**Learning-Based Concept-Hierarchy Refinement through Exploiting Topology, Content and Social Information**

Tsung-Ting Kuo, Shou-De Lin
MSLab, GINM, NTU

## Introduction

This is the readme for the source code of the "Learning-Based Concept-Hierarchy Refinement through Exploiting Topology, Content and Social Information" research. The source code, training data and other related resources can be downloaded on our website: http://www.csie.ntu.edu.tw/~d97944007/refinement/

## Software Platform

We use Java as our development platform. Please install Java 1.6.0 or above in order to execute or modify our source code.

## Hardware Platform

We suggest using AMD Opteron 2350 2.0GHz Quad-core CPU or above, with 32GB RAM or above to run the program. For the InterPro dataset, only 2GB RAM is required.

## Acknowledgement

We used these libraries in our program:

| | |
|---|---|
| *Apache Common Math 1.2 or above* | http://commons.apache.org/math |
| *Tim's Java Utilities* | http://www.csie.ntu.edu.tw/~d97944007/utility |
| *Weka 3.5.8 or above* | http://www.cs.waikato.ac.nz/ml/weka |

## Getting Started

1. Please download "Full Package" from our website (http://www.csie.ntu.edu.tw/~d97944007/refinement/) and extract it. The directories are listed as follows:

| | |
|---|---|
| *src* | Source code of our program |
| *jar* | JAR file of our program (for quick start) |
| *lib* | Directory to put required libraries |
| *yahoo* | Data files for content features |
| *html* | Data files for social features |

2. Please download prerequisite libraries listed in **Acknowledgement** section and put the JAR files in the *lib* directory.

3. Among the source codes in the *src* directory or JAR file in the *jar* directory, three Java classes are designed to be run directly:

| | |
|---|---|
| *acmtree.ACMTreeAutoRunMain* | Main class for the ACM CCS dataset |
| *acmtree.DOAJAutoRunMain* | Main class for the DOAJ dataset |
| *acmtree.InterProAutoRunMain* | Main class for the InterPro dataset |

4. All possible concept hierarchy refinement methods are defined in the *acmtree.Method* class, including:

*Random*
*Similarity_Level*
*Similarity_Sibling*
*Similarity_Children*
*Similarity_Frequency*
*Similarity_Name*
*Similarity_Page*
*Similarity_Jaccard*
*Similarity_NGD*
*Similarity_Coauthor*
*Similarity_Sequence*
*Similarity_OneNorm*
*Learning*

5. While using *Learning* as method, the features for learning are defined in the *acmtree.Feature* class, including:

*Topology*
*Content*
*Social*
*All*

6. While using *Learning* as method, except for using Topology as feature, we can set the configuration of enrichment in the *acmtree.Enrichment* class, including:

*Enable*
*Disable*

7. After execution, the results are saved in the following files:

| | |
|---|---|
| *results_acmtree.csv* | Results on ACM CCS dataset |
| *results_doaj.csv* | Results on DOAJ dataset |
| *results_gpcr.csv* | Results on InterPro dataset |

**Note**

1. Please remember to include all prerequisite libraries in the Java class path (ex. use *-cp* command line argument to run Java).

2. Please also remember to set memory-related arguments of Java Virtual Machine (ex. use *-Xms512M -Xmx30G* to run Java). We suggest using 32G RAM for running the program to avoid out of memory error. If machine with such large memory is not available, please run a*cmtree.InterProAutoRunMain* solely, because the InterPro dataset requires about 2G RAM only.

3. A command line example is shown as below. It should be noted that for some operating system, the separating symbol of the *-cp* argument is *colon*, while other is *semicolon*. Also, please change the value of *-Xmx* argument to 32G when running on the ACM CCS or DOAJ datasets. Besides, the versions of downloaded libraries might differ from those used in the following example.

   *java -Xms512M -Xmx2G -cp .:jar/acmtree.jar:lib/utility.jar:lib/commons-math-1.2.jar:lib/weka.jar acmtree.InterProAutoRunMain*

Please let us know if you have any question or suggestion. We appreciate your time for using our source code.

Tsung-Ting Kuo, Ph.D. Candidate of MSLab, GINM, NTU d97944007@csie.ntu.edu.tw
Shou-De Lin, Assistant Professor of MSLab, GINM, NTU sdlin@cise.ntu.edu.tw