



Automatic Acquisition of Concept Hierarchies in an Integrated Information Retrieval/Information Extraction Framework

Niranjan Balasubramanian



Overview

- Introduction
- Motivation
- Background
- Solution
- Evaluation
- Results
- Conclusions
- Future Work

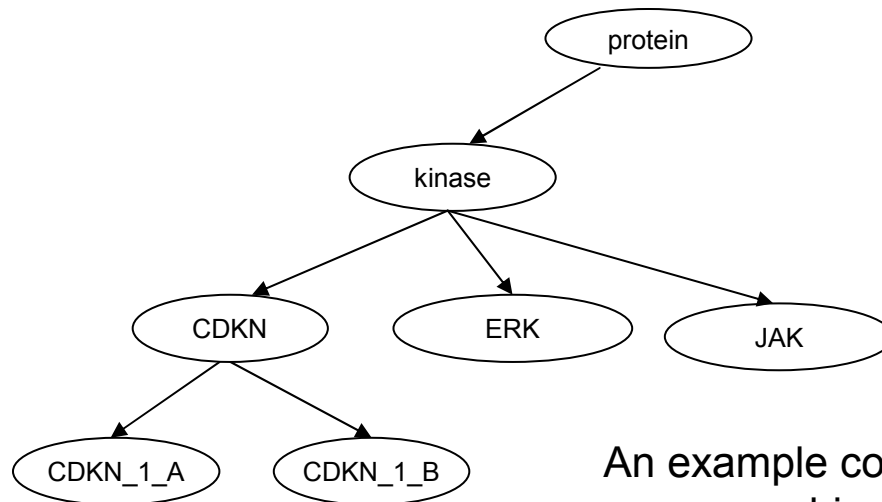


Introduction

- Information Explosion: Accessing, Organizing and Discovering information
- Accessing Information
 - Information Retrieval (IR)
- Organizing Information
 - Hierarchical categorization of documents, Yahoo![1]
 - Thesauri - UMLS[2] & WordNet[3]
- Discovering Information
 - Text mining (TM) – finding novel information from text

Concept Hierarchies

- Collection of salient concepts in a domain
- Concepts share hierarchical relationships
- Utility
 - For organizing documents hierarchically
 - For use as domain level information in a text mining framework



An example concept hierarchy for the biomedical domain



Text Mining

- Discovering novel, hidden information from text
 - Marti Hearst[4]
- Text Mining in Biomedical domain
 - Swanson [5], discovering new causal chains for diseases
 - Protein-Protein interaction[6]
- Unapparent information revelation, Srihari et al.[7]
 - Document level information
 - Document subset level information
 - Domain level information



Motivation

- Automated concept hierarchy generation
- Use of concept hierarchies as domain level information in the text mining framework by Srihari et al.[7]
- Manually constructed hierarchies
 - WordNet – A generic thesaurus
 - UMLS Metathesaurus – Biomedical domain thesaurus
- Problems with manual construction
 - Coverage
 - Corpus level usage information is lost



Concept hierarchy generation methods

- Subsumption methodology
 - Keywords or phrases as concepts in an IR framework
 - Use subsumption statistics, Sanderson and Croft [8]
 - Use topical coverage of concepts to identify hierarchical relationships[8, 9]
- Lexical and Syntactic Methods
 - Use lexical clues, lexicosyntactic patterns, Marti Hearst[10]
 - Lexical dispersion, Anick T. and S. Tipirneni[11]
 - Noun head and modifier information to organize nouns into hypernym based relationships, Woods [12]



Subsumption Methodology

- Subsumption definition [8] – a concept ‘x’ subsumes concept ‘y’
 - if $P(x|y) = 1$ and $P(y|x) < 1$
 - for lack of absolute co-occurrence
 - $P(x|y) > c$ and $P(y|x) < c$, for some empirically determined c .
- Manual evaluation by presenting hierarchical menus
- A graph theoretic approach by Lawrie, Croft [9]
 - uses conditional probabilities similar to [8] for finding topical words
 - uses an approximation algorithm to greedily find topical words that cover most of the vocabulary
- Methods for comparing hierarchies
 - using link based approach, Lawrie et al. [9]



Lexical and Syntactic Methods

- Lexical clues, Lexicosyntactic patterns
 - Hearst [10] identified ten lexical patterns that suggest hierarchical relationships
 - “such as” is suggestive of IS-A relationship
 - Noun hierarchies, Woods[12]
 - Use a lexicon that suggests relationships between terms to get subsumption axioms
 - Identify subsumption relationships between noun phrases based on
 - Subsumption axioms on phrases
 - Transitive use of axioms
 - Nouns constituting the phrase
 - Used in interactive browsing/information seeking efforts
 - Selection of concepts is topical rather than domain specific



Problem Definition

- Automatically generate concept hierarchies for domain specific concepts
- Intended use of concept hierarchies
 - as domain level information in a Text Mining framework
- Properties of concept hierarchies generated
 - Identified concepts should be domain specific and must be meaningful
 - The hierarchies should accommodate the use of pre-defined or existing knowledge bases



Testbed

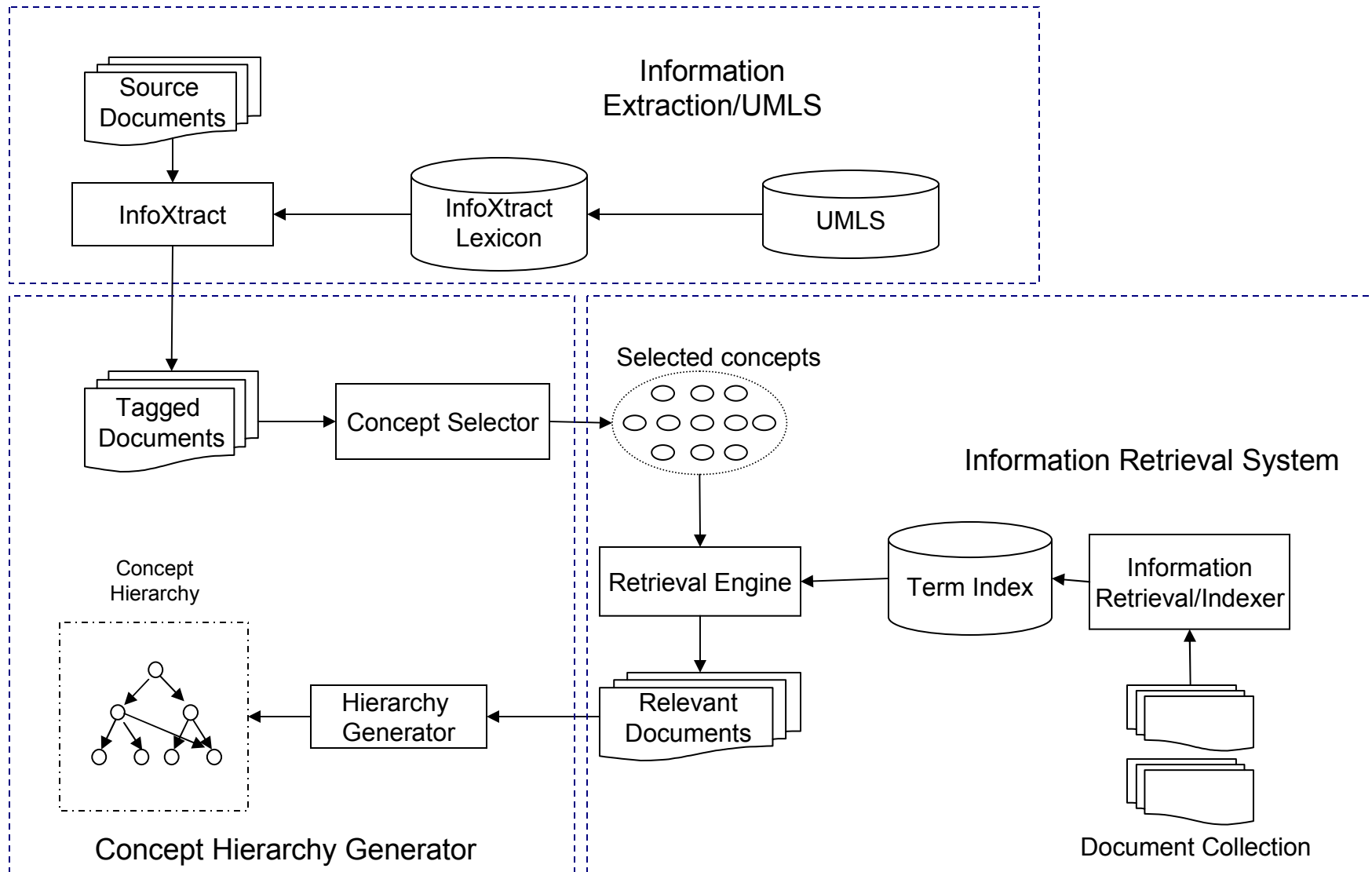
- Domain – Biomedical domain
- Collection - TREC 2003 Genomics track collection
- Source documents size – 400
- Indexed collection – 550,000 documents
- Pre-defined Domain Knowledge
 - UMLS
 - Sample UMLS concept types from UMLS
 - Amino acids, Peptides
 - Gene names
 - Genetic functions
 - Growth substances
 - Hormone substances



Solution Overview

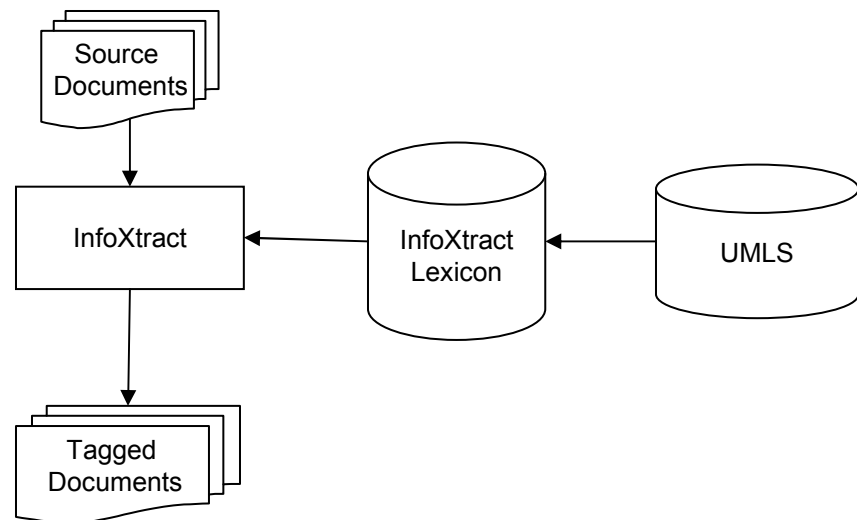
- Information Extraction, can identify meaningful concepts
- IR systems can provide with co-occurrence statistics
- An integrated IR/IE framework can be used to
 - select concepts
 - generate hierarchies

Solution Architecture



Information Extraction

- Pre-defined domain knowledge
 - Identify domain concepts
 - UMLS Metathesaurus
- InfoXtract an IE engine
 - InfoXtract lexicons
 - Add domain concepts to InfoXtract lexicons
- Select source documents
 - Document subset that covers topics relevant to the domain
- Identify concepts in the source documents
 - Tag documents using InfoXtract
 - Select concepts and relationships based on InfoXtract output





Sample Text Document

```
<DOC>
<DOCNO> 11695234 </DOCNO>
<TEXT>
<TITLE>
Castration induces apoptosis in the male accessory sex organs of
Fas-deficient lpr and Fas ligand-deficient gld mutant mice.
</TITLE>
<ABSTRACT>
The role of the Fas ligand-Fas system in castration-induced apoptosis in
the epithelia of the ventral prostate (VP), seminal vesicle (SV),
coagulating gland (CG) and epididymis (Ep) was investigated using lpr/lpr,
and gld/gld mutant mice which are deficient in Fas and Fas ligand,
respectively. The degree of apoptosis in the epithelium was quantitatively
estimated by an apoptotic ...
....
There was no
significant difference in the apoptotic index of these organs among +/+,
lpr/lpr and gld/gld mice on days 0-8 after castration. Agarose gel
electrophoresis of DNAs extracted from the VP, SV, CG and Ep of +/+,
lpr/lpr and gld/gld mice on day 4 after castration showed a ladder
pattern. The present results suggest that the Fas ligand-Fas system plays
little role in castration-induced apoptosis in the mouse male accessory
sex organs such as the VP, SV, CG and Ep.
</ABSTRACT>
</TEXT>
</DOC>
```



InfoXtract Output

```
<document id="1"><doc-info><action>add</action>
<content-provider>Unknown</content-provider>
<name>/trec/software/infoXtract_new/cdrom0/InfoXtract/docproc/input/local/116952
34</name>
<doc-size>1469</doc-size>
<orig-size>1469</orig-size>
<word-count>0</word-count>
</doc-info>
<content><sentence number="1"><tok id="1" type="atom" soff="14" eoff="22"><txt>1
1695234</txt>
</tok>
</sentence>
<sentence number="2"><tok id="2" type="atom" soff="47" eoff="57"><txt>Castration
</txt>
</tok>
<tok id="3" type="atom" soff="58" eoff="65"><txt>induces</txt>
</tok>
<tok id="4" type="atom" soff="66" eoff="75"><txt>apoptosis</txt>
</tok>
<tok id="403" type="group"><tok id="5" type="atom" soff="76" eoff="78"><txt>in</
txt>
</tok>
<tok id="351" type="group"><tok id="6" type="atom" soff="79" eoff="82"><txt>the<
/txt>
</tok>
```

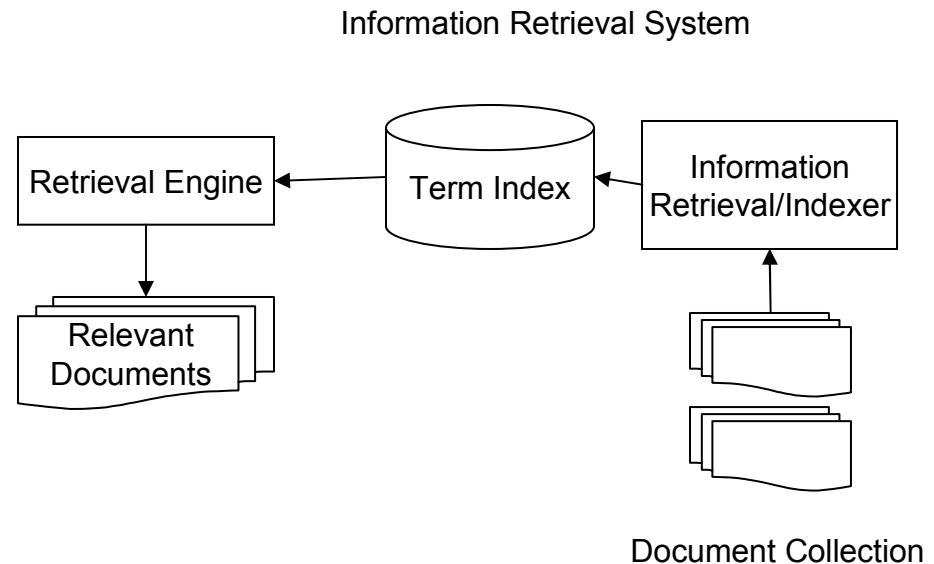


Concepts Identification

```
<DOC>
  <DOCID>5234</DOCID>
  <CONCEPTS>
    <CONCEPT ID="295" NORM="sex_organ" TYPE="NN">
      <VALUE>sex organs</VALUE>
    </CONCEPT>
    <CONCEPT ID="296" NORM="seminal_vesicle" TYPE="NN">
      <VALUE>seminal vesicle</VALUE>
    </CONCEPT>
    <CONCEPT ID="297" NORM="body_weight" TYPE="NN">
      <VALUE>body weight</VALUE>
    </CONCEPT>
    <CONCEPT ID="2" NORM="castration" TYPE="NNP">
      <VALUE>Castration</VALUE>
    </CONCEPT>
    <CONCEPT ID="3" NORM="induce" TYPE="NeGeneFun">
      <VALUE>induces</VALUE>
    </CONCEPT>
    <CONCEPT ID="4" NORM="apoptosis" TYPE="NeGeneFun">
      <VALUE>apoptosis</VALUE>
    </CONCEPT>
    <CONCEPT ID="8" NORM="accessory" TYPE="NN">
      <VALUE>accessory</VALUE>
    </CONCEPT>
  </CONCEPTS>
</DOC>
```

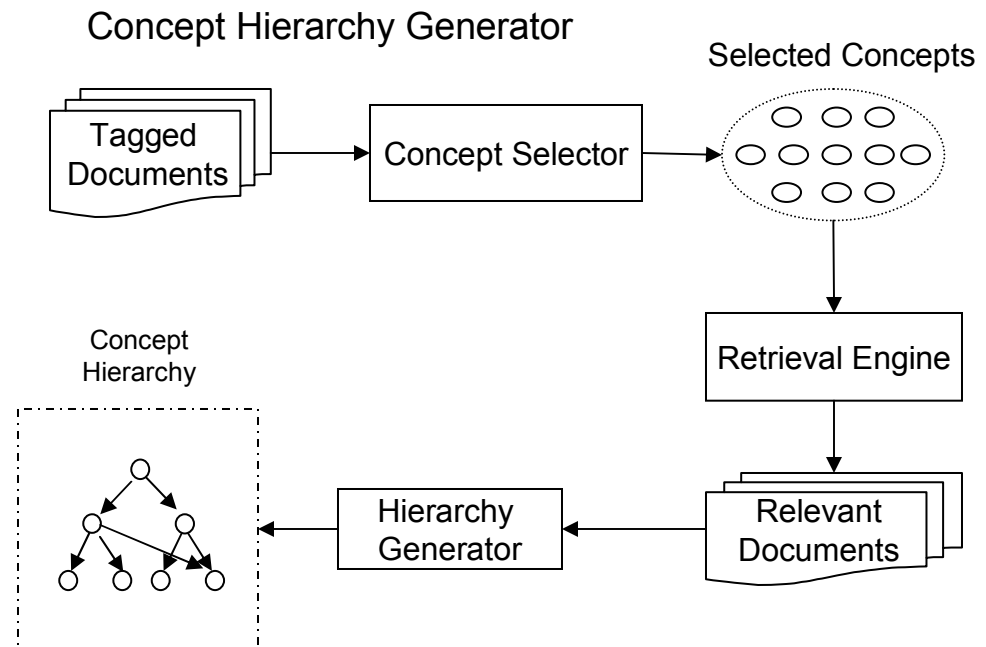
Information Retrieval

- IR system is used to
 - generate term index for obtaining co-occurrence statistics
 - Retrieve relevant documents



Concept Hierarchy Generation

- Concept hierarchy generation
 - Concept selection
 - Hierarchy generation
- Concept selection
 - Identify UMLS concepts
 - Identify new concepts
- Hierarchy generation
 - Use selected concepts to retrieve relevant documents
 - Find co-occurrence between concepts





Concept Selection

D: Set of domain (UMLS) concepts, N: Set of new concepts

NS: Set of new concepts associated with each domain concept

DS: Set of new concepts associated with each domain concept type

Procedure SelectConcepts (Input: N, NS, DS)

Collect source documents (S_d)

Identify concepts (C) and their relationship (R) in S_d

Create graph $G(C,R)$, where C is the vertex set and R the edge set

for each v in C

if v not in D

ignore

else

for each edge $r(u,v)$ in R

if u not in D

add u to N

add u to $NS(v)$

add u to DS (type (v))

end if

end for

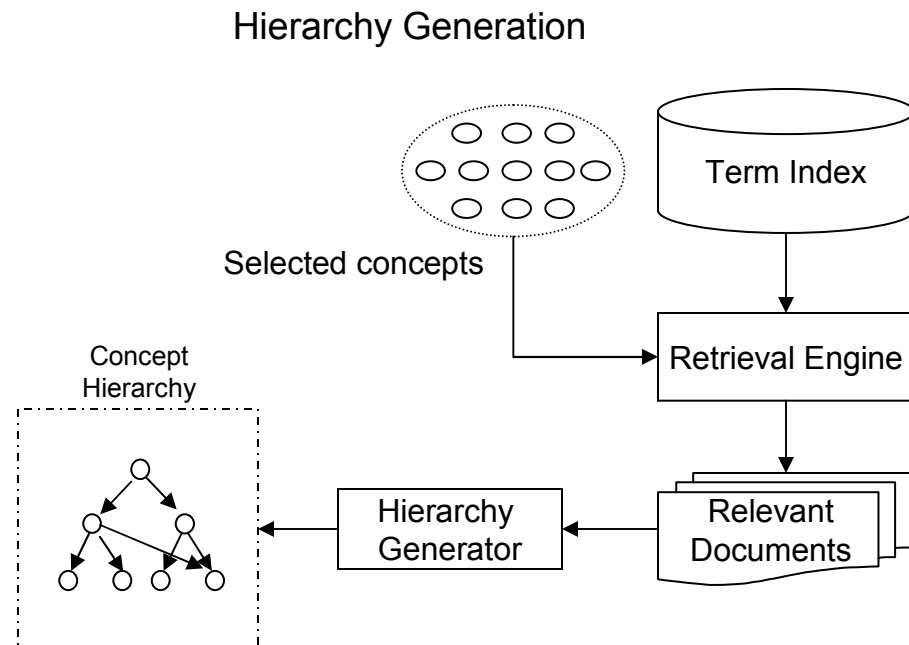
end if

end for

end of procedure

Hierarchy Generation

- Use selected concepts as query
- Retrieve relevant documents from the IR system
- For each pair of concepts find the conditional probabilities
 - $P(x|y) = \text{count}(x,y)/\text{count}(x)$
 - $P(y|x) = \text{count}(x,y)/\text{count}(y)$
- If $P(x|y) > c$, and $P(y|x) < c$ then x is assigned to be a parent of y , $c = 0.7$





Evaluation

- Previous works evaluated concept hierarchies in an IR framework
- Coverage of concepts
 - Goal of the concept selection process is to identify concepts that are related to the domain
 - Leave one out
 - Ignore concepts belonging to a single UMLS concept type
 - Use concept selection process to identify concepts belonging to the ignored concept type
 - Compute coverage



Experiments

- Grouping concepts related to each UMLS concept type - I
 - Cluster of concepts associated to concept types
 - Hierarchies generated for certain topics
- Grouping concepts related to each UMLS concept - II
 - Cluster of concepts associated to very important domain concepts
 - Better separated hierarchies
- Using all concepts - III
 - UMLS concepts + new concepts
 - Overall picture for the domain
 - Can be used for a text mining effort

Concept type grouping - I

- Concept selection coverage
 - Leave one out method
- Table shows
 - number of concepts picked versus
 - the actual number of UMLS concepts present for different types

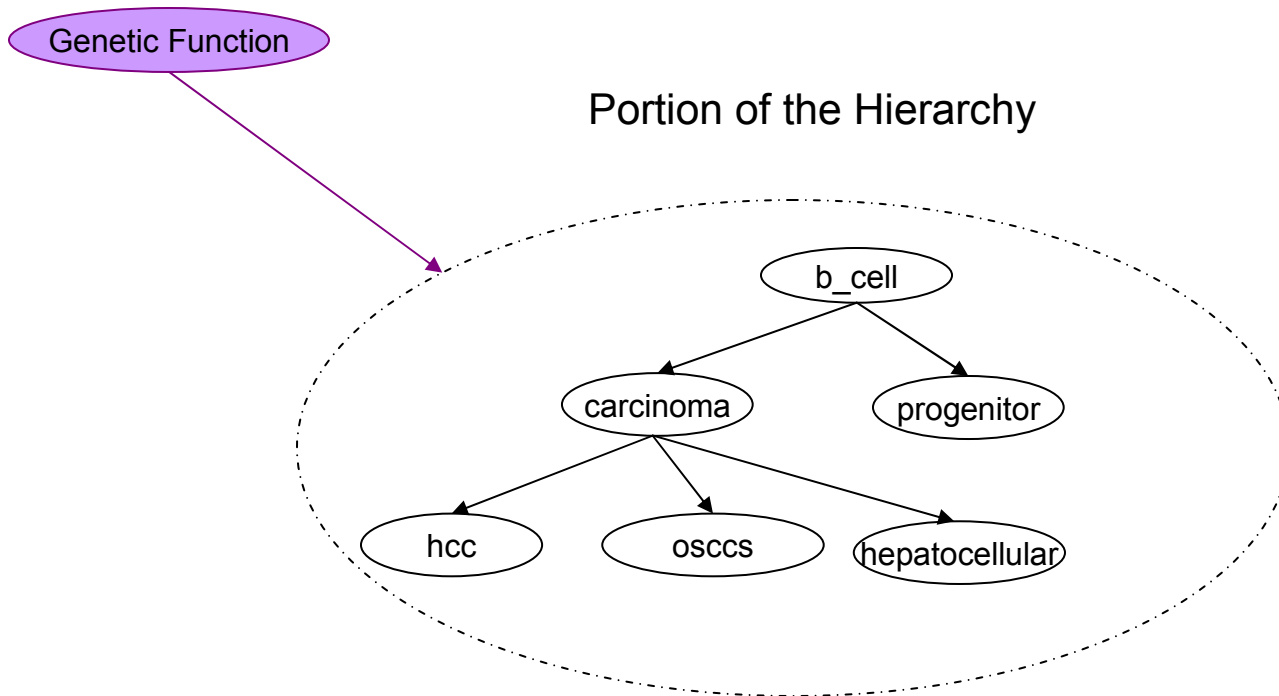
Domain Concept type	Actual Number of concepts	Concepts covered
Amino acids, Peptides	2	10
Gene Functions	50	60
Gene Name	69	70
Proteins	26	28

Concept Type Grouping - I

- Table shows
 - Number of new concepts grouped under each concept type
 - Sample new concepts

UMLS Concept type	New concepts Covered	Sample new concepts
Amino acids, Peptides	78	Glutamine, sequence
Gene Functions	1632	Growth, infection, estrogen
Gene Name	1244	Sodium dodecyl, pathway, human
Proteins	521	kinase, inhibitor

Concept type grouping I

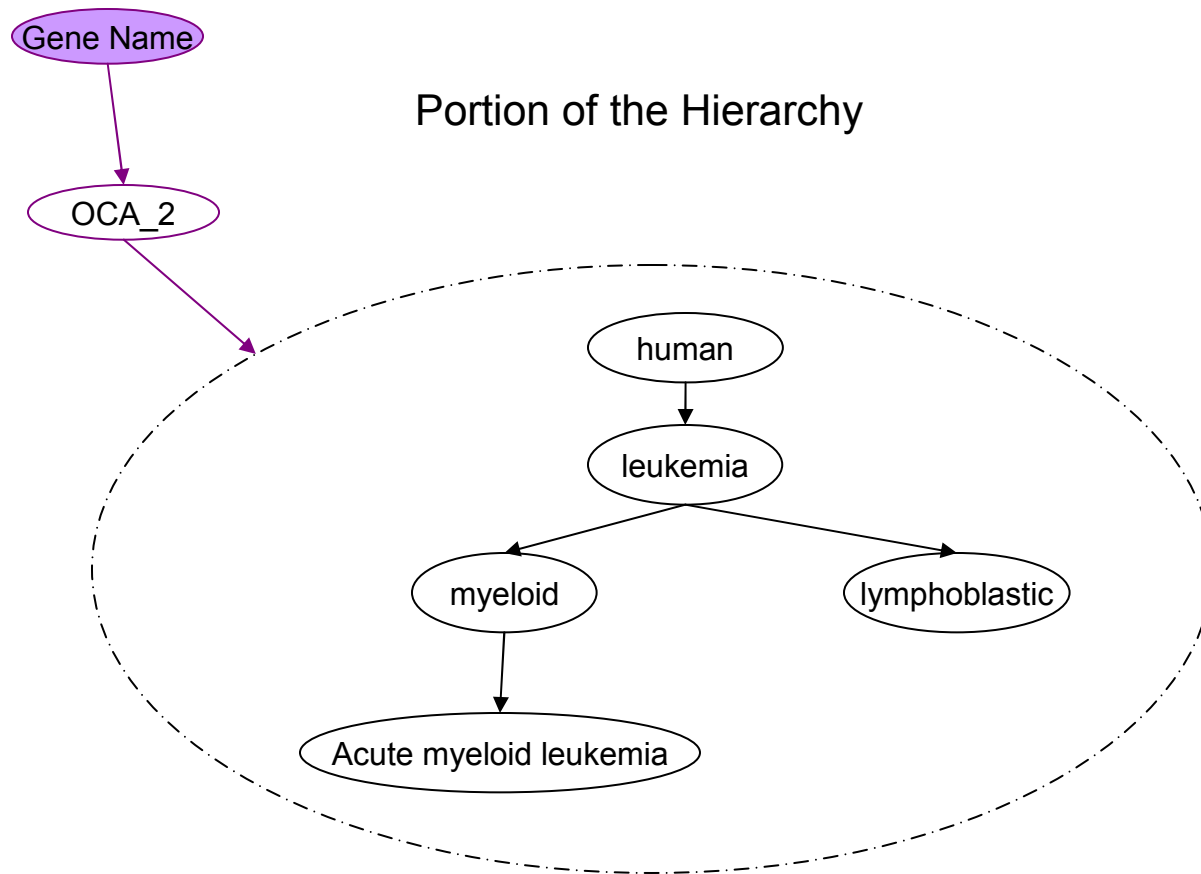


Concept Grouping - II

- Concept grouping
- Collecting all concepts relevant to each UMLS concept
- Table shows coverage of each concept

UMLS Concept	New concepts Covered	Sample new concepts
expression	847	b_cell, gene, mutation, inhibition
binding site	97	mobility, protein, promoter
inhibit	503	growth, stimulator, cancer
OCA_2	289	carcinoma, leukemia, protein

Concept grouping II





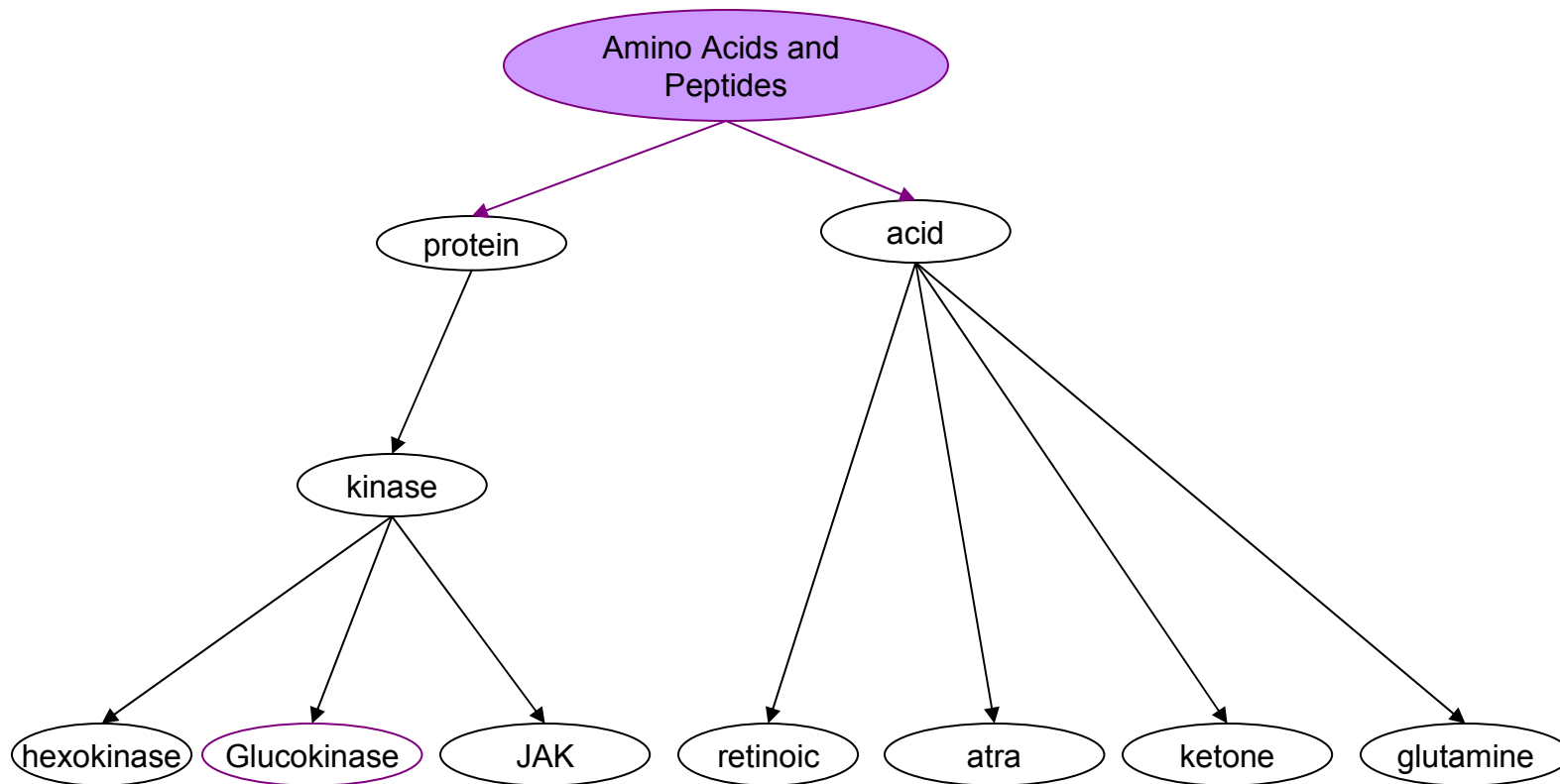
All concepts - III

- Using all concepts
- UMLS concepts
 - UMLS concepts identified by IE system
- New Concepts identified by concept selection procedure

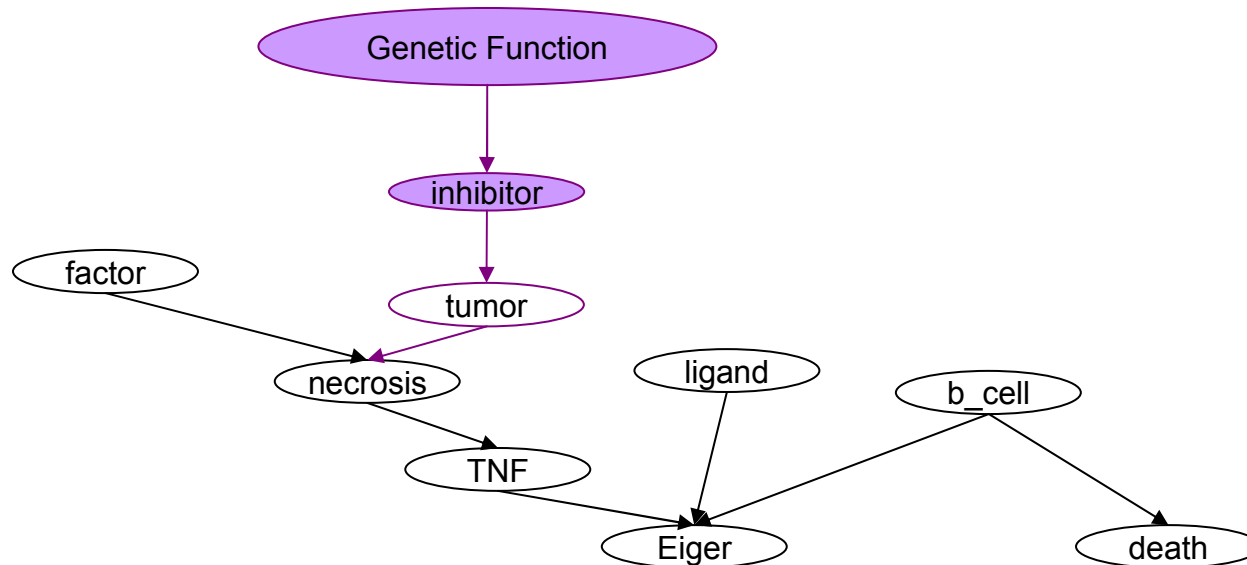
UMLS concepts	196
New concepts	1655

All concepts III

Portion of the Hierarchy



Sample concept chain



A document snippet from PubMed[8] document[13] that is relevant to the concept chain

Genetic evidence shows that [Eiger induces cell death](#) by activating the Drosophila JNK pathway. Although this cell death process is blocked by Drosophila [inhibitor-of-apoptosis protein 1 \(DIAP1\)](#), it does not require caspase activity. We also show genetically that [Eiger is a physiological ligand](#) for the Drosophila JNK pathway. Our findings demonstrate that Eiger can initiate cell death through an IAP-sensitive cell death pathway via JNK signaling.



Conclusions

- Concepts selected are domain specific
- Concepts selection method used shows good coverage
- Concepts and hierarchies can be linked to UMLS concepts and hence to the UMLS network
- The relationships generated between concepts are intuitive
- The concept hierarchies capture domain level information



Future Work

- Evaluation Methodology: Graph isomorphism with UMLS network
- Use lexical dispersion and Subsumption
- Clustering concepts and generating relationships between clusters
- Bootstrap UMLS and InfoXtract using the hierarchy generation process.
- Utility of the hierarchy in a text mining application, such as scenario identification



References

1. Yahoo! www.yahoo.com
2. WordNet, edited by Christiane Fellbaum, Cambridge, MA
3. UMLS, Unified Medical Language System, <http://www.nlm.nih.gov/research/umls>
4. Marti A. Hearst, Untangling Text Data Mining, *in proceedings of ACL 99*.
5. Don R. Swanson and N. R. Smalheiser, An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence*, 91:183-203, 1997.
6. James W. Cooper, An evaluation of Unnamed Relations Computataion for Discovery of Protein-Protein Interactions, in *SIGIR: Workshop on Text Analysis and Search for Bioinformatics*, 2003
7. Rohini K.Srihari, Miguel Ruiz, Munirathnam Srikanth, Concept Chain Graphs: A hybrid IR framework for biomedical text mining, in *SIGIR: Workshop on Text Analysis and Search for Bioinformatics*, 2003
8. M. Sanderson and W. B. Croft. Deriving concept hierarchies from text. In. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 206–213, 1999.
9. W.B.Croft, D.Lawrie and A. Rosenberg, Finding Topical Words for Hierarchical Summarization
10. Marti Hearst, *Automated Discovery of Wordnet Relations*, 1998
11. P. Anick and S. Tipirneni. The paraphrase search assistant: Terminological feedback for iterative information seeking. In M. Hearst, F. Gey, and R. Tong, editors, *Proceedings on the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 153–159, 1999.
12. W.Woods, *Conceptual Indexing: A better way to organize knowledge*, 1997
13. PubMed: National Library of Medicine, <http://www.ncbi.nlm.nih.gov/>
14. Eiger, a TNF superfamily ligand that triggers the Drosophila JNK pathway. Igaki T, Kanda H, Yamamoto-Goto Y, Kanuka H, Kuranaga E, Aigaki T, Miura M. PMID: 12065414



Acknowledgement

A special note of thanks to all those who have contributed towards this thesis

Advisor:

- Dr. Rohini K. Srihari

Committee Members:

- Dr. David Pierce
- Dr. Jian Pei

IR group members: (in no specific order)

- Munirathnam Srikanth
- Dr. Miguel E. Ruiz
- Xiayoun Wu
- Wei Dai
- Himashu Ashiya
- Dharmendra Mahi



Results

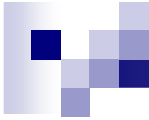
- Add InfoXtract output
- Add concept extraction output
- Test bed explanation
- Add domain concept links
- Talk abt UMLS



Background cont'd.

- Lexical dispersion

- a measure of the number of different phrases in which a word participates in
- Anick. P and Tipirneni. S [6] show a method of acquiring hierarchical relationships based on lexical dispersion
- Used to generate topical words that cover topics relevant to a document subset
- The importance of a word in a domain is modeled as the number of phrases it participates in
- Lawrie et al., compared lexical dispersion method with subsumption based methods and argue that both methods are complimentary in some respects.



- Summary of Related Work:

- Concept hierarchies can be obtained using IR frameworks based on co-occurrence of words or phrases
- Concept hierarchies can also be obtained by looking at lexical clues or lexical dispersion measure of words in phrases
- Concept selection is targeted at covering topics in the collection



Solution Overview

- Select candidate concepts
 - Use domain concepts and their relationships to identify new concepts
- Identify relevant documents
 - Use IR to identify relevant documents for candidate concepts
- Generate hierarchies based on subsumption notion
 - Use co-occurrence statistics as defined by Sanderson, Croft [2]

Sample Hierarchy

