KeyConcept: Exploiting Hierarchical Relationships for Conceptually Indexed Data

Master’s Thesis Defense

Presented by
Devanand Ravindran
University of Kansas
January 30, 2004

Committee
Dr. Susan Gauch (Chair)
Dr. Costas Tsatsoulis
Dr. Jerry James

Outline

• Motivation
• Related Work
• KeyConcept Architecture
• Exploiting Hierarchy
• Data Sets
• Experiments
• Future Work
• Conclusion
Motivation – Problem

- Search engines perform keyword search
- Index based purely on word content

“wildcats”

- What did the user really want?

Motivation - Solution

- Train and index based on word and concept
- User additionally indicates desired concept
- Use structural knowledge of training data to improve conceptual search
Related Work - I

  - Uses keyword + concept as query input
  - Classifies all documents into 20 categories – loses hierarchical information

- “Yahoo! As An Ontology – Using Yahoo! Categories To Describe Documents” - Yannis Labrou, Tim Finin.
  - Collects documents based on Yahoo! directory structure
  - Pre-classified documents collection thus made available

Related Work - II

- “Ontology-Based Web Site Mapping For Information Exploration” - Xiaolan Zhu, Susan Gauch, Lutz Gerhard, Nicholas Kral, Alexander Pretschner.
  - Uses ontologies to map user profiles to site maps

- “Collaborative Learning of Term-Based Concepts for Automatic Query Expansion” - Stefan Klink, Armin Hust, Markus Junker, Andreas Dengel.
  - Each query gets a concept assigned through relevance feedback
KeyConcept - Retrieval

Query \rightarrow \text{query/sentence} \rightarrow \text{Inverted File} \rightarrow \text{POST} \rightarrow \text{DOCs}

- Retrieval process similar for keyword and concepts

- Keyword and Concept accumulators are combined using an $\alpha$-factor
Exploiting Hierarchy

- Open Directory Project (dmoz.org) used for training documents

- ODP ontology contains hierarchical information

- Two types
  - Pruning results
  - Retrieval based on hierarchy

Exploiting Hierarchy I - Pruning Result Sets

- Search
  - Keyword: “rock”
  - Concept: arts/music/styles

- Retrieve
  - Document \( d \) belongs to arts/television/interactive
    - Level 1: \( d \) not pruned
    - Level 2: \( d \) pruned
Exploiting Hierarchy II – Hierarchy-based Retrieval

- Search
  - Keyword: “rock”
  - Concept: arts/music/styles

- Retrieve neighboring concepts in hierarchy
  - Children
  - Siblings
  - Grandchildren
  - Parent
  - Combinations

Data Sets

- Training Data
  - Open Directory Project – dmoz.org
  - Cut-off at third level of the tree
  - 2,991 concepts and 125,000 documents

- TREC Data
  - 100,000 documents from TREC’s WT2g Collection
  - 50 queries from each WT2g topic
  - Relevance judgments provided for each query

- Pruning Queries
  - TREC queries are too restrictive
  - Set of 24 queries – single-word, 2-word and 3-word length
KeyConcept Example - Input

KeyConcept
A Conceptual Search Engine

Enter Keywords: medical instruments

Enter the keywords you want to search for and select the categories you are looking for. You may select up to 10 categories.

Categories Selected:

Select Categories:
- Arts
- Business
- Computers
- Dentistry
- Home
- Insurance
- Nutrition
-便器
- 美食
- 美容
- 职业
- 职业健康与安全

Selected Categories:
- 确切
- 历史
- 手术
- 外科

Search

KeyConcept Example - Output

KeyConcept
A Conceptual Search Engine

Results:

- Consumer Health Information
  - Weight: 0.776217
  - Top 10 categories: View

- MEDICAL ASSISTANCE
  - Weight: 0.550002
  - Top 10 categories: View

- Internet Resources
  - Weight: 0.405614
  - Top 10 categories: View

Summary not available.

- KeyConcept printout/ID11510502/250.html

- KeyConcept printout/ID105180/250.html

- KeyConcept printout/ID105180/250.html
KeyConcept Example – Top Concepts

1. 7447 Top/Health/Medicine/Informatics 1.000000
2. 58346 Top/Health/Resources/Consumer 0.668753
3. 122532 Top/Health/Medicine/Directories 0.637018
4. 178733 Top/Health/Medicine/Osteopathy 0.761746
5. 7441 Top/Health/Medicine/Reference 0.754035
6. 53837 Top/Health/Resources/Professional 0.742564
7. 58443 Top/Health/Professions/Physician_Assistant 0.720177
8. 95540 Top/Health/Nursing/Internet 0.713841
9. 117578 Top/Health/Pharmacy/Drugs_and_Medications 0.685251

Experiments

- Determine baseline parameters
  - Concept matching formula
  - $\alpha$-factor

- Use Pruning on results
  - Simple pruning without conceptual retrieval
  - Pruning with conceptual retrieval

- Retrieve using hierarchical relationships
  - Parent, Children, Grandchildren
  - Combinations
Baseline Estimation – Concept Matching Formula

- Search engines use tf * idf scoring formula
  - tf = term frequency (how many times does word appear in document ?)
  - idf = inverse document frequency (how frequent is the word in the 
collection as a whole ?)

- Does cdf help ?
  - edf = Concept document frequency (how many times did the word occur 
in the concept while training ?)

- Yes it does !
  - Best precision results for tf * idf * cdf while classification

Baseline Estimation - $\alpha$-factor

- How many concepts does the user need to specify?
  - Three

- Final document score =
  \[ \alpha \cdot \text{concept score} + (1 - \alpha) \cdot \text{keyword score} \]

- What $\alpha$ yields the best precision ?
  - 0.3 or 30% importance to concept score
Pruning

- Compare effects of pruning with simple keyword search
- Contrast simple keyword search with conceptual search
- Pruning can be combined with conceptual search
- Pruning at Level 1 and Level 2

Simple Keyword Search vs. Pruning

- Best results for single-word queries
- Overall, level 2 pruning more effective than level 1 pruning
Pure Keyword vs. Conceptual Search

- Conceptual search performs better than simple keyword search for all query lengths

Pure Keyword vs. Conceptual Search with Pruning

- Keyword + Conceptual search + Pruning results = Great Precision
Overall Comparison

- Level 1 and level 2 results averaged for comparison
- And the winner is … Keyword + Conceptual + Pruning

Retrieval using Hierarchy
Retrieval using Hierarchical Relationships
- Siblings

- Adding siblings of chosen concepts doesn’t help
- Merely increases noise

Retrieval using Hierarchical Relationships
- Children

- Best results among hierarchical neighbors
- Maximum increase when user specifies only one concept
  - Weight given to children = 0.3 or 30%
Retrieval using Hierarchical Relationships
- Grandchildren

- Modest increase in precision at a weight of 0.1 for grandchildren
- Maybe combinations of the above would help…?

Retrieval using Hierarchical Combinations

- Most promising hierarchical relations chosen – children and grandchildren
- Slight increase – not significant
Conclusions

• cdf (the frequency of word occurrence in a concept) needed during conceptual classification

• $\alpha = 0.3$ yields maximum precision
  – Keyword retrieval more important than pure conceptual retrieval

• Pruning along with conceptual retrieval gives the best results
  – Precision increase from 36.70% to 63.77%

• Using children of chosen concept obtains best increase in precision (from 24.6% to 28.6%)
  – Including parent of concept showed no improvement due to sparse data in top two levels of ODP
Future Work

• Better Data
  – More content in the top two levels of the ODP ontology

• Contextualization
  – Detect user’s intent by gathering information about user’s context of search
    – Open windows, past search history etc.

• Personalization
  – Track user’s preferences and interests
  – Implement user profile in a similar hierarchy