

Dataset for Scholarly Paper Recommendation

1. Introduction

Much of the world's new knowledge today is now largely captured in digital form and archived within a digital library system. However, these trends lead to information overload, where users find an overwhelmingly large number of publications that match their search queries but are largely irrelevant to their latent information needs. We address this problem by providing recommendation results by using latent information about the user's research interests that exists in their publication list (For further details, see the paper shown in "3. Publications"). We release our dataset used in the paper. If you are interested in research on recommendation of scholarly papers, please try our dataset for your experiments!

2. Dataset

This dataset contains the following information:

- Feature vector of candidate papers to recommend
(from ACL Anthology Reference Corpus)
- Citation and reference information about each candidate papers to recommend
(from ACL Anthology Reference Corpus)
- Research interest of 15 junior researchers
(feature vectors generated from their papers in DBLP list)
- Research interest of 15 junior researchers
(feature vectors generated from their papers in DBLP list)
- IDs of papers relevant to junior and senior researchers' research interest

You can download this dataset from the following URL:

<http://www.comp.nus.edu.sg/~sugiyama/SchPaperRecData.html>

The structure of this dataset is illustrated as follows:

Scholarly PaperRecData

```
|----Doc
|      |----- index.html, README.pdf
|
|----RecCandidatePapersFV
|      |----- P00-1001_recfv.txt, P00-1002_recfv.txt, ...
|
|----Interlink
|      |----- acl.20080325.net
|
|----JuniorR
|      |----- y1
|      |      |----- y1-rlv.txt
|      |      |----- y1-1_fv.txt
|      |      |----- y1RefsFV
|      |
|      |----- y2
|      |      |----- y2-rlv.txt
|      |      |----- y2-1_fv.txt
|      |      |----- y2RefsFV
|      |
|      |
|      |----- y15
|      |      |----- y15-rlv.txt
|      |      |----- y15-1_fv.txt
```

```

|----- y15RefsFV
|
|----- SeniorR
|----- m1
|----- m1-rlv.txt
|----- m1-1
|----- m1-1_fv.txt (file, feature vector of the most recent paper)
|----- m1-1CitsFV (folder, feature vectors of citation papers are stored.)
|----- m1-1RefsFV (folder, feature vectors of reference papers are stored.)
|
|----- m1-2
|----- m1-2_fv.txt (file, feature vector of the most recent paper)
|----- m1-2CitsFV (folder, feature vectors of citation papers are stored.)
|----- m1-2RefsFV (folder, feature vectors of reference papers are stored.)
|
|
|----- m2
|----- m2-rlv.txt
|----- m2-1
|----- m2-1_fv.txt (file, feature vector of the most recent paper)
|----- m2-1CitsFV (folder, feature vectors of citation papers are stored.)
|----- m2-1RefsFV (folder, feature vectors of reference papers are stored.)
|
|----- m13
|----- m13-rlv.txt
|----- m13-1
|----- m13-1_fv.txt (file, feature vector of the most recent paper)
|----- m13-1CitsFV (folder, feature vectors of citation papers are stored.)
|----- m13-1RefsFV (folder, feature vectors of reference papers are stored.)
|
|

```

Doc

This folder contains the following two documents:

- index.html
- README.pdf

RecCandidatePapersFV

This folder stores feature vectors of candidate papers to recommend. The feature vectors are generated from 597 full papers published in ACL 2000-2006, by applying TF-IDF. The ID before “_” in each file name denotes the paper ID.

Interlink

The file “acl.20080325.net” in this folder contains citation and reference information about ACL papers.

JuniorR

This folder includes information about 15 junior researchers. Let i ($i=1, \dots, 15$) be junior researcher’s id. Each folder has the following two information: feature vector of the most recent paper ($y_{i-1_fv.txt}$) and feature vectors of reference papers in the most recent paper. The feature vector of the most recent paper is generated from the paper in the researcher’s DBLP list. The feature vectors of reference papers are stored in “ $y_iRefsFV$.” For example, when “ $i=2$,” “ $y2-1_fv.txt$ ” denotes the feature vector of the most recent paper of junior researcher “ $y2$,” and “ $y2RefsFV$ ” denotes the folder that stores the feature vectors of reference papers. The feature vectors are generated using TF scheme.

In addition, all of junior researchers from y1 to y15 have paper IDs relevant to these researchers. “y1-rlv.txt,” “y2-rlv.txt,” ... “y15-rlv.txt” include the paper IDs relevant to each researcher. These IDs correspond to those in file names in “RecCandidatePapersFV” folder described above.

SeniorR

This folder includes information about 13 senior researchers. Let j ($j=1, \dots, 13$) be senior researcher’s id. Each folder has the information about the researcher’s published paper. When “ $j=3$,” folder “m3” has the information about his published paper. “m3-1” is the most recent paper. m3-2, m3-3, ... store information about his papers published in older days. These feature vectors are generated from the papers in the researcher’s DBLP list. For example, folder “m3-1” has the following three information:

- m3-1_fv.txt: Feature vector of the most recent paper (file)
- m3-1CitsFV: Feature vector of citation papers are stored (folder)
- m3-1RefsFV: Feature vector of reference papers are stored (folder)

Similarly, “m3-2” stores information about his paper published before “m3-1”:

- m3-2_fv.txt: Feature vector of the paper (file)
- m3-2CitsFV: Feature vector of citation papers are stored (folder).
- m3-2RefsFV: Feature vector of reference papers are stored (folder).

The feature vectors are generated using TF scheme.

In addition, all of senior researchers from m1 to m13 have paper IDs relevant to these researchers. “m1-rlv.txt,” “m2-rlv.txt,” ..., “m13-rlv.txt” include paper IDs relevant to each researcher. These IDs correspond to those in file names in “RecCandidatePapersFV” folder described above.

3. Publications

Kazunari Sugiyama, Min-Yen Kan:

“Scholarly Paper Recommendation via User’s Recent Research Interests,”

The 10th ACM/IEEE Joint Conference on Digital Libraries (JCDL 2010), pp.29-38,
Gold Coast, Queensland, Australia, June 21-25, 2010.

4. Group Members

Kazunari Sugiyama (National University of Singapore)

Min-Yen Kan (National University of Singapore)