Semantic-Based Information Retrieval of Biomedical Data

Dr. Yu (Cathy) Jiao and Dr. Thomas E. Potok

Applied Software Engineering Research Group
Computational Sciences and Engineering Division
Oak Ridge National Laboratory
Oak Ridge, TN 37830
potoke@ornl.gov
Outline

• Introduction

• Background and Related Work

• MEDTHES – A Medical Thesaurus

• Application of MEDTHES – A Case Study

• Conclusion
Introduction

• The effectiveness of information retrieval is assessed by Precision and Recall.

• If the user’s search term is not the index term of a document, precision and recall decrease.

• Query expansion is one of the methods that can be used to alleviate this problem.
Introduction

• General English Language Thesauri
  – Roget’s Thesaurus
  – WordNet
  – ...

• Domain Specific Thesauri
  – NASA Thesaurus
  – Educational Resources Information Center Thesaurus
  – Astronomy Thesaurus
  – ...

• Medical Thesauri
  – Medical Subject Headings
  – Unified Medical Language System
  – Systematized Nomenclature of Medicine
  – ...

Introduction

• Common Problems

– Poor interoperability and reusability as the result of not following the standard.

– Lack of semantic similarity information.

– Require the users to possess precise knowledge of the controlled vocabulary.
Major Contributions

• Establish a medical thesaurus (MEDTHES) that follows the ANSI standard for thesaurus design.

• Provide semantic similarity measures to assist users in performing imprecise queries.

• Include synonyms of medical terms from a general English thesaurus, WordNet, in order to ease the use of MEDTHES for non-medical professionals.

• Incorporate MEDTHES into an existing mobile agent-based information search engine and demonstrate its practicality.
Background

• The ANSI/NISO Z39.19-2003 Standard
  – Title: “Guidelines for the Construction, Format, and Management of Monolingual Thesauri Abstract”
  – Rules for term selection, thesaurus structure, relation definitions, and thesaurus maintenance.
  – Three types of semantic relations: equivalence, hierarchical (broader and narrower), and related.
Background

• Medical Subject Headings (MeSH)
  – It is the standardized vocabulary developed by National Library of Medicine.
  – It contains approximate 22,000 terms (descriptors).

• Advantages
  – Comprehensive and well-maintained

• Disadvantages
  – Synonymous relationship is not clearly defined.
  – Does not follow the ANSI standard.
  – Designed for medical professional.
Background

• WordNet
  – It’s an online thesaurus that models the lexical knowledge of the English language.
  – It organizes English nouns, verbs, adjectives, and adverbs into synonym sets.

• Advantages
  – Comprehensive
  – Make fine distinctions among word meanings

• Disadvantage
  – Do not follow the ANSI standard
Semantic Similarity Functions

- The Edge Counting Algorithm
  - The number of edges along the shortest path between any two terms.

- The Leacock & Chodorow Algorithm
  - Relatedness(t₁, t₂) = -log(len(t₁, t₂)/2D)
  - Len(t₁, t₂): the number of edges along the shortest path between t₁ and t₂.
  - D: the maximum depth of the hierarchy

- The Wu & Palmer Algorithm
  - SemanticDistance(t₁, t₂) = (N₁+N₂+2*N₃)/2*N₃
  - N₁ & N₂: the length of the shortest path from t₁ and t₂ to their least common ancestor (LCA).
  - N₃: the length of the shortest path from the LCA to root.
MEDTHES

- MeSH + WordNet
- Relationship Definitions

<table>
<thead>
<tr>
<th>ANSI Relationship</th>
<th>MEDTHES Representation</th>
<th>Abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equivalence</td>
<td>USE</td>
<td>USE</td>
</tr>
<tr>
<td></td>
<td>Used For</td>
<td>UF</td>
</tr>
<tr>
<td>Hierarchical</td>
<td>Broader Term</td>
<td>BT</td>
</tr>
<tr>
<td></td>
<td>Narrower Term</td>
<td>NT</td>
</tr>
<tr>
<td>Associative</td>
<td>Related Term</td>
<td>RT</td>
</tr>
<tr>
<td></td>
<td>Subject Category</td>
<td>SC</td>
</tr>
</tbody>
</table>
MAMDAS Overview

NodeManagers
DSWorker
DSSlave

Host

Thes Server
ThesMaster

DSMaster

Host

AdminMaster

Host

HostMaster
HostMaster

Network
Application of MEDTHES

[Diagram showing a network of nodes and hosts]
Application of MEDTHES

I found the following in 1467 ms.

**********breen doppler3**********

[Navel]

**********borg bln4**********

[Umbilicus]

**********breen doppler1**********

[Bellybutton]
Correlation of the Semantic Distance Algorithms

- Test Queries: OHSUMED test collection
- TREC Data Search Engine: Zettair
- Test Corpus: TREC-9 data collection
Correlation of the Semantic Distance Algorithms

<table>
<thead>
<tr>
<th>Edge Counting</th>
<th>Leacock &amp; Chodorow</th>
<th>Wu &amp; Palmer</th>
</tr>
</thead>
<tbody>
<tr>
<td>D = [0, 1.0]</td>
<td>D = [0, 0.2]</td>
<td>D = [0, 1.0]</td>
</tr>
<tr>
<td>D = (1.0, 2.0)</td>
<td>D = (0.2, 0.6)</td>
<td>D = (1.0, 1.6)</td>
</tr>
<tr>
<td>D = (2.0, 3.0)</td>
<td>D = (0.6, 1.0)</td>
<td>D = (1.6, 1.7)</td>
</tr>
</tbody>
</table>
Conclusion

• We addressed the issue of semantic-based information retrieval by using a medical thesaurus, MEDTHES, and a mobile agent-based data search engine.

• MEDTHES follows the ANSI standard and therefore, has great interoperability.

• We incorporated three well-known semantic distance calculation algorithms into MEDTHES in order to support novice users.

• We provided quantitative guidance on the usage of the semantic distance calculation algorithms.