Improving System Ability to Identify Symptom Complexes in Free-Text Data

Cory Forbach, BS, Matthew J. Scholer, PhD, MD,
Dennis Falls, BS, Amy Ising, MSIS, Anna Waller, ScD
NC DETECT Syndrome Definition Workgroup, UNC Department of Emergency Medicine

OBJECTIVE

This paper describes a novel approach to the construction of syndrome queries written in Structured Query Language (SQL). Through the advanced application of character set wildcards, we are able to increase the number of valid records identified by our queries while simultaneously decreasing the number of false positives.

BACKGROUND

Text-based syndrome case definitions published by the Center for Disease Control (CDC)[1] form the basis for the syndrome queries used by the North Carolina Disease Event Tracking and Epidemiologic Collection Tool (NC DETECT). Keywords within these case definitions were identified by public health epidemiologists for use as search terms with the goal of capturing symptom complexes from free-text chief complaint and triage note data for the purpose of early event detection and situational awareness.

Initial attempts at developing SQL queries incorporating these search terms resulted in the return of many unwanted records due to the inability to control for certain terms imbedded within unrelated free text strings. For example, a query containing the search term “h/a”, a common abbreviation for headache, also returns false positives such as “cough/asthma”, “skin rash/allergic reaction” or “psych/anxiety”.

Simple abbreviations without punctuation, such as “ha”, were even more problematic. Global wildcards ('%') indicate that zero or more characters of any type may substitute for the wildcard.[2] The term “ha” as a synonym for "headache" appears frequently in the data, but searching this term bracketed by global wildcards returns any instance where the two letters appear together (e.g. pharyngitis, hand, hallucinations, toothache).

Using global wild cards to search for common symptoms such as headache using simple abbreviations, with or without specialized punctuation, results in the return of many unwanted false positive records. We describe here the advanced application of SQL character set wildcards to address this problem.

METHODS

In contrast to global wildcards, character set wildcards allow for targeted searches of free-text data. SQL allows users to define a set of characters for which to search, in which any one character will satisfy the query.[2] Replacing a search for "h/a" with "h[/-,]a", for example, will return all results in which "h" and "a" are separated by either a slash, a hyphen, or a comma. Using a caret ("^") indicates that all characters except for the ones listed should be allowed. Therefore, searching for "h[^a-z]a" returns all records which contain "h" and "a" separated by any single non-alphabetic character.

Character set wildcards can also be used to search for isolated alphabetic terms. By surrounding the term “ha” with the exclusionary character set “%[^a-z]ha[^a-z]%”, we eliminate false positives where "ha" appears within another word, but capture records where "ha" appears as a standalone term.

RESULTS

A search of NC DETECT emergency department data for ‘%h/a%’, the method previously used, for the time period from 12/22/2005 through 12/31/2005 yielded 213 records, eight of which are false positives. This represents a positive predictive value of 0.964. A revised search, using a combination of character set wildcards and run against the same database and date range as the previous headache query, returns 975 records with six false positives. The positive predictive value of the new method is 0.994, with 4.73 times as many true positives.

The revised query took approximately 11 seconds to run versus approximately 6 seconds for the original query but returned nearly five times the number of true positives. We feel this increase in processing time is justified by the dramatically increased number of valid records identified. The wildcard character set methodology was applied to multiple other symptom queries with similar results.

REFERENCES

1. Centers for Disease Control and Prevention (October 23, 2003). Syndrome definitions for diseases associated with critical bioterrorism-associated agents. Available at http://www.bt.cdc.gov/surveillance/syndromedef/index.asp. Accessed June 24, 2005.

2. How to: Search with Regular Expressions”, SQL Server 2005 Books Online, http://msdn2.microsoft.com/en-us/library/ms174214.aspx, Accessed June 29, 2006.

Funding provided by the NC Division of Public Health.