An Efficient Computer-Aided Enrollment and Controlling System by Using Sorted Neighborhood Method

Aye Chan Mon, Myint Myint Khaing
Computer University, Pin Lon
achanmon@gmail.com,myintkhaing06@gmail.com

Abstract

The most important field, enrollment and controlling the students' data is very essential for all universities. Nowadays, computers are being used in the field of information in different way. When it is linked from one data source to another, the most common problems arise the difference of data format, difficulty in merging the up-to-date data and in eliminating duplicate data. Tracking students who move from university to university within districts or across the state/division is a major problem in all universities. This system is intended to implement the Upper Myanmar Computer Universities enrollment and controlling students' data in effective way by using sorted neighborhood method (SNM). The effectiveness of using SNM is to track critical data of students, such as enrollment, achievement scores and other factors throughout their university life. The advantages of create key in SNM eliminate duplicate records and compresses of all universities databases in efficient way.

1. Introduction

The development of a country mostly depends on all university in that country. That is, if a country has a qualified education, then it can develop in all various fields slowly or fastly. In university, it needs to enroll and control the linking student data accurately and precisely. If a student transfers from one university to another, his data need to be traced whether the information is correct or not.

Record Linkage is the task of identifying records corresponding to the same entity form one or more data sources. Real-world data is dirty and sources of variation in identifying fields include lack of a uniform representation or format, misspellings, abbreviations and typographical errors. Record linkage can be considered as part of the data cleaning process, which is a crucial first step in the knowledge discovery process. Standardization is an essential first step in every linkage process to clean and standardized the data. Blocking or searching techniques are used to reduce the number of comparisons. A good blocking method can greatly reduce the number of record pair comparisons and achieve significant performance speed-up. The data sets are partitioned into small blocks using blocking variables. The comparison vectors generated by such detailed comparison functions are then passed to the decision model to determine the final status of record pairs [1].

The rest of the paper is organized as follows: section 2 describes the related work, section 3 represents the sorted neighborhood method, and section 4 gives the design and implementation of the system. Conclusion is then discussed in section 5.

2. Related work

Record Linkage is a key problem in information retrieval. Many techniques for solving record linkage problem have been developed.

M.Elfeky et al. have proposed three machine learning record linkage model: induction record linkage model, clustering record linkage model and hybrid record linkage model.

R. Baxter et al. proposed recently blocking methods such as bigram indexing and canopy clustering that provide scalable blocking methods while maintaining or improving upon record linkage accuracy. This method has been the direct evaluation of reduction ratio and pair completeness for some diverse indexing methods. If the blocks of records are too small, the true record pairs may be missed and the record linkage accuracy is reduced.

M.hernadez and S.Stolfo proposed the system that provides a rule programming module that is easy to program and quite good at finding duplicate records with massive amounts of data. It also presented the sorted neighborhood method that performed the data cleaning process multiple times over small windows. The sorted neighborhood method is effective in detecting duplicate records [4].

3. Standardization and Sorted Neighborhood Method

In this system, it needs to standardize the input student data before using Sorted Neighborhood Method to control the student data.

3.1. Standardization

Standardization also called data cleaning or attribute-level reconciliation. Data standardization is important preprocessing steps for successful record linkage and before such data can be loaded into data warehouses or used for further analysis. It is also need to make sure that the variables to be used as linkage keys are formatted in the same way on each file. This process is employed before performing record linkage in order to increase the probability of finding matches. Without standardization, many true matched could be wrongly designated as non-matches because the common identifying attributes do not have sufficient similarity [8].

3.2. Sorted Neighborhood Method

The Sorted Neighborhood Method (SNM) sorts the data file first and move a window size of a specific size over data file, comparing only the record that belong to this window. The Sorted Neighborhood Method sorts the records based on a sorting key (SK), and then moves a window called sliding window (SW) of fixed size sequentially over the sorted records. SNM is a standard method for detecting exact duplicates in a database.

Sorted Neighborhood Method is done in three phases:

- Create Keys: A key for each record is created by extracting relevant fields or portions of fields.
- Sort Data: Sort the records by using the created key.
- Merge: After the records in the database have been sorted, move a fixed size window through the sequential list of records for matching records with another record in the window.

If the size of the window is ‘w’ records, every new record entering the window is compared with the previous ‘w-1’ record to find matching records. The first record slides out of the window as shown in Figure1 [2].

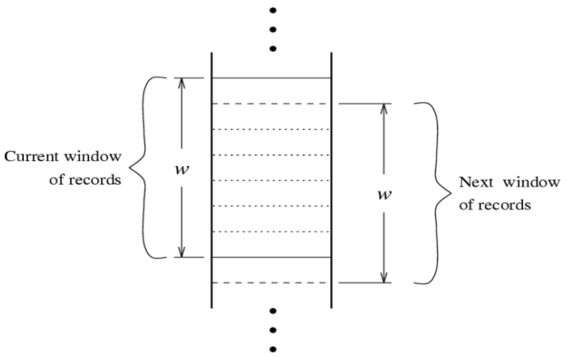


Figure 1. Merging of records using sliding

3.2.1. Selection of Key. The effectiveness of the SNM highly depends on the key selected to sort the records. A key is defined to be a sequence of a subset of attributes or substrings within the attributes, chosen from the record. Keys must provide sufficient discriminating power.

Table 1. Sample of key creation

Name	DOB	Sex		Key
Aye Aye	2.9.1987	F	→	yaf2987

3.2.2. Key Sorting. All records are sorted according to the alphabetic manner of created key. Therefore, all equivalent or matching records will appear close to each other. If the data was not sorted, a record may be near the beginning of the array of records and a duplicate record may be near the end of the array of records.

3.2.3. Merging. After the records in the database have been sorted, moves a fixed size window through the sequential list of records for matching records with another record in the window. When a duplicate key is found, the first record is deleted from the window and then the user can define a new fixed size window.

3.2.3.1 Merging (Equational Theory). The comparison of records, during the merge phase, to determine their equivalence is a complex inferential process that considers much more information in the compared records than the keys used for sorting. The more information there is in the records, the better inferences can be made [3].

As an example, here is a simplified rule in English that exemplifies one axiom of equational theory relevant to student databases:

Given two records, r1 and r2
  IF the last name of r1 equals the last name of r2,
    AND the first names differ slightly,
    AND the address of r1 equals the address of r2

```
THEN
    r1 is equivalent to r2.
```

3.2.3.2 Comparison Function. The basic idea of string comparison is to be able to compare pairs of strings such as 'Mon,Min' that contain minor typographical error. String comparison in record linkage can be difficult because lexicographically "near by" record look like "matches" when they are in fact not. A string comparator function returns a value depending on the degree of the match of the two strings.

4. Design and Implementation of the System

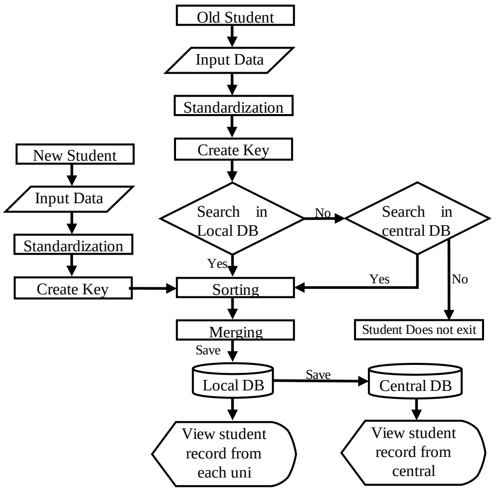


Figure 2. Design of the enrollment system

4.1 System Design

There are two major sections in this system. The first section is for enrollment only in one university throughout the academic year and the next section is for transfer student. The system is intended to enroll the students from the Upper Myanmar Computer Universities. This is well serves the approximation of the number of students (about 80000 students). When first year student is enrolling in the university, student's data are converted into standardized format. Then, by using SNM, student's data are sorted with a created key and stored in local and central databases. If old student enrolls in this university, the system will recreate a key to sort and merge the latest data. Finally, his/her data will be stored in the local and central databases. For transfer student, his data can be search in central database. And confirm that student has permission to enroll in the desire university. After confirming the necessary data, this student are allowed to do the enroll process as an old student in desire university. If student data is not found in the database error message will show. When entering data, user must enter student name, date of birth, NRC no, father name and mother name. The NRC field of the student is not use in key creation but it is use in merging phase to compare records.

4.2 Description of System Implementation

4.2.1 Data Standardization Main task is the conversion of raw input data into well-defined consistent forms. Strategies for standardization name fields which are used in this thesis:

- In name standardization, all letters are converted into lower case and remove certain character (like punctuations) as shown in Table 2.
- In gender attribute, it is replaced by one of the standard format such as Male or M or 1 is replaced by "m" and Female or F or 2 is replaced by "f" as shown in Table 3.
- For date of birth, it's converted into formatted form (day-month-year) as shown in Table 4.

Table 2. Example of name standardization

	Record 1	**Record 2**	**Record 3**
Name	aye chan mon	su nandar aung	cherry oo

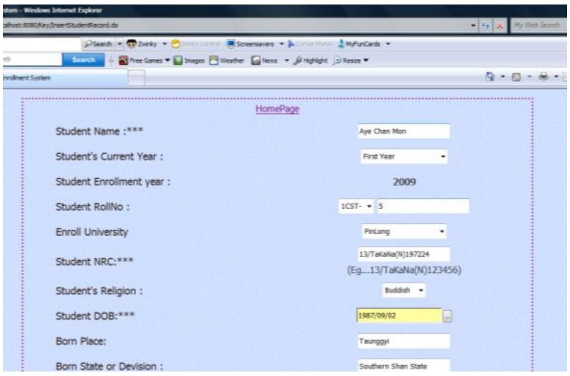
Table 3. Example of gender standardization

Original	**Replacement**
Male	M
Female	F

Table 4. Example of DOB standardization

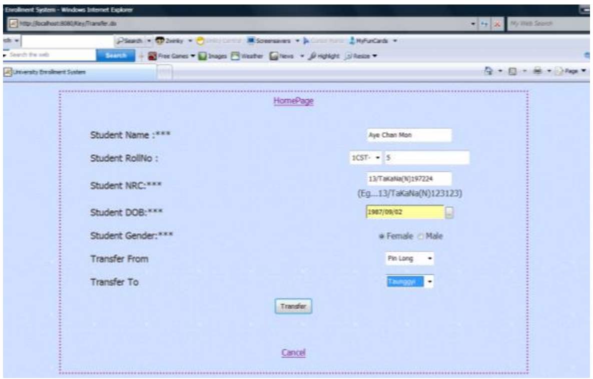
Original	**Replacement**
8.7.1987	8.7.1987
8.7.03	8.7.2003

4.2.2. Using Sorted Neighborhood Method (SNM) The name of the student and their associated data such as year he/she is attending, DOB, university enrollment number, etc are stored in database. It's created by the first three consonants of a last name are concatenated with the first letter of the first name field, followed by a gender field and day, month, last two numbers of the year of birth and followed by last number of the university enrollment number . If the name is only one word then it will take consonants of three words of the name. University enrollment number is auto increase. In the sorting phase, all records are sorted according to the alphabetic manner of created key. In the merging phase, this paper use “edit-distance” functions over some fields as a last attempt for merging a pair of records. In this paper, edit distance is use in student name, student nrc no, father name and mother name field. If three fields distance are more than 2 that two records are not same. Once matching weights of individual attributes of two records are calculated, the next step is to combine them as a one record. User can enroll in university from first year to master and add new student information to the database as shown in Figure 3. For transfer student, user must enter student name, roll number, NRC no, date of birth, gender and need to choose which university to transfer from which university as shown in Figure 4. In Figure 5, user can view the student record by year to year. If student is first year then the status will be new. When student enroll next year than status will change to old. When student is transfer from one university to another, his status will change to Transfer.



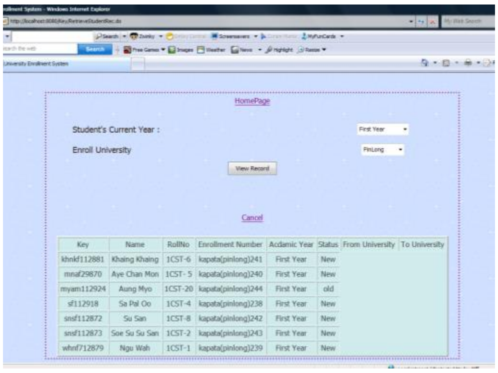
Student Name :***	Aye Chan Mon
Student's Current Year :	First Year
Student Enrollment year :	2009
Student RollNo :	1CST- 5
Enroll University	PinLong
Student NRC:***	13/TaKaNa(N)197224 (Eg...13/TaKaNa(N)123456)
Student's Religion :	Buddish
Student DOB:***	1987/09/02
Born Place:	Taunggyi
Born State or Devision :	Southern Shan State

Figure 3. Data entry for student enrollment



Student Name :***	Aye Chan Mon
Student RollNo :	1CST- 5
Student NRC:***	13/TaKaNa(N)197224 (Eg...13/TaKaNa(N)123123)
Student DOB:***	1987/09/02
Student Gender:***	● Female ○ Male
Transfer From	Pin Long
Transfer To	Taunggyi

Figure 4. Data entry for transfer student



Student's Current Year : First Year

Enroll University PinLong

Key	Name	RollNo	Enrollment Number	Acdamic Year	Status	From University	To University
khnkf112881	Khaing Khaing	1CST-6	kapata(pinlong)241	First Year	New		
mnaf29870	Aye Chan Mon	1CST- 5	kapata(pinlong)240	First Year	New		
myam112924	Aung Myo	1CST-20	kapata(pinlong)244	First Year	old		
sf112918	Sa Pal Oo	1CST-4	kapata(pinlong)238	First Year	New		
snsf112872	Su San	1CST-8	kapata(pinlong)242	First Year	New		
snsf112873	Soe Su Su San	1CST-2	kapata(pinlong)243	First Year	New		
whnf712879	Ngu Wah	1CST-1	kapata(pinlong)239	First Year	New		

Figure 5. Key Creation and Sorting Keys

4.3. Evaluation Graph

Comparison of duplicate data elimination with key and without key is shown in graph. x-axis of the graph shows number of students and y axis shows the number of duplicate records for university students' data. When student is enrolling in one year his/her data is save in the database as one record. Without the creation of key, his/her data can not be merged when he/she is enrolling next year and it is difficult to consider that student is existing one or new one. With the creation of key, duplicate data can be found easily and can merge the student data easily.

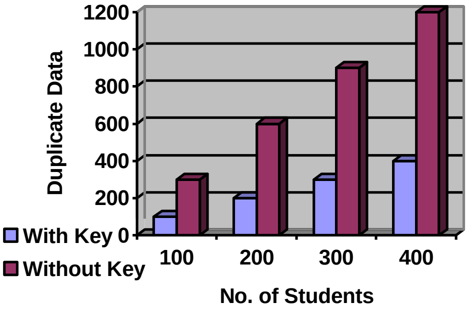


Figure 6. Evaluation graph with key and without key

5. Conclusion

Data mining techniques have been applied to the problem of possible duplicate record detection. Sorted Neighborhood Method provides Key Creation, Sorting and Merging. SNM is powerful, mathematically elegant and effective way to help university for student’s data management. The system can easily search, sort and control students’ information. The choice of keys for sorting, their order and the extraction of relevant information from a key field are the knowledge intensive activity that must be explored and carefully evaluated prior to running a data cleaning process. Student's enrollment system using SNM helps university to fulfill students' enrollment needs.

6. References

[1] Concept:Record Linkage htpp://mchpappserv.cpe.umanitoba.ca/viewConcept.php?

[2] L.Gu and R.Baxter, "Adaptive Filtering for Efficient Record Linkage", SIAM international conference on data mining, Orlando, 2004.

[3] M. Hernandez and S. Stolfo, "Real World Data is Dirty: Data Cleaning and the Merge/Purge Problem", Journal of Data Mining and Knowledge Discovery, 2(1), pages 9-37, 1998

[4] M. Hernandez and S. Stolfo, "The Merge/Purge Problem for Large Databases", In Proc. of the ACM SIGMOD Conf., pages 127-138, 1995.

[5] Program Review, "Safeguarding Student Social Security Numbers in the UW System", August 2005.

[6] R.Baxter, P.Christen, and T.Churches, "A Comparison of Fast Blocking Methods for Record Linkage", In Proc. Of ACM SIGKDD'03 Workshop on Data Cleaning, Record Linkage, and Object Consolidation, pages 25-27, Washington, DC, USA, August 2003

[7] T. Jackson, E.Aheam, "Unique Student Identifiers", QTA – A brief analysis of a critical issue in special education, May 2004

[8] Zin War Tun, Nilar Thein, "An Approach of Standardization and Searching Based on Hierarchical Bayesian Clustering (HBC) for Record Linkage System," c5,pp.54-60, Fifth International Conference on Creating, Connecting and Collaborating through Computing (C5 '07), 2007.