Web service for 19th century Irish personal name matching

Phattara Wangrungarun and Dr Adam Winstanley (supervisor)

Erasmus Mundus – MSc in Dependable Software Systems (DESEM) – Maynooth University – 2014/15

Abstract

Before the first Irish civil registration on 1864, census materials were mostly lost or incomplete. So genealogical research uses parish records and also some 'census substitute' documents, such as land ownership and tenancy records. However, some of these documents may not contain enough information in identify individuals. Some of them contains a name and address, whereas others might contain only a name.

Record linkage is one method to gather scattered information among many documents. It uses a person's name as a reference to link that person's information between many documents. With patience, a more complete information about that person can be obtained.

Therefore linking or matching a person's name is important in the process. Unfortunately, in the 19th century, in Ireland, there was no standard spelling of names, handwriting could be difficult to read and contractions or abbreviations were often used. The names with the same pronunciation and for the same individual could be written in many different ways. Moreover, names in the Irish language which are equivalent to English names were used, for example, Irish version of 'Smith' could be 'Gowan'. A further complication is that historical and genealogical research often requires large quantities of names to be matched.

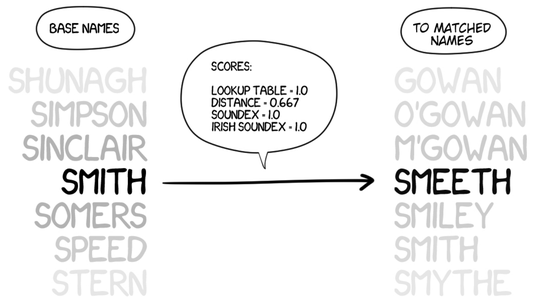
To handle these name variations, various solutions have been created to find matching different names that refer to the same person. However, for our extent knowledge, there is yet no public system which encodes those solutions together and provides a service of bulk name matching.

Thus, we developed a web service system using Ruby on Rails framework to achieve our goal. The system is initially encoded with 4 matching algorithms, Levenshtein distance, soundex, Irish soundex, and lookup table.

Research Questions

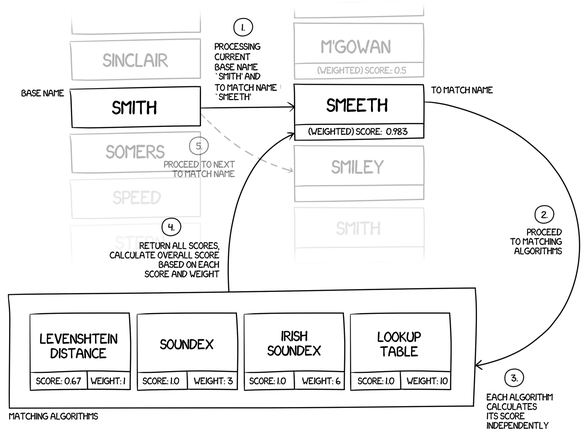
- Can we provide a web service to match names, where matching can be a complicated process because of the way people record their names.
- Can the web service act as a platform system for general names or words matching system so that it can be extended to other languages as well.

Initial Idea



Base name 'SMITH' comparing to *to-match name* 'SMEETH'. Scores of each matching algorithms are presented in the bubble above the arrow.

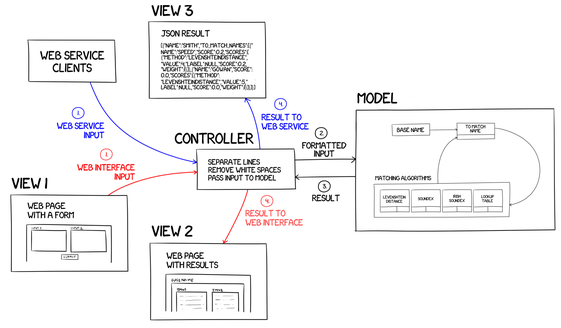
Actual System



One matching cycle consists of 5 steps.

1. Processing current *base name* 'SMITH' and *to-match name* 'SMEETH'
2. Proceed to matching algorithms.
3. Each algorithm calculates its score indepedently.
4. Calculate *overall weighted score*.
5. Matching cycle for 'SMITH' and 'SMEETH' is finished with *overall weighted score* 0.983.

MVC



Web Service

```
{
  "base_names":"Smith",
  "to_match_names":"Smythe\r\nO'Gowan",
  "matching_algorithms":{
    "1":{"name":"LookupTable", "weight":"10"},
    "2":{"name":"LevenshteinDistance", "weight":"1"},
    "3":{"name":"Soundex", "weight":"3"},
    "4":{"name":"IrishSoundex", "weight":"6"}
  },
  "threshold":"0",
  "standard_list":""
}
```

```
$ curl -H "Accept: application/json" -H "Content-type:
    application/json" -X POST -d @sample.json
    http://localhost:4001/match.json
```

```
[
  {
    "base_name": "SMITH",
    "to_match_names": [
      {
        "to_match_name": "SMYTHE",
        "overall_weighted_score": 0.983,
        "scores": [
          {
            "method": "LookupTable",
            "value": "Matched",
            "label": "1897",
            "score": 1,
            "weight": 10
          },
          {
            "method": "LevenshteinDistance",
            "value": 2,
            "label": null,
            "score": 0.667,
            "weight": 1
          },
          {
            "method": "Soundex",
            "value": "S530 <=> S530",
            "label": null,
            "score": 1,
            "weight": 3
          },
          {
            "method": "IrishSoundex",
            "value": "S530 <=> S530",
            "label": "SMYTHE",
            "score": 1,
            "weight": 6
          }
        ]
      },
      {
        "to_match_name": "O'GOWAN",
        "overall_weighted_score": 0.5,
        ...
```

Evaluation

We tested our system by matching 12,944 names.

Matching algorithms	Response speed (ms)
Levenshtein distance	1,337
Soundex	2,024
Irish soundex	2,456
Lookup table	24,293
All 4 algorithms	28,786

Table 1: Response speed for each matching algorithms.

Matching algorithms	Memory usage (bytes)
Levenshtein distance	48,518,621
Soundex	53,066,150
Irish soundex	69,534,598
Lookup table	244,302,744
All 4 algorithms	373,544,727

Table 2: Memory usage for each matching algorithms.

References

[1] Adam Winstanley.
Identifying People on the Morpeth Roll.
July 2014.
Postgraduate Diploma in Genealogical, Palaeographic & Heraldic Studies 2013-14.

[2] Robert Edwin Matheson.
Varieties and synonymes of surnames and christian names in Ireland.
1901.
accessed May 4th, 2015.

[3] Vincent Ramdhanie.
What is a 'web service' in plain English?
October 2008.
accessed May 5th, 2015.

Maynooth University
National University of Ireland Maynooth

European Commission
ERASMUS MUNDUS