

Graph Connectivity Measures for Unsupervised Parameter Tuning of Graph-Based Sense Induction Systems

Ioannis Korkontzelos, Ioannis Klapaftis and Suresh Manandhar
Department of Computer Science
The University of York
Heslington, York, YO10 5NG, UK

5 June 2009

Outline

- 1 Introduction
- 2 Graph Connectivity Measures
- 3 Collocational graphs for WSI
- 4 Evaluation & Results

Parameter Estimation

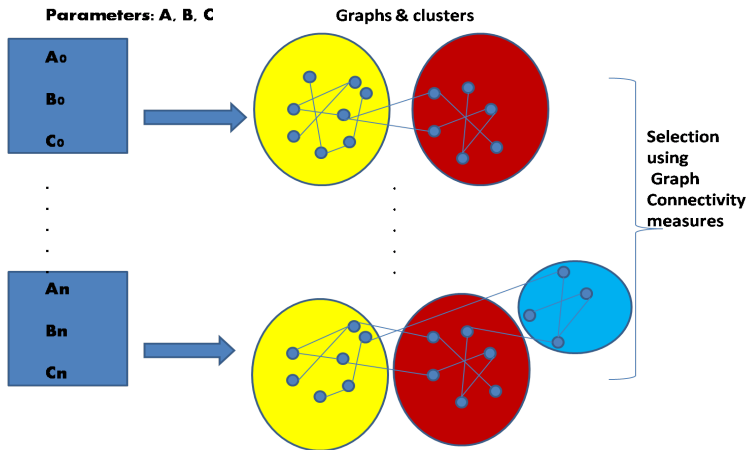
Problem

- Many unsupervised methods rely on several free parameters to perform their task. E.g. The WSI systems on SemEval-2007 WSI task.
- The same set of parameters might not be appropriate for all datasets, applications, domains etc.

Common solutions

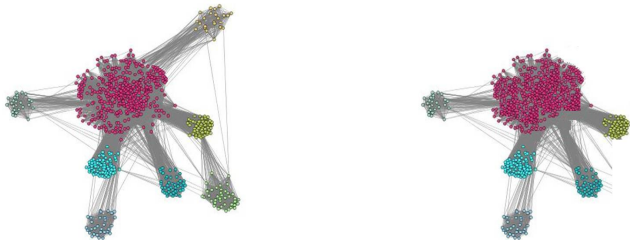
- Select parameters empirically: Ad hoc & not reliable
- Perform a supervised tuning of the parameters.
 - But what if there is no training data?

Unsupervised parameter estimation for graph-based approaches



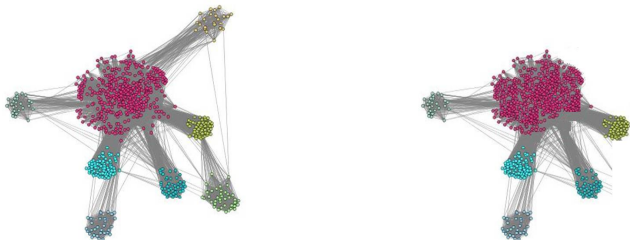
Unsupervised parameter tuning for graph-based approaches

- Noun *chance*, 8 clusters
- Noun *chance*, 6 clusters
- Which solution to select ?



Unsupervised parameter tuning for graph-based approaches

- Assess each cluster of each solution.
- Treat each cluster as a subgraph of the original unclustered graph.
- Evaluate the connectivity in each cluster.



Graph connectivity measures for unsupervised parameter tuning

- Graph connectivity measures evaluate the degree of connectivity in a graph.
- They have been applied to WSD (Navigli & Lapata, 2007)
- Our target is to test whether they can be used as a means of evaluating clustering solutions.

Average Degree

- The average number of edges per vertex.

$$\text{deg}(u) = |\{(u, v) \in E : v \in V\}|$$

- Sum of the degree of each vertex divided by the total number of vertices.
- The higher the average degree the more connected the graph is.
- Weighted version takes into account edge weights.

Clustering coefficient

- The clustering coefficient of a vertex in a graph quantifies how close its neighbors are to being a clique.
- The average number of edges between the neighbours of each vertex of a cluster divided by the maximum number of edges that can exist between these neighbours.

$$cc(u) = \frac{|E(\Gamma(i))|}{\binom{|\Gamma(i)|}{2}}$$

where $\Gamma(i)$ provides the set of vertices adjacent to vertex i , and $E(\Gamma(i))$ provides the edges that exist between the $\Gamma(i)$ neighbours of i .

- Weighted version.

Graph entropy

Graph entropy

- Entropy measures the amount of information (alternatively the uncertainty) in a random variable.
- Here, high entropy indicates that many vertices are equally important.
- Low entropy indicates that only few vertices are relevant.

$$en(u) = - \sum_{u \in V} p(u) \log_2 p(u) \quad p(u) = \left\{ \frac{deg(u)}{2|E|} \right\}_{u \in V}$$

- Weighted version.

Edge density

- Global measure, it evaluates the graph as a whole.
- Number of edges divided by the maximum possible edges (complete graph).

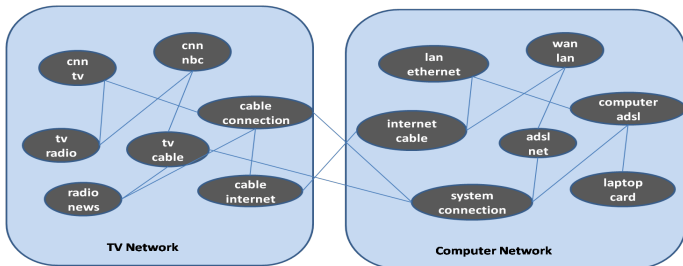
$$ed(G(V, E)) = \frac{|E|}{A(V)} \quad A(V) = \binom{|V|}{2}$$

- Weighted version.

Collocational graphs for WSI

(Klapaftis & Manandhar, 2008)

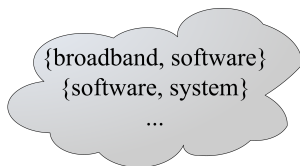
- Use senses based on collocations.
- Cluster graph of related collocations.



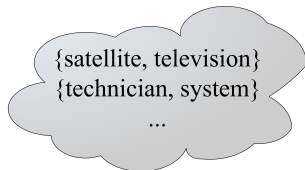
Collocational graphs for WSI

(Klapaftis & Manandhar, 2008)

- Use senses based on collocations.
- Cluster graph of related collocations.



**computer
network**



**television
network**

Identifying contextually relevant words

Charles Redmond, a NASA spokesman, said the agency discovered the virus on Monday on the collection of computer networks collectively called Internet and expected 100 university centers to be infected by today.

- Red - unrelated to network
- Green - related to network

- Given a reference corpus (BNC) and the target word corpus
- Corpora comparison using log-likelihood (Dunning, 1993)
- Keep words with a log-likelihood higher than a threshold (p_1)

Extracting Collocations

e.g. collocation: {computer, software}

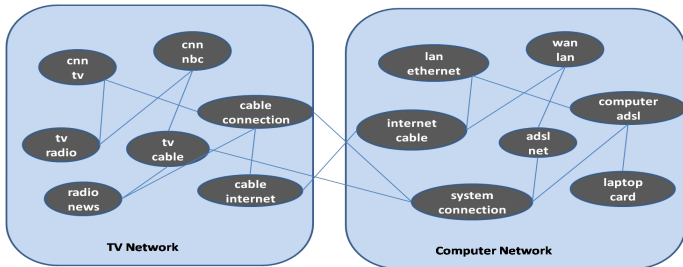
- Generate Collocation pairs at the paragraph level.
- Filter collocations below a given frequency threshold (**p2**).
- Weights aims to measure “*mutual attraction*” of co-occurring words.

$$w(x, y) = \text{average}\{p(x|y), p(y|x)\} \quad (1)$$

- Calculations take place on the whole SemEval corpus, 27132 paragraphs to deal with data sparsity.
- Collocations below a given threshold are removed (**p3**).

Creating Collocational Graph

- Each collocation is represented as a vertex.
- Two vertices are connected, if they co-occur in one or more paragraphs.
- Clustering the collocational graph using Chinese Whispers (Biemann, 2006).



Parameters

Parameter	Range
Log-likelihood threshold for selecting contextual words (p1)	5, 10, 15
Frequency threshold for collocations (p2)	4, 6, 8, 10
Weight threshold for collocations (p3)	0.2, 0.3, 0.4

Evaluation on nouns of SemEval-2007 WSI task

System	Unsupervised Evaluation				Sup. Recall
	FSc.	Pur.	Ent.	# Cl.	
I2R	68.0	88.4	29.7	3.1	86.8
Col-Sm-org	78.0	88.6	31.0	5.9	86.4
UMND2	67.1	85.8	37.6	1.7	84.5
UPV_SI	69.9	87.4	30.9	7.2	82.5
UOY	65.8	89.8	25.5	11.3	81.6
UBC-AS	80.8	83.6	43.5	1.6	80.7
1c1w-MFS	80.7	82.4	46.3	1.0	80.9
1c1inst	6.6	100	0	73.1	N/A

Unsupervised & supervised evaluation

Graph Connectivity Measure	Unsupervised Evaluation				Sup. Recall
	FSc	Pur.	Ent.	# Cl.	
Average Degree	79.2	87.2	34.2	<u>3.9</u>	84.8
Average Weighted Degree	77.1	87.8	32.0	5.5	84.2
Average Cluster Coefficient	72.5	88.8	28.5	9.1	83.9
Average Weighted Cluster Coefficient	65.8	88.4	28.0	9.6	84.1
Graph Entropy	67.0	89.6	25.9	12.3	83.8
Weighted Graph Entropy	72.7	89.4	28.1	9.6	84.1
Edge Density	47.8	91.8	19.4	18.4	84.8
Weighted Edge Density	53.4	90.2	23.1	15.5	83.7
<i>Supervised Tuning Baseline</i>	78.0	88.6	31.0	5.9	86.4

System	Bound type	Unsupervised Evaluation				Sup. Recall
		FSc.	Pur.	Ent.	# Cl.	
Col-Sm	MaxR	79.3	90.5	26.6	7.0	88.6
Col-Sm	MinR	62.9	89.0	26.7	12.7	78.8
Col-Sm	MaxF	83.2	90.0	28.7	4.9	86.6
Col-Sm	MinF	43.6	90.2	22.1	17.6	83.7
Col-Sm-org	sup.	78.0	88.6	31.0	5.9	86.4

Summary

- Graph connectivity measures are able to identify useful differences regarding the quality of the induced clusters for different parameter combinations.
- They improve the worst performing parameter setting by large margins in both evaluation schemes
- But they are far below the best performing parameter setting.
- All of them estimate a set of parameters which are above the Most Frequent Sense (MFS) baseline.
- **Average degree** and **weighted average degree** estimate a set of parameters that performs closely to a set of parameters estimated in a supervised manner.

End of presentation

Thank you for your attention.

Any questions?