Wikitology
Wikipedia as an Ontology
Zareen Syed, Tim Finin and Anupam Joshi
University of Maryland
Baltimore County
zarsyed1@umbc.edu, finin@cs.umbc.edu, joshi@cs.umbc.edu
Outline

• Introduction and motivation
• Wikipedia
• Methodology and Experiments
• Evaluation
• Future Work Directions
• Conclusion
Introduction

• Identifying the topics and concepts associated with a document or collection of documents is a common task for many applications and can help in:
  – **Annotation** and categorization of documents in a corpus.
  – Modelling **user interests**
  – **Business intelligence**
  – Selecting **Advertisements**
Motivation

• **Problem:** describe what an analyst has been working on to support collaboration

• **Idea:**
  – track documents she reads
  – map these to terms in an ontology
  – aggregate to produce a short list of topics

• intro  • wikipedia  • experiments  • evaluation  • next  • conclusion
Approach

• Use Wikipedia articles and categories as ontology terms
• Categories as Generalized Concepts
• Articles as Specialized Concepts
• How to map the documents she reads to the ontology terms?
  – Use document to Wiki-article similarity for the mapping
• How to aggregate to get a shorter list?
  – Use spreading activation algorithm for aggregation
What’s a document about?

• Two common approaches:
  
  (1) **Statistical Approach**
  Select words and phrases using TF-IDF that characterize the document

  (2) **Controlled Vocabulary or Ontology**
  Map document to a list of terms from a controlled vocabulary or ontology

• First approach is **flexible** and does not require creating and maintaining an ontology

• Second approach can tie documents to a rich knowledge base
Wikitology!

- Using Wikipedia as an ontology offers the best of both approaches
- Each article is a concept in the ontology
- Terms linked via Wikipedia’s category system and inter-article links
- It’s a consensus ontology created, kept current and maintained by a diverse community
- Overall content quality is high
- Terms have unique IDs (URLs) and are “self describing” for people
- Underlying graphs provide structure: categories, article links
Wikipedia Graph Structures

- Wikipedia Category graph is a thesaurus

- Wikipedia Page links graph is similar to WWW Network
Methods

• **Goal**: given one or more documents, compute a ranked list of the top N Wikipedia **articles** and/or **categories** that describe it.

• **Basic metric**: document similarity between Wikipedia article and document(s)

• **Variations**:
  – role of categories
  – eliminating uninteresting articles
  – use of spreading activation
  – using similarity scores for weighing links
  – number of spreading activation pulses
  – individual or set of query documents, etc, etc.
Spreading Activation

• In **associative retrieval** the idea is that it is possible to retrieve relevant documents if they are associated with other documents that have been considered relevant by the user.

• The documents can be represented as nodes and their associations as links in a network.
Spreading Activation

Start with an initial set of activated nodes
At each pulse/iteration, spread activation to adjacent nodes
Spreading Activation

Some nodes will have higher activation than others

Constraints
- Distance
- Fan out
- Path constraints
- Activation threshold
Method 1

Using Wikipedia Article Text and Categories to Predict Concepts

Input
Query doc(s) similar to
Cosine similarity

Similar Wikipedia Articles
0.8
0.2
0.1
0.3
0.2
0.1
0.3
0.2
Method 1

Using Wikipedia Article Text and Categories to Predict Concepts

Input

Query doc(s)

similar to

Cosine similarity

0.2
0.1
0.3
0.2
0.8

Similar Wikipedia Articles

Wikipedia Category Graph
Method 1

Using Wikipedia Article Text and Categories to Predict Concepts

Output

Rank Categories
1. Links
2. Cosine similarity

Input

Query doc(s)

similar to

Cosine similarity

0.2
0.1
0.3
0.8
0.2
0.3
0.9

Similar Wikipedia Articles

Wikipedia Category Graph
Method 2

Using Spreading Activation on Category Links Graph to get Aggregated Concepts

Output
Ranked Concepts based on Final Activation Score

Input
Query doc(s)

Similar to
Cosine similarity

Input Function \( I_j = \sum_i O_i \)

Output Function \( O_j = \frac{A_j}{D_j * k} \)
• Can we predict concepts that are **NOT** present in the category hierarchy?
• Use the **article concepts**!
• But **How**?
Method 3
Using Spreading Activation on Article Links Graph

Input
Query doc(s)
Similar To

Threshold: Ignore Spreading Activation to articles with less than 0.4 Cosine similarity score

Edge Weights: Cosine similarity between linked articles

Spreading Activation

Wikipedia Article Links Graph

Node Input Function
\[ I_j = \sum_i O_i w_{ij} \]

Node Output Function
\[ O_j = \frac{A_j}{k} \]

Output
Ranked Concepts based on Final Activation Score
Preliminary Experiments

- An initial informal evaluation compared results against our own judgments
- Downloaded articles from internet and predicted concepts
- Using Single Document and Group of Related Documents

Prediction for Single Test Document

<table>
<thead>
<tr>
<th>Test Document Title</th>
<th>Method 1 Ranking Categories Directly</th>
<th>Method 2 Spreading Activation Pulses=2</th>
<th>Method 2 Spreading Activation Pulses=3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weather Prediction of thunder storms (CNN)</td>
<td>“Weather_Hazards”</td>
<td>“Weather_Hazards”</td>
<td>“Meterology”</td>
</tr>
<tr>
<td></td>
<td>“Winds”</td>
<td>“Current_events”</td>
<td>“Nature”</td>
</tr>
<tr>
<td></td>
<td>“Severe_weather_and_convection”</td>
<td>“Types_of_cyclone”</td>
<td>“Weather”</td>
</tr>
</tbody>
</table>

More pulses -> More Generalized Concepts
# Preliminary Experiments

## Prediction for Set of Test Documents

### Test Document Titles in the Set: (Wikipedia Articles)
- Crop_rotation
- Permaculture
- Beneficial_insects
- Neem
- Lady_Bird
- Principles_of_Organic_Agriculture
- Rhizobia
- Biointensive
- Intercropping
- Green_manure

### Concept not in the Category Hierarchy
- Organic_farming
- Sustainable_agriculture
- Organic_gardening
- Agriculture
- Companion_planting

<table>
<thead>
<tr>
<th>Method 1</th>
<th>Method 2 (2 pulses)</th>
<th>Method 3 (2 pulses)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranking Categories Directly</td>
<td>Spreading Activation on Category links Graph</td>
<td>Spreading Activation on Article Links Graph</td>
</tr>
<tr>
<td>Agriculture</td>
<td>Skills</td>
<td>Organic_farming</td>
</tr>
<tr>
<td>Sustainable_technologies</td>
<td>Applied_sciences</td>
<td>Sustainable_agriculture</td>
</tr>
<tr>
<td>Crops</td>
<td>Land_management</td>
<td>Organic_gardening</td>
</tr>
<tr>
<td>Agronomy</td>
<td>Food_industry</td>
<td>Agriculture</td>
</tr>
<tr>
<td>Permaculture</td>
<td></td>
<td>Companion_planting</td>
</tr>
</tbody>
</table>

- intro • wikipedia • experiments • **evaluation** • next • conclusion •
Evaluation

- Select wikipedia articles randomly and predict their categories and links
- Sort the results based on Average Similarity

**Average Similarity**

```
Query doc(s) similar to

Cosine similarity

0.5 + 0.9 + 0.7 + 0.2 + 0.8

5
```
Evaluation

- If our system predicts a category three levels higher in hierarchy than the original category we consider our prediction to be correct.

Observation

Articles are linked often with super and sub categories both.

Diagram:

- Medicines
- Medical Treatments
- Antibiotics
- Tetracyclin
- Oxytetracyclin
Spreading activation with two pulses worked best

Only considering articles with similarity > 0.5 was a good threshold
• Spreading activation with one pulse worked best
• Only considering articles with similarity > 0.5 was a good threshold

![Graph showing the relationship between average similarity threshold and precision, average precision, recall, and F-measure. The graph indicates that precision and average precision increase with higher similarity thresholds, while recall and F-measure decrease.]

Similar Documents, N = 5
Spreading Activation pulses=1

• intro • wikipedia • experiments • evaluation • next • conclusion •
Prediction Accuracy

• Issues:
  – To what *extent* the concept is represented in Wikipedia For eg. we have a category related to the fruit apple but not for mango
  – Presence of links between *semantically* related concepts
  – Presence of links between irrelevant articles (term definitions, country names)

• Possible Solutions:
  – Use *Average Similarity Score* to measure the extent of concept representation within Wikipedia
  – Use existing *semantic relatedness measures* to handle presence or absence of semantically related links
Potential Applications

- **Recommending** categories and links for new Wikipedia articles
- Introducing *new* Wikipedia categories
- **Automating** the process of building a Wiki from a corpus
Future Work

• Classifying links in Wikipedia using Machine learning techniques
  – To Predict semantic type of article
  – To control flow of spreading activation
• Exploit parallel execution on cluster
• Refining Wikipedia ontology
• Bridging the gap between Wikipedia and formal ontologies
Document Expansion with Wikipedia Derived Ontology Terms

• Expansion of each TREC document using Wikitology terms
• We are still working on refining the methodology

* In Collaboration with Paul McNamee, John Hopkins University Applied Physics Laboratory

Doc: FT921-4598 (3/9/92)
... Alan Turing, described as a brilliant mathematician and a key figure in the breaking of the Nazis' Enigma codes. Prof IJ Good says it is as well that British security was unaware of Turing's homosexuality, otherwise he might have been fired 'and we might have lost the war'. In 1950 Turing wrote the seminal paper 'Computing Machinery And Intelligence', but in 1954 killed himself ...

Turing_machine, Turing_test, Church_Turing_thesis, Halting_problem, Computable_number, Bombe, Alan_Turing, Recusion_theory, Formal_methods, Computational_models, Theory_of_computation, Theoretical_computer_science, Artificial_Intelligence
Conclusion

• We tested the idea of using Wikitology for describing documents and proposed different methods using the Wikipedia article text, category links and article links

• Suggested improvements

• Using average similarity to judge the accuracy of prediction

• Easily extendable to other wikis and collaborative KBs, e.g., Intellipedia, Freebase
References


References


• URL:http://arxiv.org/abs/cs/0606097v1


Thank you

Questions and Suggestions?