NEW PHONETIC NAME SEARCH ALGORITHM

ALAN FRANK
Medford, Massachusetts

Telephone books work because the listings are generated in a way that ensures that the company has the correct spelling of your name, and when your friends look you up, they'll also generally have a pretty good idea of how it's spelled.

These assumptions don't hold for a patient database at a large urban hospital. Names may be entered on the basis of a hasty telephone call or in an emergency room; they may be looked up under similar circumstances, usually by someone who isn't confident of the exact name. The clerks are not always well trained, and the patients may not be native English speakers.

Thus, many institutions implement a fancy lookup system of some sort, allowing for a phoneticization of the patient name and combining it with the patient's sex and date of birth. Special processing may also be accorded to middle names, hyphenated names, maiden names, and common nicknames. We will deal here only with algorithms to represent names phonetically.

The classic algorithm in this regard is the Soundex system, invented in 1918 by Odell and Russell, and described by Donald Knuth in Sorting and Searching, Volume 3 of his series The Art of Computer Programming. This algorithm works as follows:

1. Replace every letter by its number according to the following: 0 AEHIOUWY, 1 BFPV, 2 CGJKQSXZ, 3 DT, 4 L, 5 MN, 6 R
2. Replace adjacent occurrences of any digit by single occurrences
3. Remove all non-leading zeroes
4. Reduce to four characters
5. Replace the leading digit by the first letter of the name

The name ECKLER would go through the steps 022406, 02406, 0246, E246.

This lets us find many names which we would otherwise miss; for example, misspellings such as EGGLER or ECKLARD would be found. But many dissimilar names are grouped together (Knuth gives as examples LISSAJOUS/LUKASIEWICZ=L222 and KNUTH/KANT=K53) and many similar names are encoded differently, as discussed below.

Based on an extensive analysis of the type of mistakes actually made at one institution, a revision of the Soundex algorithm was developed. Tests showed that it was significantly better than the original Soundex method at finding misspelled names. One important facet of the new system is that it sometimes generates several codes for a name; as a result, there are often more names with a given code than in the original system. However, as the methods to deal with that problem are in the software engineering field, they will not be discussed here.

The steps of the revised algorithm are as follows:

1. If the name ends in S, encode it both with and without the final S. This enables WILLIAM/WILLIAMS and WEEKS/WEEKES to match
2. If the name begins WR, remove the initial W. This makes names such as WRONSKY easier to find
3. If the name begins KN, encode it both with and without the initial K. This enables KNOWLES/NOLES to match, without losing the ability to match KNOOP/KENOOP
4. If DG appears in the middle of a name, encode it as both J and DG. This enables ROGERS/RODGERS to match, without losing MADGAN/MADAGAN
5. If GH appears, treat it as K if followed by a vowel or as silent otherwise. This gives us BLIGH/BLY, NEIGHBORS/NABORS, and LANGHORNE/LANKHORNE. Names in which GH is pronounced as F do not appear in the subject database
6. Replace G by C
7. Replace every letter other than C, F and X by its number or symbol according to the following: + AOU, - EIY, 1 BPV, 2 JSZ, 3 DT, 4 L, 5 MN, 6 R, 7 KQ, 8 H, 9 W. Note that the former 2 category is split up into hard and soft letters, in order to reduce the number of false hits such as BUCK/BASS
8. Replace X by 2 if it's in an initial position and by 7 otherwise. This way, XENAKIS/ZENAKIS and FOWKES/FOX both continue to match
9. Replace adjacent occurrences of any character by single occurrences
10. If C is followed by 4, 5, 6 or +, replace it by a 7; otherwise, encode the name both with the C replaced by 7 and by 2. This will preserve matches such as ECCLES/EKKLES and MCGILL/MCKELL while avoiding CLEON/SLOAN
11. If F is followed by 6, replace it by a 1; otherwise, encode the name both with the F replaced by 2 and by 1. This will preserve matches such as STEFAN/STEPHEN and MAVROULES/MAFROULES, while also finding CLAFF/CLASS (note that "eff" and "ess" sound alike)
12. Again replace duplicate occurrences of any number by single occurrences; also, replace 72 by 7
13. Remove all non-leading occurrences of +, -, 8 and 9
14. Replace the leading character according to the following table: + O, - O, 1 B, 2 S, 3 D, 4 L, 5 M, 6 R, 7 K, 8 H, 9 W. By not simply going back to the original first letter of the name, *we can find matches such as CAPLIN/KAPLAN, MORRIS/NORRIS* and ALLMAN/ULLMAN
15. Reduce to four characters