Visual Clustering and Detecting Community Structure in a Web Usage Network of E-learners

Leyla Zhuhadar, Ph.D.
Research Scientist
Western Kentucky University, USA.
How can we detect a community of similar users based on the structure of a huge social network?

How can we present this interconnection among communities visually to analyze our users’ behaviors?

Our final goal is building a community-based recommendation system.
The knowledge retained by Cyberlearners when they use search engines (keywords search) to find learning resources could be compared to 

*Episodic Memory*

– or knowledge based on a particular concept.

The semantic search not only increases the knowledge base on a particular concept, it also helps the searcher discover other words, notions, or ideas *related* to the original item searched–what we call 

*Semantic Memory.*

Our goal is to help Cyberlearners to grasp multiple concepts and build–what we call 

*Mental Encyclopedia*.
What is the HyperManyMedia Repository?

Proposed Architecture

1. Semantic Domain (hierarchy)
2. Learner Activity
3. Pruning Algorithm (extracting user ontology profile)
4. Learner Semantic Profile (hierarchy)
5. Semantic Domain Clusters (top keywords in each cluster)
6. Mapping Algorithm (similarity to clusters)
7. Re-ranking Learner Search Algorithm (Reorder documents)
8. Personalized Semantic Search
9. Semantic terms-recommendation Algorithm (Semantic terms)
10. Semantic Term Recommendations

HMM Architecture
Information Visualization Toolkit: http://prefuse.org/

Overall Modeling and Searching Techniques

Generic Search
- Using standard keyword search (TF/IDF)

Metadata Search
- Using Metadata Search Fields for (Language, College, Course, etc.)

Personalized Semantic Search
- Using OWL/RDF

User Relevance Feedback
- Updating User Profiles based on his/her Relevance Feedback

Recommender System
- Updating User Profiles based on similar Users

Cross-Language Search
- Providing the user with MLIR (English/Spanish)

Visual Search
- Using Prefuse to Visualize HyperManyMedia ontology

- Vector Space Model (VSM)
- Metadata Techniques
- Ontology (Protégé)
- Rocchio’s Algorithm
- K-Nearest Neighbor & Fast XOR bit Operation Method
- Thesaurus-based Approach & Corpus-based Approach
- Prefuse Graph Model
What is the HyperManyMedia Repository?

WKU, Distance Learning HyperManyMedia Platform

Home Back

Search Normal Search | Metadata Search | Cross Language Search | Visual Ontology Search

Engineering

Semantic Search

Sub Classes:
- Database Internet and Systems Integration Technologies
- Hydrology
- Flood Plain Management

Synonyms:
- Engineering
- Ingenieria
What is the HyperManyMedia Repository?

The HyperManyMedia Repository is a platform that provides access to a variety of lecture materials, including presentations and problem-solving sessions. It offers search features to help users find relevant content based on keywords and specific course topics. The repository covers a range of subjects, such as urban transportation planning, logistical and transportation planning methods, database Internet and system integration technologies, and introduction to computers and engineering problem solving.
Evaluation of Recommender Search Engine

Non-Personalized Semantic vs. Personalized Semantic Search Results

Top-n-Recall and Top-n-Precision:

- Top – n Recall = \frac{\text{number of relevant retrieved documents within top n results}}{\text{total number of relevant documents}}
- Top – n Precision = \frac{\text{number of relevant retrieved documents within top n results}}{n}

Average Percentage of Improvement in Top-n Recall
Average Percentage of Improvement in Top-n Precision
10 profiles with three sizes of queries (1, 2, and 3 keywords)

• This is great! But is our cognitive system able to deal with this vast amount of resources?

• The most difficult question raised here: Is our conceptual recognition of these learning resources able to find what we really want?
Lexical access during speech perception can be successfully modeled as a process mediated by identification of individual primitive elements, the phonemes from a small set of primitives. We need only 44 phonemes to code all the words in English (Wilson, 1980).

(Biederman, 1987) estimated that as few as 36 geons could produce millions of unique objects.
How is it possible to visualize a graph which represents knowledge and reasoning of a massive, ambiguous, and vast document set using minimum vocabulary?

- **Zipf's laws**– (Human Behavior and the Principle of Least Effort, 1949): building a small set of vocabulary that represents the whole domain of our repository.

- **Collocation Concept** (Manning, 1999): compound phrases (e.g., Introduction to Literature: as a course name).
The Power of Visual Recognition

- We call this representation “top-down” cognitive representation.
- It starts with a knowledge driven by the Cyberlearner who knows what he/she is looking for.
- Visually finds his/her learning resource with three clicks!
The Power of Visual Recognition
Comparison between All the Models

Cyberlearners' Preferences (2006-2011)

- Generic Search: 1
- Metadata Search: 3
- Semantic Search: 7
- Personalized Semantic Search: 12
- Personalized Search with User Relevance: 5
- Collaborative Filtering: 6
- Cross-Language Search: 8
- Visual Search: 58
I am a Cyberlearner and need help to find a learning resource!

But, I really don’t know what type of help I need!
I am a Cyberlearner and need help to find a resource!

But I really don’t know what type of help I need!
1. How can we visually analyze a Logfile of recent activities of users to learning resources?
2. How can we find communities in this social network (Logfile)?
3. How can this finding of communities in this social network benefit the users?
Methodology

- Proposing a bottom-up approach (No pre-knowledge).
- Data-driven approach: archived activities of logfile for the last 6 years of Cyberlearners visited HMM (~800,000).
- Looking underneath the structure of HMM social networks.
• Community structure: tendency for vertices to divide into groups with dense connections within groups and only sparse connections between them [3,4]

• Communities are of interest because they often correspond to collections of pages on a single topic on the web [5]

• Reminder! Simplistic approach (analogy between language and perception still holds)
Three algorithms were deployed on HyperManyMedia's logfile extracted during the period of (2/1/2011- 8/1/2011):

- Fruchterman and Reingold
- Yi Fan Hu
- ForceAtlas2

After filtering out some data based on the conditions:

- (users' visits >= 10 & length of accumulated sessions >= 30 minutes)
- Deleting outliers

The network consisted of 8,510 Vertices and 23,079 Edges (# of edges between learning resources)

**Modularity measurement** was used to visualize the network structure [6].

\[ Q = \frac{1}{2m} \sum_{k=0}^{n} \left[ A_{ij} - \frac{k_ik_j}{2m} \right] \delta(c_i, c_j); \]

where \( A_{ij} \) represents the weight of the edge between \( i \) and \( j \), \( k_i = \sum_j A_{ij} \) is the sum of the weights of the edges attached to vertex \( i \), \( c_i \) is the community to which vertex \( i \) is assigned, the \( \delta \)-function \( \delta(u, v) \) is 1 if \( u = v \) and 0 otherwise and \( m = \frac{1}{2} \sum_{ij} A_{ij} \).
Fruchterman and Rheingold

Force Atlas 2

Yi-Fan Hu

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Modularity</th>
<th># of Communities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fruchterman and Rheingold</td>
<td>0.606</td>
<td>14</td>
</tr>
<tr>
<td>Force Atlas 2</td>
<td>0.610</td>
<td>14</td>
</tr>
<tr>
<td>Yi-Fan Hu</td>
<td>0.607</td>
<td>15</td>
</tr>
</tbody>
</table>

**Community Centrality:** certain vertices their situation within the network have the power to make substantial contributions to the overall modularity of the network.

- Vertices diameters indicate the community centrality.
- **Community Centrality = Centrality measure = vector magnitude |xi|**.

---

**Figure 2**

Static phase:
- Detecting communities (Modularity measure)
- Discovering Hubs in each community (Community Centrality)
- Defining edges linked to Hubs (thicker=higher correlation between the hub and a resource)
- Order Top 5-9 resources (future recommendation)
Recap: Semi-automatic Approach

**Dynamic phase:**
- New user → Create User Profile
- Mapping stage:
  - User Profile → Hubs
- Detect the **Best Hub**
- Recommendation (5-9 top new courses)
- Recalculate for each single access to HMM
I am a **Cyberlearner** and need help to find a resource!

But I really don’t know what type of help I need!
I am a **Cyberlearner** and need help to find a resource!

But I really don’t know what type of help I need!
Pros:

- detect a community of similar users based on the structure of a huge social network
- present the interconnection among communities visually to analyze our users’ behaviors?
- Discovering Hubs in each community (Community Centrality)
- Future Work: build recommendation based on Community Centrality.

Caveats:

- No assessments in HMM: Re-examine the design of HyperManyMedia platform by incorporating new assessment methodologies that could transform students’ knowledge, and success such as scaffolding theory and computational modeling to identify instructional conditions that cause robust learning.


Thanks for your Attention!

Leyla Zhuhadar, Ph.D.
Email: leyla.zhuhadar@wku.edu