

Labelling Topics Using Unsupervised Graph-based Methods

Nikolaos Aletras and Mark Stevenson
 {n.aletras, m.stevenson}@dcs.shef.ac.uk



Task: The aim of the task is to associate labels with automatically generated topics.

1. Retrieving Text Information 2. Creating Text Graph

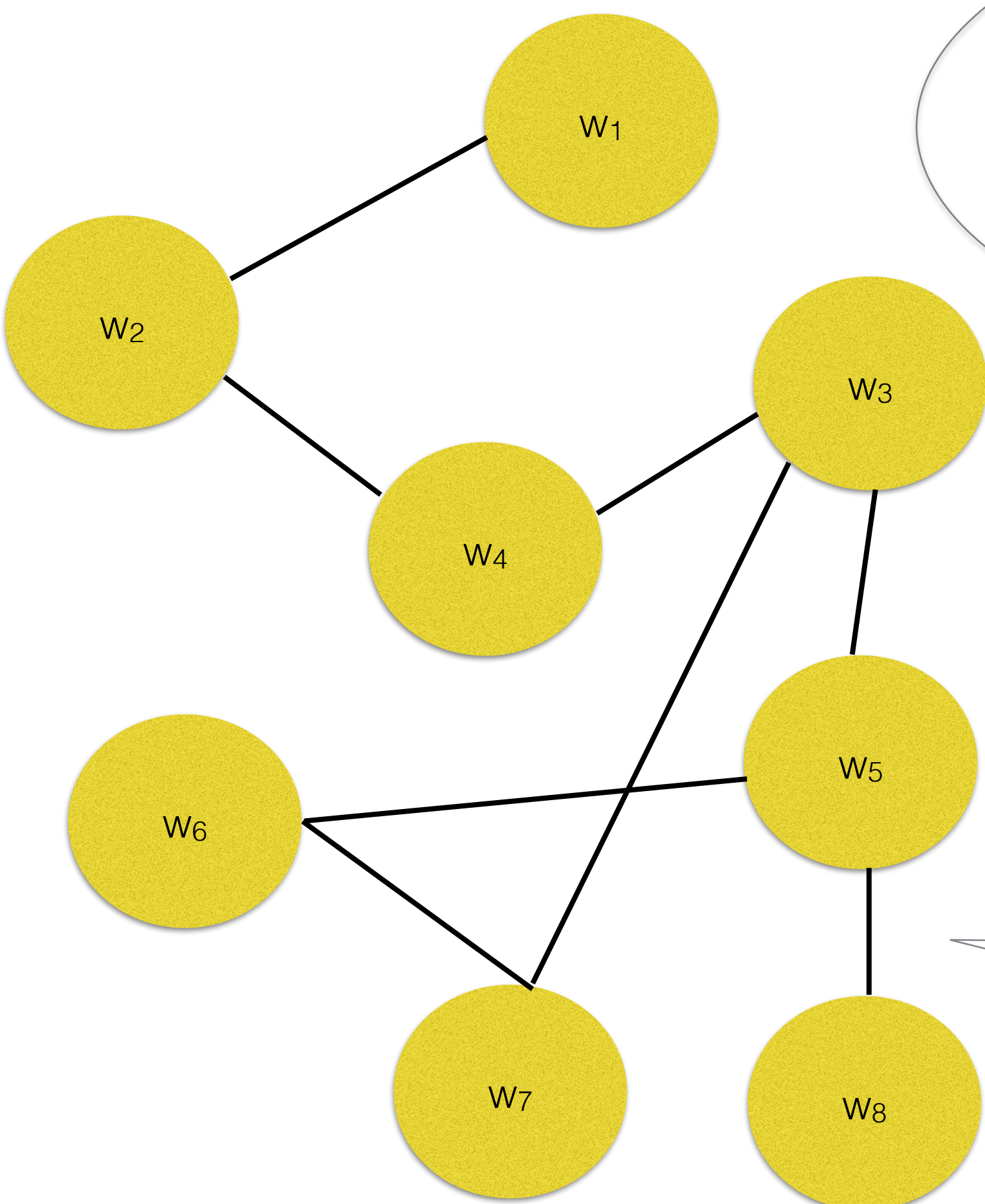
{VMWARE,
 SERVER, VIRTUAL, ORACLE,
 UPDATE, VIRTUALIZATION, APPLICATION,
 INFRASTRUCTURE, MANAGEMENT,
 MICROSOFT}

TOPIC

SEARCH
 ENGINE

RESULTS
 METADATA

'Description': 'Microsoft will accelerate your
 journey to cloud computing with an
 agile and responsive datacenter built from your existing
 technology investments.',
 'DisplayUrl': 'www.microsoft.com/en-us/server-cloud/
 datacenter/virtualization.aspx',
 'ID': 'a42b0908-174e-4f25-b59c-70bdf394a9da',
 'Title': 'Microsoft | Server & Cloud | Datacenter |
 Virtualization ...',
 'Url': 'http://www.microsoft.com/en-us/server-cloud/datacenter/
 virtualization.aspx',
 ...

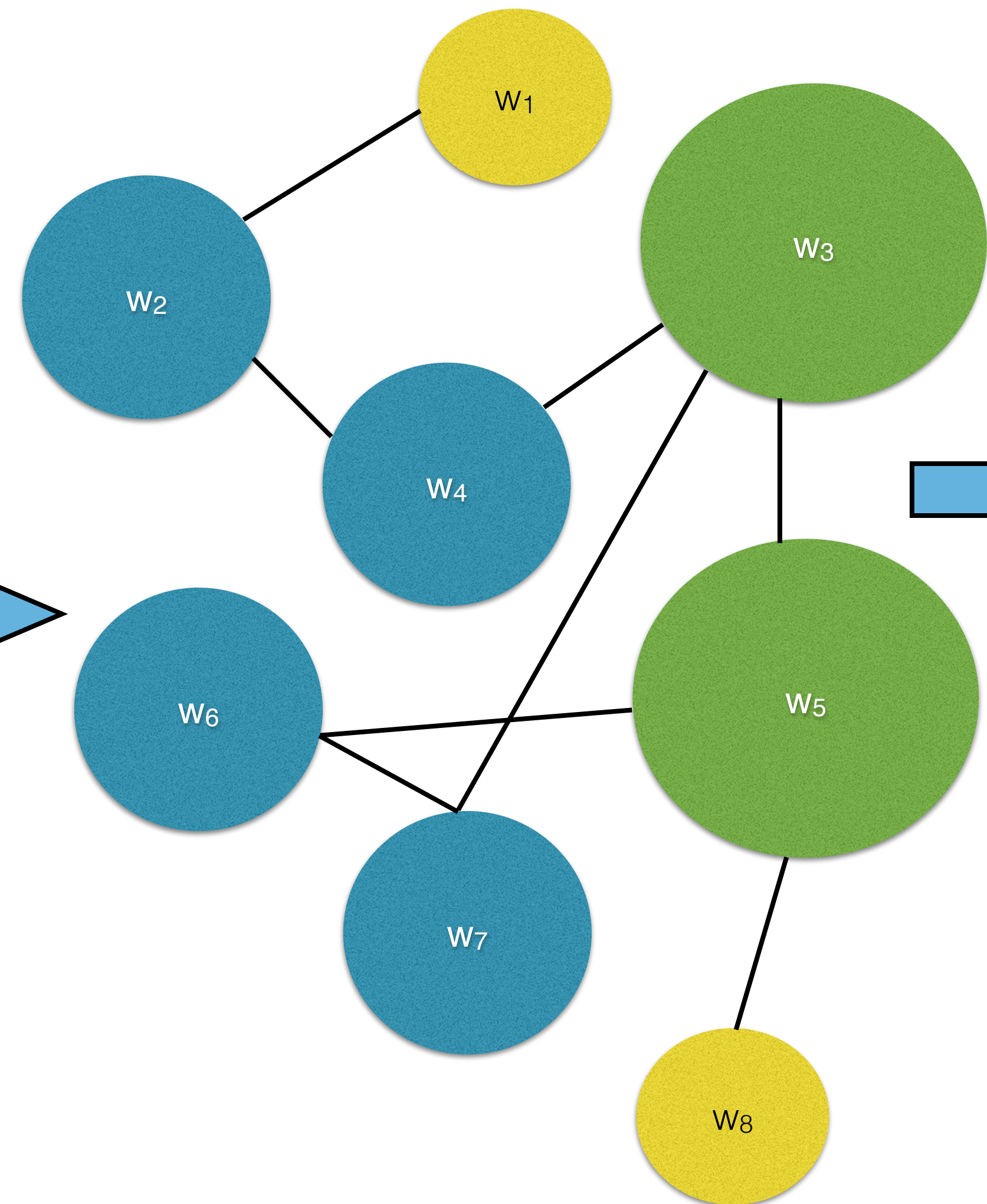


$$Pr(w_i) = d \cdot \sum_{w_j \in C(w_i)} \frac{sim(w_i, w_j)}{\sum_{w_k \in C(w_j)} sim(w_j, w_k)} Pr(w_j) + (1 - d)w$$

PageRank

Edge weight:
 1 - Unweighted Graph (PR)
 Normalised PMI - Weighted Graph (PR-NPMI)

3. Identifying Important Terms



4. Scoring Candidate Labels

Candidate Label: $L = \{w_1, w_2, \dots, w_m\}$

Scoring Function:

$$Score(L) = \sum_{i=1}^m Pr(w_i)$$

Candidate Labels:

HARDWARE VIRTUALIZATION
 VIRTUAL MACHINE
 VIRTUAL PC

Data

#Topics: 228

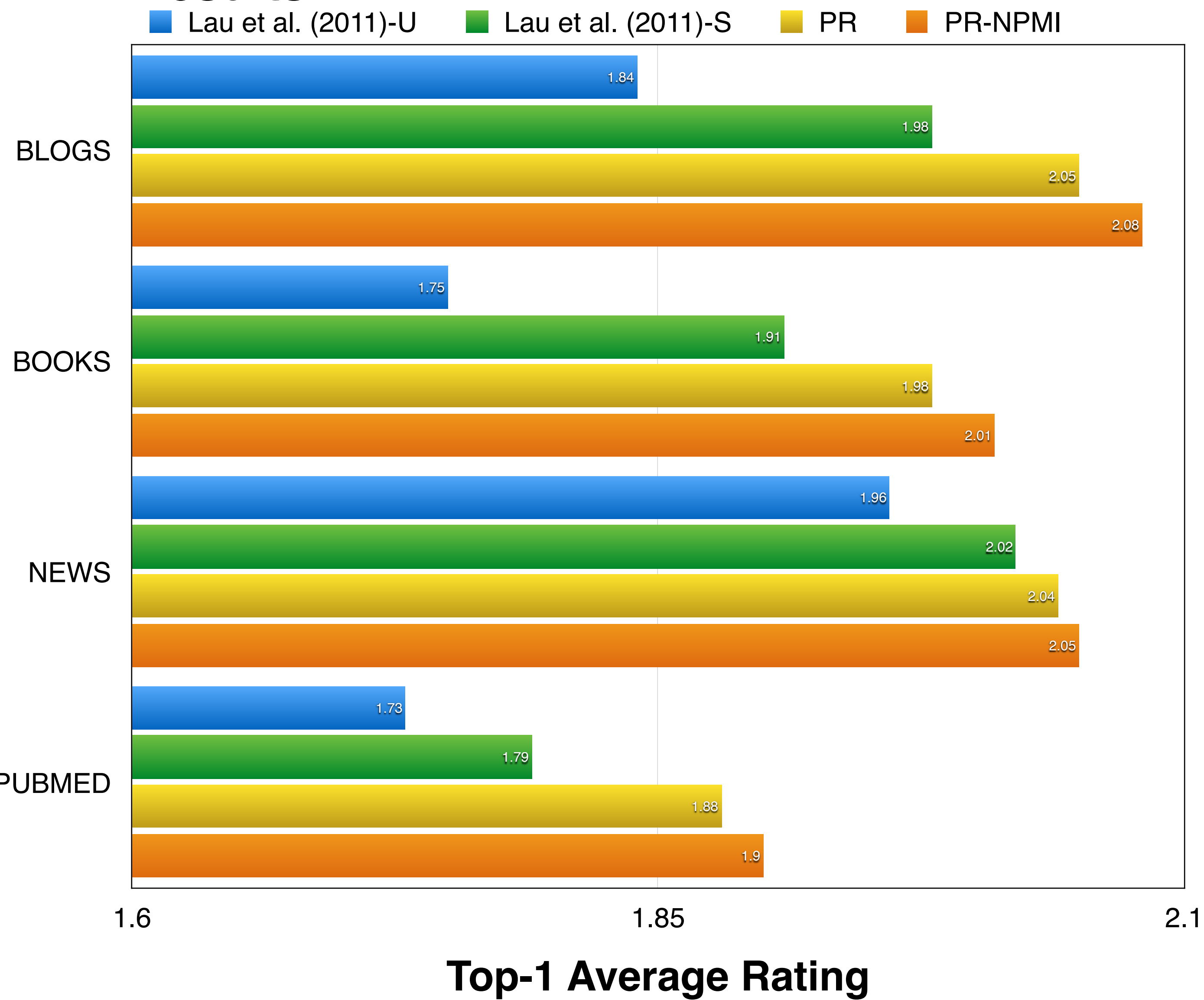
#Candidate Labels: ~6,000

Domains: BLOGS, BOOKS,
 NEWS, PUBMED

Candidate labels rated by humans
 (0-3)

Published by Lau et al. (2011).

Results



Conclusion

- Unsupervised graph-based method to associate textual labels with automatically generated topics.
- Our method performs consistently better than previous supervised methods.