

## College of Information Science and Technology



Drexel E-Repository and Archive (iDEA)

<http://idea.library.drexel.edu/>

Drexel University Libraries

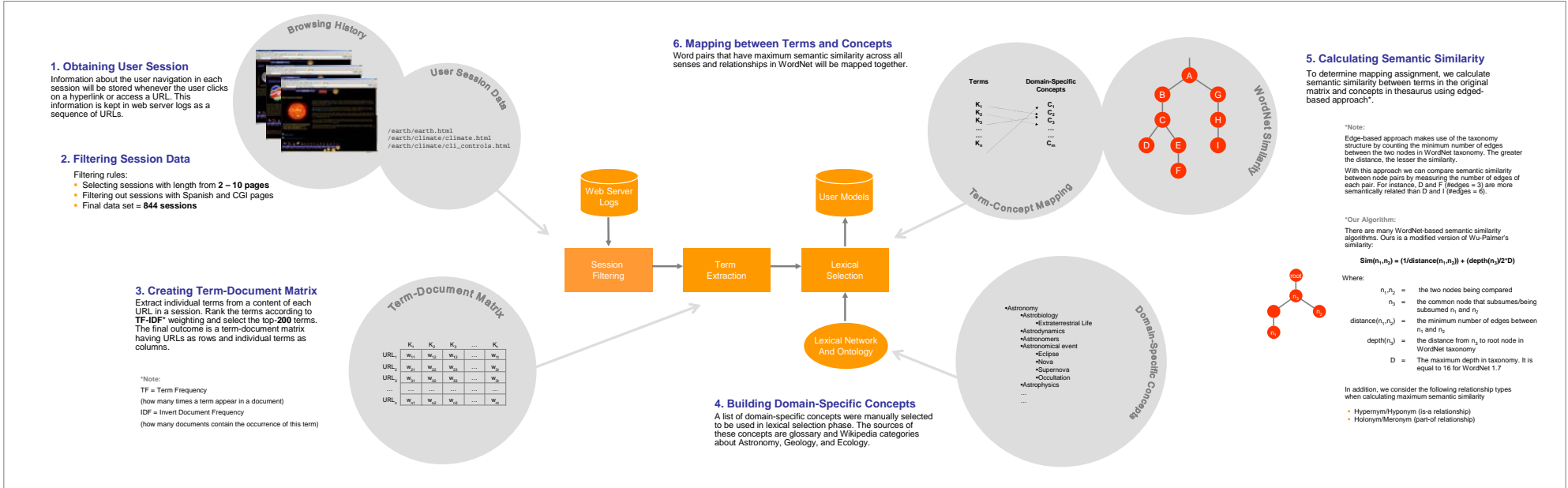
[www.library.drexel.edu](http://www.library.drexel.edu)

The following item is made available as a courtesy to scholars by the author(s) and Drexel University Library and may contain materials and content, including computer code and tags, artwork, text, graphics, images, and illustrations (Material) which may be protected by copyright law. Unless otherwise noted, the Material is made available for non profit and educational purposes, such as research, teaching and private study. For these limited purposes, you may reproduce (print, download or make copies) the Material without prior permission. All copies must include any copyright notice originally included with the Material. **You must seek permission from the authors or copyright owners for all uses that are not allowed by fair use and other provisions of the U.S. Copyright Law.** The responsibility for making an independent legal assessment and securing any necessary permission rests with persons desiring to reproduce or use the Material.

Please direct questions to [archives@drexel.edu](mailto:archives@drexel.edu)

# Using Semantic Similarity to Improve User Modeling in Web Personalization Systems

Palakorn Achananuparp and Hyoil Han  
College of Information Science and Technology, Drexel University



## Problems and Hypothesis

Polysemy (one word, multiple meanings) and synonymy (multiple words, one meaning) are well-known problems in keyword-based approach in information retrieval and filtering. This is also a big issue for web personalization system because they prevent user interests and preferences to be accurately captured by the system.

As a result, we formulate our main hypothesis as follow:

- We can improve the accuracy of user model by incorporating semantic content over keyword-based approach
- Domain-specific concepts are better than individual terms as representation of user interests
- Semantic similarity can be used as a measure to map between terms and domain-specific concepts

So the follow-up questions are:

- What are the best sources for domain-specific concepts, e.g. domain ontology, thesaurus?
- Which semantic similarity techniques we should use?

Approach	Technique	Pros/Cons	Example
Statistical	Probability of co-occurrence in a text corpus	Accurately reflect human performance <i>Type of relatedness is not captured</i> <i>Different word senses are treated as one</i>	Latent Semantic Analysis (LSA) Point-wise Mutual Information using IR (PMI-IR)
Taxonomical	The contents and relations of a hierarchy of terms	Provide relationships between terms <i>Require a comprehensive knowledge base</i> <i>Taxonomy structure is subjective</i>	WordNet-based semantic similarity
Hybrid	Taxonomy + statistical properties	Enhance taxonomy with corpus statistics for similarity & relatedness values	WordNet-based measure + probability-based information content

Figure 2: Semantic Systems Comparison (source: Kaur & Hornof, 2005)

## Abstract

Personalization is a process by which the users are presented with web resources customized to their interests. Critical to the personalization process is the user model which is the system's representation of the user characteristics and preferences. However, a keyword-based user model used in most web personalization systems does not consider the semantics of the content.

In this study, we propose a method to improve user modeling in web personalization systems by incorporating the semantic content. To achieve that, we map keywords extracted from web pages' contents to concepts in domain ontology using semantic similarity between terms in WordNet taxonomy.

## Methodology

### Session Data

We obtained the total of 844 user sessions from The Window to the Universe website (<http://www.windows.ucar.edu>). The number of URLs in each session ranges from 2 – 10 URLs. Note that these sessions were generated by web crawler.

### Semantic Similarity Technique

We chose WordNet-based semantic similarity technique since it is simple to implement, requires less computational power, and provides various types of relationships between words. Moreover, it is reasonably accurate in determining semantic similarity as compared with human judgment (our implementation gave 86% correlation with human judgment).

### Domain-Specific Concepts

Initially, we planned to use domain ontology (astronomy) as a source of concepts. However, many ontologies we found are work-in-progress and are not comprehensive enough. Thus, we manually created a thesaurus containing 181 topics related to astronomy, geology, and ecology (content domains in the website). The sources of these topics are web glossary and Wikipedia categories.

## Discussion & Future Works

We are at the stage of generating and evaluating mapping results. We have processed a few sessions so far and some of the results look promising while some need further refinements.

Source Terms	Target Concepts
Metal	Metallic Element
Nickel	
Iron	
Monkey	Primates
Gorilla	
Chimpanzee	
Orangutan	
Sink	Craters
Fault	
Collection	Galaxy

Figure 3: Examples of Term-Concepts Mapping

For the future works, we plan to:

- Adjusting cut-off level for semantic similarity
- Adjusting weighting scheme for the mapped terms
- Refining concepts in thesaurus
- Performing user evaluation of the user models