Considerations for searching SNOMED CT using Vector Space Model Algorithms

Jay Kola,
Noesis Informatica Ltd, UK
Outline

• Why SNOMED CT search matters?
• Introduction to Vector Space Model (VSM) algorithm
• Configuring SNOMED CT search with VSM engines like Lucene
  – What is broken?
  – Why chickens can not always win!
  – Why Cone is not a gender
• Configuring a browser (e.g. Snolex) to handle issues with VSM algorithms
  – Fixing what is broken
Why SNOMED search matters?

• Long topic – being discussed at a tutorial at this conference!

• For input field based data entry systems, search results influence what is record
  – An appropriate match not within the first ~20 results is likely to be not picked
  – Users are more likely to pick the first few top hits
  – Garbage In Garbage Out principle
Why SNOMED search matters? (2)

• In SNOMED CT search is more tricky because:
  – Synonyms – more than one concept can have the same synonym
  – Counts matter? – Limited search space, so sometimes users know number of matches to expect.
  – Search expansion using ‘known synonyms’ – kidney & renal
  – Importance of special characters in terms - ^, %, etc

• Multiple Versions & Editions
  – International edition comes out every Jan & July; National editions come out every Apr/May & Oct/Nov
  – Keeping them all in sync while maintaining inter version dependencies is tricky
  – Not all editions are in English – Swedish, Spanish, etc
SNOMED CT Browsers

• All browsers can be good, all browsers can be broken!
  – In my experience, browsers are use case specific. So if your users are ‘happy’, then you are okay!

• Browsers are cheap to build – dime a dozen!
  – No SNOMED CT browser wars! No ‘one SNOMED CT browser to rule them all!
  – Learn by building ‘quick & dirty’ browsers

• Secret sauce to building SNOMED CT browser
  – Apache Lucene
  – IHTSDO Developer Toolkit anyone?
Apache Lucene

• Lucene is an open source ‘information retrieval’ software library published by the Apache Software Foundation.

• Makes it very easy to index and search a collection of documents or SNOMED CT concepts

• Almost the ‘go to’ library for implementing search these days

  – Hopefully means that SNOMED CT browsers won’t have to rely on SQL’s LIKE query to return results!

    • Word order agnostic search possible in Lucene -- ‘Pneumonia Acute’

  – Built my Lucene based SNOMED CT browser in 2 hours in 2008; but still haunted by questions about it…
Vector Space Model

• Lucene’s secret sauce is:
  – Term frequency/inverse document frequency (tf/idf) – score calculation
  – Vector Space Model algorithm – score comparision
Vector Space Model (2)

- Lucene’s secret sauce is:
  - Term frequency/inverse document frequency (tf/idf) – weight calculation
  - Vector Space Model algorithm – score comparison
- Term frequency: Number of times a term occurs in a document
  - Higher is better
- Inverse document frequency: Number of times the term appears in all the documents in the collection
  - Lower is better
- VSM allows the scores to be compared for multiple query terms – based on a mathematical function (Cosine similarity)
VSM influence

• When you search for ‘ast’
  - asthmatic > astigmatism > aster > asthma
VSM influence (2)

• When you search for ‘ast’
  – asthmatic > astigmatism > aster > asthma

• The lower the idf, the more the weight – or higher the ranking of the result
  – asthmatic > astigmatism > aster > asthma
  – 16 < 33 < 39 < 202

• Similarly when you search for ‘asthma’
  – Asthmatic > asthma
VSM Influence (3)

• Search for ‘gall’

  – Gallus > gallon > gallium > ... gallstone
VSM Influence (4)

• Search for ‘gall’
  – Gallus > gallon > gallium > ... gallstone

• The higher the term frequency, the better the ranking...

• Gallus = Gallus gallus (chicken), two ‘gall’ in a term is better than one (gallon, gallbladder).

• So chicken always wins!
But... Chickens can’t always win!

• A search engine should rank more common words higher – not too many care about chickens in clinical practice!
  – So shouldn’t ‘asthma’ and ‘gallstone’ which are more common diagnosis be ranked higher?

• Answer: Lucene supports boosting of matches
  – Index time boost vs Search time?
  – Clinical term usage frequencies can be used to make gallstones or asthma appear higher
Clinical term frequencies

• Using clinical term frequencies means:
  – Asthma > Astigmatism > Aster
  – Gallstone > Gallon > Gallus
Clinical term frequencies

• Using clinical term frequencies means:
  – Asthma > Astigmatism > Aster
  – Gallstone > Gallon > Gallus

• Works nicely, until the user tells you otherwise
  – Confused about why ‘Asthma’ is ahead of ‘Aster’
  – More predictable to have ‘Aster’ ahead of ‘Asthma’
  – ‘The user is always right, even when they are wrong’ – Alan Rector, Medical Informatician
Search contexts...

• By this point, the ‘quick and dirty’ search engine is no longer ‘cheap’ to build/change...

• Search ranking preference depends on user profile
  – Clinicians – prefer less chickens
  – Mappers – prefer more consistency in results
  – Newbies to SNOMED CT – prefer all merged results + visual clues (e.g. Search categories)
  – Collaboration anyone? – please get in touch!

• Now we finally move to SNOMED CT specific considerations...
Concept vs Description

- What is a Lucene document in SNOMED CT?

<table>
<thead>
<tr>
<th>Concept</th>
<th>Myocardial Infarction</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Id</td>
<td>• 22298006</td>
</tr>
<tr>
<td>• Fully Specified Name</td>
<td>• Myocardial Infarction</td>
</tr>
<tr>
<td>• Preferred Term</td>
<td>(disorder)</td>
</tr>
<tr>
<td>• Synonym</td>
<td>• Myocardial Infarction</td>
</tr>
<tr>
<td></td>
<td>• Heart attack</td>
</tr>
<tr>
<td></td>
<td>• …</td>
</tr>
</tbody>
</table>

- Closer to the Lucene ‘document’ idea
- All descriptions for a concept together contribute to relevance of a concept when it is being ranked.
- Disadvantages – since all descriptions contribute, quite hard to control ranking…
Better control over matches

Most of the time, users seem to care about ‘terms’—aka Descriptions

Disadvantages: How do you deal with ‘exact’ term matches for different concepts?

Search for ‘fundus’ brings back multiple matches—different hierarchies
  - Fundus of stomach (body part)
  - Fundus of gallbladder (body part)
  - Part of eye....
## Concept vs Description (3)

### Myocardial Infarction

- 22298006
- Myocardial Infarction (disorder)
- ...

### Solutions:

- Display Fully Specified Names
- Display categories (if relevant)
  - Fracture (morphological abnormality)
  - Fracture (finding)
Concept vs Description (4)

• What if the same concept has multiple ‘matches’ for the search term?
  – Asthma
  – Asthmatic

• Solution… Merge and replace with Preferred Term?
  – So matches would only show ‘Asthma’
Concept vs Description (5)

- What if the same concept has multiple ‘matches’ for the search term?
  - Asthma
  - Asthmatic

- Solution... Merge and replace with Preferred Term?
  - So matches would only show ‘Asthma’

- Search for ‘mastectomy’
  - Mastectomy – Excision of breast tissue (procedure)
  - Mastectomy – Simple Mastectomy (procedure)
  - So matches would look like...
    - Excision of breast tissue
    - Mastectomy
Jag talar inte engelska

• SNOMED CT is published in a few different languages – Swedish, Spanish, Danish...
• But most of the guidance published is English centric (or en dialects).
• So guidance says – It should be possible to search for Sjogren’s disease either with:
  – Sjögren’s disease or
  – Sjogren’s disease
• Normalising all diacritics and non-en alphabets to ASCII characters works well for English speaking world, but not for non English speaking world
Ilium is not Ileum

• VSM itself does not have issues with non-English languages but:

• Out of the box settings in Lucene might ‘normalise’ to ASCII resulting in:
  – Confounding results
  – Erroneous results

• English equivalent of returning
  – ‘reflex’ as a match for ‘reflux’
  – ‘Ante’ for ‘Anti’
  – ‘Ilium’ for ‘Ileum’
But Cone (kon) is not Gender (kön)!

• å, ä, and ö are alphabets in the Swedish – not diacritics!
• Search for ‘kön’ should return
  – 263495000 |kön| (gender).
  – Not 421504000 |kon| (cone),
• Search for ‘aska’ should return
  – 225867006 |preferens gällande hantering av aska| (225867006 | preference for disposal of ashes|),
  – Not 257494002 |åska| (257494002 |thunder|)

Credit: Mikael Nystrom for examples
Compound words

• Some languages use compound words, to combine two words into one:
• Blåscancer = Blåsa + cancer (bladder cancer)
• Distinct from English where word combinations are separated by hyphens – sugar-free, post-coordination
• Not the same as agglutination in other languages – words combined to form new words
• Searching compound words – all matches for ‘cancer’
  – Not just starting with ‘cancer’
  – Compound words with xxxxxcancer too...
  – No out of the box setting for VSM and Lucene
Summary

• Lucene and other VSM implementations need to be ‘configured’
• As non English SNOMED CT editions become more common, there will be greater need for better non-English resources (e.g. Stop words) and guidance
• Remember, accuracy of search matter – influences data entry
• It is more important to build a ‘more complete’ search experience than building one ‘quickly’...
• Question: Does IHTSDO need to update the Developer Toolkit to bring it up to date to the world of ‘Lucene’?
What has not been covered

• Other similarity scoring algorithms – word length normalisation

• Handling stop words – break your browser by typing ‘to be or not to be’?

• Advanced features – stemming, phonetic matches (Soundex, Metaphone…)

• Presentation of search results, grouping, etc.
Try some of the enhancements in...

https://snolex.com
The journey continues…

• Share experiences & resources…

• Collaboration…

• Questions…

@jaykola

jay@noesisinformatica.com