The Personal Name Problem and a Data Mining Solution

Clifton Phua
Monash University, Australia

Vincent Lee
Monash University, Australia

Kate Smith-Miles
Deakin University, Australia

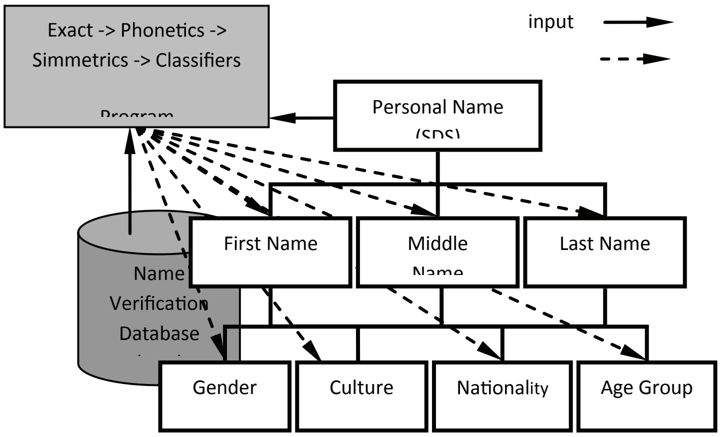
INTRODUCTION

Almost every person has a life-long personal name which is officially recognised and has only one correct version in their language. Each personal name typically has two components/parts: a first name (also known as given, fore, or Christian name) and a last name (also known as family name or surname). Both these name components are strongly influenced by cultural, economic, historical, political, and social backgrounds. In most cases, each of these two components can have more than a single word and the first name is usually gender-specific. (see Figure 1).

There are three important practical considerations for personal name analysis:

- Balance between manual checking and analytical computing. Intuitively, a small proportion of names should be manually reviewed, the result has to be reasonably accurate, and each personal name should not take too long to be processed.
- Reliability of the verification data has to be examined. By keeping the name verification database's updating process separate from incoming names, it can prevent possible data manipulation/corruption over time. However, the incompatibility of names in databases can also be caused by genuine reasons as such as cultural and historical traditions, translation and transliteration, reporting and recording variations, and typographical and phonetic errors (Borgman and Siegfried, 1992).

Figure 1. Hierarchy chart on the inputs, process, and outputs of the name verification task.



- Domain knowledge has to be incorporated into the entire process. Within the Australian context, the majority of names will be Anglo-Saxon but the minority will consist of many and very diverse groups of cultures and nationalities. Therefore the content of the name verification database has to include a significant number of popular Asian, African, Middle Eastern, and other names.

Figure 1 illustrates the input, process, and output sections. Input refers to the incoming names and those in the verification database (which acts like an external dictionary of legal names). Process/program refers to the possible four approaches for personal name analysis: exact-, phonetical-, similarity matching are existing and traditional approaches, while classification and hybrids are newer techniques on names-only data. Output refers to the insights correctly provided by the process. For simplicity, this paper uses first name to denote both first and middle names; and culture to represent culture and nationality. While the scope here explicitly seeks to extract first/last name and gender information from a personal name, culture can be inferred to a large extent (Levitt and Dubner, 2005), authenticity and age group can be inferred to a limited extent.

BACKGROUND

In this paper, we argue that there are four main explanations when the incoming first and last name does not match any name in the verification database exactly. First, the personal name is not authentic and should be manually checked. Second, it is most likely due to an incomplete white list. It is impossible to have a name verification database which has every possible name, especially rare ones. Third, the incoming name does not have any variant spelling of name(s) in the database (i.e. Western European last names). Fourth, there are virtually millions of potential name combinations or forms (i.e. East Asian first names).

The last three reasons are problems which prevent incoming personal names from being verified correctly by the database. Without finding an exact match in the name verification database, the personal name problem in this paper refers to scenario where ordering and gender (possibly culture, authenticity, and age group) cannot be determined correctly and automatically for every incoming personal name. Therefore, additional processing is required.

There are three different and broad application categories of related work in name matching (Borgman and Siegfried, 1992):

1. **Information retrieval:** Finding exact or variant form(s) of incoming name in verification database with no changes to the database. This present work is the most similar to ours where an incoming personal name is used as a search key to retrieve first/last name and gender information.
2. **Name authority control:** Mapping the incoming name upon initial entry to the most commonly used form in database. Current publications on author citation matching within the ACM (Association of Computing Machinery) portal database are examples of this (Feitelson, 2004). Unlike author names where the first names are usually abbreviated, many personal names in large databases have complete first and last names.
3. **Record linkage/duplication detection:** Detecting duplicates for incoming multiple data streams at input or during database cleanup. Recent publications focused on supervised learning on limited labelled data (Tejada *et al*, 2002) and on approximate string matching (Bilenko *et al*, 2003). Unlike their matching work which uses comparatively smaller data sets and has other informative address and phone data. Intelligently matching incoming names-only data with a comprehensive verification database seems like a harder problem.

Other specific applications of personal name matching include art history (Borgman and Siegfried, 1992), name entity extraction from free text (Cohen and Sarawagi, 2004; Patman and Thompson, 2003; Bikel *et al*, 1999), genealogy, law enforcement (Wang *et al*, 2004), law (Navarro *et al*, 2003; Branting, 2003), and registry identity resolution (Stanford ITS, 2005). Name matching has been explicitly or implicitly researched under databases, digital libraries, machine learning, natural language processing, statistics, and other research communities; and also known as identity uncertainty, identity matching, and name disambiguation.