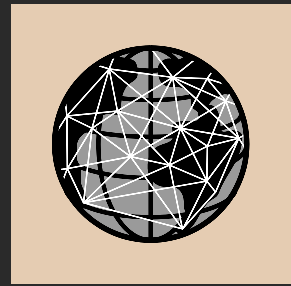
Kernel-Based Link Analysis

Takahiko Ito

Nara Institute of Science and Technology

Motivation

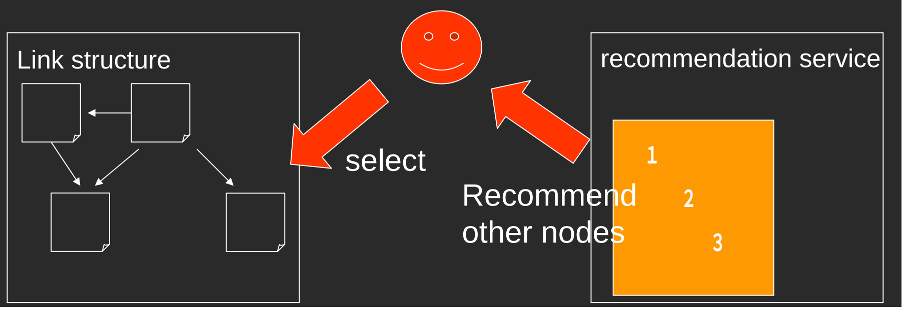
- WWW or citations are represented by a huge graph
  - Node: web page, paper
  - Edge: hyper link, citation



- Methods to explore graph data are desired

Recommendation service

1. Users select favorite nodes (root nodes) – papers / web pages
2. based on links around of root nodes, the system recommend other nodes that may interest the users



To recommend pages

- Link Analysis measures:
  - Measures for analyzing the relationship among nodes in graphs.
  - However, classical link analysis measures have some limitations, if they are applied to recommendation services
- We proposed a new link analysis measure based on kernel methods.

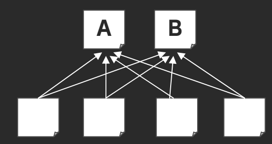
Table of contents

- Introduce link analysis measures.
- Propose a new link analysis measure for recommendation services applying kernel methods in graphs.

Co-citation/bibliographic coupling "relatedness"

Co-citation coupling [Small et al., 1973]

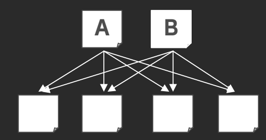
defines relatedness as the number of papers jointly citing the given pair of papers



Co-citation coupling (A,B) = 4

Bibliographic coupling [Kessler,1963]

defines relatedness as the number of common citations made by two papers



Bibliographic coupling (A,B) = 4

Computing co-citation/bibliographic coupling

Given adjacency matrix A of a citation graph,

- (i, j)-element of A^TA
  - ➔ Co-citation relatedness between nodes i and j
- (i, j)-element of AA^T
  - ➔ Bibliographic relatedness between nodes i and j

HITS "importance"

HITS [Kleinberg, 1999]

assigns two scores to each node:

Authority score:

Nodes cited by many nodes receive a high authority score

Hub score:

Node citing many authoritative nodes receive a high hub score.

Fact: equivalence of HITS and eigenvector computation

Given an adjacency matrix A of a citation graph, it is well known that

HITS authority vector = principal eigenvector of $A^T A$

HITS hub vector = principal eigenvector of AA^T

Application of link analysis measures to recommendation service

- Importance measures recommend popular (important) nodes.
  - ➔However, system may return nodes with different topic to root nodes
- Relatedness measures recommend nodes on the same topic to root nodes
  - ➔ However, system may return low quality nodes

Proposed link analysis measure

- We propose the measure that is an interpolationbetween importance and relatedness
  - ➔ System can recommend pages not only popular but same topic.
  - ➔ In addition, a parameter can control the bias

This property allow each user to adjust the induced link analysis measures to suit user's objectives by tuning of a parameter.

Table of contents

- Introduce link analysis measures.
- Propose a new link analysis measure for recommendation services applying kernel methods in graphs.

Neumann kernels [Kandola et al., 2003]

- Original Neumann kernels compute document relatedness, but *not* on the basis of citations.
- They use graphs induced by the content of documents:

  Edge between nodes (documents) has a weight based on the number of common terms in their contents.

Definition:

$NK_\beta(XX^T) = XX^T + \beta(XX^T)^2 + \beta^2(XX^T)^3 + \ldots$ (document relatedness)

$NK_\beta(X^TX) = X^TX + \beta(X^TX)^2 + \beta^2(X^TX)^3 + \ldots$ (term relatedness)

where X is a document-by-term matrix, and β is a weighting parameter of matrices.

Neumann kernels for link analysis

Neumann kernels in this work

- are applied directly to citation graphs.
- i.e., use adjacency matrix A of a citation graph in place of document-by-term matrix X.

Definition:

$NK_\beta(AA^T) = AA^T + \beta(AA^T)^2 + \beta^2(AA^T)^3 + \ldots$

$NK_\beta(A^TA) = A^TA + \beta(A^TA)^2 + \beta^2(A^TA)^3 + \ldots$

What do $(AA^T)^n$ and $(A^TA)^n$ in these series represent?

Meaning of $(A^TA)^n$

- (i, j)-element of $(A^TA)^n$ = number of paths of length n between nodes i and j in a co-citation graph.
- Increasing n from 1 towards ∞ changes $(A^TA)^n$ from relatedness to importance.

After n=5, all rows of $(A^TA)^n$ give an identical ranking C>D>B>A.
This ranking also matches the HITS authority ranking.

$(A^TA)^n$ tends towards HITS importance as $n\to\infty$

Theorem.

Given the co-citation matrix A^TA,

$$\left(\frac{A^TA}{\lambda}\right)^n \to xx^T \quad \text{as } n \to \infty$$

where

λ is the principal eigenvalue of matrix A^TA, and
x is its principal eigenvector (HITS authority vector),
N.B., every row/column of xx^T gives the same ranking of nodes as HITS authority.

Corollary.

Given any two nodes i and j with Authority(i) > Authority(j), there is an integer m s.t.

$(A^TA)^n [i,k] > (A^TA)^n [j,k]$ for all n>m and for any node k.

To sum up, Neumann kernel is

- Computing a weighted sum of path weights between nodes.
- And it is a "mixture" of relatedness and importance.

$$NK_{\beta}(A^TA) = A^TA + \beta(A^TA)^2 + \beta^2(A^TA)^3 + \beta^3(A^TA)^4 + \dots$$

Relatedness ⟷ Importance

Small β ➔ NK is biased towards relatedness

- Special case at β=0:

  $NK_{\beta}(A^TA)$ reduces to the co-citation coupling matrix

Large β ➔ NK is biased towards importance

Summary

We showed that

$NK_{\beta}(A^TA)$

interpolation between co-citation coupling and HITS authority scores

$NK_{\beta}(AA^T)$

interpolation between bibliographic coupling and HITS hub scores

Co-citation coupling / Bibliographic coupling — Neumann kernels — HITS

small ⟷ β ⟷ large

Experiments

Compare

- Neumann kernels

with

- HITS

Dataset:

Citation graph consisting of 2687 papers on natural language processing

Neumann kernel with large β (β=.005)

Neumann kernel gives the same ranking as HITS

NK	HITS	Title
1	1	Building a large annotated corpus of English: the Penn Treebank
2	2	A stochastic parts program and noun phrase parser for unrestricted text
3	3	Statistical decision-tree models for parsing
4	4	A new statistical parser based on bigram lexical dependencies
5	5	Unsupervised word sense disambiguation rivaling supervised methods
6	6	Word-sense disambiguation using statistical models of Roget's
7	7	The mathematics of statistical machine translation: parameter estimation

Neumann kernel with small β (β=.001)

The titles of papers show that most of the high-ranked papers are related to the root paper

NK	HITS	Title
1	1	Building a large annotated corpus of English: the Penn Treebank
2	771	Empirical studies in discourse
3	50	Attention, intentions, and the structure of discourse
4	76	Assessing agreement on classification tasks: the Kappa statistic
5	201	The reliability of a dialogue structure coding scheme
6	604	Message Understanding Conference (MUC) Tests of Discourse Processing
7	1061	Effects of variable initiative on linguistic behavior in human-computer spoken natural language dialogue

Comparison between Neumann kernels and HITS (quantitative evaluation)

- The difference between Neumann kernels and HITS authority ranking
  - Making each of paper one by one as the root node
  - Using K-min distance [Fagin et al., 2003]:
    - If two ton-n lists have similar rankings ⇒ small
    - If two top-n lists have similar rankings ⇒ large

MAX:100 MIN: 0

	Neumann kernels (λ)						
	λ=0.0001	0.001	0.003	0.004	0.0045	0.0048	0.005
HITS	89.9	89.9	88.7	86.2	81.7	73.3	20.4

Conclusions

- Neumann kernels on citation graphs provide a new link analysis measure that is feasible for recommendation services.

References

- M. M. Kessler. Bibliographic coupling between scientfic papers. *American Documentation* 14:10-25, 1963
- J. Kandola, J. Shawe-Taylor, and N. Cristianini. Learning semantic similarity. In *Proc. NIPS* 15, 2003.
- H. Small. Co-citation in the scientific literature: a new meature of the relationship between two documents. *J. Am. Soc. Information Science*, 24:265-269, 1973.
- S. White and P. Smyth. Algorithms for estimating relative importance in networks. In *Proc. ACM SIGKDD*, 2003.