A Case Study in Name Matching

Ronald J. Leach
Howard University

We examined variants of a well-documented habitational surname of English origin and evaluated the performance of several common name search algorithms to determine their efficiency in identifying equivalences of many known variants of the original surname. The surname had over fifty known variants. A new algorithm to include habitational information was developed as part of the analysis of search algorithms. The new algorithm had improved performance on the data set, achieving 92% success in name matching when the best two matches were used. The new algorithm is easily automated and holds promise as a search technique for much larger data sets.

Introduction

During the late sixteenth and early seventeenth centuries, the name *Shirecliffe* was prominent in Ecclesfield, a small town near Sheffield in Yorkshire, England. Formerly, the area was classified as part of the much larger region of Hallamshire. Several members of the Shirecliffe family were buried in crypts at the front of St. Mary's Cathedral in Ecclesfield, indicating their family's stature. The Shirecliffe family manor house, Whitley Hall, remains in use today as a small hotel, although it is no longer owned by the Shirecliffe family. Whitley Hall still has a "priest's hole," to allow Catholic worship to take place in relative secrecy during the religious warfare in England (Jones and Jones, 2003; Gatty, 1847 445-448). Members of this family emigrated to Massachusetts on the Mayflower in 1620 and to St. Mary's County, Maryland in 1636 (Shurtleff, 1912; Shurtleff, 1976).

Names 54:4 (December 2006): 321–330
ISSN:0027-7738

The Dictionary of American Family Names describes this name as being a habitational name of English origin from Shirecliff in Sheffield, South Yorkshire (Tucker, 2003). The first known occurrence of a person with (essentially) this surname is Nicholas de Shirecliffe in 1334 (Gatty, 1847, 445).

There have been many variants on the *Shirecliffe* surname. A recent book states that there were more than fifty variants of this surname during the period 1334 and 1795 (Jones and Jones, 2002). We have been able to find 50 of the supposed variants of this surname in written historical and genealogical records. The surnames are: de Shercliffe, Scherkleif, Schirecliff, Schyreclyff, Schyrteclyff, Sharkley, Shearcliffe, Sheircliefe, Sheircliff, Sheircliffe, Shercliff, Shercliffe, Shereclyffe, Sherecliff, Sherecliffe, Sherkliff, Sherley, Shertcliffe, Shertliffe, Shiercliff, Shiercliffe, Shierclyff, Shierclyffe, Shircliff, Shircliffa, Shircliffe, Shircliffo, Shirclyff, Shirclyffe, Shirecliff, Shirecliffe, Shireclyf, Shirkcliff, Shirliffe, Shirtcliff, Shirtcliffe, Shirtleff, Shirtliff, Shirtliffe, Shortlife, Shortliffe, Shriecliffe, Shurkey,
Shurtlef, Shurtleff, Shurtliff, Shurtliffe, Shurtlifh, Sureclife, and Surecliff.

An examination of the genealogical records available from the major commercial genealogy web site, ancestry.com, indicates records exist for each of the 50 previously found surname variants except *Sureclife* and *Surecliff*. This suggests that nearly all surname variants have been recognized by many descendants who were doing genealogical research.

The focus of this paper is determine to what degree standard techniques for searching computerized databases could determine all the variants of the *Shirecliffe* surname. Our concern, of course, is with finding as many of the surname variants as possible while finding as few as other surnames that were "false positives" as possible. The issue studied here is obviously one part of a much larger question of determining to what degree the name searching techniques work for other surnames not necessarily of English origin.

The paper is organized as follows. This introductory section is followed by a brief discussion of the Soundex and Daitch-Mokotoff encoding systems. A new algorithm for name searching using habitational information is given, followed by the analysis, conclusions, and future work.

Soundex and Daitch-Mokotoff Encoding

The most common way of encoding names is the Soundex system, which was originally patented by Richard Russell in 1918, well before the advent of computers. During the Depression, Soundex-based indexes of surnames were prepared for several censuses to provide birth information for the newly created Social Security System.

Soundex provides a way to conduct searches to find everything that "sounds somewhat like" the entered words. The basic principle is that the degree to which two things "sound alike" can be quantified with some degree of accuracy, even if they are spelled differently, such as *Smyth* and *Smythe*.

The Soundex encoding of a name starts with the first letter of the name to be encoded, then applies a set of simple numerical substitutions to most consonants. In general, vowels are ignored and only the first four consonants are used.

A Soundex encoding consists of four characters, with the first letter of the surname followed by a digit, normally in the range 1 to 6. The letters B, F, P, and V are encoded as a 1; C, G, J, K, Q, S X, and Z are encoded as a 2; D and T are encoded as a 3; L is encoded as a 4; M and N are encoded as a 5; and R is encoded as a 6. If there are not enough consonants, one or more 0's are added to produce a four character output. The Soundex encoding of Smith and Jones are S530 and J520, respectively.

The Soundex encodings of the *Shirecliffe* variant surnames had the following distribution.

Soundex encoding	Number of Occurrences	Percentage
S620	1	2%
S624	32	64%
S632	4	8%
S634	11	22%
S640	1	2%
S641	1	2%

A Soundex-based search of the 50 variant surnames using *Shirecliffe* as the starting point would thus be able to find 64% of the known variants, with 86% found if we used the two most common variant encodings.

The Daitch-Mokotoff (D-M) Soundex System (Mokotoff, 1985; Daitch, 1986) was designed to improve the accuracy of Soundex for primarily Jewish surnames that are of eastern European origin. The D-M encoding is not limited to Jewish names. The D-M encoding is six characters long, instead of the four characters used for Soundex. Thus, names that sound the same initially, but differ at the end, are coded differently, giving a smaller set of database matches to be searched. For example, the names *Anders* and *Anderson* have identical Soundex, but different D-M, encodings. The six characters are all digits.

Moreover, the use of digits means that letters map into ten possible D-M codes rather than the seven of Soundex (for all but the first letter).

Unfortunately, D-M coding suffers from the disadvantage that some letter combinations have more than one possible encoding because they represent different sounds. For example, the surname Jones has two D-M encodings, 164000 and 464000. The surname Smith has only the D-M encoding 463000.

An easy-to-use program for computing D-M and Soundex encoding, with a link to a discussion by Gary

Mokotoff, can be found at Stephen Morse's web site stevenmorse.org/soundex.html. The D-M algorithm is far too complex to describe here because of the different options in the coding chart and the rules that describe its operation.

The D-M encodings of the *Shirecliffe* variant surnames had the following distribution.

D-M encoding	Number of Occurrences	Percentage
300000	1	2%
493587	1	2%
493870	12	24%
494870	4	8%
495870	31	62%
498000	1	2%
498700	1	2%

The accuracy of the highest encoding is the same as before, with only 62% of surname variants matched by the most frequent encoding, and 86% by the two most common ones.

A New Algorithm Based On Habitational Information

It is clear that neither Soundex nor D-M searching were entirely satisfactory on this data, with many variants not found by the two most frequent choices. The term "precision" is often used in conjunction with this type of name searching, especially in the information retrieval research community. The "precision" of a search is defined to be the number of names relevant to the desired search, divided by the total number of names examined. (The related term, "recall," is the number of names relevant to the desired search, divided by the total number of possible names, whether examined by the search or not. We do not discuss "recall" in this paper,

because we do not know the size of the universe of possible surnames.)

We observe that there is additional information that can be used to search for variants of the Shirecliffe surname – it is a habitational name. It is easy to separate each of the surname variants into a political region (the "shire") and a topographical name (the "cliffe"). This separation leads, in turn, to an obvious algorithm. We apply Soundex or D-M encoding to each part of the surname and combine the results. The next two tables give the Soundex encodings of each half of a surname, followed by the frequency of the habitational Soundex-based encoding.

Soundex first half	Frequency	Percentage	Soundex second half	Frequency	Percentage
S600	34	68%	C410	33	66%
S620	1	2.08%	K000	2	4%
S630	15	31.25%	K410	2	4%
			L000	2	4%
			L100	11	22%

Combined Soundex	Frequency	Percentage
D000, S600C410	33	66%
S600K000	1	2%
S620K400	1	2%
S630K410	2	4%
S630L000	2	4%
S630L100	11	22%

It is just as easy to do the same thing for the D-M encoding. The results are given in the next two tables. For

simplicity we have ignored the data for alternate D-M encoding, since it will not affect the analyses given here.

D-M first half	Frequency	Percentage	D-M second half	Frequency	Percentage
490000	34	68%	500000	1	2%
493000	15	30%	580000	1	2%
495000	1	2%	587000	35	70%
			800000	2	4 %
			870000	11	22%

Combined D-M encoding	Frequency	Percentage
490000500000	1	2%
490000580000	1	2%
493000587000	35	70%
493000800000	2	4%
493000870000	11	22%

Analysis

As reported earlier, we used standard Soundex and Daitch-Mokotoff encodings of 50 variants of a single habitational surname and found relatively poor "precision," in the sense that we needed to use two different encodings to find more than 80% of the names. Using a single Soundex or D-M search would find 64% of the known variants, with 86% found if we used two encodings.

Using habitational information, we devised a simple algorithm based on splitting the surname into two parts. The algorithm led to results that were slightly better. We obtained matches of 66% for a single match using habitational information with Soundex, with 88% matches using the most common encodings of variants.

We did even better with the habitational information in a modified D-M encoding, with 70% matching using a single match and 92% using the two most common encodings of variants. This is more satisfactory. Perhaps this is the best possible, especially in view of the inherent difficulty in name matching (Swart 1989)

Another small experiment was carried out using the "Onomastics Plugin" from the web site www.blueshoes.org to examine relationships between a few of the variants of the Shirecliffe surname. This web site uses the "semantic web" and some knowledge representation techniques to provide a dynamic way of processing information. People are able to add information to this web site, making the conclusions made about relationships more reliable over time.

At present, the state of name matching leaves quite a bit to be desired. For example, the web site's estimated probabilities (expressed as percentages) of the surnames *Shercliff, Shercliffe, Sheircliffe, Shirecliff,* and *Shurtlef,* being direct variants of *Shirecliffe* were 36.5%, 40%, 41%, 80%,and 0%, respectively. We never found probabilities greater than 80% and, in some cases, the probability was estimated as 0!

Conclusion

One must be extremely careful when reasoning from analytical results on a small data set. However, the use of habitational information into other common approaches to name matching has some promise.

It is very hard to find similarities between names common to different cultural or linguistic groups, especially if the names have been transformed into a third language such as English. Often the writing or spelling of the names does not indicate their correct pronunciation.

We note that there is a body of work on name matching algorithms using habitational, occupational, and religious information, but that much of the recent work has not been released for general publication due to security issues.

Future work

Name searching techniques have become a major emphasis of several governments due to the current international climate. As such, there is both a need to go beyond simplistic approaches and a need to make sure that there is common basis for evaluation of new approaches, such as the surname set described herein. We offer the surname set as a well-defined test bed for other researchers in this field. It has the advantage of having nearly all surname variants known in advance, as opposed to more general data sets (Hanks and Tucker 2000; Tucker 2002).

The Levenshtein algorithm, which computes the distance between two character strings as the minimal number of insertions, deletions, or substitutions needed to transform one string into the other, can be very helpful in quantizing some estimates. We intend to extend our approach to other surnames not necessarily of English origin, but for which the number of variants is relatively well understood.

We are also looking at this and other, larger, data sets to see if some of the trends in the evolution of surnames described in (Galbi 2002) can be used to aid in name matching.

Acknowledgement

This research was partially supported by the National Science Foundation under grant number 0324818.

References

Blueshoes Corporation. 2005. www.blueshoes.org/en/plugins/onomastics/example_compare_name_pair/

Daitch, Randy. 1986. "Jewish soundex--A revised format". *Avotanu* 1:19-26.

Galbi, Douglas A. 2002. "Long Term Trends in the Frequencies of Given Names" *Names* 50: 275-288.

Gatty, Alfred. 1847. *Hallamshire: The History and Topography of the Parish of Sheffield of the County of Yorkshire, Second Edition*. Sheffield, England: Pawson and Brailsford.

Hanks, Patrick, and D. Kenneth Tucker 2000. "A Diagnostic Database of American Personal Names" Names 48:59-69.

Jones, Joan and Mel Jones. 2003. *Whitely Hall: an Illustrated History*. Rotherham, England: Green Tree Publications.

Mokotoff, Gary. 1985. "Proposal for a Jewish soundex code". *Avotaynu* 1:5-10.

Mokotoff, Gary. 1997. "Soundexing and Genealogy. http://www.avotaynu.com/soundex.html.

Morse, Stephen. 2005. stevenmorse.org/census/soundex.html

Shurtleff, Benjamin, Descendants of William Shurtleff, 2 vols, 1912.

Shurtleff, Roy L., Descendants of William Shurtleff: 1976 rev. ed. San Francisco. R. L. Shurtleff, 1976.

Swart, E. R. 1989. "A Computer Simulation of the Ineradicable Uncertainty in Genealogical Research". *Family History*, 11:118.

Tucker, D. K. 2002. "Distribution of Forenames, Surnames, and Forename-Surname Pairs" *Names* 50: 105-132.

Tucker, Patrick, ed. 2003. *Dictionary of American Family Names*, Oxford: Oxford University Press.