Next Generation Data Classification and Linkage

Role of Probabilistic Models and Artificial Intelligence

Gayan Prasad Hettiarachchi
Department of Physics
Osaka University
Osaka, Japan

Dhammika Suresh Hettiarachchi
Department of Electrical Engineering and Information Systems
The University of Tokyo
Tokyo, Japan

Nadeeka Nilmini Hettiarachchi
Department of Population Genetics
National Institute of Genetics
Mishima, Japan

Azusa Ebisuya
Graduate School of Economics
Osaka University
Osaka, Japan

Abstract— **Data classification and linkage is the task of identifying information corresponding to the same entity from one or more data sources. Methods used to tackle data classification and linkage problems fall into two broad categories. One commonly used method is deterministic models, in which sets of often very complex rules are used to classify pairs of entities as links. The other is the probabilistic model, in which statistical or probabilistic approaches are used to classify pairs. However, these models fail to deliver when there are lots of missing values, typographical errors, non-standardized entities, etc. To this end, intelligent routines making use of artificial neural networks, genetic algorithms and clustering algorithms can provide the next generation models for data classification and linkage. An introduction to data linkage, impact on humanity and community, current models, associated pitfalls, new directions and issues both technical and social for next generation data classification and linkage systems are discussed using an example prototype. A new model for linkage is proposed, where it is highlighted that not only the relationships between attributes of different entities, but also identification of relationships within the attributes of an entity is important in handling missing values and can provide better accuracy.**

Keywords— Big data; classification; data linkage; machine learning; phonetic matching; probabilistic models; string comparison

I. Introduction

The quality of data residing in a data source gets degraded and leads to misinterpretation of information due to a multitude of factors. Such factors vary from poor design (update anomalies due to lack of normalization), lack of standards for recording data, to typographical errors (lexicographical errors, character transpositions). Data of such poor quality could result in many damages being caused, for example, in a business application; sending products and invoices to the wrong customer, sending wrong products or bills to customers, inability to locate customers, generating wrong statistics, generating wrong predictions, etc. In such situations, it is important to identify duplicates and merge them into a single entity, i.e., identify whether two or more entities are approximately the same and produce a single entity by making best use of information contained in redundant locations/entities [1]. Real world entities of interest include individuals, families, organizations, geographic regions, etc., while applications of data linkage are in areas such as marketing, customer relationship management (CRM), law enforcement, fraud detection, epidemiological studies and administration, just to name a few [2].

Methods used to tackle data linkage problems fall into two broad categories. One commonly used method is deterministic models, in which sets of often very complex rules or production systems are used to classify pairs of records as links (that is relating to the same entity). The other is the probabilistic model, in which statistical or probabilistic approaches are used to classify record pairs [3].

Although numerous attempts are being made to address the issue of data linkage, many inherent drawbacks are found in those approaches. “Missing values” is one of the major problems encountered by researchers focusing on frequency based matching, which is none other than probabilistic models [4]. Therefore, the requirement arises for a novel methodology for data linkage, which sets sight beyond mere probabilistic and deterministic models.

Recent developments in computational techniques enable researchers to move from classical probabilistic models to newer and advanced approaches using maximum entropy, machine-learning techniques such as artificial neural networks (ANN), genetic algorithms (GA) and phonetic matching. The aim of this paper is to introduce how this can be achieved and bring to light technical and social issues that need to be considered in the process.

II. Impact on Humanity and Community

With the rapid advancements in science and technology, globalization and the web, vast amounts of information is added to data resources each second all around the globe. Whether it is medical data pertaining to patients, personal information of clients, administrative information, data used by law enforcement agencies, or news reports of global and local events, analysis of this big data becomes a major requirement in all aspects. This requirement is further complicated by the fact that data is replicated or intentionally stored at multiple data resources. Finding links in data,

978-1-4799-7193-0/14/$31.00 ©2014 IEEE

connecting them together to represent single entities, mining the big data and producing useful information have been troubling academic researchers and industry giants alike for decades.

An example, where data linkage can be useful in a humanitarian perspective would be the linking of patient information. Information pertaining to a particular patient may be stored in different isolated data sources over a period of time, for example, data sources belonging to different hospitals or medical clinics. There are occasions when the patients' complete medical history is of utmost importance in elucidating a current illness and prescribing treatment. In producing the complete history of a patient, different data sources will have to be accessed and data pertaining to the patient will have to be uniquely identified and linked. This can become more complicated, especially in developing countries, where the health care systems are still progressing or at an infant stage. In such countries, the information pertaining to a patient is stored using the patient name or a locally designed identifier, in contrast to a nationally valid health insurance number, for example. In such a situation, correctly identifying whether two or more records belong to the same patient in the absence of a nationally valid unique identifier can be very difficult. Even in the presence of a nationally valid unique identifier, combining multiple records belonging to a single patient to provide a complete history of the patient in a simple and summarized manner is extremely difficult, if not impossible. A complex and intelligent data linkage system can help make the initial steps towards achieving such a goal.

Another extreme but plausible example where data linkage can be of help in humanitarian efforts is during a natural or man-made disaster. During a disaster, it is often the case that members belonging to a single family are disoriented and dislocated. Members of the same family maybe located at different relief camps, like in the Great East Japan Earthquake and Tsunami in 2011. If a system can be in place that can sort through the members present at different relief camps, and group and link together members belonging to a single family, not only it could help individuals be relieved from anxiety related stress but also help the relief workers focus their invaluable time and efforts on other relief related tasks, which would otherwise be consumed by search efforts to locate individuals that are present at a different location.

Apparently zettabytes of data generated by ubiquitous sensors, mobile and computing devices during a major humanitarian crisis like war, earthquake or famine should be analyzed in order to find answers that could help minimize negative effects of another future event and to maximize the current humanitarian response. A curated database that focuses on zero replication, quality and timeliness is a mandatory pre-requisite to an effective response. Replication may lead to redundant humanitarian efforts resulting in wastage of resources. In order to minimize replication, a sound record linkage system that can extract only the relevant data from multiple electronic sources is necessary.

One of the major drawbacks of current data linkage systems is the lack of capability for extensibility. Most systems are custom built to address a particular problem domain, for example, medical record linkage [3]. Software that are tailor made to handle medical record linkage will not support and cannot be used for any other classification and linkage problem, such as linking of administrative records. Therefore, to start with, the requirement is evident for a data linkage framework that can be easily applied to different problem domains.

Missing values of attributes is a major problem faced by systems based on probabilistic and deterministic models. Enforcing a generic framework for data linkage with capability to handle missing values is of utmost importance. Therefore, the requirement arises for a novel methodology for data linkage that can utilize machine learning algorithms for handling missing values.

The identifying characteristics or attributes of entities are not always clearly defined and represented as well organized records ready for comparison, especially when data are stored as raw text files, for example, newspaper articles related to certain events. Being able to obtain attributes of interest from the raw data, for comparison with other entities would be of value for reducing human intervention and thereby boosting performance while minimizing human error, for example, typographical errors.

Phonetic string/text matching plays a major role in any data classification and linkage system. Even though there are different methods in the literature that address the issue of phonetic string matching, their performance and accuracy can vary depending on the language and its language components, making it clear that there is no exact algorithm for deriving the likely sound of a string. [5]. Development of strong phonetic matching algorithms are essential and useful, however, an algorithm that can change its decision rules based on the language would be ideal. To this end, a statistical approach can be adopted to match string attributes. In the literature there are table books that provide phonemes, phoneme strings and spellings that include the corresponding sounds, but these sources do not provide statistics of the likelihood of correspondence [6]. An alternative approach is to use a software dictionary that provides the spellings and pronunciation of words. The sounds or the pronunciation given in the dictionary provides an alternative approach to phonetic matching. The purpose of phonetic matching can be directly substituted by this method. However, the generic framework can include algorithms for phonetic matching, distance based matching and dictionary based matching. An efficient phonetic string matching algorithm undeniably facilitates global humanitarian organizations working across diverse cultures, languages and demographics.

These were the factors that sparked motivation in carrying out this preliminary research, since all of the ideal requirements described above are yet to be established and experimental research of the same are scarce. In addition, the fact that the research tries to cater to a broad user community ranging from pure academic to general use in humanitarian and nonprofit organizations speaks of the importance of the study in helping local and international communities.

III. Data linkage models

Before moving on to the research problems at hand, it is useful to explore two of the probabilistic data linkage models most widely used today.

A. Newcombe's model

Newcombe's model was based on two basic but important decision rules. The first was that the relative frequency of occurrence of a value such as a surname among matches and non-matches could be used in computing a weight or score associated with the matching of two records. The second was the scores calculated over different fields such as surname, first name, age, etc. They could be added to obtain an overall matching score [7], [8]. More specifically, emphasis was on odds ratios that are shown below,

$$\log_2(p_L) – \log_2(p_F) \ (1)$$

where pL is the relative frequency among matches (links) and pF is the relative frequency among non-matches (non-links). Since the true matching status is often not known, an approximate for the above odds ratio was introduced.

$$\log_2(p_R) – \log_2(p_R)^2 \ (2)$$

where p_R is the frequency of a particular string (first name, initial, birthplace, etc.). Whenever a large universe file is matched with itself, the second ratio provides a very good approximate of the first one [7].

B. Fellegi and Sunter model

Fellegi and Sunter introduced a formal mathematical foundation for record linkage in 1969. The proposed methodology was designed to match two files *A* and *B* by considering all the possible records that can be generated through the cross product of the two files [3]. The idea is to classify pairs in a product space *A* X *B* into *M*, the set of matches, and *U*, the set of non-matches [2]. Fellegi and Sunter, making use of rigorous concepts introduced by Newcombe, came up with ratios of probabilities of the form,

$$R=P(\gamma\varepsilon\Gamma|M)/P(\gamma\varepsilon\Gamma|U) \ (3)$$

Where γ is an arbitrary agreement pattern in a comparison space given by Γ. For instance, the comparison space might consist of eight patterns representing simple agreement or disagreement (binary values) on three attributes such as, the person name, street name, and city. The ratio *R* or any monotonically increasing function of *R*, such as the natural logarithm is referred to as a matching weight (score) [9].

C. Problems associated with these models

The linkage models described above can perform well when there are little typographical errors and other forms of non-homogeneity between the files being matched. The methods may not work well due to failures of the assumptions used in the models, lack of sufficient variables for matching, sampling or lack of overlap between files, and extreme variations such as typographical errors and missing values [2]. Each of the following types of errors provides examples of situations where pairs of entities will not have homogeneous identifying characteristics and renders the aforementioned probabilistic models inadequate, demanding for a novel methodology for data classification and linkage.

- Records that are not standardized, for example names, addresses, etc.,.
- Records with a lot of missing values.
- Records that do not have easily comparable fields or unprocessed raw text files.
- Records having a lot of typographical errors.

D. Methods to overcome such drawbacks

The lack of standardized fields leads to difficulty in comparing entities. A good solid approach for standardizing attribute values is a primary requirement as an initial step before moving onto the more rigorous and complicated classification and linking tasks. A good data linkage system would contain procedures for comparing all the attribute values against a set standard and making necessary modifications to the values to conform to that standard.

Missing values stand out as one of the major issues that has a direct and crucial impact on the accuracy of data linkage. One approach would be to introduce a model that can deal with missing values instead of ignoring them, i.e., try to predict missing values based on some training dataset. In order to accomplish this, it is required to identify what attributes of an entity defines or best describes the missing attribute, i.e., identification of the independent variables. This may be performed using a machine learning/artificial intelligent procedure such as GAs. Once the independent variables are identified, training can be performed using a training dataset for which the value of the dependent variable is available. This can be performed using, for example, ANNs. The trained ANN and its parameters (weights) maybe used for predicting the missing values of the attribute of interest based on the independent variables. However, it should be noted, that prediction of missing values is not always possible for all kinds of attributes, for example, simply put, prediction of a person's name based on other attributes would be impossible, let alone irrational. But prediction of a person's monthly income, for example, could be possible based on attributes such as, occupation, age, highest qualification, etc.

Handling of records that do not have easily comparable fields or unprocessed raw text is complicated. This falls under the category of natural language processing (NLP) and text summarization and is still a very complicated and slowly progressing field. Some of the aspects of NLP that would be essential in data linkage are text summarization, named entity recognition, co-reference resolution and relationship extraction. Statistical machine learning algorithms such as decision trees and statistical models using soft probabilistic decisions could play the role of a starting point to further design, improve, and make additions upon.

One obvious approach to avoid typographical errors would be to compare attribute values as they are entered into the system against the entries of a dictionary. By employing such an approach it is possible to notify of any misspellings during data entry. However, this approach alone does not suffice for all the aspects a data linkage system is expected to cover. For example, when a large quantity of data is imported from an already existing file, it is expected that a linkage system would be able to handle already present typographical errors. Approximate string matching comes in handy in such situations, enabling matching of two strings even in the presence of typographical errors. Approximate string matching can be performed in several ways, such as, distance based approaches, phonetic or sound based matching or a combination of both that revealed an improvement in accuracy when applied on text containing Sinhalese words and names [5]. However, phonetic matching algorithms that are designed to work with a particular language, i.e., makes use of language components, such as, phonemes and consonants pertaining to a particular language might fail or result in a reduction in accuracy in the face of a different language containing different language components.

IV. Proposal for a new data linkage model and a generic framework

As discussed earlier, it is necessary for next generation data linkage systems to be able to handle tasks pertaining to different problem domains. Therefore, it is required for a system to be extensible and provide generic functions to allow users to build upon it application specific requirements with minimal modifications. To this end, the design of a framework containing generic classes that address the implementations of requirements mentioned earlier, together with an application programming interface (API) which allows users to interact with the framework, is essential. A basic design diagram of the system is shown in Fig. 1.

The framework need to include classes for tasks such as, data cleaning and standardization, classification and prediction, NLP, string comparison, and linking. The API will provide an interface for users to interact with the framework and make use of the classes in the framework to implement their data linking applications. The classes may have different implementations to facilitate a wide variety of requirements and allow users to test out different techniques and improve the accuracy of the linking task. The general flow of a next generation data linking application is shown in Fig. 2.

In order to validate the suggestions and proposals made for a next generation data linkage model that makes use of both probabilistic and artificial intelligent routines, a prototype was built. However, at this point the routines for the text analysis task illustrated in Fig. 2 are not complete. Routines and classes for other tasks shown in Fig. 2 are available, although there is ample room for improvements and further additions. In the following section, an example application built on top of the framework will be introduced and a comparison will be provided between the results produced by probabilistic routines only and the results produced by the new model for data linkage.

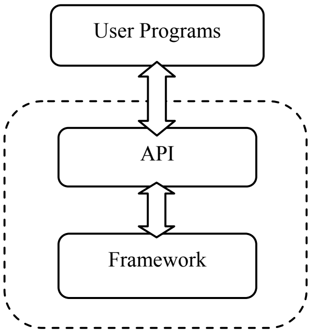


Fig. 1. An abstract design diagram of a next generation data linkage system. The framework provides classes that implement the requirements described in the earlier section that are sought after in next generation data classification and linkage systems. The API provides users with the facility to modify and fine-tune the functionality of the classes to implement their application specific requirements. The idea is to provide a system that can be easily extended to different problem domains.

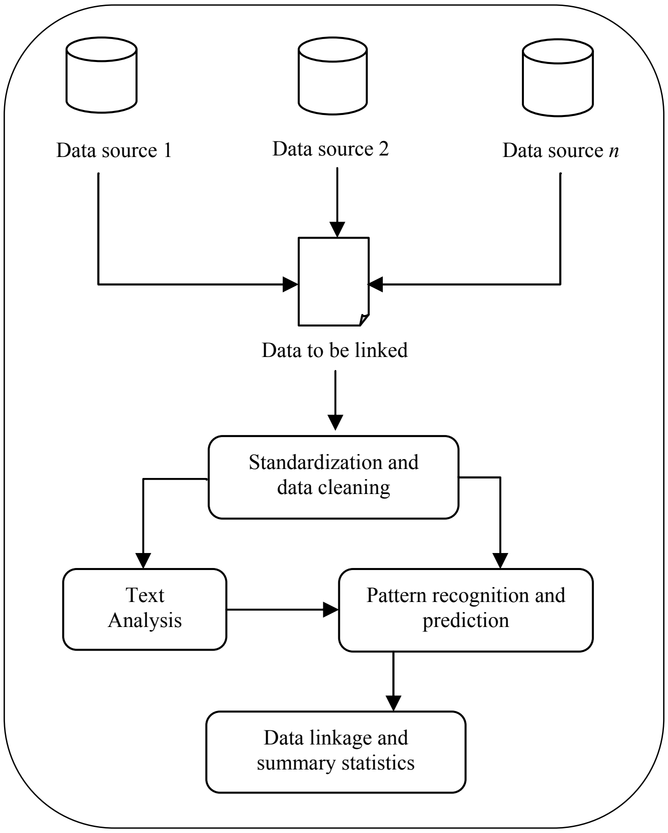

Data source 1
Data source 2
Data source *n*
Data to be linked
Standardization and data cleaning
Text Analysis
Pattern recognition and prediction
Data linkage and summary statistics

Fig. 2. General flow of a next generation data linkage application. The process is expected to be facilitated by the classes provided in the framework with necessary adjustments by users based on their application specific requirements.

V. A test application

As a test, an application was built upon the framework to link data of newspaper articles related to human rights violations. This is one area where duplication is quite evident and data linkage proves to be quite useful in finding links between articles of different newspapers or of the same newspaper published on different days. For example, such a linking process can provide useful statistics and complete information of incidents for interested parties.

The newspaper articles of human rights violations, for example, contains key information such as, incident date, incident type, victim name, perpetrator name, perpetrator type, city, age, etc. The task of the text analysis routines would be to locate this information, for example using mining techniques for named entity recognition, co-reference resolution, and relationship extraction. As mentioned earlier these methods are incomplete at this point. Therefore, for the test application the key information that represent an incident were manually extracted from newspaper articles. The attributes and their brief descriptions are provided in Table I. The dataset contained 1500 records pertaining to 400 distinct incidents collected from different newspapers and some records contained missing values of some attributes. The task is to identify which records belong to the same incident or entity and finally make use of the information of multiple records that belong to the same incident and generate a single record to represent that incident providing a complete understanding. In achieving this, first the data set was subject to standardization. Second it was analyzed for any relationships between attributes that could support in reducing the number of missing values. Based on the results, missing values of some attributes were replaced. Finally, clustering and linking were carried out.

The dataset was subject to standardization and cleaning, where standardization of names and addresses, removal of garbage characters, converting all words to lower case, updating the dictionary, etc., were carried out. These operations are provided with a view to minimize the impact of typographical errors on the overall performance and accuracy of the system. A dictionary permanently stores any new words found in the dataset for future referencing in the alphabetic order. The dictionary is implemented as a binary search tree and whenever a new word is found it is allocated to the correct location of the search tree, thereby maintaining a list of alphabetically ordered words at all times.

Next the dataset was analyzed to find any relationships between attributes that could be used for predicting the missing values of the dataset. As the initial step in this process, attributes having a lot of missing values are identified. Next, it is required to identify an optimal set of attributes that best describes the behavior of a particular attribute of interest. This is accomplished by using a randomized feature selection technique based on GAs. Candidate feature subsets are generated using genetic algorithms, while the selection of the best subset is based on the predictive accuracy on a test dataset. Calculation of the predictive accuracy was performed using a three layer ANN. For each candidate subset, the ANN was trained in the supervised mode using back propagation learning rule [10]. The training and testing datasets are automatically generated using the initial dataset. For example, half of the records for which the dependent variable of interest contains a value, is chosen to be the training dataset. Afterwards, the trained network was tested for its predictive accuracy using the other half. The framework provides ANN classes for designing and implementing neural networks ranging from supervised networks to self-organizing maps (SOM). The candidate subset with the highest predictive accuracy is selected to perform the pattern recognition task of the attribute of interest and predict the missing values. As an example, the prediction of the missing values of *perpetrator category* is described below.

The randomized feature selection technique together with the routines for comparing the predictive accuracy of those feature sets revealed that the *perpetrator category* (mob, terrorist, mafia, etc.) may be predicted using *incident type* and *number of victims*. Using this information, a three layer ANN was designed with two nodes in the input layer and five nodes in each of hidden and output layers as shown in Fig. 3 to predict the missing values of *perpetrator category*. First the ANN was trained in the supervised mode using the records for which the values of *perpetrator category* is available. Next, the missing values were predicted using the trained network. If there are five major classes of perpetrators, for example, the five neurons in the output layer would suffice to uniquely identify each of the classes. If there are more than five classes, combinations of the output values can identify the classes distinctly. A screenshot of the prediction of missing values of *perpetrator category* is provided in Fig. 4.

TABLE I. THE ATTRIBUTES AND THEIR DESCRIPTIONS

Attribute	**Brief description**
Paper	Name of the newspaper
Paper Date	Date of the newspaper
Incident Date	Date the incident took place
Location	Location of the incident
Category	Violation type
Number of victims	Number of victims involved
Victim Name	Name of victim if applicable
Age	Age of the victim
Perpetrator Category	Perpetrators if known

In a similar fashion, dependencies between other variables of interest may be obtained and a prediction of the missing values can be performed. It was also revealed that the *paper date* alone can predict the missing values of the *incident date*. However, it will not be discussed at this point.

Once prediction of missing values is complete, clustering of entities can be performed. The order in which attributes are considered in the clustering process can depend on a weight assigned to each attribute based on its impact, for example, attributes with lesser missing values receive higher priority in the clustering process. In order to perform clustering, the framework provides functionality to support, *k*-nearest neighbor classification (*k*NN) [11], *k*-means clustering, SOM, etc. However, to use SOM in clustering, there is a requirement to have a training dataset to train a network, in order that it can classify new entities later. For this particular application, a training dataset is not available, in which entities are already classified into groups.

Therefore, as one approach, *k*NN was used. In the *k*NN approach, each entity was arranged in a two-dimensional map according to an encoding technique on its string attributes and numerical attributes as shown in the cartoon illustration in Fig. 5. The *k*NN algorithm is used on the two dimensionally arranged entities to cluster them using the two encoded values obtained for string-type attributes and numerical-type attributes.

In a second approach, initial clusters are identified based on the attribute with the highest weight. The same process can be followed on all the attributes, according to the ranking of the weights, generating lower level clusters as the process continues. In this approach, standard blocking is applied for string attributes and the sorted nearest neighbor technique is applied on numerical attributes [12]. A cartoon image of the process is shown in Fig. 6. Once all the candidate attributes are compared, a single lowest level cluster provides the entities that are similar to each other.

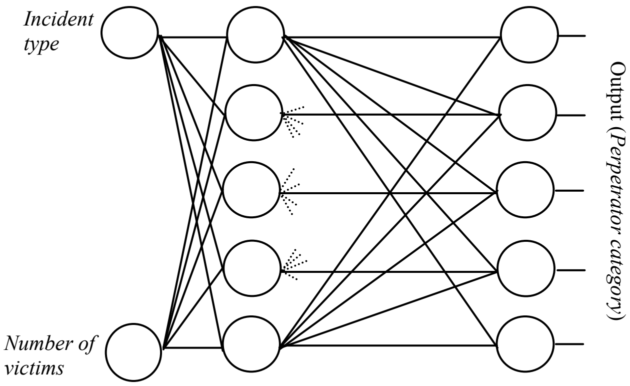
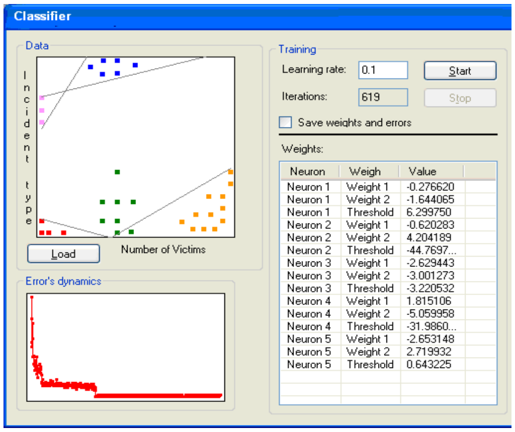


Fig. 3. Two layer ANN with two input nodes and five nodes in the hidden and output layers. The input nodes accept the values of *Incident type* and the *Number of victims* and predict the *Perpetrator cateogry* based on the training. The dotted lines in the nodes of the second layer indicates that each node is connected to all five nodes in the output layer. The *Perpetrator category* is identified by analyzing the outputs of the five output nodes.



Learning rate: 0.1

Iterations: 619

☐ Save weights and errors

Neuron	Weigh	Value
Neuron 1	Weight 1	-0.276620
Neuron 1	Weight 2	-1.644065
Neuron 1	Threshold	6.299750
Neuron 2	Weight 1	-0.620283
Neuron 2	Weight 2	4.204189
Neuron 2	Threshold	-44.7697...
Neuron 3	Weight 1	-2.629443
Neuron 3	Weight 2	-3.001273
Neuron 3	Threshold	-3.220532
Neuron 4	Weight 1	1.815106
Neuron 4	Weight 2	-5.059958
Neuron 4	Threshold	-31.9860...
Neuron 5	Weight 1	-2.653148
Neuron 5	Weight 2	2.719932
Neuron 5	Threshold	0.643225

Fig. 4. Screenshot illustrating the prediction of *Perpetrator category* based on the *Incident type* and the *Number of victims*.

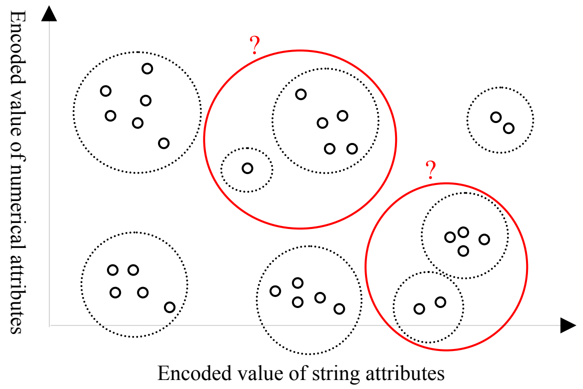

Encoded value of numerical attributes
Encoded value of string attributes
?
?

Fig. 5. A cartoon illustration of *k*NN clustering based on two encoded values extracted for string-type attributes and numerical-type attributes for each instance (each record) of an entity.

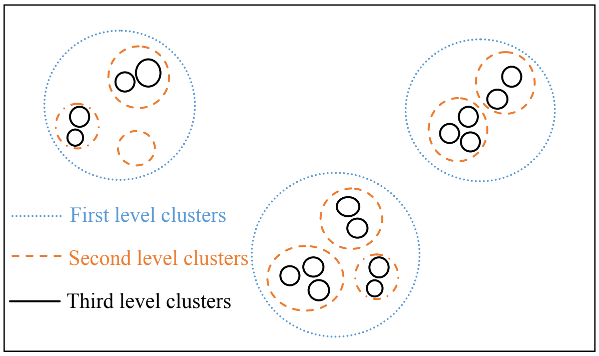


Fig. 6. A cartoon illustration of clusters generated by comparing individual attributes of each entity. First level clusters are generated by comparing a single attribute for all the entities and finding matches and grouping them together. Next, within the first level clusters, values of a second attrbiute is compared to generate the second level clusters. The sequence in which the attributes are considered can be based on a weight assigned to them as explained in the text.

As a final step, instances of an entity grouped into a single cluster needs to be processed in order to produce a single instance that provide a complete picture of a single entity. This process includes operations such as, extracting information, calculating probabilities to represent the frequency of a value for a particular attribute within a cluster, associate a credibility measure of an instance in the frequency calculation, and generating a single complete record to represent an entity. In the test application, the credibility measure was represented by the credibility of the newspaper from which the incident was extracted. The credibility of the newspaper was based on a general ranking of the newspapers. Depending on the application, users can define their own credibility measures to incorporate in their decision rules. It is assumed that the lowest level clusters that are created as a result of the clustering process would ideally contain duplicate instances pertaining to a single entity or incident in the test application. In some erroneous situations they would, in addition, contain records pertaining to other incidents. However, the clustering process is designed in a way to minimize the error rate as much as possible. Once linkage completes, the user is provided with a facility to automatically import linked data into a file type of choice for easy reference and portability.

VI. Results and Discussion

In order to highlight the use and applicability of the new methodology and the framework for data linkage, the results of the clustering task of the aforementioned dataset will be introduced below and a comparison is provided with respect to accuracy between the results of the proposed methodology and results from probabilistic models alone.

A. Summary statistics and future research

A summary of the three different methods used to find links in the dataset and their accuracies are provided in Table II. The probabilistic models (Method 1), makes use of the concepts introduced in section II. Method 2 and Method 3 differ only in the way they carry out the clustering task according to the methods shown in Fig. 5 and Fig. 6, respectively. As brought out earlier, the dataset contains 1500 records pertaining to 400 distinct incidents. Therefore, for each incident, there are multiple records, some having missing values, some containing slightly different information, and typographical errors. The three methods are employed to see which method(s) gives acceptable accuracy in identifying the 400 incidents distinctly and cluster the records pertaining to the same entity in the presence of such imperfections in the dataset.

According to method 1, out of the 400 entities, 298 entities were correctly identified. The process produced a total of 421 distinct entities, indicating that the unidentified 102 entities and their redundant records were incorrectly clustered into 123 groups. One possible reason for this low accuracy and the high error rate could originate from the fact that the dataset contains a lot of imperfections and the probabilistic models are not equipped with techniques to handle such issues.

Method 2, out of the 400 entities, correctly produced 384, demonstrating a marked increase in accuracy in comparison to method 1. The unidentified 16 entities were incorrectly clustered into 10 groups. This might be possible if the *k*NN algorithm incorrectly groups two entities in close proximity (shown by the circles with black dotted lines) into one group as shown by the red circle with a question mark in Fig. 5. In order to prevent such situations, it is necessary to implement a solid encoding technique that could identify even the slightest differences in attribute values and separate them in the two-dimensional map. Further studies on the encoding techniques and the *k*NN are pointers for future research.

Method 3, out of the 400 entities, correctly produced 391. It has slightly better accuracy than method 2. The unidentified 9 entities were incorrectly clustered into 10 groups. In this clustering method, where clusters are formed step by step based on individual attributes, there is a tendency for an instance (record) of an entity to be clustered into a wrong group, especially when a missing value is encountered at the beginning of the clustering process. For example, if *Person name* is missing and it is the initial attribute of the clustering process and it cannot be predicted using the concept of the new methodology, then that instance could be clustered into a wrong group during clustering based on the second attribute, for example *age*. This is possible because two different entities can have the same or similar values for the age attribute. This is one drawback of the clustering process that considers the attributes sequentially and scans the records for similarities sequentially. This is also a pointer for future research. It would be interesting to investigate how artificial intelligent routines, for example SOM, handles such situations and whether they can provide better results.

TABLE II. SUMMARY OF THE LINKING APPROACHES

Method	Approach	Accuracy
Method 1	Standardization, Probabilistic models in section II	298 correct entities
Method 2	Standardization, Predict missing values, *k*NN	384 correct entities
Method 3	Standardization, Predict missing values, Clustering based on individual attributes	391 correct entities

The low accuracy and the high error rate of method 1 could originate from the high rate of imperfections in the dataset and the fact that method 1 is not completely equipped with techniques to handle such situations. Methods involving only probabilistic models can fail in the presence of a lot of missing values. Method 2 and 3, on the other hand, not only investigates for relationships between entities, but also looks for relationships between attributes of the same entity and exploit those in minimizing the imperfections of the dataset before clustering leading to higher accuracy as observed.

B. *Technlogical and social issues*

- With the increase in accuracy comes a decrease in performance. Machine learning and artificial intelligent routines demand a lot of resources in terms of processing time and space. It is a compromise that needs to be made in the search for better accuracy. However, fine-tuning of these intelligent algorithms can lead to lesser use of resources. Especially, when the files are large, and training and learning from a subset of the dataset is performed as suggested by the proposed methodology, resource requirements could increase exponentially. A detailed investigation with respect to resource usage between probabilistic models and the new methodology is suggested as future research.
- Finding a balance between the use of probabilistic models and machine learning is also a factor that needs to be considered thoroughly. As brought out in the earlier sections, it is clear that a combination of probabilistic models and artificial intelligence at different stages of the linking process is the order of next generation linkage systems. But where do we draw the line between the two? Is it depending on the amount of missing values or depending on the type of attributes or problem domain that is being dealt with, that determines the extent to which artificial intelligent routines become useful? These are open issues that needs to be addressed.
- Confidentiality becomes one of the issues that needs to be taken into account, especially when experimental research is carried out. For example, researchers need to be supplied with large amounts of real world data to play around with and test their methodologies and ideas for improvement. When general-purpose public-use files are involved, proper precautions are necessary in order to make sure that individually identifiable information is not released for public use. Therefore, when large amount of data is released to the public, there is a requirement to make sure that the data is analytically useful, but at the same time preserves confidentiality. This is a very fine line that needs to be maintained, since when different techniques are used to preserve the confidentiality of entities, the data could render useless in testing data linkage models, since the representation of data deviates from the actual situation. This can, especially affect artificial intelligent routines because they rely on accurate and large datasets for training. Therefore, data perturbation, masking, or additive noise, techniques that can be used for preserving confidentiality, needs to be weighed against the ability to preserve validity. These questions need attention in the future in terms of designing new algorithms for protection of confidentiality that can walk the fine line between privacy and usability.
- Data classification and linkage is almost always not 100% accurate. When clustering and linkage is used in

administration, law enforcement, etc., cases where accuracy is of utmost importance, to what extent can the users' base their actions on the obtained results is a daunting question.

- Data classification and linkage have to work with data from multiple sources. What are the implications of having to handle data from multiple data sources but with different recording standards? Sometimes, different aspects or attributes of an entity could be stored in different data sources. How to identify that particular records containing different attributes on different data sources belong to the same entity when there are no bridging files between the data sources?

These are some of the bitter realities that data linkage has to face in the search for an ultimate solution, and pose not only technological but also social issues. There is no clear cut answer, but present methodologies and new concepts will have to be combined and evolved.

VII. Conclusion

One of the major objectives of this work is to develop a novel methodology for data classification and linkage. The proposed methodology differs from existing linkage models in many ways. The most highlighted difference, apart from extensibility and cost-effectiveness is the ability to adopt the best of both probabilistic models and computational machine learning/artificial intelligence into its decision rules. The realization of this methodology into a practical system consisted of implementing components for, standardization and cleaning, pattern recognition and prediction, linking and summary statistics. The successful implementation of these modules was supported by the classes provided in the framework which can be accessed through an API. In addition, the framework classes were designed and developed in a reusable fashion to support future development of different linkage applications. A test application was developed on top of the framework and the proposed methodology provide better accuracy in clustering and linking in comparison to the use of only probabilistic models.

There is still room for improvement in the methodology and the framework with respect to concepts pertaining to text analysis, improvements in machine learning tasks, computational resource usage, and mechanisms to cover social issues that need to be considered in data linkage, to name a few. It is expected to carry out further concept development and experimental research along these lines in the search for an ultimate data classification and linkage system. In the process, many hurdles with respect to technological and social issues will have to be tackled and walk a fine line between technical interests and social interests.

Acknowledgment

The authors would like to thank the Asia Foundation (Sri Lanka Section) for providing a sample dataset on human rights violations published by local newspapers. This work was partially supported by the memorial funding program of the graduate school of economics, Osaka University.

References

[1] G. P. Hettiarachchi, D. Attygalle, D. S. Hettiarachchi, and A. Ebisuya, "A Generic Statistical Machine Learning Framework for Record Classification and Linkage," IJIIP: International Journal of Intelligent Information Processing, vol. 4, No. 2, pp. 96-106, 2013.

[2] W. E. Winkler, "Advanced Methods for Record Linkage", Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 467-472, 1994.

[3] T. Churches, "Secure Health Data Linkage and Geocoding: Current Approaches and Research Directions", National e-Health Privacy and Security Symposium, Brisbane, 2006.

[4] D. R. Wilson, "Beyond probabilistic record linkage: Using neural networks and complex features to improve genealogical record linkage", Proceedings of the 2011 Joint International Conference on Neural Networks, pp. 9-14, 2011.

[5] G. P. Hettiarachchi, D. Attygalle, "SPARCL: An Improved Approach for Matching Sinhalese Words and Names in Record Clustering and Linkage," Proceedings of the IEEE Global Humanitarian Technology Conference (GHTC), pp. 423-428, 2012.

[6] N. Sandro, "The effect of lexicographical information costs on dictionary making and use", in Lexikos (AFRILEX-reeks/series 18), pp.170–189, 2008.

[7] W. E. Winkler., "The State of Record Linkage and Current Research Problems," Statistical Societyof Canada, Proceedings of the Survey Methods Section, pp. 73-80, 1999.

[8] T. Blakely, C. Salmond, "Probabilistic Record Linkage and a Method to Calculate the Positive Predictive Value", International Journal of Epidemology, vol. 31, pp. 1246-1252, 2002.

[9] M. A. Jaro., "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," Journal of the American Statistical Association, vol. 89, pp. 414-420, 1989.

[10] J. M. Zurada, Introduction to artificial neural systems. JAICO Books, 2002.

[11] D. Coomans, D. L. Massart, "Alternative *k*-nearest neighbor rules in supervised patter recognition: Part 1. *K*-Nearest neighbor classification by using alternative voting rules", Analytica Chimica Acta, vol. 136, pp. 15-27, 1982.

[12] R. Baxtor, P. Christen, T. Churches, "A Comparison of Fast Blocking Methods for Record Linkage", Proceedings of the Workshop on Data Cleaning, Record Linkage and Object identification, Washington DC, 2003.