Mining Heterogeneous Network

Clustering and Ranking

Jiatu Shi

Data Mining Lab@UESTC

April 13, 2015

Outline

Outline

A Little Story

One day, Prof.Han asked his students to find out his rank in computer scientists from the DBLP database(¿1.8M papers, ¿0.7M authors, ¿10K venues, ¿70K terms). As he feel honor with his outstanding contribution in data mining.

A Little Story

But, Prof.Han felt frustrated when he saw the following results.

Table 1: A set of conferences from two research areas

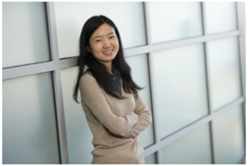
DB/DM	{SIGMOD, VLDB, PODS, ICDE, ICDT, KDD, ICDM, CIKM, PAKDD, PKDD}
HW/CA	{ASPLOS, ISCA, DAC, MICRO, ICCAD, HPCA, ISLPED, CODES, DATE, VTS }

Table 2: Top-10 ranked conferences and authors in the mixed conference set

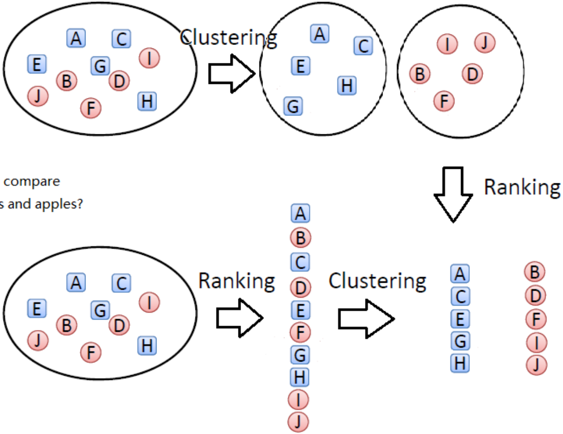
Rank	Conf.	Rank	Authors
1	DAC	1	Alberto L. Sangiovanni-Vincentelli
2	ICCAD	2	Robert K. Brayton
3	DATE	3	Massoud Pedram
4	ISLPED	4	Miodrag Potkonjak
5	VTS	5	Andrew B. Kahng
6	CODES	6	Kwang-Ting Cheng
7	ISCA	7	Lawrence T. Pileggi
8	VLDB	8	David Blaauw
9	SIGMOD	9	Jason Cong
10	ICDE	10	D. F. Wong

A Little Story

So Prof.Han had a talk with Dr.Sun.



How to compare oranges and apples?



A Little Story

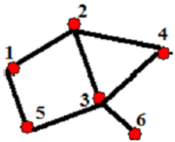
Prof.Han felt happy when he saw the result found by Dr.Sun.

Table 3: Top-10 ranked conferences and authors in DB/DM set

Rank	Conf.	Rank	Authors
1	VLDB	1	H. V. Jagadish
2	SIGMOD	2	Surajit Chaudhuri
3	ICDE	3	Divesh Srivastava
4	PODS	4	Michael Stonebraker
5	KDD	5	Hector Garcia-Molina
6	CIKM	6	Jeffrey F. Naughton
7	ICDM	7	David J. DeWitt
8	PAKDD	8	Jiawei Han
9	ICDT	9	Rakesh Agrawal
10	PKDD	10	Raghu Ramakrishnan

Basic Concepts of Network

A network/graph: G = (V, E), where V: vertices/nodes, E: edges/links



$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Adjacency matrix:

A_{ij} = 1 if there is an edge between vertices i and j; 0 otherwise

Weighted graph:

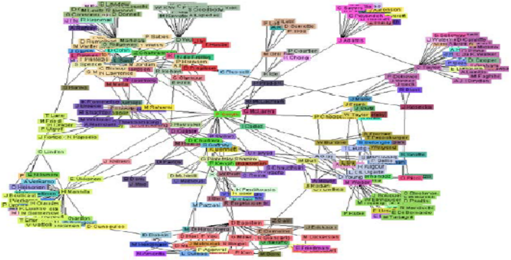
Edges having weight (strength), usually a real number

Directed network (directed graph): if each edge has a direction

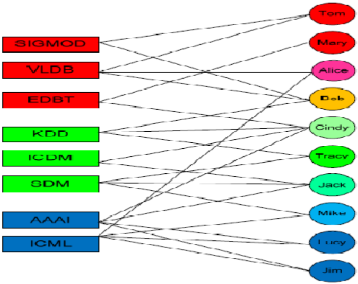
Labeled graph:

Edges have a label (e.g., creation date)

Basic Concepts of Network



Co-author Network



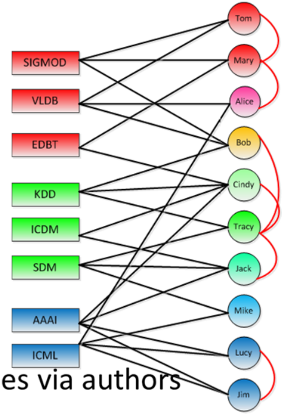
Conference-Author Network

- Homogeneous networks
  - Single object type and single link type (one-mode data)
  - Web: a collection of linked Web pages
- Heterogeneous or multi-typed networks
  - Multiple object and link types
  - Medical network: patients, doctors, diseases, treatments
  - Bibliographic network: publications, authors, venues

Outline

Weight Matrix and Conditional Rank

- A case study on bi-typed DBLP network
- Links exist between
  - Conference (X) and author (Y)
  - Author (Y) and author (Y)
- A matrix denoting the weighted links
  - $W = \begin{pmatrix} W_{XX} & W_{XY} \\ W_{YX} & W_{YY} \end{pmatrix}$
- Goal:
  - Clustering and ranking conferences via authors



$$G = \langle \{X \cup Y\}, W \rangle$$

$$\forall x \in X, \vec{r}_X(x) \geq 0, \sum_{x \in X} \vec{r}_X(x) = 1, \ \ and$$

$$\forall y \in Y, \vec{r}_Y(y) \geq 0, \sum_{y \in Y} \vec{r}_Y(y) = 1,$$

$$X' \subseteq X \quad G' = \langle \{X' \cup Y\}, W' \rangle$$

$$\vec{r}_{X|X'}(x) = \frac{\sum_{j=1}^{n} W_{XY}(x,j)\vec{r}_{Y|X'}(j)}{\sum_{i=1}^{m}\sum_{j=1}^{n} W_{XY}(i,j)\vec{r}_{Y|X'}(j)}$$

- Ranking as the feature of the cluster
  - Ranking is conditional on a specific cluster
    - E.g., VLDB's rank in Theory vs. its rank in the DB area
  - The distributions of ranking scores over objects are different in each cluster
- Clustering and ranking are mutually enhanced
  - Better clustering: rank distributions for clusters are more distinguishing from each other
  - Better ranking: better metric for objects is learned from the ranking
- Not every object should be treated equally in clustering!

Ranking Methods

$$\begin{cases} \vec{r}_X(x) = \dfrac{\sum_{j=1}^{n} W_{XY}(x,j)}{\sum_{i=1}^{m}\sum_{j=1}^{n} W_{XY}(i,j)} \\ \vec{r}_Y(y) = \dfrac{\sum_{i=1}^{n} W_{XY}(i,y)}{\sum_{i=1}^{m}\sum_{j=1}^{n} W_{XY}(i,j)} \end{cases}$$

- Simple Ranking
  - Proportional to # of publications of an author / a conference
  - Considers only **immediate neighborhood** in the network

What about an author publishing 100 papers in very weak conferences?

Ranking Methods

- Authority Ranking:
  - More sophisticated "rank rules" are needed
  - **Propagate** the ranking scores in the network over different types
- **Rule 1**: Highly ranked authors publish *many* papers in highly ranked conferences

$$\vec{r}_Y(j) = \sum_{i=1}^{m} W_{YX}(j,i)\vec{r}_X(i) \qquad \vec{r}_Y(j) \leftarrow \frac{\vec{r}_Y(j)}{\sum_{j'=1}^{n} \vec{r}_Y(j')}$$

- **Rule 2**: Highly ranked conferences attract *many* papers from *many* highly ranked authors

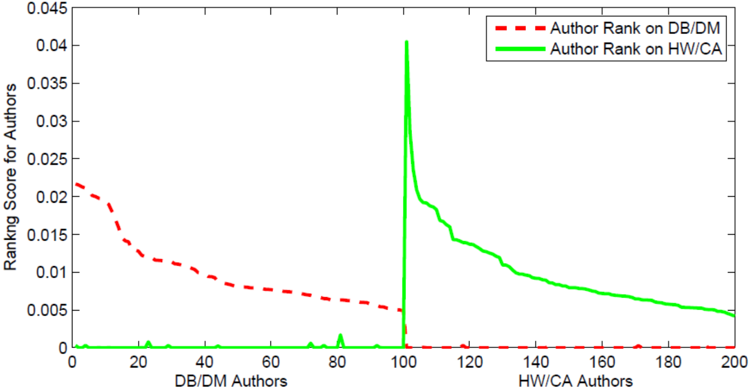
$$\begin{cases} \vec{r}_X = \dfrac{W_{XY}\vec{r}_Y}{\|W_{XY}\vec{r}_Y\|} \\ \vec{r}_Y = \dfrac{W_{YX}\vec{r}_X}{\|W_{YX}\vec{r}_X\|} \end{cases}$$

$$\vec{r}_X(i) = \sum_{j=1}^{n} W_{XY}(i,j)\vec{r}_Y(j) \qquad \vec{r}_X(i) \leftarrow \frac{\vec{r}_X(i)}{\sum_{i'=1}^{m} \vec{r}_X(i')}$$

- **Rule 3**: The rank of an author is enhanced if he or she co-authors with *many* highly ranked authors

$$\vec{r}_Y(i) = \alpha \sum_{j=1}^{m} W_{YX}(i,j)\vec{r}_X(j) + (1-\alpha) \sum_{j=1}^{n} W_{YY}(i,j)\vec{r}_Y(j).$$

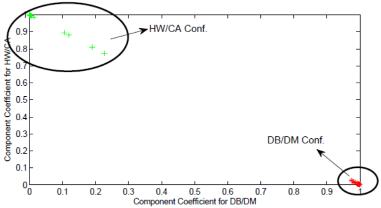
Mixture Model

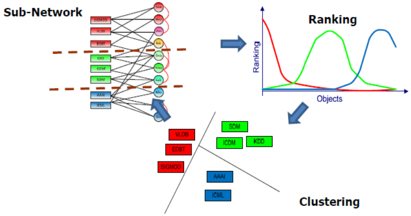


Conditional Rank as Cluster Feature

$$p_{x_i}(Y) = \sum_{k=1}^{K} \pi_{i,k} p_k(Y), \text{ and } \sum_{k=1}^{K} \pi_{i,k} = 1.$$

Algorithm Framework





$\Theta_{m\times K} = \{\pi_{i,k}\}$ *EM Algorithm*

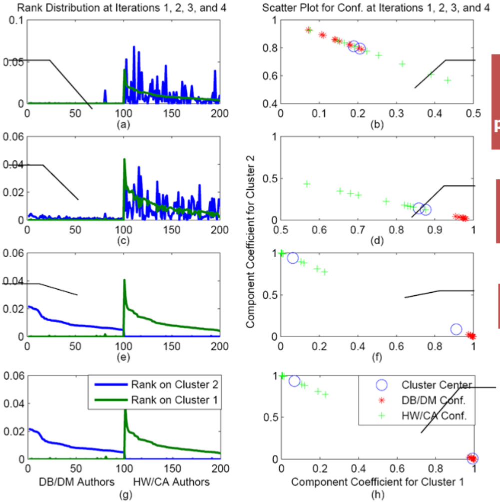
- Initialization
  - Randomly partition
- Repeat
  - Ranking
    - Ranking objects in each sub-network induced from each cluster
  - Generating new measure space
    - Estimate **mixture model coefficients** for each target object
  - Adjusting cluster
- Until stable

Visualization of Clustering

Initially, ranking distributions are mixed together

Improved a little

Improved significantly



Two clusters of objects mixed together, but preserve similarity somehow

Two clusters are almost well separated

Well separated

Stable

Clustering Result

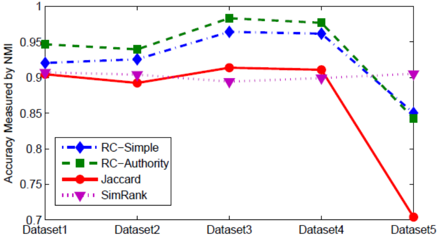


Figure 4: Accuracy of Clustering

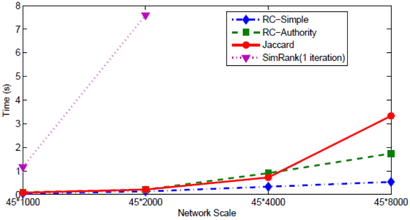


Figure 5: Efficiency Analysis

$$p(i,j) = \frac{n(i,j)}{N}$$
$$p_1(j) = \sum_{i=1}^{K} p(i,j)$$
$$p_2(i) = \sum_{j=1}^{K} p(i,j)$$

$$NMI = \frac{\sum_{i=1}^{K}\sum_{j=1}^{K} p(i,j)\log\left(\frac{p(i,j)}{p_1(j)p_2(i)}\right)}{\sqrt{\sum_{j=1}^{K} p_1(j)\log p_1(j)\sum_{i=1}^{K} p_2(i)\log p_2(i)}}$$

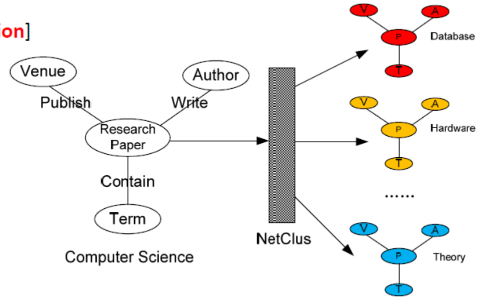
Complexity Analysis

- At each iteration, |E|: edges in network, m: number of target objects, K: number of clusters
  - Ranking for sparse network
    - ~O(|E|)
  - Mixture model estimation
    - ~O(K|E|+mK)
  - Cluster adjustment
    - ~O(mK^2)
- In all, linear w.r.t. |E|
  - ~O(K|E|)

Outline

Star Network

- Beyond bi-typed information network
  - A Star Network Schema [**richer information**]
- Split a network into different layers
  - Each represented by a **network cluster**



Conference	Rank Score
SIGMOD	0.315
VLDB	0.306
ICDE	0.194
PODS	0.109
EDBT	0.046
CIKM	0.019
...	...

Author	Rank Score
Michael Stonebraker	0.0063
Surajit Chaudhuri	0.0057
C. Mohan	0.0053
Michael J. Carey	0.0052
David J. DeWitt	0.0051
H. V. Jagadish	0.0043
...	...

Term	Rank Score
database	0.0529
system	0.0322
query	0.0313
data	0.0251
object	0.0138
management	0.0113
...	...

Table 1: Ranking Description for Net-Cluster of Database Research Area

Refinement of Methods

$$w_{x_i x_j} = \begin{cases} 1, \text{ if } x_i(x_j) \in A \cup C \text{ and } x_j(x_i) \in D, \\ \quad \text{and } x_i \text{ has link to } x_j \\ c, \text{ if } x_i(x_j) \in T \text{ and } x_j(x_i) \in D \text{ and } x_i(x_j) \\ \quad \text{appears c times in } x_j(x_i), \\ 0, \text{ otherwise.} \end{cases}$$

$$p(x|G) = p(T_x|G) \times p(x|T_x, G)$$

$$P(C|T_C, G) = W_{CD} D_{DA}^{-1} W_{DA} P(A|T_A G)$$

$$P(A|T_A, G) = W_{AD} D_{DC}^{-1} W_{DC} P(C|T_C, G)$$

$$p(d|G_k) = \prod_{x \in N_{G_k}(d)} p(x|T_x, G_k)^{W_{d,x}} p(T_x|G_k)^{W_{d,x}}$$

$$P_S(X|T_X, G_k) = (1 - \lambda_S) P(X|T_X, G_k) + \lambda_S P(X|T_X, G)$$

Simple Ranking

$$p(x|T_x, G) = \frac{\sum_{y \in N_G(x)} W_{xy}}{\sum_{x' \in T_x} \sum_{y \in N_G(x')} W_{x'y}}$$

Authority Ranking

$$P(Y|T_Y, G) = W_{YZ} W_{ZX} P(X|T_X, G)$$

Refinement of Results

- The network cluster for database area: Conferences, Authors, and Terms
  - Better clustering and ranking than RankClus

	NetClus (A+C+T+D)	PLSA (T+D)
Accuracy	**0.7705**	0.608

Table 6: Accuracy of Paper Clustering Results

	RankClus $d(a) > 0$	RankClus $d(a) > 5$	RankClus $d(a) > 10$	NetClus $d(a) > 0$
NMI	0.5232	0.8390	0.7573	**0.9753**

Table 7: Accuracy of Conference Clustering Results