduced enormously and the complexity of data decreases gradually this returns a Dataframe where every row stores the index of rows in Source file and represent a block at last comes pairwise matching We started working on the basic outline of big data ,to have an idea of how it would work and what exactly we do because of the end of the day the data needed to be simple maneuver around as a result it determines the Similarity score of every data in a given dataframe. These algorithms compare entities in pairs to check which of the two entities has higher priorities in terms of quantity, or if any two entities are similar. The two most common features are to use the mean of the components or to use Weighted Sum of the components, and by using Thresholds.

CHAPTER 5

Conclusion and Further Work

Conclusion

Element goals is the procedure that resolves substances and identifies connections. The pipelines perform substance goals as they process approaching character records in three stages: perceive, resolve, and relate.

Future Scope

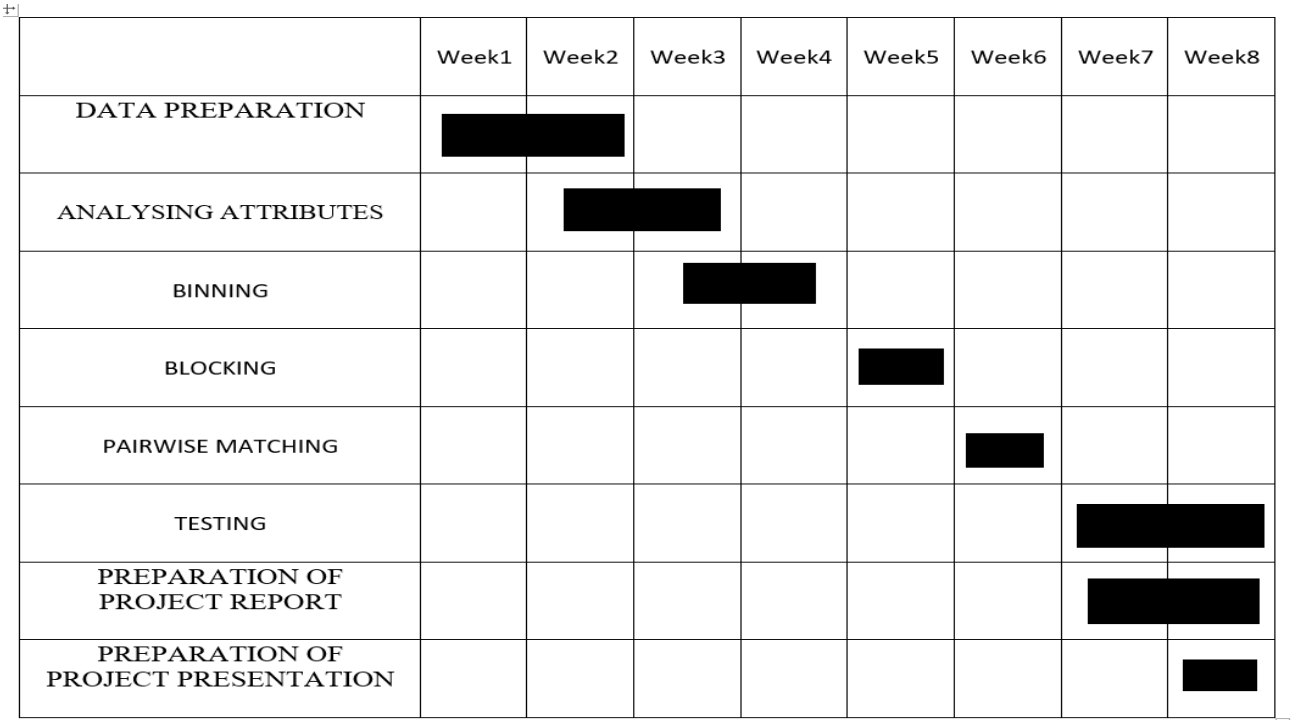
The most important one for parallel ER is Entity-based criteria, which is quite efficient. It can have many applications in future in public health data, web searching, census data, comparison shopping and in many more fields. The technologies that are used are :-

Blocking
Binning
Pairwise Matching

TABLE 5.1 SHOWING DETAILS ABOUT PROJECT PLANNING AND MANAGEMENT

ACTIVITY	STARTING WEEK	NO OF WEEKS
DATA PREPARATION	1st week of January	2
ANALYSING ATTRIBUTES	3rd week of January	2
BINNING	2nd week of February	2
BLOCKING	1st week of March	1
PAIRWISE MATCHING	2nd week of March	1
TESTING	3rd week of March	2
PREPARATION OF PROJECT REPORT	2nd week of April	2
PREPARATION OF PROJECT PRESENTATION	1 week of May	1

GANTT CHART IS SHOWN BELOW



ENTITY RESOLUTION

NAME: SHASHIKANT SINGH

1806068

Entity resolution (ER) is the task of disambiguating records that correspond to real world entities across and within datasets. The applications of entity resolution are tremendous, particularly for public sector and federal datasets related to health, transportation, finance, law enforcement, and antiterrorism.Element goals is the procedure that resolves substances and identifies connections. The pipelines perform substance goals as they process approaching character records in three stages: perceive, resolve, and relate.The goal is to construct, for a pair of records, a "Comparison Matrix" of similarity scores of each component attribute.

Individual contribution and findings: My part of the project was to prepare Dataset .I have collected the suitable dataset from Kaggle and Cleaned the data, ensuring that the data set is capable of providing valid answers when the data is analyzed like missing values, out of range values, nulls, and whitespaces that obfuscate values, as well as outlier values that could skew analysis results. In this task, schema attributes are matched (e.g. contact number vs phone number), and compound attributes like addresses are normalized. Data normalization involves converting all strings to upper or lower case and removing whitespace .I have used python programming language for the preparation.

Individual contribution to project report preparation: I have prepared the chapter 2.2 in the report and also prepared the table content and helped in overall preparation of the report.

Individual contribution for project presentation and demonstration: In the presentation part I prepared the topic of Data preparation and going to introduce the data preparation,workflow of the project and the goal of the project at the time of presentation.

Full Signature of Supervisor:

…….

Full signature of the student:

SHASHIKANT SINGH

……..

ENTITY RESOLUTION

NAME: ANANYA

1806458

Entity resolution (ER) is the task of disambiguating records that correspond to real world entities across and within datasets. The applications of entity resolution are tremendous, particularly for public sector and federal datasets related to health, transportation, finance, law enforcement, and antiterrorism.Element goals is the procedure that resolves substances and identifies connections. The pipelines perform substance goals as they process approaching character records in three stages: perceive, resolve, and relate.The goal is to construct, for a pair of records, a "Comparison Matrix" of similarity scores of each component attribute.

Individual contribution and findings: My part of the project was to analyze various attributes that were present in our document .The attributes that were present were as follows : constant , boolean , Numeric with length =1, Numeric Continuous, Uniformly spaced, Non- Uniformly spaced, Only text,Text with Number & special character,Uniformly Distributed and data. To find the various data types given in the dataset I used python programming language to write the code .The first step was to remove all the duplicate data , then find the basic difference between digits and alphabets . I used the concept of regex data type for alphabets and the Sapiro-Wilk test to check the distribution of numerical data. By analysis attribute we can easily group the same type of data together.

Individual contribution to project report preparation: I have prepared the chapter 2.3 in the report and also prepared the gantt chart and the table associated with it.

Individual contribution for project presentation and demonstration: In the presentation part I prepared the topic of analysis of attributes and going to introduce the project and difficulty in ER at the time of presentation.

Full Signature of Supervisor: Full signature of the student:

ANANYA

…… ……

ENTITY RESOLUTION

NAME: SAKSHI RAI

1806059

Entity resolution (ER) is the task of disambiguating records that correspond to real world entities across and within datasets. The applications of entity resolution are tremendous, particularly for public sector and federal datasets related to health, transportation, finance, law enforcement, and antiterrorism.Element goals is the procedure that resolves substances and identifies connections. The pipelines perform substance goals as they process approaching character records in three stages: perceive, resolve, and relate.The goal is to construct, for a pair of records, a "Comparison Matrix" of similarity scores of each component attribute.

Individual contribution and findings: My part of the project was Binning.Here Numerical values are converted into its categorial counterparts using this process. This makes it more useful as the data is clustered together. When the possible data points are more than the observed data points, Binning as a process is more convenient to use.I have taken the Path or Name of the File. I have used python programming language and used pandas.cut() for sorting and segments.I have also imported bin_config_file which stores the number of bins for individual columns in a dictionary.and finally It returns a Dataframe with all the numeric column with a bin number specified to every value.

Individual contribution to project report preparation: I have prepared the chapter 2.4 in the report and also prepared the acknowledgement and helped in preparing the overall report.

Individual contribution for project presentation and demonstration: In the presentation part I prepared the topic Binning and going to introduce the attribute analysis and Binning at the time of presentation.

Full Signature of Supervisor:

…….

Full signature of the student:

SAKSHI RAI

…….

ENTITY RESOLUTION

SAMDARSHI KUMAR

1806060

Entity resolution (ER) is the task of disambiguating records that correspond to real world entities across and within datasets. The applications of entity resolution are tremendous, particularly for public sector and federal datasets related to health, transportation, finance, law enforcement, and antiterrorism.Element goals is the procedure that resolves substances and identifies connections. The pipelines perform substance goals as they process approaching character records in three stages: perceive, resolve, and relate.The goal is to construct, for a pair of records, a "Comparison Matrix" of similarity scores of each component attribute.

Individual contribution and findings: my part of the project was blocking. It is clustering similar entities into blocks so that pairwise comparisons are executed only between co-occurring entities. In these blocks, all comparisons are executed within the same block. This makes it easy to compare data as the size of data is reduced enormously and the complexity of data decreases gradually.i used python language i used python programming language i made a function in python which takes input source,row_pair_agg and cutoff source takes the binned data frame of a source as input. row_pair agg determines how row pair aggregation is done. It can take average or weighted average and last cutoff It takes a cutoff of similarity to reduce the number of blocks. (Default: 80%).I also imported block_config_file(python file) which stores the weights of column for row pair aggregation and at last this returns a Dataframe where every row stores the index of rows in Source file and represent a block.

Individual contribution to project report preparation: I have prepared the chapter 2.5 in the report and also prepared the abstract of report.

Individual contribution for project presentation and demonstration: In the presentation part I prepared the topic of blocking and going to demonstrate blocking and conclusion.

Full Signature of Supervisor:

…….

Full signature of the student:

……SAMDARSHI……..

ENTITY RESOLUTION

MANISHA GODARA

1806486

Entity resolution (ER) is the task of disambiguating records that correspond to real world entities across and within datasets. The applications of entity resolution are tremendous, particularly for public sector and federal datasets related to health, transportation, finance, law enforcement, and antiterrorism.Element goals is the procedure that resolves substances and identifies connections. The pipelines perform substance goals as they process approaching character records in three stages: perceive, resolve, and relate.The goal is to construct, for a pair of records, a "Comparison Matrix" of similarity scores of each component attribute.

Individual contribution and findings: The pairwise matching including the data cleaning were done by me and with the help of my team members.We started working on the basic outline of big data ,to have an idea of how it would work and what exactly we do because of the end of the day the data needed to be simple maneuver around.I got my inspirations from different tutorial by Dr. Getoor and different websites . In pairwise matching which determines the Similarity score of every data in a given dataframe. These algorithms compare entities in pairs to check which of the two entities has higher priorities in terms of quantity, or if any two entities are similar. The two most common features are to use the mean of the components or to use Weighted Sum of the components, and by using Thresholds.

Individual contribution to project report preparation: Creating chapter 5 and chapter 2.6 pairwise matching algorithm.

Individual contribution for project presentation and demonstration: Collect the data needed for the completion of the document and introduce pairwise matching and Conclusion in presentation.

Full Signature of Supervisor: Full signature of the student:

……. ……Manisha……..

References

1. www.datacommunitydc.org/blog/2013/08/entity-resolution-for-big-data

2. www.districtdatalabs.com/basics-of-entity-resolution

3. www.ibm.com/support/knowledgecenter/en/SS2HSB_8.1.0/com.ibm.iis.ii.overview.doc/top ics/eas_con_entityresolution.html

Plagiarism Scan Report

Page 1

SmallSEOTools

PLAGIARISM SCAN REPORT

Words	915	Date	May 13,2021
Characters	6042	Excluded URL	

2%	98%	1	49
Plagiarism	Unique	Plagiarized Sentences	Unique Sentences

Content Checked For Plagiarism