Classification of Web Comics

Chandra Sekhar Mallarapu
Naresh Singh
Nehal Bandi

December 16, 2011

Motivation

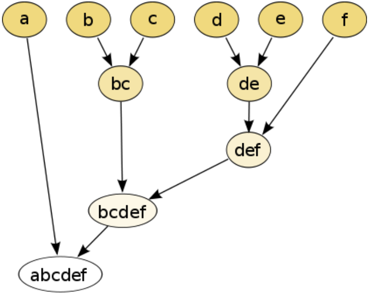
- Lots of web comics available online
- We would like to determine similarities between comics and group them.
- Its a cool thing to do
- It allows one to find out comics of interest

Dataset

- About 20000 documents from 9 comic series
- Downloaded following comic series from OhNoRobot.com
  –Nukees
  –College Roomies From Hell
  –Questionable Content
  –Sheldon
  –Goats
  –General Protection Fault
  –Diesel Sweeties
- XKCD comics. Available from their website
- Calvin and Hobbes

Approach

- Clustering is a natural solution to this problem
- It also makes sense to create a heirarchy of clusters



- Moreover, simultaneously clustering together both comic series and individual documents, and also creating a heirarchy will tell us similarity between series and documents
- This helps group a series with related documents using similarities of comics and vice-versa

Heirarchical Co-clustering

- Given a set of *m* comic documents $D = D_1, D_2, \cdots, D_m$ and a set of *n* series $S = S_1, S_2, \cdots, S_n$
- Also given a *mxn* document-series relationship matrix X, with x_{ij} representing the relation between *i*-th document in D and *j*-th series in S
- HCC simultaneously generates a heirarchical clustering of D and S based on X

HCC Algorithm

Algorithm 1 HCC Algorithm Description

```
Create an empty heirarchy $H$
$List \leftarrow Objects\ in\ A + Objects\ in\ B$
$N \leftarrow size[A] + size[B]$
for $i = 0$ to $N - 1$ do
    $p, q$=PickUpTwoNodes($List$)
    $o$=Merge($p, q$)
    Remove $p, q$ from $List$ and add $o$ to $List$
    Add $List$ to $H$ as next layer
end for
```

Merging Nodes

- Cluster Heterogeneity Measurement(CH) is used for the clustering heterogeneous types
- If we want to cluster $P \subseteq D$ having r rows, and $Q \subseteq S$ having t columns, caculate

$$CH(P,Q) = \frac{1}{rt} \sum_{i \in P, j \in Q} (x_{ij} - \mu)^2$$

  where μ is the max of the entries in the matrix X
- Calculate $CH(P,Q)$ for all possible pairs from present clusters, and choose that pair which has least cluster heterogeneity

Co-Clustering Words and Documents

- To co-cluster documents and series, we need to build the relaitonship matrix between documents and series
- We build that by obtaining information from the results of co-clustering words and documents
- W is the set of words from all the documents
- Create a word-document relationship matrix X, with the documents representing the columns and the rows representing the words.

$$x_{ij} = tfidf(w_i, d_j)$$

- Co-cluster words and documents by using the HCC algorithm described earlier

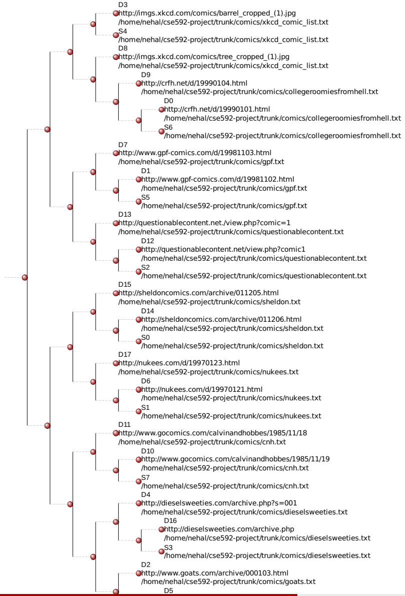
Document-Series Relationship Matrix

- Let X be the realtionship matrix between series and documents
- Let $K = |W| + |D|$, where W=set of words and D=set of documents
-
- For node N_i created in iteration i of the HCC algorithm run for co-clustering words and documents, using nodes $N1$ and $N2$ present from previous iteration
- $K = K - 1$
- For each document d_i in $N1$,
  For each unique series k that the documents in $N2$ belong to,

  $$x_{ik} = x_{ik} + K$$

  Do the same reversing $N1$ and $N2$

HCC Dendrogram



D3
http://imgs.xkcd.com/comics/barrel_cropped_(1).jpg
/home/nehal/cse592-project/trunk/comics/xkcd_comic_list.txt
S4
/home/nehal/cse592-project/trunk/comics/xkcd_comic_list.txt
D8
http://imgs.xkcd.com/comics/tree_cropped_(1).jpg
/home/nehal/cse592-project/trunk/comics/xkcd_comic_list.txt
D9
http://crfh.net/d/19990104.html
/home/nehal/cse592-project/trunk/comics/collegeroomiesfromhell.txt
D0
http://crfh.net/d/19990101.html
/home/nehal/cse592-project/trunk/comics/collegeroomiesfromhell.txt
S6
/home/nehal/cse592-project/trunk/comics/collegeroomiesfromhell.txt
D7
http://www.gpf-comics.com/d/19981103.html
/home/nehal/cse592-project/trunk/comics/gpf.txt
D1
http://www.gpf-comics.com/d/19981102.html
/home/nehal/cse592-project/trunk/comics/gpf.txt
S5
/home/nehal/cse592-project/trunk/comics/gpf.txt
D13
http://questionablecontent.net./view.php?comic=1
/home/nehal/cse592-project/trunk/comics/questionablecontent.txt
D12
http://questionablecontent.net/view.php?comic1
/home/nehal/cse592-project/trunk/comics/questionablecontent.txt
S2
/home/nehal/cse592-project/trunk/comics/questionablecontent.txt
D15
http://sheldoncomics.com/archive/011205.html
/home/nehal/cse592-project/trunk/comics/sheldon.txt
D14
http://sheldoncomics.com/archive/011206.html
/home/nehal/cse592-project/trunk/comics/sheldon.txt
S0
/home/nehal/cse592-project/trunk/comics/sheldon.txt
D17
http://nukees.com/d/19970123.html
/home/nehal/cse592-project/trunk/comics/nukees.txt
D6
http://nukees.com/d/19970121.html
/home/nehal/cse592-project/trunk/comics/nukees.txt
S1
/home/nehal/cse592-project/trunk/comics/nukees.txt
D11
http://www.gocomics.com/calvinandhobbes/1985/11/18
/home/nehal/cse592-project/trunk/comics/cnh.txt
D10
http://www.gocomics.com/calvinandhobbes/1985/11/19
/home/nehal/cse592-project/trunk/comics/cnh.txt
S7
/home/nehal/cse592-project/trunk/comics/cnh.txt
D4
http://dieselsweeties.com/archive.php?s=001
/home/nehal/cse592-project/trunk/comics/dieselsweeties.txt
D16
http://dieselsweeties.com/archive.php
/home/nehal/cse592-project/trunk/comics/dieselsweeties.txt
S3
/home/nehal/cse592-project/trunk/comics/dieselsweeties.txt
D2
http://www.goats.com/archive/000103.html
/home/nehal/cse592-project/trunk/comics/goats.txt
D5

References

- Jingxuan Li et al, *HCC: A Hierarchical Co-clustering Algorithm*
- T. Eckes et al, *An error variance approach to two-mode hierarchical clustering*