National College of Ireland

Use of Machine Learning Techniques for Integrating Source Data from Diverse Data Sources

MSc Research Project
Data Analytics

Shanthi Marie Teres Alexis
x15000966

School of Computing
National College of Ireland

Supervisor: Dr. Paul Hayes

National College of Ireland
Project Submission Sheet – 2015/2016
School of Computing

Student Name:	Shanthi Marie Teres Alexis
Student ID:	x15000966
Programme:	Data Analytics
Year:	2016
Module:	MSc Research Project
Lecturer:	Dr. Paul Hayes
Submission Due Date:	29/08/2016
Project Title:	Use of Machine Learning Techniques for Integrating Source Data from Diverse Data Sources
Word Count:	4,514

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	11th September 2016

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Contents

Use of Machine Learning Techniques for Integrating Source Data from Diverse Data Sources

Shanthi Marie Teres Alexis
x15000966
MSc Research Project in Data Analytics

11th September 2016

"That's been one of my mantras - focus and simplicity. Simple can be harder than complex: You have to work hard to get your thinking clean to make it simple. But it's worth it in the end because once you get there, you can move mountains."- Steve Jobs

Abstract

With the advancements of both the computational power of machine learning models and the complexity of data being generated for analysis, the process of identifying the integration flow of information before analysis is becoming more challenging. Implemented models require information such as source data hierarchy and cardinality information and correlation of subject matter as used in ontology. This research project focuses on the integration phase of database elements to its matching component in the data warehouse using only the source data elements and its content information. It questions the capability of these elements and the profile information of itself and its contents in automating the matching process to its designated target element in a data warehouse.

1 Introduction

Metadata information for mapping has been successfully used in the multimedia domain (Nin et al.; 2014) and in file processing (Li et al.; 2005). In the process of data integration, many research has been conducted and most of these research use an extensive amount of information and complex models (Bernstein et al.; 2011). The objective of this research project is to evaluate the contribution of database elements and its contents' information in matching to its corresponding element in a data warehouse. The research question is based on the hypothesis that the profile information of a database element and its content may in fact be able to contribute to the integration process in an automated approach.

The current bottleneck of data processing is due to the process of identifying useful information to integrate. By addressing this research question- the answer to what data should be integrated may also be answered. Furthermore, this research will utilize techniques of analysis for a dataset that contains categorical and numerical data. The descriptive analysis phase in processing such dataset will provide future researchers consideration to be made as they venture into their respective research. Finally, this research will outline the challenges and prepare consideration as this evaluations are carried out.

The structure of this research paper is as follows: an overview of existing research and approaches carried out which will be discussed in Section 2. Section 3 will discuss the selection process for the techniques and data used in this research. Next, Section 4 will then explain the implementation process followed by a short review of the evaluation process in Section 5. Last but not least, Section 6 will explain the findings and gaps where future research should be taken into consideration.

2 Related Work

Automated data integration is an area of research mentioned as early as 1967 (Marron and de Maine; 1967) when discussing 'information explosion' and introducing a system to "automatically decode the compressed information on an item-by-item basis when it is required". Fast forward 49 years where data lake, data pool and data fog are the hype -yet the same dilemma of optimal data storage and data retrieval persists. This is especially demanding due to the bottle neck of the available abundance of rich but raw data becoming actual information; and the current pace this information is derived at. As of 2015, the number of models developed to automate schema mapping had exceeded 34 with at least 71 scientific publications. These were documented by (Sutanta et al.; 2016) which gives a summary of the models ,approaches and challenges faced. Documented as the most recurring academically researched approach is the Linguistic based models followed closely by the Composite method and third is the Ontology based models. These approaches can be further classified as rule-based or learner-based approaches. The rule-based approach is driven by engineered rule that have been observed and specified such as those used in ontology classification and traditional software engineering. The learner-based approach is the current trend for research; utilizing machine learning concepts that are trained and may possibly be self-trained.

This research topic explores the application of machine learning specifically in the mapping of schema which has been researched by at least 17 identified models which may be divided into models that utilize only one specific machine learning method such as Delta (Clifton et al.; 1998), Semint (Li and Clifton; 1994), Autoplex (Berlin and Motro; 2001) and Dumas (Bilke and Naumann; 2005), to those that use a combination of models. These combination models can be further classified into two distinct types. First, the use of multiple learner-based approach; demonstrated by the models Dike (Palopoli et al.; 1999), XClust (Lee et al.; 2002), and Hybrid-RDR (Anam et al.; 2015) models. Secondly, a combination of learner-based and rule-based approaches as tested in the LSD (Doan et al.; 2001), Rondo (Melnik et al.; 2003), iMap (Dhamankar et al.; 2004) and SCM (Hoshiai et al.; 2004) models.

As this research proposes the utilization of the n-Gram model; in depth reference to the Dike, XClust and Hybrid-RDR models is discussed.

The Dike model applied the clustering technique to derive the similarities of schema objects on a strength of similarity as well as measurements of dissimilarity. The mapping process consisted of using lexicographical reference points to create a set of basic inter-schema properties to calculate the strength in similarity. Subsequently the distances between the objects are weighted and inter-schema properties are further refined and weighted again. Finally, an overall weighted mean of similarity is calculated where an acceptable threshold will remove objects considered dissimilar. The verification of accuracy of the output was on-going as the research was published and thus the level of

accuracy is unknown.

The XClust model similar to Dike uses the clustering technique with a difference that Dike is implemented on a relational database whilst XClust uses the XML path context i.e. document type definitions (DTD) to measure similarities. Similarity calculated by XClust is based on the level that two schema objects were designed in the XML. XClust like its predecessor models i.e. Cupid (Madhavan et al.; 2001) and LSD is able to take into account various cardinality of a schema mapping e.g. 1-1 and 1-n. Performance of the XClust is calculated by the percentage of wrong clusters that are derived after the incorrect threshold is surpassed.

The latest model known at the time of this research is the Hybrid-RDR that applies the J48 decision tree (a rewritten version of C4.5 in Java) to classify data. Nevertheless, other machine learning classification algorithms such as Naive Bayes or any other decision tree may be used. The learner-based technique implemented is the Ripple-Down Rule (RDR) that consist of the Censor Production Rules (CPR). The RDR CPR is an expert initiated and self learning rule constructing component that creates rules in the form 'If a case satisfies condition then do action unless the case does not satisfy the censor condition c' (Anam et al.; 2015, p.2) which is useful when there are many validations to be considered. Unlike other models where a large number of data is required, Hybrid-RDR is capable of retrieving a single classification with a small amount of data. The models researched have one thing in common that is the calculation of similarity distance in identifying the classification of mapping. For this research topic, the calculation of distance by dissimilarity will be evaluated using the Partitioning Around Medoids (PAM) algorithm that applies the Gowers dissimilarity coefficient which is discussed further in Section 3.

3 Methodology

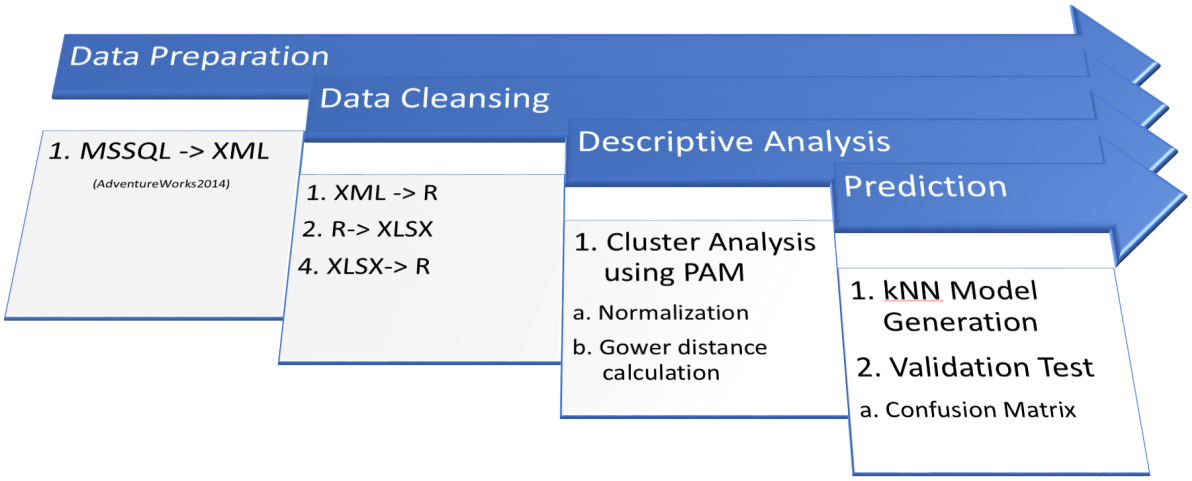

Figure 1: Overview of methodology

For this research, the main stages of study were mainly focused on the selection of data, data pre-processing and cleansing stage and finally the descriptive analysis and prediction stage. This initial research will utilize a small, with minimal attributes and consisting of a predefined matching that is accurate to ascertain the proof of concept. For

the same reason, the model to analyze the data should be able to function with limited number of input as well as limited attributes with a variety of data types.

The technology considered for processing and cleansing the data is R, Python and Weka. These three technologies are the most the widely used open source tools. In addition to the project time line factor, these were the ideal choices for the machine learning technology. The descriptive and analysis stage required algorithms that would be able to process test and predict data that contains both numeric and categorical data. These components need to be able to process multiple data type as well.

3.1 Data Selection

The data selection is driven by a need to obtain real-world simulation data that consists of an existing matching destination. The Adventure-Works2014 database aside from being easily accessible included an existing reliable mapping process to the Adventure-Works Data warehouse. Other alternatives considered were the North-wind database, and Purchase Orders data set made available by Hybrid-RDR research. Whilst the North-wind data set contains the desired volume and attribute features, it lacked a predefined matching destination. The data set made available by the Hybrid-RDR consisting of four XDR schema that may be downloaded initially from Biztalk[1] and now available at Database Group Leipzig[2]. The option to use this data set whilst being ideal in volume size and having an existing matching outcome was deterred by the time constraint of having to understand the transformation effort required for XDR formats. Thus leaving Adventure-Works data set as the selected data set. The data integration to validate the matching between source and destination is made available along with the database at CodePlex[3].

For this research the scope of the data set for matching is limited to person profile information only found in the Person table of AdventureWorks2014. This table is selected as it contains the various data type and has an ideal combination of fields that are both similar and dissimilar. Below is the summary of the elements of the selected dataset.

Data set Summary			
Tagged Target Fields	Element Name	Content Data Type	Count
DIM_CUSTOMER_BirthDate	birthdate	date	18205
DIM_CUSTOMER_FirstName	firstname	character	18313
DIM_CUSTOMER_Gender	gender	character	18369
DIM_CUSTOMER_LastName	lastname	character	18341
DIM_CUSTOMER_MaritalStatus	maritalstatus	character	18227
DIM_CUSTOMER_MiddleName	middlename	character	10544
DIM_CUSTOMER_NameStyle	namestyle	number	18294
UNK	businessentityid	number	18198
UNK	modifieddate	date	18367

UNK denotes fields that do not have a matching target.

[1] www.biztalk.org

[2] http://dbs.uni-leipzig.de/en/bdschemamatching

[3] https://msftdbprodsamples.codeplex.com/releases/view/125550

3.2 Machine Learning Model Selection

The proposition to use the n-Gram and k Nearest Neighbour(kNN) for this research is based on a series of factors considering the available data to schema match versus the consideration to also account for the lack of schema data to match and the possibility of limited data to train. In comparison to the existing work the volume of data is surmountable. For example, the Hybrid-RDR uses decision tree models to create a representation of rules that data points need to qualify before matching to a certain outcome. Therefore, the bigger the volume of data available, the more credible the outcome predicted will be. This underlying fact applies to almost all supervised and semi-supervised machine learning techniques.

Research to identify the various approaches to automate schema matching conducted (Rahm and Bernstein; 2001) in this research can be considered an individual matcher approach that is both an instance and element content based matcher. Trailing on this ripple of classification in Figure 2, the ideal models for this research are linguistic based.

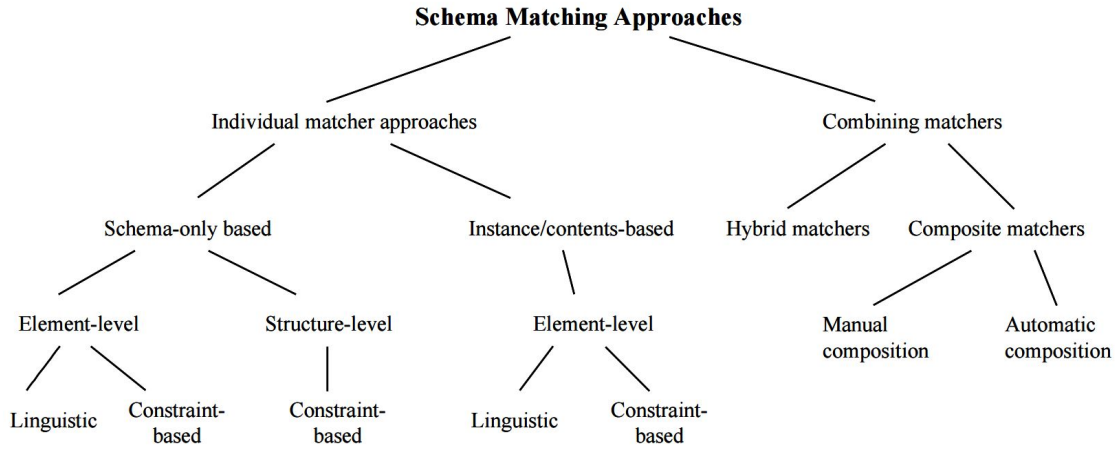

Figure 2: Classification of schema matching approaches
(Rahm and Bernstein; 2001)

Machine learning techniques that perform notably well for text analysis include Decision Tree, Neural Network (NN), Support Vector Machine (SVM), Naive Bayes, Hidden Markov chain (HM), Natural Language Processing (NLP) being the most solicited techniques. The characteristics of this domain would make it necessary that the model is able to accommodate new and untrained instances. This need narrows the selection of techniques to unsupervised models. In addition text analysis need three main elements: feature selection, classification system and similarity measure by (Vega; 2008). Possible methods considered were NLP's n-gram and Named Entity Recognition (NER), artificial Neural Network (aNN), text/web mining's Self-Organizing Map (SOM) and Naive Bayes. These methods are available in all the considered technology libraries. With consideration to the available time, the identification of the challenges faced were scoped to using only 2 of the techniques.

To re-cap , this research will need to be able to automate the matching of incoming data based on element name and element value. From the filtered list of techniques and so the final selection of technique is n-Gram for matching the element names along with its element values.

3.2.1 Element Profile Matching

Analysis on the element and its content will begin with implementing the n-Gram on the element's content. Alternate methods considered are lexical matching (Bernstein et al.; 2006), tokenization and synonym look-up using WordNet researched in (Anam et al.; 2016). n-Gram is a simple algorithm based on the Markov model that is able to process small to large amounts of data smoothly. This ease of scalability coupled with its simplicity are the main benefit of using n-Gram. Using n-Gram the input in the form of a string (i.e. an anagram) is compared to a subset of sequential characters i.e. n number of characters (hence the term n-gram). Motivation for using this tool in automating n-gram is based on promising results by (Karasneh et al.; 2009) that similarly used an attribute name matcher and a data type matcher on a relational database. This in addition to attribute name and data type, included matching for the overall schema, primary and foreign keys as well as other constraints. Application of n-gram in this research approaches the matching from a similarity of element name and content profile ranking. This means, the elements' profile information such as length of element name, average, maximum and minimum length of element content, actual n-gram breakdown are calculated. Various methods of application of n-grams real-world logic of its mechanics have been explored by (Brown et al.; 1992) and explains the intricate details of application. Another research by (Kondrak; 2005) accounts for how the similarity and distance may account for grouping matching in a mathematical approach. The outcome of this research will be to identify the similarity and dissimilarity distances of the elements in the source data file. Armed with this model, the hypothesis is that the applied model will be able to identify the possibly matched target column. As the next step after a successful classification, this research will apply a model to test the theory that elements, element content and profile information which in this research consists of both categorical and numeric data. A survey of technique of matching techniques using combination of data type was conducted by (Lee and Kim; 2010) using SVM, kNN and Bayesian Network models concluded that Bayesian Network classifiers performs better; being robust to noise and data loss as well as its computing efficiency. The only limitation that it poses to this research is that the data volum. As documented by (Pernkopf; 2004) for datasets with small sample sizes, kNN will perform better than Bayesian Network. Hence this research will use the kNN model. /par In the descriptive analysis stage the PAM algorithm will be used. PAM is an alternate clustering algorithm to k-means that from the given data set will calculate the dissimilarity matrix with the calculation reference point of clustering being its medoids. A thorough and in depth research conducted by (Park and Jun; 2009) entails the strengths and benefit of this algorithm the main factor for this research being the capability to calculate for heterogeneous data. In applying this algorithm, the distance metric used is the Gower dissimilarity coefficient¿ Unlike the widely used Euclidean and Manhattan metrics, the Gower is a calculation of dissimilarity that factors in distances of inertia, variance and weighted increase of variance (Gower and Legendre; 1986).

3.3 Training Data Preparation

The scope of data extracted from AdventureWorks2014 database is limited to basic customer profile information i.e. name (first, middle and last), gender, birth date and marital status. To add some randomness to the data , two fields i.e. business identity id and modified date were extracted as unknown data class. A SQL statement was executed to extract the fields in XML format.The result is then saved as an XML which is parsed

using R. The parser script will identify the element name and element value. Additional processing is done to identify the data type of the element. The schema information that may be available in the XML is not used to imitate the real-world scenario of a disperse source data i.e. in the case of a non-XML format that will not provide any meta data of the element being integrated. A simple library file is created to test the element value whether character, numeric or date. This code maybe extended to test if an element value fits email formats , phone number formats and any easily identifiable format that is unique.

Classification flags are inserted that identify the destination columns. This information is derived from the AdventureWorks Multi Dimensional Datawarehouse MSSQL Server Integration Services (SSIS) available on-line.In view that the AdventureWorks is a widely documented and used sample database, the accuracy of the mapping is reliable and correct.

The next step of the data preparation is to ensure that all the data point are suitable for analysis meaning the sparsity of data is coded for or in this case removed. As the method being use is n-gram, the element values and element names need to be cleaned beforehand by removing any space, special character and lower case all the letters to uniform standard.

Finally, the data set was stratified to a training and test data set.

3.4 Limitations and Strengths

The elements used in this research are small and share many traits. As this method of finding similarity based on the elements profile information - there will be elements that will be wrongly tagged due to the closeness of similarity i.e. firstname and middlename or dateofbirth and modifieddate. With more test, perhaps more profile information regarding the elements may surface which could prove useful in the calculation of distance. A re-cap of the objective which is to identify profile information and models that will aid in the matching of source data to the target. This approach of using distance to test which profile and model combination works best will be the strongest findings from this research. In addition, this approach is tested using the English language and a singular regional locale. A variety of languages and element names are not included in the data. The content of elements in the data set do not contain duplicates which are usually present in real-world data. Duplicate data will impact the calculation of the average content lengths and sway the models to a particular length. As for regional data, lengths of element content may be impacted. The content length for a particular region may be extremely long e.g. middle east first names that could consists of up to 2 or 3 names in comparison to the norm. Although this data is valid, it will need to be processed separately. In this research the data type identified is in the simplest combination of date, character and number. By being able to add information or categorizing the information further i.e. telephone number , social security number based on identifiable patterns, the approach will deliver a more reliable calculation of distances.

4 Implementation

A total of 18,484 customer profiles were extracted from the AdventureWorks2014's Person table. The selected field were BusinessIdentityID, FirstName, MiddleName, LastName, Namestyle, BirthDate, MaritalStatus, Gender and ModifiedDate. These records were

coerced into XML using the XML AUTO function in MSSQL Server. This completed the process of preparing an XML format input file. Next step is to upload the XML file for cleaning before performing the analysis. The parsing process was done in R during the importing step. A successful parsing depends on knowing the record schema tag which is fed into the parsing script manually. The parse then chunks this record and does a string breakdown into R. Results of this parser include not only the parsed record broken into the varies information of field names, and field types, but also records where field names vary between fields with one letter parsed to fields mashed together. The planned approach is to write a C DLL to identify the data types per record. Alternately, the assertive package could be used in R to identify character type and even categorize type of data. Option to use C as the conversion logic is due to the ease of customising the code based on region. The assertive package in R accommodates only the UK and US regions. Due to time constraints, the process of cleansing and data type identification was done in Excel. Also, tagging to the target schema was formulated in Excel. Finally, a cleansed and tagged data set summarized below is prepared.

Next,the file is uploaded in a comma separated format into Microsoft R Open(MRO) where profile information of the element and element content is calculated and made into matrices. The profile information derived is as Figure 3.

Data: prelim_class_profile

	row.names	fieldnames	field_datatype	avg_field	avg_fieldnames	max_field	max_fieldnames	min_field	min_fieldnames
1	6	birthdate	date	10.0000000	9	10	9	10	9
2	8	businessentityid	number	4.5746236	16	5	16	0	16
3	1	firstname	character	5.9397696	9	11	9	0	9
4	2	gender	character	0.9992923	6	1	6	0	6
5	3	lastname	character	5.5420097	8	17	8	0	8
6	4	maritalstatus	character	0.9995611	13	1	13	0	13
7	5	middlename	character	1.0051214	10	10	10	0	10
8	7	modifieddate	date	9.9702728	12	10	12	0	12
9	9	namestyle	number	0.9995627	9	1	9	0	9

Figure 3: Profile information of element and element contents

Then, the dataset is segregated by field; and the contents of each field from all rows are concatenated. The results being a long string of character (either character or numerical respective to the field being processed) is produced as seen in Figure 5. The string is then cleansed to remove spaces and punctuations if any exists. Finally, the string will be processed through the n-grams algorithm. This process is an exploratory process and will involve breaking the n-grams into either 2, 3 or 4 character sequences.

The results of the n-gram, the example here used the 2 character sequence n-gram, in Figure 5 is combined with the profile information and a final data set as Figure 4 is prepared. Descriptive analysis using PAM whilst applying the Gower distance calculation is done. Through this combination of factors the number of clusters can be identified using a scree plot. Finally, the kNN algorithm can be run with the number of clusters to train a model that will predict outcome on against the test dataset. As a validation step, a confusion matrix is produced to evaluate the accuracy and performance of the model. The process for n-gram is then run for it should be executed for the 3 and 4 character sequence.

```
[1] "ALVAREZWRIGHTROBINSONRUSSELLKUMARVANCEHILLHUGHESPATELCHENBUTLERRANARUBIOJONESBARNESRAMIREZMARTINJAIBAILEYALONSOWUPOSTIJENKINSMARTINEZCOLEMANLINADAMSXUA$
> n2bday
   [1] "A_L"  "L_V"  "V_A"  "A_R"  "R_E"  "E_Z"  "Z_W"  "W_R"  "R_I"  "I_G"  "G_H"  "H_T"  "T_R"  "R_O"  "O_B"  "B_I"  "I_N"  "N_S"  "S_O"$
  [22] "R_U"  "U_S"  "S_S"  "S_E"  "E_L"  "L_L"  "L_K"  "K_U"  "U_M"  "M_A"  "A_R"  "R_V"  "V_A"  "A_N"  "N_C"  "C_E"  "E_H"  "H_I"  "I_L"$
  [43] "H_U"  "U_G"  "G_H"  "H_E"  "E_S"  "S_P"  "P_A"  "A_T"  "T_E"  "E_L"  "L_C"  "C_H"  "H_E"  "E_N"  "N_B"  "B_U"  "U_T"  "T_L"  "L_E"$
  [64] "R_A"  "A_N"  "N_A"  "A_R"  "R_U"  "U_B"  "B_I"  "I_O"  "O_J"  "J_O"  "O_N"  "N_E"  "E_S"  "S_B"  "B_A"  "A_R"  "R_N"  "N_E"  "E_S"$
  [85] "A_M"  "M_I"  "I_R"  "R_E"  "E_Z"  "Z_M"  "M_A"  "A_R"  "R_T"  "T_I"  "I_N"  "N_J"  "J_A"  "A_I"  "I_B"  "B_A"  "A_I"  "I_L"  "L_E"$
 [106] "A_L"  "L_O"  "O_N"  "N_S"  "S_O"  "O_W"  "W_U"  "U_P"  "P_O"  "O_S"  "S_T"  "T_I"  "I_J"  "J_E"  "E_N"  "N_K"  "K_I"  "I_N"  "N_S"$
 [127] "A_R"  "R_T"  "T_I"  "I_N"  "N_E"  "E_Z"  "Z_C"  "C_O"  "O_L"  "L_E"  "E_M"  "M_A"  "A_N"  "N_L"  "L_I"  "I_N"  "N_A"  "A_D"  "D_A"$
 [148] "S_X"  "X_U"  "U_A"  "A_R"  "R_U"  "U_N"  "N_A"  "A_N"  "N_D"  "D_E"  "E_R"  "R_S"  "S_E"  "E_N"  "N_S"  "S_U"  "U_N"  "N_W"  "W_A"$
 [169] "S_O"  "O_N"  "N_H"  "H_A"  "A_L"  "L_L"  "L_W"  "W_A"  "A_R"  "R_D"  "D_S"  "S_H"  "H_E"  "E_N"  "N_M"  "M_U"  "U_R"  "R_P"  "P_H"$
 [190] "L_O"  "O_P"  "P_E"  "E_Z"  "Z_T"  "T_H"  "H_O"  "O_M"  "M_A"  "A_S"  "S_F"  "F_L"  "L_O"  "O_R"  "R_E"  "E_S"  "S_M"  "M_O"  "O_O"$
 [211] "E_W"  "W_R"  "R_I"  "I_G"  "G_H"  "H_T"  "T_Z"  "Z_H"  "H_A"  "A_N"  "N_G"  "G_P"  "P_E"  "E_R"  "R_E"  "E_Z"  "Z_G"  "G_A"  "A_R"$
 [232] "I_A"  "A_Z"  "Z_H"  "H_A"  "A_N"  "N_G"  "G_L"  "L_E"  "E_E"  "E_S"  "S_A"  "A_R"  "R_A"  "A_H"  "H_A"  "A_L"  "L_L"  "L_L"  "L_O"$
 [253] "G_G"  "G_O"  "O_E"  "E_L"  "L_Z"  "Z_H"  "H_A"  "A_N"  "N_G"  "G_L"  "L_U"  "U_O"  "O_B"  "B_U"  "U_T"  "T_L"  "L_E"  "E_R"  "R_P"$
 [274] "I_C"  "C_E"  "E_C"  "C_O"  "O_O"  "O_P"  "P_E"  "E_R"  "R_R"  "R_A"  "A_N"  "N_A"  "A_R"  "R_U"  "U_S"  "S_S"  "S_E"  "E_L"  "L_L"$
 [295] "L_O"  "O_N"  "N_S"  "S_O"  "O_X"  "X_U"  "U_A"  "A_N"  "N_A"  "A_N"  "N_D"  "D_M"  "M_I"  "I_T"  "T_C"  "C_H"  "H_E"  "E_L"  "L_L"$
 [316] "L_V"  "V_A"  "A_R"  "R_E"  "E_Z"  "Z_F"  "F_O"  "O_S"  "S_T"  "T_E"  "E_R"  "R_H"  "H_A"  "A_Y"  "Y_E"  "E_S"  "S_C"  "C_O"  "O_X"$
 [337] "E_L"  "L_S"  "S_O"  "O_N"  "N_W"  "W_R"  "R_I"  "I_G"  "G_H"  "H_T"  "T_G"  "G_A"  "A_R"  "R_C"  "C_I"  "I_A"  "A_R"  "R_A"  "A_J"$
 [358] "E_D"  "D_W"  "W_A"  "A_R"  "R_D"  "D_S"  "S_B"  "B_R"  "R_O"  "O_W"  "W_N"  "N_T"  "T_A"  "A_Y"  "Y_L"  "L_O"  "O_R"  "R_M"  "M_A"$
```

Figure 4: Concatenated string before and after the 2 n-gram processing

	row.names	fieldnames	x	field_datatype	avg_field	avg_fieldnames	max_field	max_fieldnames	min_field	min_fieldnames
453	2367	birthdate	6_5	date	10	9	10	9	10	9
454	2371	birthdate	1_3	date	10	9	10	9	10	9
455	2373	birthdate	1_9	date	10	9	10	9	10	9
456	2374	birthdate	9_7	date	10	9	10	9	10	9
457	2382	birthdate	9_6	date	10	9	10	9	10	9
458	2384	birthdate	2_0	date	10	9	10	9	10	9
459	2389	birthdate	1_9	date	10	9	10	9	10	9
460	2393	birthdate	1_1	date	10	9	10	9	10	9
461	2394	birthdate	1_2	date	10	9	10	9	10	9
462	2402	firstname	C_H	character	5.93977	9	11	9	0	9
463	2404	firstname	A_R	character	5.93977	9	11	9	0	9
464	2420	firstname	R_C	character	5.93977	9	11	9	0	9
465	2429	firstname	A_S	character	5.93977	9	11	9	0	9
466	2432	firstname	I_N	character	5.93977	9	11	9	0	9
467	2442	firstname	H_A	character	5.93977	9	11	9	0	9
468	2444	firstname	N_N	character	5.93977	9	11	9	0	9
469	2446	firstname	O_N	character	5.93977	9	11	9	0	9
470	2451	firstname	N_C	character	5.93977	9	11	9	0	9
471	2462	firstname	U_L	character	5.93977	9	11	9	0	9

Figure 5: Final dataset that will be processed kNN

5 Evaluation

Confusion matrix is a contingency table that calculates overall accuracy based on two components. There are two elements: "Actual" and "Predicted" that calculate the average values of the outcomes i.e.: Negative results (N) or Positive results (P). This validation technique is widely used for prediction models and accurately provides a summary of the classifier's performance.

5.1 Two Character Sequence n-Gram

Figure 6 shows the result of the identified clusters for the initial case study that consisted of 3 main elements i.e.: birthdate, firstname and gender and Figure 7 for 7 elements i.e.: birthdate, firstname, gender,lastname, middlename, namestyle and maritalstatus. The kNN performed well in identifying the various data types and the matches. As expected with applying the additional elements there exists an overlap of elements (i.e. cluster 7 and 9)

The confusion matrix for these cases as seen in Figure 8 rightly shows one hundred percent accuracy.

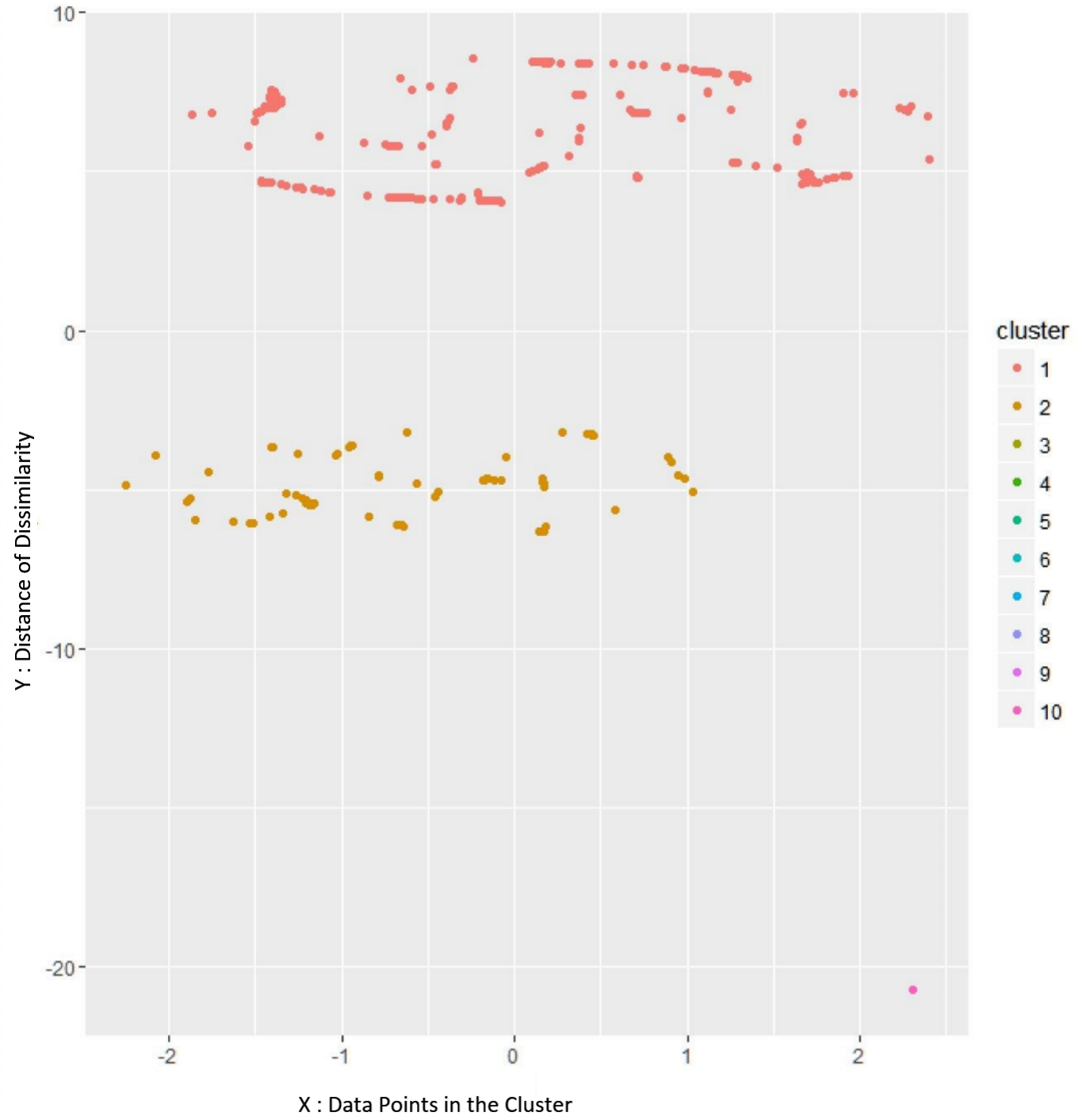

Figure 6: GGPlot of Cluster Proximity

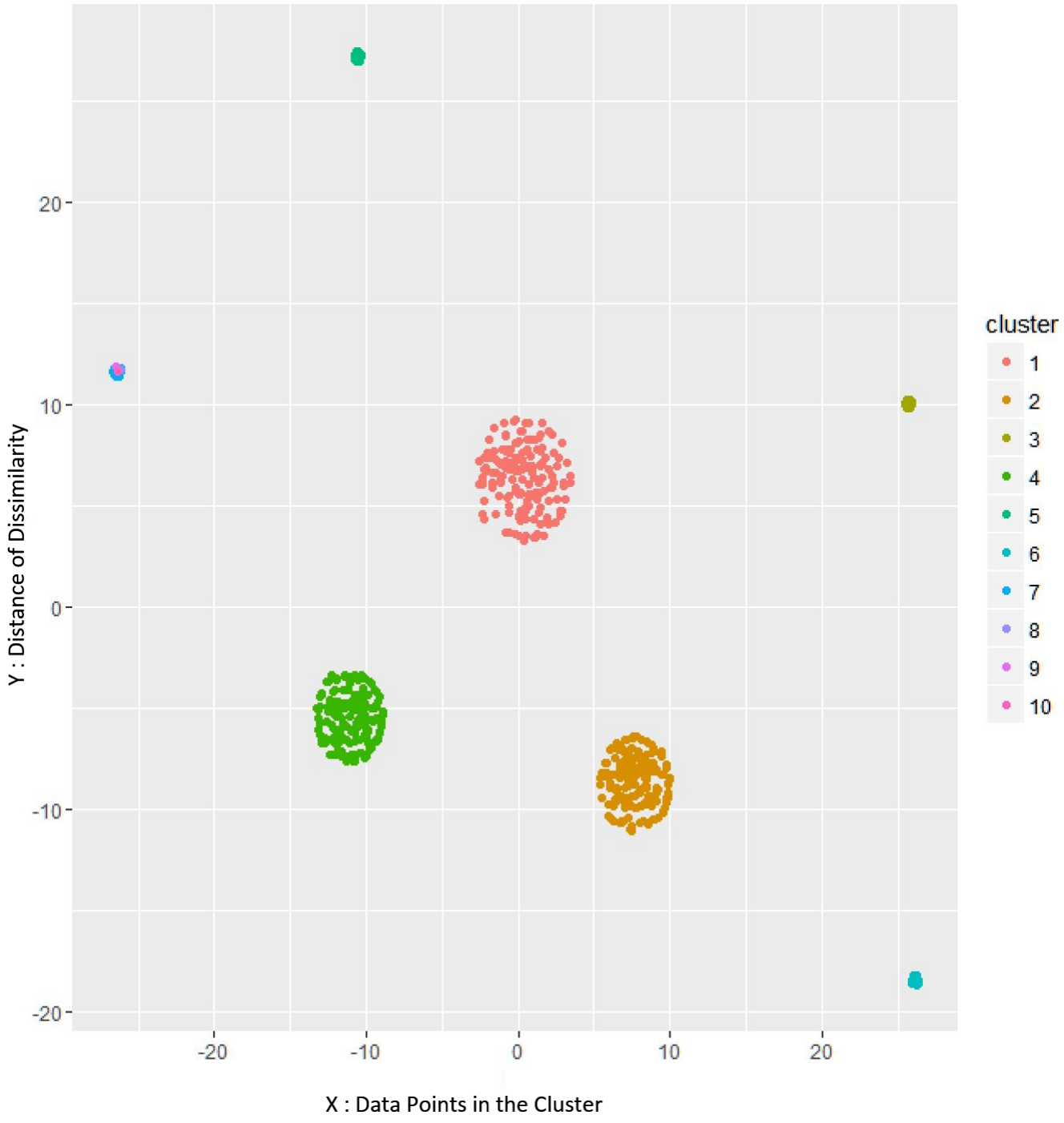

Figure 7: GGPlot of Cluster Proximity

```
=== Summary ===

Correctly Classified Instances         879               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0.0082 %
Root relative squared error              0.0114 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)      12.5    %
Total Number of Instances              879

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   <-- classified as
 487   0   0   0   0   0   0   0 |   a = DIM_CUSTOMER_BirthDate
   0 340   0   0   0   0   0   0 |   b = DIM_CUSTOMER_FirstName
   0   0  52   0   0   0   0   0 |   c = DIM_CUSTOMER_Gender
   0   0   0   0   0   0   0   0 |   d = DIM_CUSTOMER_LastName
   0   0   0   0   0   0   0   0 |   e = DIM_CUSTOMER_MaritalStatus
   0   0   0   0   0   0   0   0 |   f = DIM_CUSTOMER_MiddleNAme
   0   0   0   0   0   0   0   0 |   g = DIM_CUSTOMER_NameStyle
   0   0   0   0   0   0   0   0 |   h = UNK
```

Figure 8: Confusion Matrix Summary for 3 Elements

```
=== Summary ===

Correctly Classified Instances         475               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0.0059 %
Root relative squared error              0.0126 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)      12.5    %
Total Number of Instances              475

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   <-- classified as
 186   0   0   0   0   0   0   0 |   a = DIM_CUSTOMER_BirthDate
   0 116   0   0   0   0   0   0 |   b = DIM_CUSTOMER_FirstName
   0   0  17   0   0   0   0   0 |   c = DIM_CUSTOMER_Gender
   0   0   0 106   0   0   0   0 |   d = DIM_CUSTOMER_LastName
   0   0   0   0  18   0   0   0 |   e = DIM_CUSTOMER_MaritalStatus
   0   0   0   0   0  15   0   0 |   f = DIM_CUSTOMER_MiddleNAme
   0   0   0   0   0   0  17   0 |   g = DIM_CUSTOMER_NameStyle
   0   0   0   0   0   0   0   0 |   h = UNK
```

Figure 9: Confusion Matrix Summary for 7 Elements

5.2 Discussion

In the process of testing, the research faced some challenges due to limited memory on the machine and so not the full extent of data was analysed. Hence, the experiments are still on-going with plans to move into the cloud environment. Nevertheless, based on this result we can clearly say that if the dissimilarity of elements are clear, the use of machine learning does help immensely in the identification of the target destination. The overall results can be argued is currently over-fitted for the task at hand which is undeniable. Continued test that will be performed will be able to provide more insight to the use of element and the the profile information especially as the element profile information start overlapping. This with future test , should start reducing the accuracy rate and show that the classifiers are not performing as well.

6 Conclusion and Future Work

Results from this research indicate that the element and contents profile information is useful but not across all cases. In the dissimilarity of data type, it is useful in identifying which are the possible matches where the information is very disperse. In the similarity calculation , we find that information that are closely similar such as names which are broken down i.e. the overlap of clusters 7 and 9 in Figure 7 will not find profile information useful for mapping to the target destination. Although, the answer to this question bring us to question- in a data warehouse, do we really need to be segregating the difference of names? In an environment where closely similar information such as these are not analyzed, data warehouse designers may want to consider bringing in the information as a whole. The results in this research exhibited indications of over-fitting with a perfect accuracy rate. To validate this, a cross-validation needs to be run.

During the implementation of this research, the challenges faced can be classified as technical and time-line challenges. First and foremost, many of the research papers whilst explaining the research results and findings, failed to provide the gap between attaining the data set as is and preparation steps to be analytically ready. This challenge resulted in wayward analysis on some very wrong data presentations which resulted is many errors and unintelligible results. Next, the time taken to explore the multiple options in using any single analysis package for the analysis can be overwhelming. This being an advantage and a disadvantage during this research because of the many possibilities and fine-tuning in process that effects the results. Furthermore, the number of packages available that are similar but having their own strengths and weaknesses - requires extensive hands-on trial and error before being able to create in-depth analysis to answer the research question. As the quote goes:

> *"For the things we have to learn before we can do them, we learn by doing them."- Aristotle*

The continuation of this research should first explore the distance calculation used to identify similarity and dissimilarity of matches. This research only explored the Gower distance and the Partition Around Medoids (PAM) of the R package Cluster. The heart of the analysis is identifying a good distance metric. In addition, this research required a package to be able to manage both categorical and numeric type data in the calculation of distance. For future research, the processing of this data types separately and identifying a calculation of unifying them may provide a better distance calculation. The performance

of segregated and combined calculation versus distance calculation that encompass all data types should also be researched in the identification of a better matching approach. Subsequently to that, this research would be able to then be extended using elements from different regions and languages.

Acknowledgements

In performing this research project , I am very thankful and appreciative for the endless guidance and help provided by Dr. Paul Hayes, Mr. Michael Bradford and Dr. Simon Caton. I am also thankful for the divine powers that be and to my family for their constant support and vote of confidence.

References

Anam, S., Kim, Y. S., Kang, B. H. and Liu, Q. (2015). Schema mapping using hybrid ripple-down rules, *the Thirty-Eighth Australasian Computer Science Conference, ACSC*, pp. 17–26.

Anam, S., Kim, Y. S., Kang, B. H. and Liu, Q. (2016). Adapting a knowledge-based schema matching system for ontology mapping, *Proceedings of the Australasian Computer Science Week Multiconference*, ACM, p. 27.

Berlin, J. and Motro, A. (2001). Autoplex: Automated discovery of content for virtual databases, *International Conference on Cooperative Information Systems*, Springer, pp. 108–122.

Bernstein, P. A., Madhavan, J. and Rahm, E. (2011). Generic schema matching, ten years later, *Proceedings of the VLDB Endowment* **4**(11): 695–701.

Bernstein, P. A., Melnik, S. and Churchill, J. E. (2006). Incremental schema matching, *Proceedings of the 32Nd International Conference on Very Large Data Bases*, VLDB '06, VLDB Endowment, pp. 1167–1170.
URL: *http://dl.acm.org/citation.cfm?id=1182635.1164235*

Bilke, A. and Naumann, F. (2005). Schema matching using duplicates, *21st International Conference on Data Engineering (ICDE'05)*, IEEE, pp. 69–80.

Brown, P. F., deSouza, P. V., Mercer, R. L., Pietra, V. J. D. and Lai, J. C. (1992). Class-based n-gram models of natural language, *Comput. Linguist.* **18**(4): 467–479.
URL: *http://dl.acm.org/citation.cfm?id=176313.176316*

Clifton, C., Housman, E. and Rosenthal, A. (1998). Experience with a combined approach to attribute-matching across heterogeneous databases, *Data Mining and Reverse Engineering*, Springer, pp. 428–451.

Dhamankar, R., Lee, Y., Doan, A., Halevy, A. and Domingos, P. (2004). imap: discovering complex semantic matches between database schemas, *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, ACM, pp. 383–394.

Doan, A., Domingos, P. and Halevy, A. Y. (2001). Reconciling schemas of disparate data sources: A machine-learning approach, *ACM Sigmod Record*, Vol. 30, ACM, pp. 509–520.

Gower, J. C. and Legendre, P. (1986). Metric and euclidean properties of dissimilarity coefficients, *Journal of classification* **3**(1): 5–48.

Hoshiai, T., Yamane, Y., Nakamura, D. and Tsuda, H. (2004). A semantic category matching approach to ontology alignment, *Proceedings of the 3rd International Workshop on Evaluation of Ontology Based Tools (EON 2004). CEUR-WS Publication*, Citeseer, pp. 437–447.

Karasneh, Y., Ibrahim, H., Othman, M. and Yaakob, R. (2009). Integrating schemas of heterogeneous relational databases through schema matching, *Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services*, ACM, pp. 209–216.

Kondrak, G. (2005). N-gram similarity and distance, *International Symposium on String Processing and Information Retrieval*, Springer, pp. 115–126.

Lee, M. L., Yang, L. H., Hsu, W. and Yang, X. (2002). Xclust: clustering xml schemas for effective integration, *Proceedings of the eleventh international conference on Information and knowledge management*, ACM, pp. 292–299.

Lee, N. and Kim, J.-M. (2010). Conversion of categorical variables into numerical variables via bayesian network classifiers for binary classifications, *Computational Statistics & Data Analysis* **54**(5): 1247–1265.

Li, W.-J., Wang, K., Stolfo, S. J. and Herzog, B. (2005). Fileprints: Identifying file types by n-gram analysis, *Proceedings from the Sixth Annual IEEE SMC Information Assurance Workshop*, IEEE, pp. 64–71.

Li, W.-S. and Clifton, C. (1994). Semantic integration in heterogeneous databases using neural networks, *VLDB*, Vol. 94, pp. 12–15.

Madhavan, J., Bernstein, P. A. and Rahm, E. (2001). Generic schema matching with cupid, *VLDB*, Vol. 1, pp. 49–58.

Marron, B. A. and de Maine, P. A. D. (1967). Automatic data compression, *Commun. ACM* **10**(11): 711–715.
URL: *http://doi.acm.org/10.1145/363790.363813*

Melnik, S., Rahm, E. and Bernstein, P. A. (2003). Rondo: A programming platform for generic model management, *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, ACM, pp. 193–204.

Nin, J., Tous, R. and Delgado, J. (2014). Variable linkage for multimedia metadata schema matching, *Multimedia tools and applications* **68**(3): 845–861.

Palopoli, L., Saccá, D., Terracina, G. and Ursino, D. (1999). A unified graph-based framework for deriving nominal interscheme properties, type conflicts and object cluster similarities, *Cooperative Information Systems, 1999. CoopIS'99. Proceedings. 1999 IFCIS International Conference on*, IEEE, pp. 34–45.

Park, H.-S. and Jun, C.-H. (2009). A simple and fast algorithm for k-medoids clustering, *Expert Systems With Applications* **36**(2): 3336–3341.

Pernkopf, F. (2004). Bayesian network classifiers versus k-nn classifier using sequential feature selection, *AAAI*, pp. 360–365.

Rahm, E. and Bernstein, P. A. (2001). A survey of approaches to automatic schema matching, *The VLDB Journal* **10**(4): 334–350.

Sutanta, E., Wardoyo, R., Mustofa, K. and Winarko, E. (2016). Survey: Models and prototypes of schema matching, *Intl Journal of Electrical and Computer Engineering (IJECE)* **6**(3).

Vega, J. (2008). Intelligent methods for data retrieval in fusion databases, *Fusion Engineering and Design* **83**(2): 382–386.