Considerations for searching SNOMED CT using Vector Space Model Algorithms like Lucene

Presenter: Jay Kola, Noesis Informatica Pte Ltd

Audience

Implementers and software developers designing SNOMED CT search functionality in clinical systems; business analysts specifying requirements for search.

Objectives

- Understand the basics of Vector Space Model based search engines/libraries like Lucene.
- Identify issues with SNOMED CT search functionality implemented naïvely (out of the box settings) using Vector Space Model (VSM) [1].
- Functionality required to address issues with VSM implementations and extensions, with a focus on searching non-English SNOMED CT extensions.

Abstract

There is a growing number of popular open source search engines/libraries that use the Vector Space Model algorithm as the basis of information retrieval (e.g. Lucene, SOLR, etc.). As a result, a number of software developers are starting to design and implement SNOMED CT search functionality in clinical applications using these libraries. While the VSM algorithms are reasonably efficient and accurate for searching generic documents, using these libraries with default (out of the box) configurations leads to unintuitive and some times erroneous results. Since some of these libraries also have extensions that enable broad-brush functionality, the inconsistencies in search results are compounded by extensions and become difficult to fix.

In this presentation we cover our experience of implementing SNOMED CT search based on VSM in Snolex, an online SNOMED CT browser [2], currently used by over 250 users to browse the Swedish and International editions of SNOMED CT. The presentation will explore how features of the VSM like term frequency – inverse document frequency (tf/idf), weighting and ranking of search results, correspondence between VSM documents and SNOMED CT components, usage of clinical terms frequencies, etc. influence search results. These features will be discussed with clinical examples of how naïve implementations cause issues. We will include demonstrations of naïve search implementations, with the corrected implementation in Snolex – to highlight differences in results, rather than to showcase Snolex as product.

We also discuss some of the specific functionality we had to implement in Snolex for users searching the Swedish SNOMED CT Edition (containing both English and Swedish terms). As the number of non-English speaking IHTSDO member countries continues to grow, there is a need to provide better guidance for searching SNOMED CT in non-English languages. Our experiences of searching the Swedish SNOMED CT edition using Lucene extensions and issues faced using ASCII character normalisation, tokenizing, stop words and synonyms will enable attendees to avoid potential issues in their own implementations and improve accuracy & consistency between various VSM enabled SNOMED CT search tools.

References

1. A Vector Space Model for indexing - http://www.cs.uiuc.edu/class/fa05/cs511/Spring05/other_papers/p613-salton.pdf
2. Snolex – A browser, demonstrator and refset manager for SNOMED CT (https://snolex.com)