Graph-based Clustering for Computational Linguistics: A Survey

Zheng Chen Heng Ji
BLENDER: Cross-lingual Cross-document IE Lab
Department of Computer Science
The Graduate Center and Queens College
The City University of New York

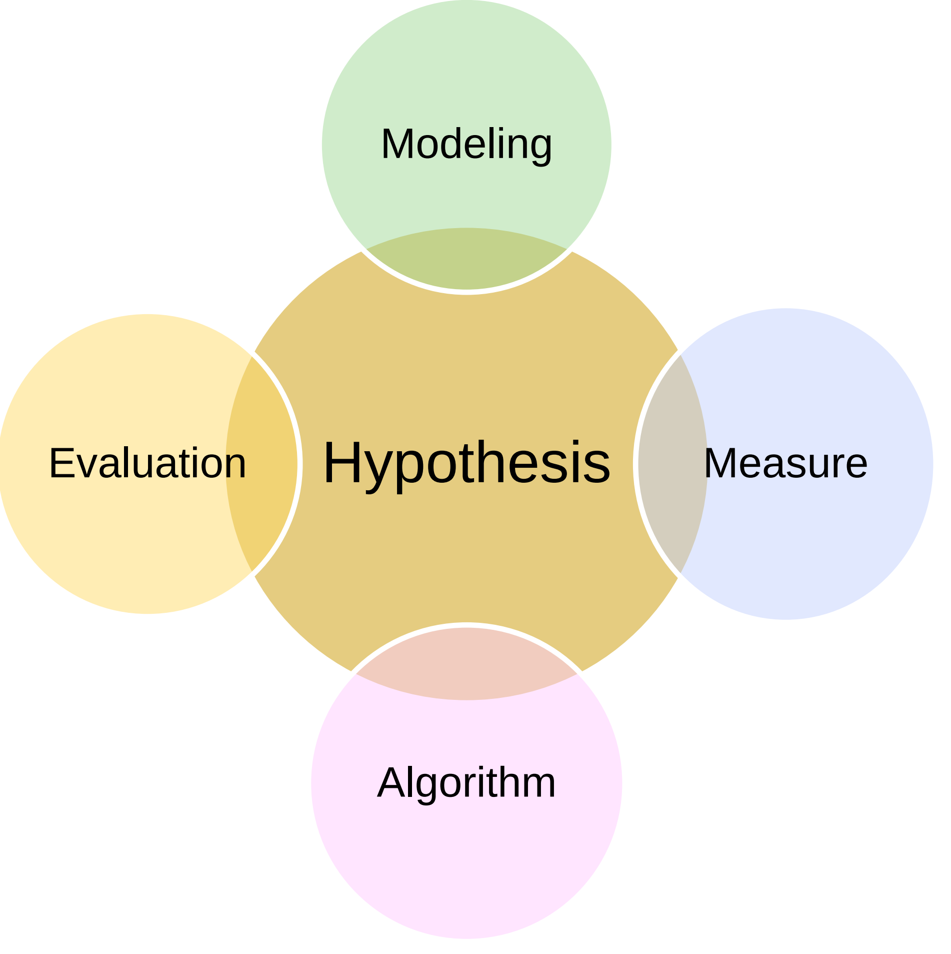
July, 2010

Motivations

- Long standing, wide applications in various areas
- Gaining interests from Computational Linguistics, applied in various NLP problems
- Bridging theories and applications, especially for computational linguistics

Outline (Part I: Theory)

Graph-based Clustering Methodology (a five-part story)



Outline (Part II: Applications)

- Coreference Resolution
- Word Clustering
- Word Sense Disambiguation

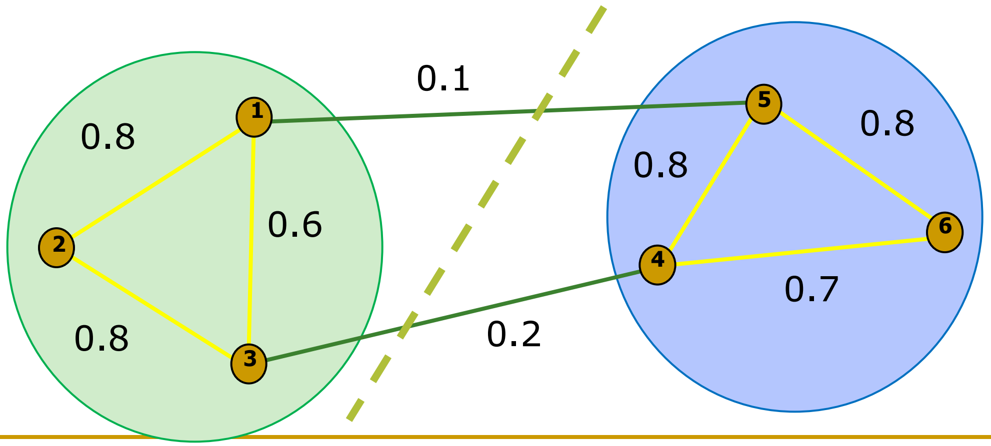
Part I
Graph-based Clustering Methodology

Clustering in Graph Perspective

$X = \{x_1, \dots, x_N\}$: a set of data points

$S = (s_{ij})_{i,j=1,\dots,N}$:the similarity matrix in which each element indicates the similarity $s_{ij} \geq 0$ between two data points x_i and x_j.

- Hard clustering problem: split the data points into several non-overlapping clusters such that points in the same cluster are similar and points in different cluster are dissimilar.
- Graph representation of data points



Hypothesis

ways:

connected

to each other

graph will likely

odes have been

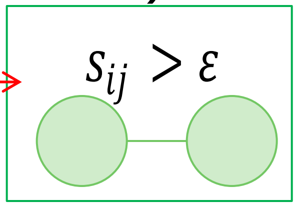
all pairs of

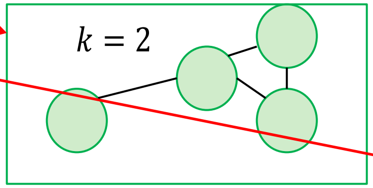
hs are likely to be

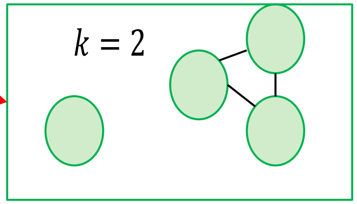
Modeling:

transforming a problem into graph structure

- Determine the meaning of nodes, edges
- Compute the edge weights
- Graph construction (Luxburg,2006)
  - ***The ε-neighborhood graph***
  - ***k-nearest neighbor graph***
  - ***mutual k-nearest neighbor graph***
  - ***The fully connected graph***







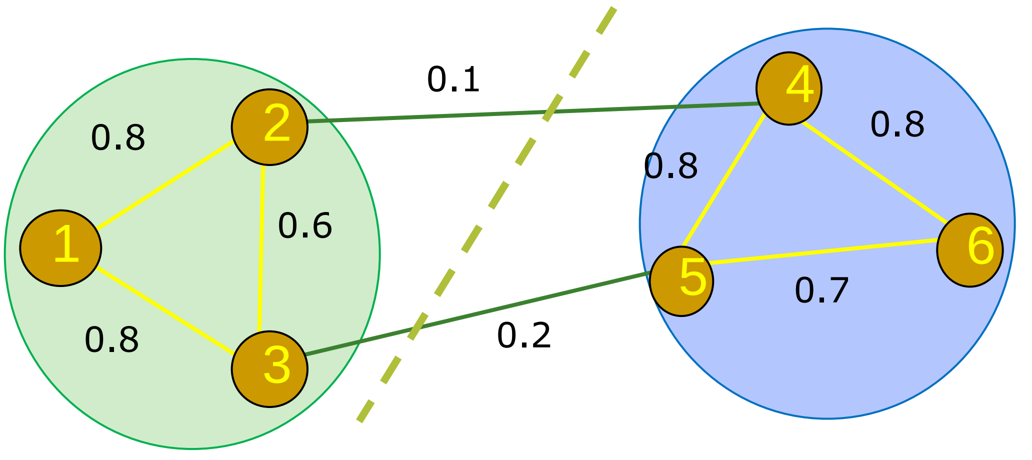
- Which graph should be chosen and how to choose parameters? (no theoretical justifications)

Measure: Objective function that rates clustering quality

Measures	Comments
1. intra-cluster density *2. inter-cluster density*	1. optimizing one is equivalent to optimizing the other 2. both favor clusters containing isolated vertices
3. ratio cut (Hagan and Kahng, 1992) *4. normalized cut* (Shi and Malik, 2000)	1. ratio cut is suitable for unweighted graph, for weighted graph, normalized cut is recommended 2. both favor clusters with equal size.
5. performance (Brandes et al., 2003)	
6. expansion *7. conductance* *8. bicriteria* (Kannan et al., 2000)	1. expansion treats all nodes as equally important 2. conductance gives more importance to nodes with high degrees and edge weights 3. neither enforces qualities pertaining to inter-cluster weights, but bicriteria does
9. modularity (Girvan and Newman, 2002)	1. requires global knowledge of the graph's topology, *local modularity* (Clauset ,2005) 2. resolution limit problem, *HQCut* (Ruan and Zhang,2008) 3. only measures existing edges in the graph but does not explicitly take non-edges into consideration, *Max-Min Modularity* (Chen et al.,2009)

An Example for Quality Measure

Goal: cluster the graph into 2 sub-graphs



Try all possible combinations:

$$maximize \sum_{i=1}^{2} intra\_density(C_i)$$

Algorithm: optimizing the measure

Category		Algorithms	optimized measure	running complexity
divisive	cut-based	*Kernighan-Lin algorithm* (Kernighan and Lin, 1970)	*intercluster*	$O(\|V\|^3)$
		cut-clustering algorithm (Flake et al., 2003)	*bicriteria*	$O(\|V\|)$
	spectral	*unnormalized spectral clustering* (Luxburg, 2006)	*ratiocut*	$O(\|V\|\|E\|)$
		normalized spectral clustering I (Luxburg, 2006; Shi and Malik, 2000)	*ncut*	$O(\|V\|\|E\|)$
		normalized spectral clustering II (Luxburg, 2006; Ng, 2002)	*ncut*	$O(\|V\|\|E\|)$
		iterative conductance cutting (ICC) (Kannan et al.,2000)	*conductance*	$O(\|V\|\|E\|)$
		geometric MST clustering (GMC) (Brandes et al., 2007)	*pluggable(any quality measure)*	$O(\|V\|\|E\|)$
		modularity oriented (White and Smyth,2005)	*modularity*	$O(\|V\|\|E\|)$
	multilevel	*multilevel recursive bisection* (Karypis and Kumar, 1999)	*intercluster*	$O(\|V\|logK)$
		multilevel K-way partitioning (Karypis and Kumar, 1999)	*intercluster*	$O(\|V\| + KlogK)$
	random	*Markov Clustering Algorithm (MCL)* (Dongen, 2000)	*performance*	$O(m^2\|V\|)$
	shortest path	*betweenness* (Girvan and Newman, 2003)	*modularity*	$O(\|V\|\|E\|^2)$
		information centrality (Fortunato et al., 2004)	*modularity*	$O(\|V\|\|E\|^3)$
agglomerative		*modularity oriented* (Newman, 2004)	*modularity*	$O(\|V\|\|E\|)$

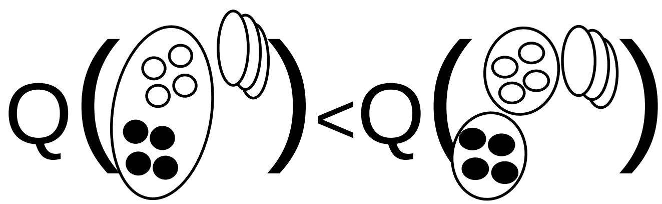
Evaluation:

rating system clustering on gold clustering

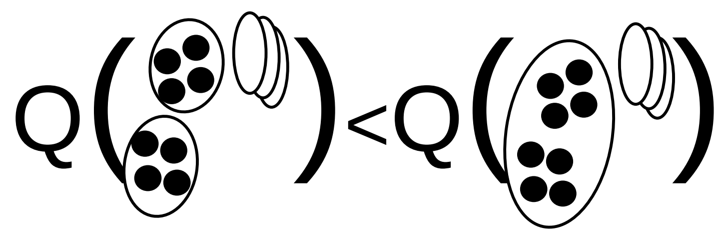
- Are there any formal constraints (properties, criteria) that an ideal evaluation measure should satisfy?

Four Formal Constraints (Amigo et al., 2008) :

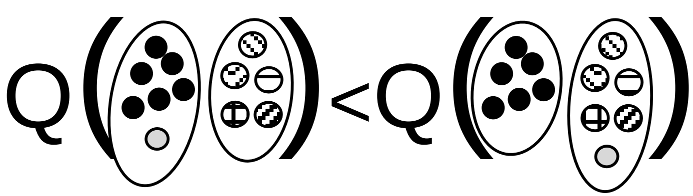
homogeneity



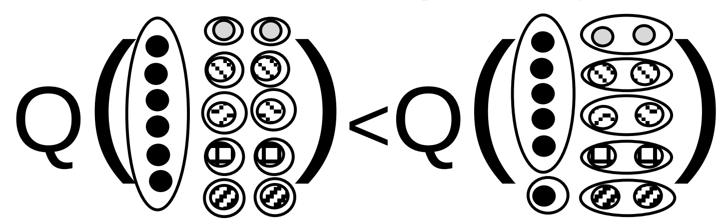
completeness



rag bag



cluster size vs. quantity



- Do the evaluation measures proposed so far satisfy the constraints?

Evaluation Measures

Measure Family	Measures	Comment
Measures Based on Set Mapping	Purity (Zhao and Karypis, 2001) Inverse purity F-measure	purity and inverse purity are easy to cheat, F-measure has a “matching” problem
Measures Based on Pair Counting	Rand index (Rand, 1971) Adjusted rand index (Hubert and Arabie, 1985) Jaccard Coefficient (Milligan et al., 1983) Folks and Mallows FM (Fowlkes and Mallows, 1983)	
Measures Based on Entropy	Entropy Mutual information (Xu et al., 2003) Variation of information (VI) (Meila, 2003) V-Measure (Rosenberg and Hirschberg, 2007)	entropy is easy to cheat, VI and V capture homogeneity and completeness
Measures Based on Editing Distance	Editing distance (Pantel and Lin, 2002)	
Measures for Coreference Resolution	MUC F-measure (Vilain et al.,1995) B-Cubed F-measure (Bagga and Baldwin, 1998) ECM F-measure (Luo, 2005)	

Summary of Part I

- Hypothesis serves as a basis for the whole graph clustering methodology
- Modeling acts as the interface between the real application and the methodology
- Quality measures and graph clustering algorithms construct the backbone of the methodology
- Evaluation deals with utility

Part II Applications:

- Coreference Resolution
- Word Clustering
- Word Sense Disambiguation

Coreference Resolution

- Entity coreference resolution

John Perry, of Weston Golf Club, announced his resignation yesterday.

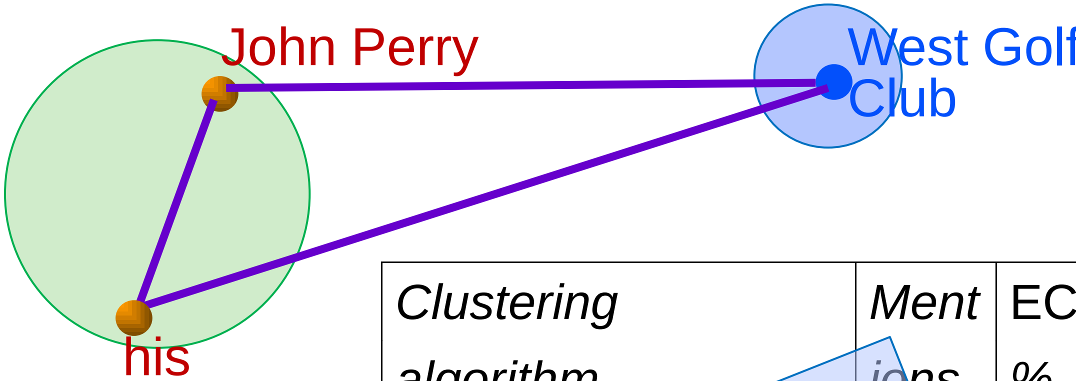
- Event coreference resolution

EM1 *An explosion in a cafe at one of the capital's busiest intersections killed one woman and injured another Tuesday.*

EM2 *The explosion comes a month after*

EM3 *a bomb exploded at a McDonald's restaurant in Istanbul, causing damage but no injuries .*

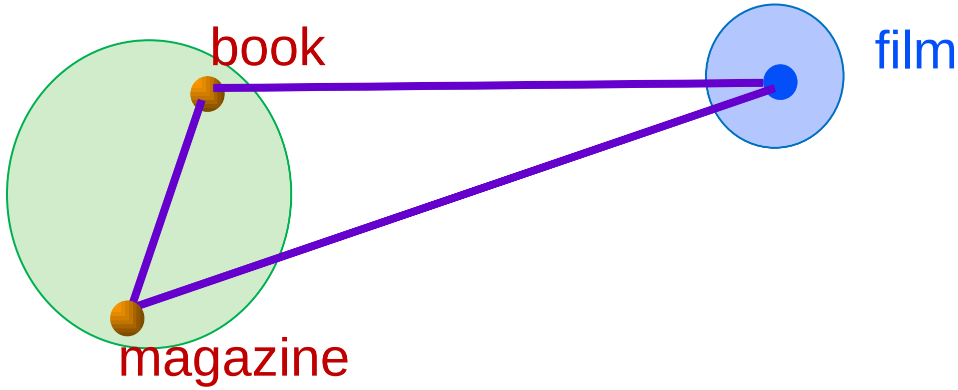
Graph-based clustering Approach



Graph-based clustering

Clustering algorithm	*Mentions*	ECM-F %	MUC *score* MUC P%	MUC R%	MUC F%
BESTCUT(Nicolae and Nicolae,2006)	key	**82.7**	**91.1**	**88.2**	**89.63**
Belltree (Luo et al., 2004)	key	77.9	88.5	89.3	88.90
Link-Best (Ng and Cardie, 2002)	key	77.9	88.0	90.0	88.99

Word Clustering



Matsuo et al. (2006):Graph-based word clustering using web search engine

Graph-based clustering

	PMI F%	Jaccard F%	χ^2 F%
Newman	0.182	0.181	0.480
Average-link	0.179	0.173	0.164

Word Sense Disambiguation

Agirre et al. (2007) :Two graph-based algorithms for state-of-the-art WSD

Graph-based clustering

Two baselines

Supervised

Unsupervised,non-graph based

	Sup.	Unsupervised		
	Rec.	Entr.	Pur.	FS
Vr	59.9	50.3	58.2	**44.1**
Vr_opt	**64.6**	**18.3**	**78.5**	35.0
Pr_fr	64.5	18.7	77.2	34.3
Pr_fx	62.2	25.4	72.2	33.3
1ex-1hub	40.1	0.0	100.0	14.5
MFS	54.5	53.2	52.8	28.3
S3LS-best	**72.9**	**19.9**	**67.3**	**63.8**
kNN-all	70.6	21.2	64.0	60.6
kNN-BoW	63.5	22.6	61.1	57.1
Cymfony (10%-S3LS)	**57.9**	**25.0**	**55.7**	**52.0**
Prob0 (MFS-S3)	54.2	28.8	49.3	46.0
clr04 (MFS-Sc)	48.8	25.8	52.5	46.2
Ciaosenso (MFS-Sc)	48.7	28.0	50.3	48.8
duluth-senserelate	47.5	27.2	51.1	44.9

S3AW task

	recall
kuaw	70.9
Pr_fr	70.7
Vr_opt	70.1
GAMBL	70.1
MFS	69.9
LCCaw	68.6

Conclusions

- Graph: elegant, with solid mathematical foundations

- Non-graph clustering algorithm: act greedily towards the final clustering

  Graph clustering algorithm: seek global "optimal" by optimizing some quality measure

- Issue of running complexity and scalability