College of Information Science and Technology

Drexel E-Repository and Archive (iDEA)
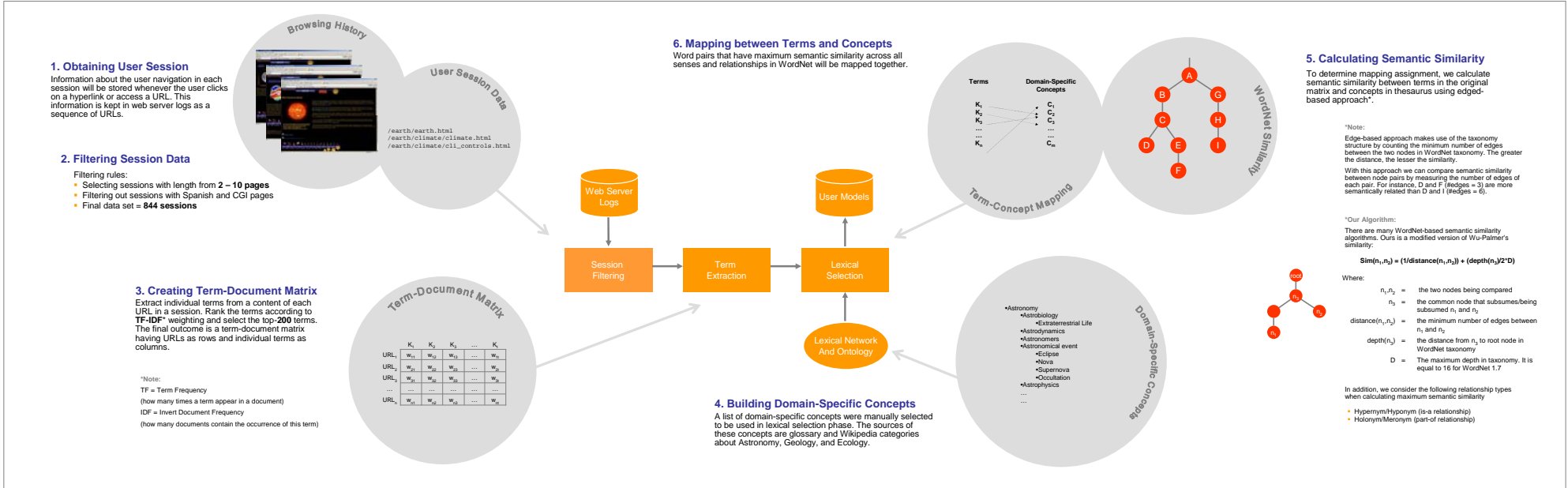http://idea.library.drexel.edu/

Drexel University Libraries
www.library.drexel.edu

Please direct questions to 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

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.

Figure 1: Overall Process

1. Obtaining User Session
Information about the user navigation in each session will be stored whenever the user clicks on a hyperlink or access a URL. This information is kept in web server logs as a sequence of URLs.

2. Filtering Session Data
Filtering rules:
- Selecting sessions with length from **2 – 10 pages**
- Filtering out sessions with Spanish and CGI pages
- Final data set = **844 sessions**

3. Creating Term-Document Matrix
Extract individual terms from a content of each URL in a session. Rank the terms according to **TF-IDF*** weighting and select the top-**200** terms. The final outcome is a term-document matrix having URLs as rows and individual terms as columns.

*Note:
TF = Term Frequency
(how many times a term appear in a document)
IDF = Invert Document Frequency
(how many documents contain the occurrence of this term)

4. Building Domain-Specific Concepts
A list of domain-specific concepts were manually selected to be used in lexical selection phase. The sources of these concepts are glossary and Wikipedia categories about Astronomy, Geology, and Ecology.

6. Mapping between Terms and Concepts
Word pairs that have maximum semantic similarity across all senses and relationships in WordNet will be mapped together.

5. Calculating Semantic Similarity
To determine mapping assignment, we calculate semantic similarity between terms in the original matrix and concepts in thesaurus using edged-based approach*.

*Note:
Edge-based approach makes use of the taxonomy structure by counting the minimum number of edges between the two nodes in WordNet taxonomy. The greater the distance, the lesser the similarity.

With this approach we can compare semantic similarity between node pairs by measuring the number of edges of each pair. For instance, D and F (#edges = 3) are more semantically related than D and I (#edges = 6).

*Our Algorithm:
There are many WordNet-based semantic similarity algorithms. Ours is a modified version of Wu-Palmer's similarity:

$$Sim(n_1, n_2) = (1/distance(n_1, n_2)) + (depth(n_3)/2*D)$$

Where:
n_1, n_2	the two nodes being compared
n_3	the common node that subsumes/being subsumed n_1 and n_2
$distance(n_1, n_2)$	the minimum number of edges between n_1 and n_2
$depth(n_3)$	the distance from n_3 to root node in WordNet taxonomy
D	The maximum depth in taxonomy. It is equal to 16 for WordNet 1.7

In addition, we consider the following relationship types when calculating maximum semantic similarity:

- Hypernym/Hyponym (is-a relationship)
- Holonym/Meronym (part-of relationship)

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 / Type of relatedness is not captured / Different word senses are treated as one	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 / Require a comprehensive knowledge base / Taxonomy structure is subjective	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)

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 / Nickel / Iron	Metallic Element
Monkey / Gorilla / Chimpanzee / Orangutan	Primates
Sink / Fault	Craters
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