Experiment results for <Integrating Multiple Heterogeneous Data Sources for Discovery Alternative Clustering Solutions>

(1) WebKB dataset

1) Source

This directory contains a selection of the WebKB dataset
http://www-2.cs.cmu.edu/~webkb/

2) Preprocessing

We use the preprocessing steps provided at
Sen et.al <Collective Classification in Network Data>

3) Dataset Description

a) Webpage contents: The 877 webpages gathered from 4 different universities(*Cornell, Texas, Washington, Wisconsin*) can be classified into 5 topics(*course, faculty, project, staff, student*). After stemming and removing stop-words in the original corpus, we are left with a vocabulary of size 1703 unique words. All words with document frequency less than 10 were removed.
Each webpage in the dataset is described by a 0/1 valued word vector indicating the absence/ presence of the corresponding word in the dictionary.

School	course	faculty	project	staff	student	total
Cornell	42	32	19	19	83	195
Texas	34	31	18	1	103	187
Washington	66	27	21	9	107	230
Wisconsin	76	35	22	10	122	265

b) Webpage links: There're links between the webpages from the same university. The link matrix consists of 1608 links. The (i,j) elements of the link matrix is set to be 1 if webpage i and j are linked and 0 otherwise. To make the matrix symmetric, (j,i) element can be set to be same with (i,j) element. Therefore we can get a sparse matrix with 0/1 valued elements.

School	Cornell	Texas	Washington	Wisconsin
Cornell	304	0	0	0
Texas	0	328	0	0
Washington	0	0	446	0
Wisconsin	0	0	0	530

4) Discussion

a) Multiple clustering solutions: As is shown above, webpages can have **two clustering solutions**. The fist one is based on different universities(*Cornell, Texas, Washington,*

Wisconsin). The second one is based on different topics((*course, faculty, project, staff, student*)).

b) Multiple data sources: There're **two sources** of information for webpage clustering. The first source is link matrix, which can be viewed as a similarity matrix between webpages for clustering on different universities. However, it's not a reliable similarity matrix between webpages if we want to do clustering on different topics because the webpages of different topics tend to link together.

5) Experiments

Our aim is to improve the performance of clustering on different universities through integrating multiple heterogeneous data sources(webpage content and webpage links) and finding alternative clustering, which should have both good clustering quality and novelty.

Case 1: single source: webpage content

Webpage content can be transformed into a similarity matrix by applying polynomial kernel with the following parameter settings:
degree = 3, coef0 = 1.

Then spectral clustering can be applied to do clustering. We set K = 4, which is equal to the number of universities.

NMI	Labels of Universities	Labels of Topics
Predicted Labels(web content)	0.3644	0.2355

Case 2: single source: webpage links

Link matrix can be used as similarity matrix between webpages. Therefore, spectral clustering can be applied by setting K=4

NMI	Labels of Universities	Labels of Topics
Predicted Labels(web links)	0.03257	0.03440

Case 3: multiple sources: webpage content + webpage links

NMI	Labels of Universities	Labels of Topics	Objective Value
Predicted Labels(web content and web links)	0.4044	0.2253	34.29

	Web content	Web links
Weights	0.8120	0.1880

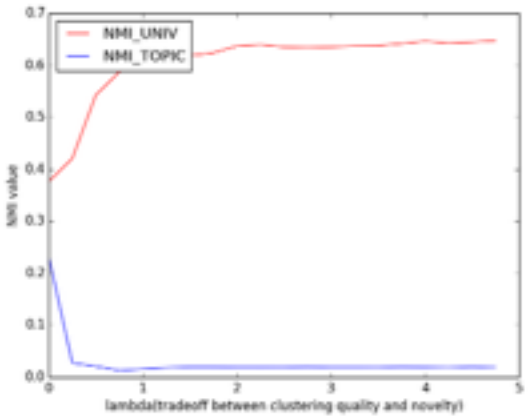
Case 4: single source: webpage content and alternative clustering

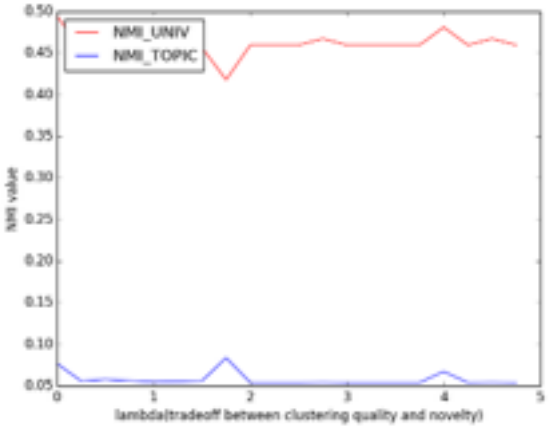
Case 5: single source: webpage links and alternative clustering

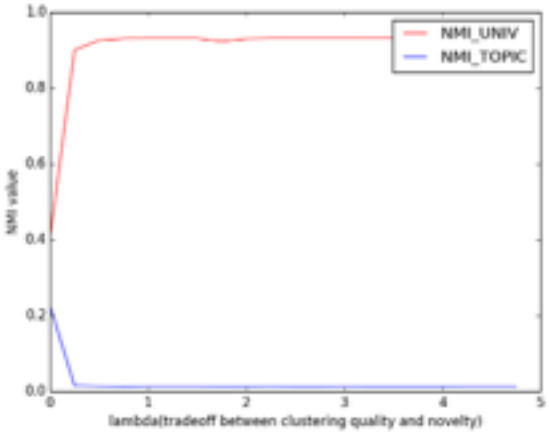
Case 6: multiple sources: webpage content + webpage links and alternative clustering

	Web content	Web links
Weights	0.8122	0.18878

The following three figures show the change of NMI value as the value of lambda is varied. They corresponds to Case 4, 5, 6 respectively.







6) Conclusions

According to the experiment results, we have the following observations