Algorithms Tested in Social Web

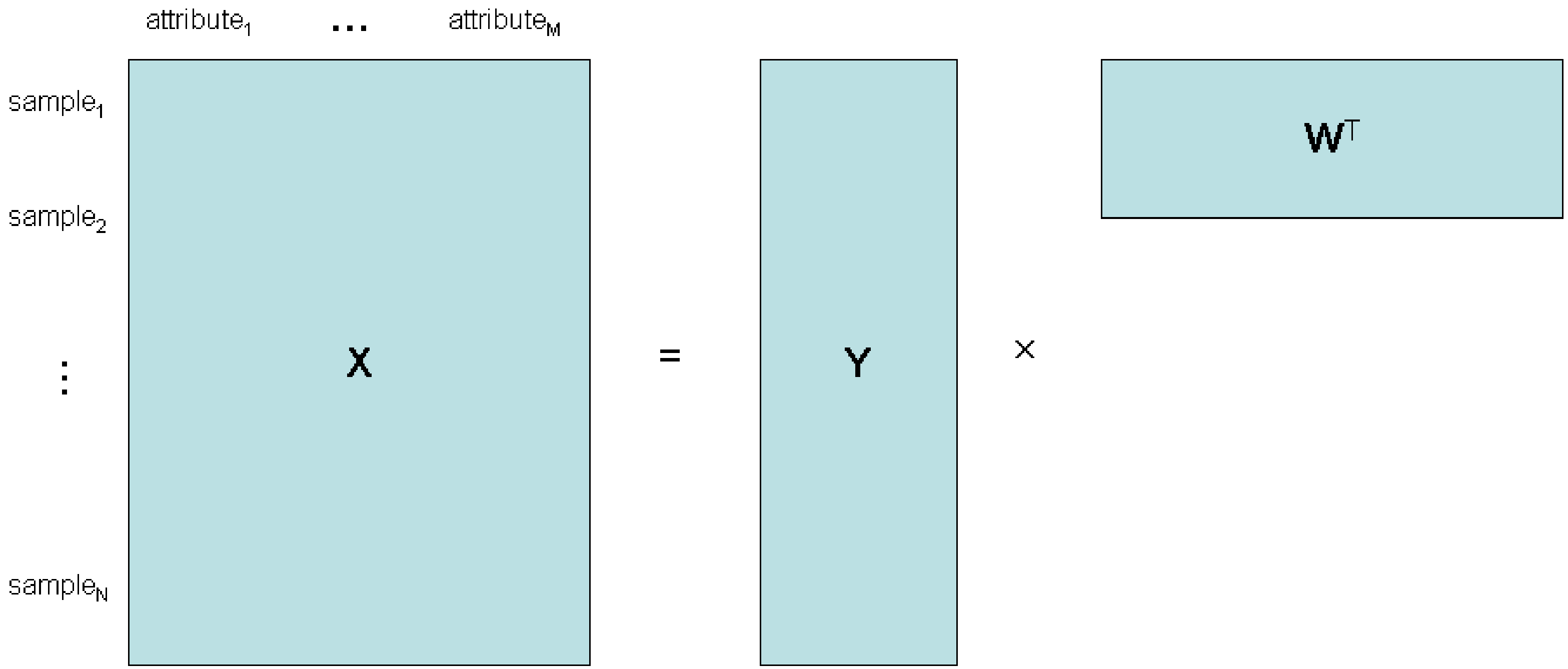
Lauri Lahti, Oskar Kohonen and Zhirong Yang

26 February 2008

Data sets

- 296 links from the popular page on del.icio.us
- 13120 Tags
- 64484 Users
- Link recommendation using links x tags
- User recommendation using users x links

Overview



$\mathbf{Y}$: $N \times R$ ($R \ll M$), rows as compact representation of the samples.

$\mathbf{W}$: $M \times R$, columns as hidden concepts (clusters).

Clustering Links

$\mathbf{X}$: link-user matrix

Latent Semantic Analysis (LSA)

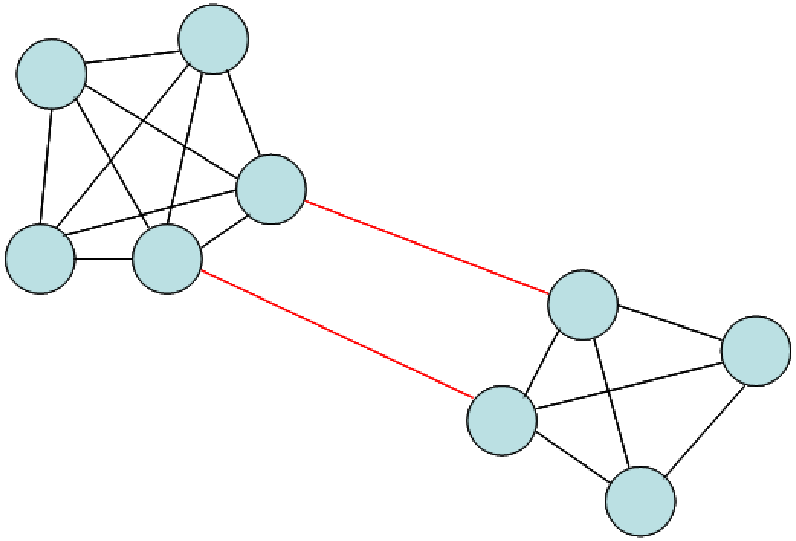
$$[\mathbf{Y}, \mathbf{W}] = \arg\min \|\mathbf{X} - \mathbf{Y}\mathbf{W}^T\|_F^2,$$

where $\|A\|_F^2 = \sum_i \sum_j A_{ij}^2$.

$$[\mathbf{U}, \mathbf{S}, \mathbf{W}^T] = \mathsf{svd}(\mathbf{X}, R),$$

and $\mathbf{Y} = \mathbf{U}\mathbf{S}$.

Spectral Graph Clustering by Normalized Cut (Ncut)



$$Ncut(A,B) = \frac{cut(A,B)}{assoc(A,V)} + \frac{cut(A,B)}{assoc(B,V)}$$

where $cut(A,B) = \sum_{u \in A, v \in B} W_{uv}$,

and $assoc(A,V) = \sum_{u \in A, t \in V} W_{ut}$.

Approximated Solution of Ncut

Given the similarity (adjacency) matrix $\mathbf{G}$, the columns of Y can be approximated by a generalized eigen decomposition:

$$(\mathbf{D}-\mathbf{G})\mathbf{y}=\lambda\mathbf{D}\mathbf{y}, \tag{1}$$

where $\mathbf{D}$ is a diagonal matrix with $D_{ii}=\sum_j G_{ij}$. Then $\mathbf{Y}$ contains the eigenvectors associated with the second to the $(R+1)$-th smallest eigenvalues of $\mathbf{D}^{-\frac{1}{2}}(\mathbf{D}-\mathbf{G})\mathbf{D}^{-\frac{1}{2}}$.

Ncut Finds the Laplace-Beltrami Operator

minimize

$$\sum_{i=1}^{N}\sum_{j=1}^{N} G_{ij}(y_i - y_j)^2$$

subject to

$$\mathbf{y}^T\mathbf{D}\mathbf{y} = 1$$

.

This actually finds a mapping $y = f(\mathbf{x})$ that minimizes $\frac{\|\nabla f\|}{\|f\|}$. Therefore, the Ncut solution is also called the *Laplacian Eigenmap*.

Locality Preserving Projection (LPP)

LPP is the linearized version of Laplacian Eigenmap. That is, we set $y = f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ for a sample $\mathbf{x}$ and solve

$$\mathbf{X}^T(\mathbf{D} - \mathbf{G})\mathbf{X}\mathbf{w} = \lambda\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{w}.$$

Then, $\mathbf{W}$ consists of the $\mathbf{w}$'s which are associated with the second to $(R + 1)$-th eigenvalues, and $\mathbf{Y} = \mathbf{X}\mathbf{W}$.

Clustering Users

$\mathbf{X}$: user-link matrix

LSA runs as well, but LPP does not because $\mathbf{G}$ is too large in this case.

Non-negative Matrix Factorization (NMF)

LSA: $[\mathbf{Y}, \mathbf{W}] = \arg\min \|\mathbf{X} - \mathbf{Y}\mathbf{W}^T\|_F^2$.

NMF: $[\mathbf{Y}, \mathbf{W}] = \arg \min_{\mathbf{Y}\geq 0, \mathbf{W}\geq 0} \|\mathbf{X} - \mathbf{Y}\mathbf{W}^T\|_F^2$.

$$\frac{\partial \|\mathbf{X} - \mathbf{Y}\mathbf{W}^T\|_F^2}{\partial Y_{ij}} = 2\left[\mathbf{Y}\mathbf{W}^T\mathbf{W}\right]_{ij} - 2\left[\mathbf{X}\mathbf{W}\right]_{ij}$$

$$\frac{\partial \|\mathbf{X} - \mathbf{Y}\mathbf{W}^T\|_F^2}{\partial W_{ij}} = 2\left[\mathbf{W}\mathbf{Y}^T\mathbf{Y}\right]_{ij} - 2\left[\mathbf{X}^T\mathbf{Y}\right]_{ij}$$

$$Y_{ij}^{\text{new}} = Y_{ij}\frac{[\mathbf{X}\mathbf{W}]_{ij}}{[\mathbf{Y}\mathbf{W}^T\mathbf{W}]_{ij}} \qquad W_{ij}^{\text{new}} = W_{ij}\frac{\left[\mathbf{X}^T\mathbf{Y}\right]_{ij}}{[\mathbf{W}\mathbf{Y}^T\mathbf{Y}]_{ij}}$$

Projective Non-negative Matrix Factorization (P-NMF)

With $\mathbf{Y} = \mathbf{X}\mathbf{W}$, P-NMF finds $\mathbf{W} \geq 0$ that minimizes

$$\mathcal{J} = \|\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{W}^T\|_F^2.$$

$$\frac{\partial \mathcal{J}}{\partial W_{ij}} =$$

$$-4\left[\mathbf{X}^T\mathbf{X}\mathbf{W}\right]_{ij} + 2\left[\mathbf{W}\mathbf{W}^T\mathbf{X}^T\mathbf{X}\mathbf{W}\right]_{ij} + 2\left[\mathbf{X}^T\mathbf{X}\mathbf{W}\mathbf{W}^T\mathbf{W}\right]_{ij}$$

$$W_{ij}^{\text{new}} = W_{ij}\frac{2\left[\mathbf{X}^T\mathbf{X}\mathbf{W}\right]_{ij}}{\left[\mathbf{W}\mathbf{W}^T\mathbf{X}^T\mathbf{X}\mathbf{W}\right]_{ij} + \left[\mathbf{X}^T\mathbf{X}\mathbf{W}\mathbf{W}^T\mathbf{W}\right]_{ij}}$$

Results - Link recommendation

- Data set likely too small
- LSI works fairly well
- LPP works fairly badly

Results - User recommendation

- Both methods work
- LSI still works well and is robust
- P-NMF works fairly well, especially with euclidean distance
  - Works well for "power user" with lots of links
  - Users with few links get squeezed into same component