

# **Learning to Group Web Text Incorporating Prior Information**

Yu Cheng, Kunpeng Zhang, **Yusheng Xie**, Ankit Agrawal, Weikeng Liao, Alok Choudhary

Dept. of Electrical Engineering and Computer Science  
Center for Ultra-Scale Computing and Security  
Northwestern University

[{ych133, kzh980, yxi389, ankitag, wkliao, choudhar}@eecs.northwestern.edu](mailto:{ych133, kzh980, yxi389, ankitag, wkliao, choudhar}@eecs.northwestern.edu)

# **Outline**

**Introduction**

**Related work**

**Semi-supervised clustering with pair-wise constraints and label**

**Experiments results**

**Conclusion**

# Web Text

- Online text becomes available in a variety of genres.
- Blog & news feeds, forum, customer reviews, book & movie summaries.
- Social media sites like Facebook, Twitter and Google Buzz allow users to post short messages.

The Facebook logo, consisting of the word "facebook" in white lowercase letters on a dark blue rectangular background.The Twitter logo, featuring the word "twitter" in a light blue, lowercase, rounded font with a white outline.The Google Buzz logo, with the word "Google" in its multi-colored font and "buzz" in a grey font, followed by a small colorful globe icon.

# Grouping Web Text

- There are a large number of messages published each day (Google News, Twitter)
- Users often face the problem of information overload.
- Grouping similar text or messages can make information more manageable.
- Systems that can cluster similar text belong to the same topic.

# **Outline**

**Introduction**

**Related work**

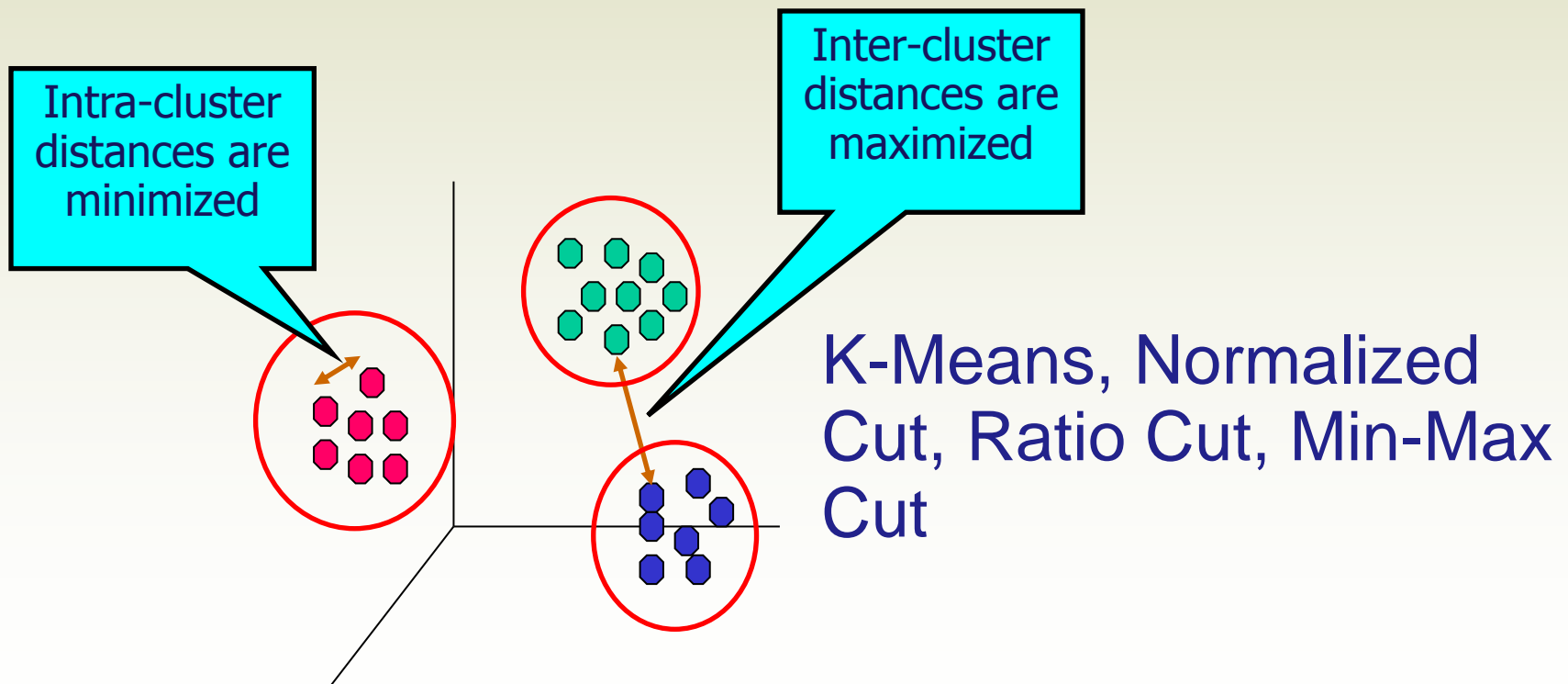
**Semi-supervised clustering with pair-wise constraints and label**

**Experiments results**

**Conclusion**

# Unsupervised clustering

Clustering algorithms are generally used in an unsupervised fashion.



# Semi-supervised Clustering

Input:

A set of unlabeled objects

A small amount of domain knowledge (we call prior knowledge)

Output:

A partitioning of the objects into  $k$  clusters

Objective:

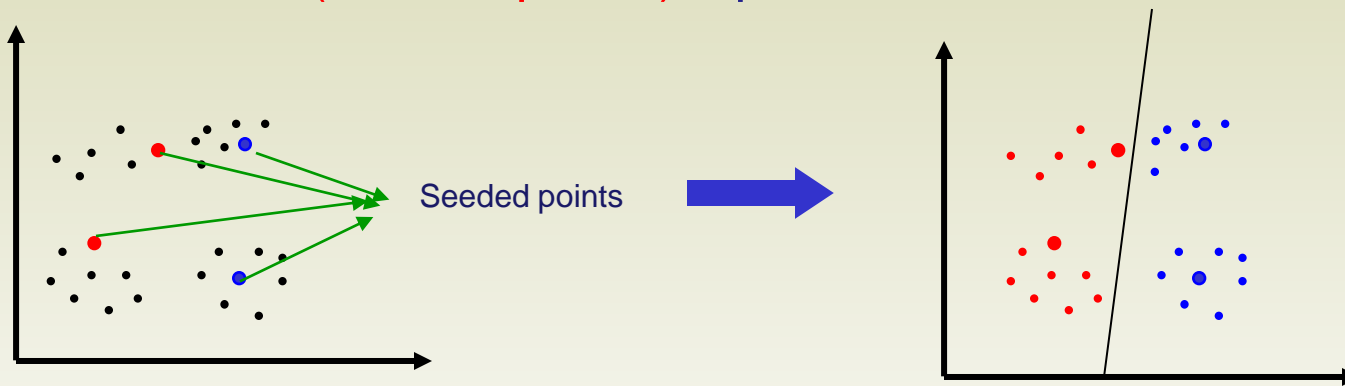
Maximum intra-cluster similarity

Minimum inter-cluster similarity

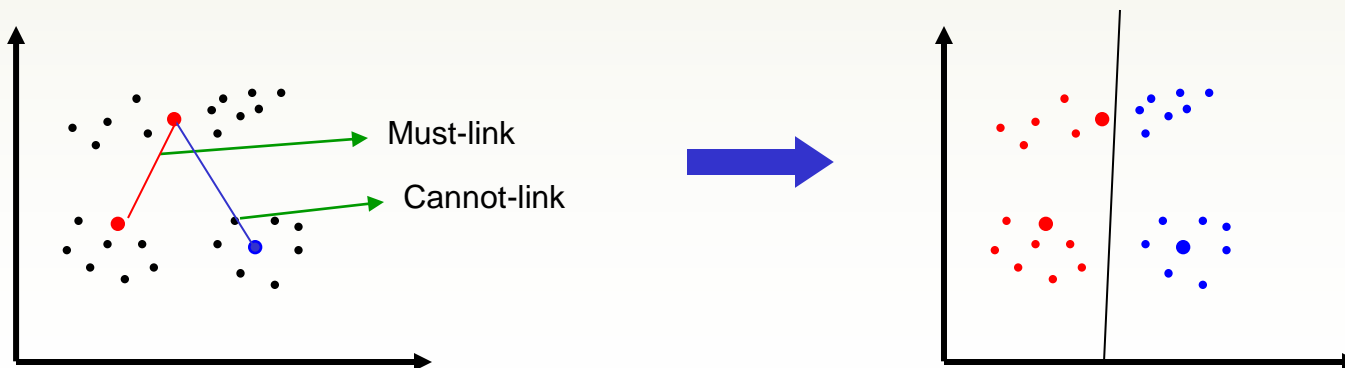
High consistency between the partitioning and the prior knowledge

# Semi-supervised Clustering

According to different given domain knowledge: Users provide **class labels (seeded points)** a priori to some of the document

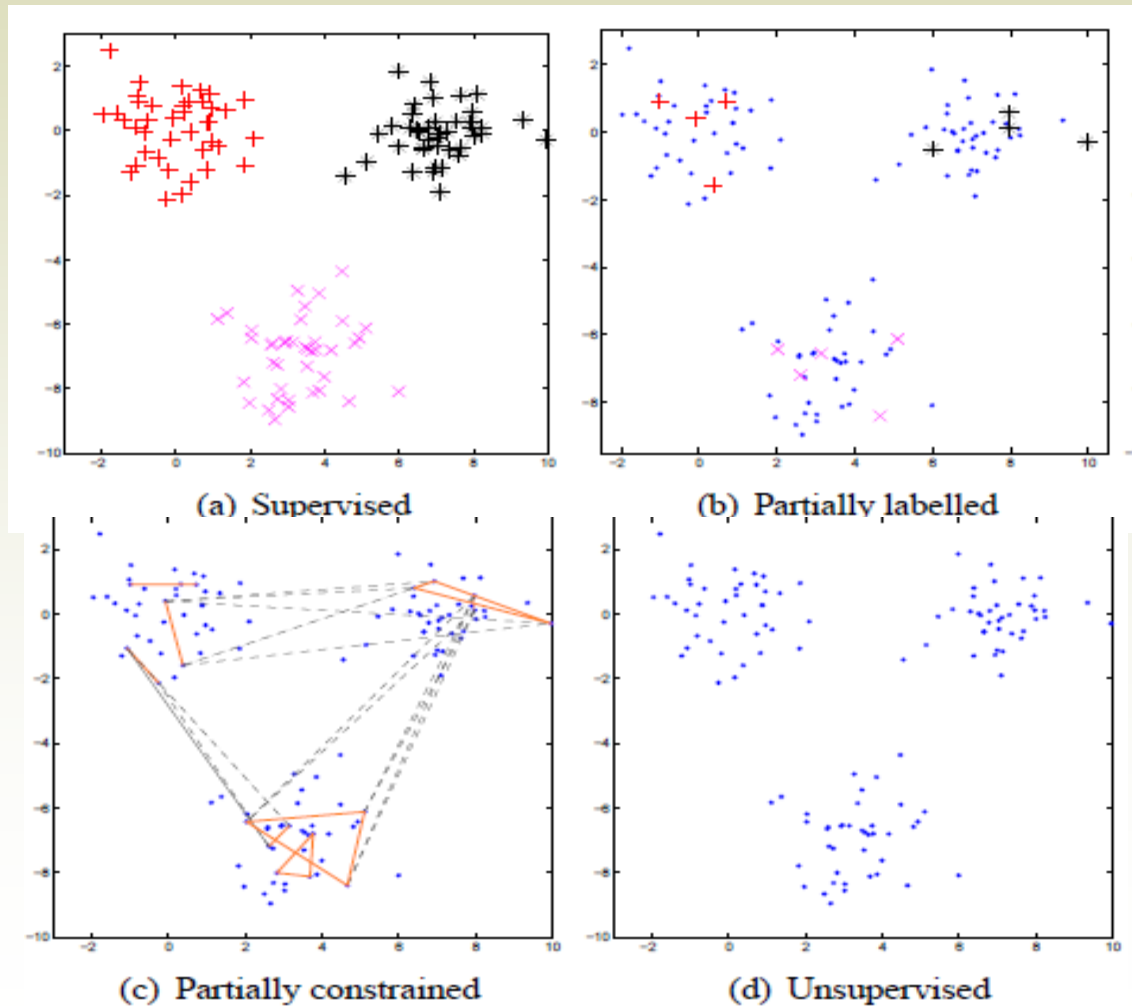


Users know about which few documents are related (**must-link**) or unrelated (**cannot-link**)





# Summaries



# Outline

**Introduction**

**Related Work**

**Semi-supervised clustering with pair-wise constraints and label**

**Experiments results**

**Conclusion**

# The Main Work

## Semi-supervised clustering

Improve existing naive clustering methods  
Using labeled data to guide clustering of unlabeled data

## The pair-wise constrains information

It is easier for the user to distinguish whether two objects are in the same group other than give the labels.  
Giving the algorithms to ask some questions, but ask it wisely.

# Model based clustering

## unsupervised clustering

Dataset: unlabeled dataset  $\mathcal{X}^u$ , data density  $P(x | \theta)$

Objective: Minimize the objective function  $\Theta_u = -\sum_{x \in \mathcal{X}^u} \log P(x | \theta)$

## supervised clustering

Dataset: labeled dataset  $\mathcal{X}^l$ , data density  $P(x_i, y_i | \theta)$

Objective: Minimize the objective function  $Q_l = -\sum_{x_i \in \mathcal{X}^l} \log P(x_i, y_i | \theta)$

## semi-supervised clustering?

Dataset: unlabeled dataset  $\mathcal{X}^u$  and labeled dataset  $\mathcal{X}^l$

Objective: select the parameters by minimizing

$$\Theta = \alpha \Theta_l + \beta \Theta_u, \quad \alpha + \beta = 1$$

**Method: Expectation-Maximization (EM)**

## General Case

Dataset: we can decompose it into three parts  
unlabeled data  $\chi^u$  labeled data  $\chi^l$  data with  
constraints  $\chi^c$

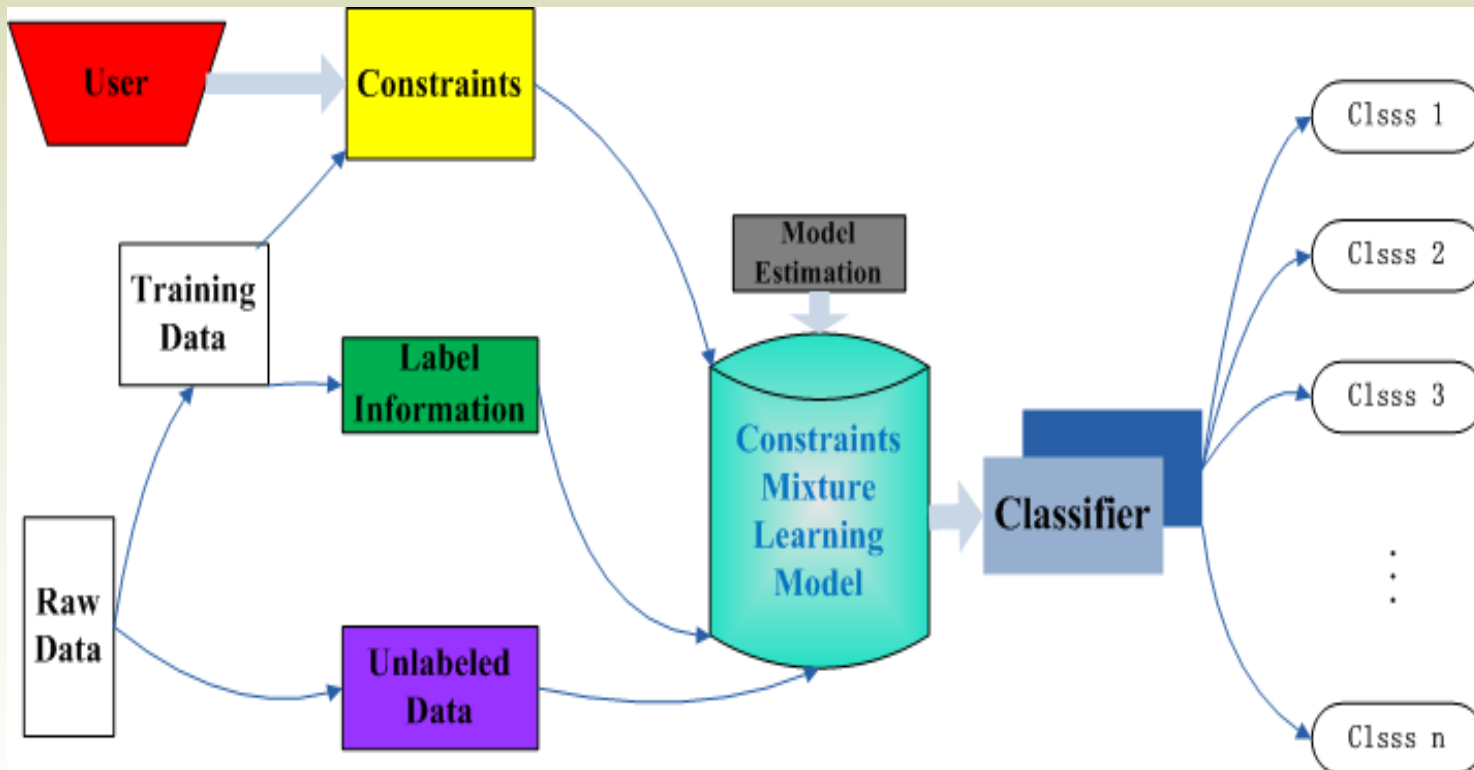
We define the following joint object function, which is a convex  
combination of  $\Theta_u$   $\Theta_l$   $\Theta_c$ ,

Objective: select the parameters by minimizing the

$$\Theta = \alpha\Theta_u + \beta\Theta_l + \gamma\Theta_c,$$
$$\alpha + \beta + \gamma = 1$$

Method: Expectation-Maximization (EM)

# The overall framework



# **Outline**

**Introduction**

**Related Work**

**Semi-supervised clustering with pair-wise constraints and label**

**Experiments**

**Conclusion**

## Experiment Setting

Datasets:

News-500: We collected 500 news documents of 4 different topics from Google News

FBS-500: 500 posts with 7 different categories from Facebook

Compared methods:

Unsupervised: K-Means

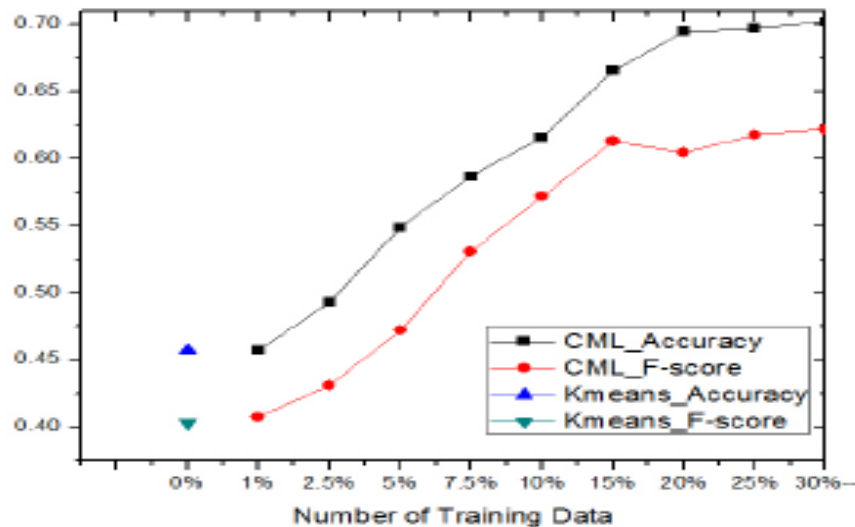
Semi-supervised: Semi-supervised K-Means and Transductive SVM



## Results(comparing with K-Means)

### THE ACCURACY AND F-SCORE OF K-MEANS AND CML ON FBS-500

Evaluation Method	Accuracy	F-score
K-Means	0.457	0.403
CML(5% training)	0.616	0.501
CML(7.5% training)	0.649	0.578
CML(10% training)	0.691	0.627



## Results(comparing with SK-Means and T-SVM)

### THE ACCURACY OF NEWS-500 DATA SET WITH DIFFERENT PERCENTAGE OF TRAINING SAMPLES

Method	1%	2.5%	5%	7.5%	10%
SK-means	0.479	0.513	0.545	0.580	0.629
Transductive SVM	0.442	0.481	0.501	0.521	0.576
CML Model	0.459	0.492	0.549	0.632	0.682

### THE ACCURACY OF FBS-500 DATA SET WITH DIFFERENT PERCENTAGE OF TRAINING SAMPLES

Method	1%	2.5%	5%	7.5%	10%
SK-means	0.486	0.502	0.526	0.540	0.575
Transductive SVM	0.403	0.431	0.458	0.485	0.524
CML Model	0.466	0.494	0.516	0.579	0.625

# **Outline**

**Introduction**

**Related Work**

**Semi-supervised clustering with pair-wise constraints and label**

**Experiments**

**Conclusion**

## Conclusion

### Grouping Text based on the same topic

- there are a large number of web data published every day
- make the messages manageable for the users

### Contribution

- A general framework for semi-supervised clustering work with label and pair wise constraints.
- Clustering the web text using the proposed framework.

### The Future Work

- Considering active learning while sampling.
- Applying the algorithm to other domain (eg. Image categorization, Bioinformatics (gene/protein clustering))

**Thank You!**

**Dept. of Electrical Engineering and Computer Science  
Center for Ultra-Scale Computing and Security  
Northwestern University**