Kernel-Based Link Analysis

Masashi Shimbo  Computational Linguistics Lab, Nara Institute of Science and Technology  shimbo@is.naist.jp

Link analysis measures

- PageRank [Brin and Page 1998]
- HITS [Kleinberg 1999]
- L
- vNK

Importance of individual nodes

- Authority vector = dominant eigenvector of \(A^T A\)
- Hub vector = dominant eigenvector of \(A A^T\)

Kernels over graph nodes

- Diffusion kernels [Kondor and Lafferty 2002]
- von Neumann kernels [Kandola et al. 2003]
- Regularized Laplacian [Smola and Kondor 2003]

Questions

Can we interpret these kernels as link analysis measures?
Are these kernels related to existing link analysis measures?
How does the character of these kernels change with the parameter? 

Other contributions

- Intermediate between HITs and Laplacian-based kernels
- Use \(L = D - A^T A\) in place of \(L\). Resulting kernel still positive semidefinite.
- Parameter sensitivity analysis
- von Neumann kernels are parameter sensitive — the change of their character from relatedness to importance is often abrupt. It is possible to estimate the change point with the help of the derivative of the kernel matrix wrt parameter \(\beta\), which has a simple closed form.
- Computational consideration
- Resulting kernel matrices can be dense even if the original (co-)citation graph is sparse. However, it is possible to compute a row of the kernel matrices one by one, which leads to a significant saving of required memory if only a few nodes are of interest.

Experimental result

Dataset: Citation graph of 2280 papers on Natural Language Processing.
Top-10 papers with respect to the root paper:
- Real-world example of the above quiz.
- Top-10 papers are identical for RL and vNK (but rankings are different).
- vNK ranks a non-discourse paper (having top HITS ranking) above the root paper.

Quiz. Which of the two nodes, \(x\) or \(z\), is more related to \(y\)?

Co-citation coupling “\(x\) and \(z\) are equally related to \(y\).”
\(x\) has many incoming edges from nodes not citing \(y\); \(z\) has no such edges. The number of these edges has no effect on co-citation counts.

Regularized Laplacian
- \(RL(z,y) > RL(z,y)\) as desired.
- \(L = D - A^T A\) = Laplacian of co-citation graph
- \(D = \text{diagonal degree matrix with } D(i,i) = \sum_j A(i,j)\)

Use of \((-L)\) in place of \(A^T\); discounting of important nodes
- Alternative heuristics to discount important nodes lead to anomaly when \(\beta\) is increased; e.g.,

\[
\text{Divide edge weights by the in-degree of nodes and apply von Neumann kernel} \\
\Rightarrow z \text{ is more related to } x \text{ than to } y!
\]