A Survey of Eigenvector Methods for Web Information Retrieval

Wenjun Lang

October 18, 2017

Information Retrieval

- Small Document Information Retrieval.
- Web information retrieval.

Small Document Information Retrieval

Jane Smith

Address Line 1 Address Line 2, City, State Zip * (212) 256-1414 * jane.smith@gmail.com

CAREER OBJECTIVE

Administrative Assistant with 6+ years of experience working directly for the President of 3M Inc., a Fortune 500 company. Possesses impeccable written and verbal communication skills and excellent interpersonal skills.

CORE COMPETENCIES

- Customer Service
- Cost Efficient
- Detailed and Organized
- Supplier Relationship

PROFESSIONAL EXPERIENCE

3M INC., New York, NY

Administrative Assistant, Apr 2006 – present

- Read and analyze incoming memos, submissions, and reports to determine their significance and plan their distribution.
- Conduct research, compile data, and prepare papers for consideration and presentation by executives, committees and boards of directors.
- Coordinate and direct office services, such as records, departmental finances, budget preparation, personnel issues, and housekeeping, to aid executives.
- Prepare invoices, reports, memos, letters, financial statements and other documents, using word processing, spreadsheet, database, or presentation software.

FLORIDA DEPARTMENT OF SOCIAL SERVICES, Orlando, FL

Rehabilitation Counselor, Aug 2004 – May 2006

- Confer with clients to discuss their options and goals so that rehabilitation programs and plans for accessing needed services can be developed.
- Prepare and maintain records and case files, including documentation such as clients' personal and eligibility information, services provided, narratives of client contacts, and relevant correspondence.
- Develop and maintain relationships with community referral sources, such as schools and community groups.
- Analyze information from interviews, educational and medical records, consultation with other professionals, and diagnostic evaluations to assess clients' abilities, needs, and eligibility for services.

EDUCATION

FLORIDA STATE UNIVERSITY, Orlando, FL

Bachelor of Art in English, May 2004

- GPA: 3.3/4.0
- Published in school's newspaper editorial
- Summer Internship for the New York Times

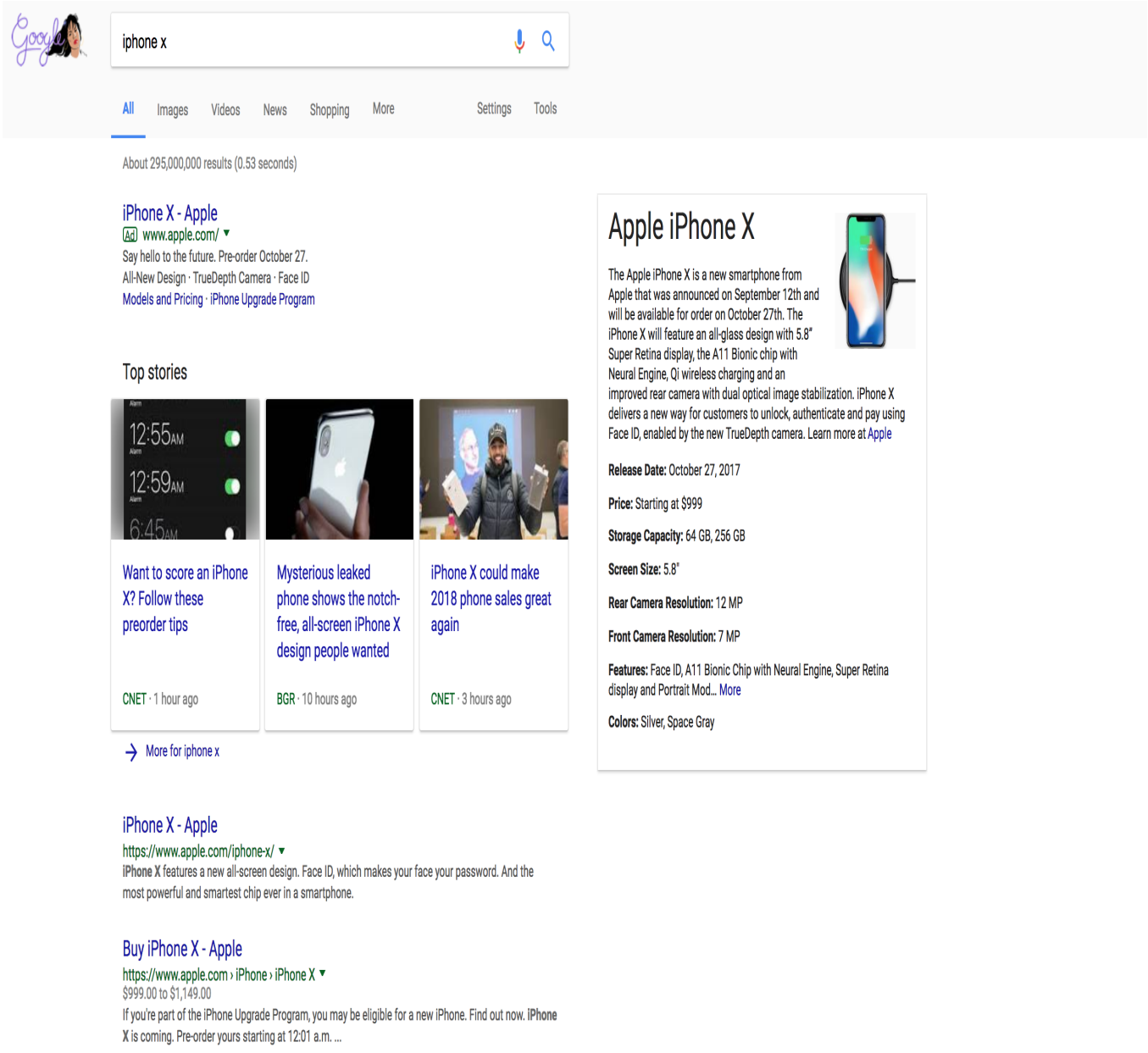
ADDITIONAL SKILLS

- Proficient in Microsoft Office and Adobe Illustrator CS5
- Bilingual Spanish and English
- Certified CPR and First Aid

AWARDS AND HONORS

- Employee of the Month for 3 consecutive months in H&M
- Won the "Writer's Digest" 2002 Award
- Awarded an employee travel award due to "Performance Excellence" 2 years in a row through 3M Inc.

Web Information Retrieval



Small Document Information Retrieval

Latent semantic indexing(LSI)

- Derived from SVD.
- Capture latent semantic associations.
- Cluster documents and terms into concepts.

LSI on web information retrieval

- The computation and storage of SVD is costly.
- Susceptible to redundant documents, broken links, and some poor quality documents.

Web Information Retrieval

Goals

- Impervious to redundant documents, broken links, and some poor quality documents.
- Impervious to spamming.
- High priority on accuracy and speed.

Methods

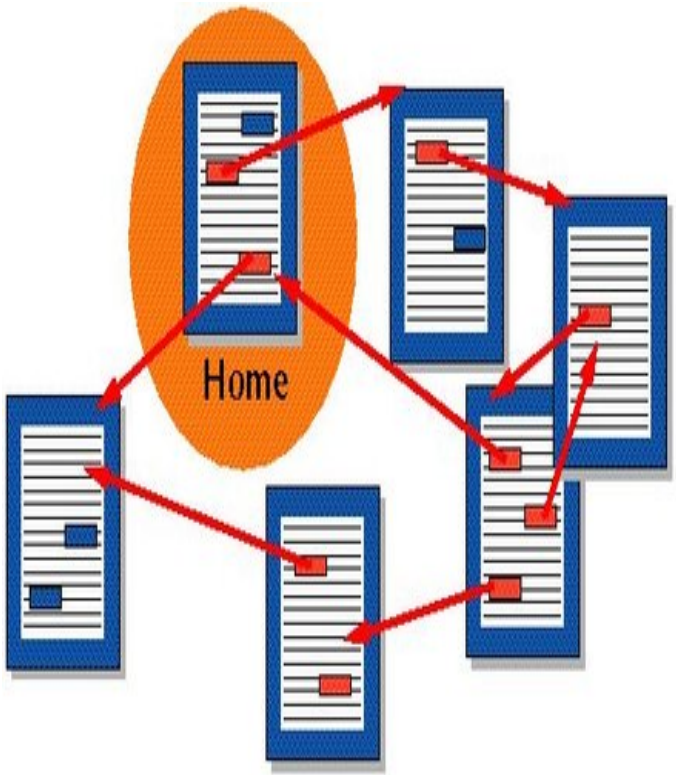
- Hypertext Induced Topic Search(HITS)
- PageRank
- Stochastic approach for link structure analysis(SALSA)

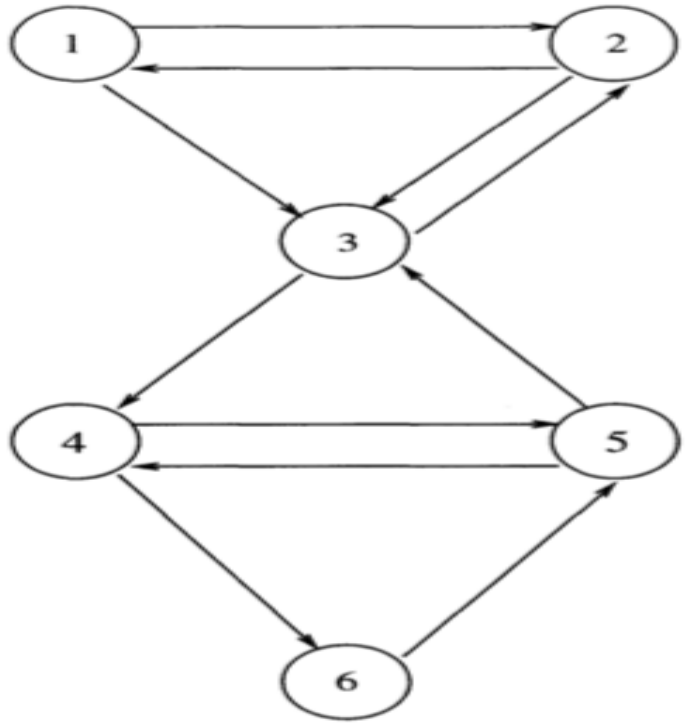
Similarity

- Makes extensive use of Web's unique hyperlink structure.

Hyperlink structure

- Several documents being connected by hyperlinks.

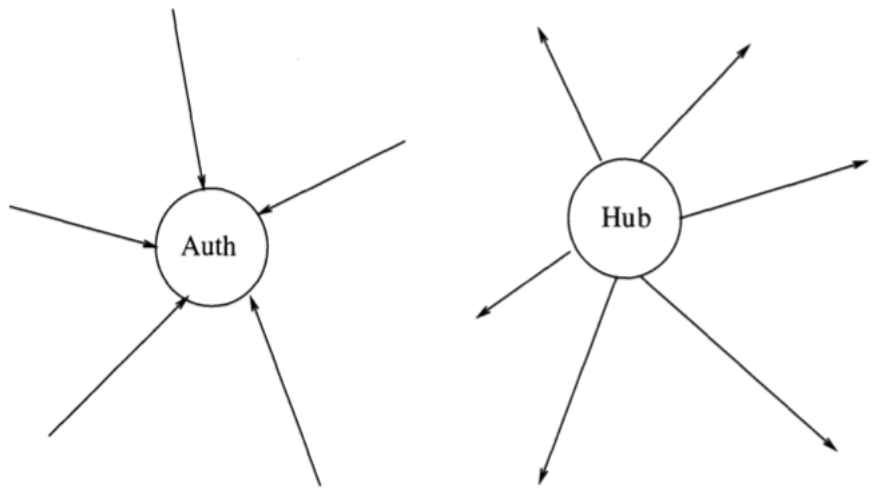




Hypertext Induced Topic Search(HITS)

Authorities and Hubs

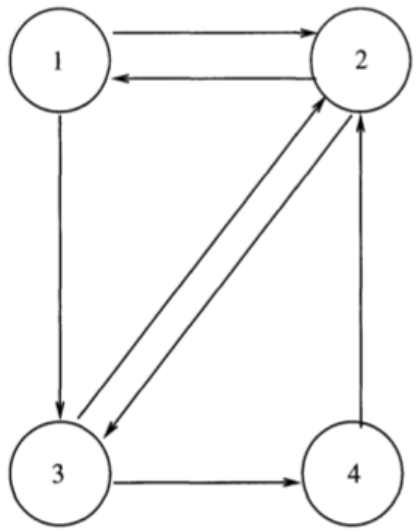
- An authority is a document with several inlinks.
- A hub is a document with sevaral outlinks.
- A document(webpage) can be both an authority and a hub.



HITS

Ideas

- Good authorities are pointed to by good hubs.
- Good Hubs point to good authorities.
- Provide an authority score and a hub score for each webpage.



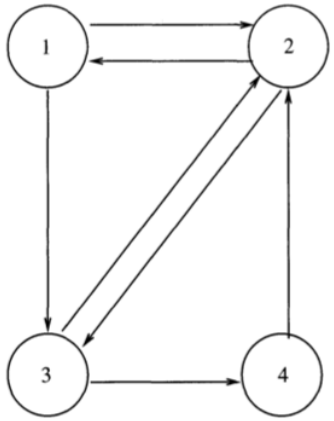
HITS

Computation of authority score and hub score

- Authority score for node i: x_i. Hub score for node i: y_i.
- Assign an initial score $x_i^{(0)}$ and $y_i^{(0)}$ and iteratively solve:

$$x_i^{(k)} = \sum_{j:e_{ji} \in E} y_j^{(k-1)} \quad y_i^{(k)} = \sum_{j:e_{ij} \in E} x_j^{(k)}$$

HITS



$$\Longrightarrow \quad L = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

$$x^{(k)} = L^T y^{(k-1)} \quad y^{(k)} = L x^{(k)}$$

$\Longrightarrow$

$$x^{(k)} = L^T L X^{(k-1)}$$

$$y^{(k)} = L L^T y^{(k-1)}$$

HITS

$$x^{(k)} = L^T L X^{(k-1)}$$
$$y^{(k)} = L L^T y^{(k-1)}$$

- $L^T L$: authority matrix. LL^T: hub matrix
- Convergence: with normalization, $x^{(k)}$ and $y^{(k)}$ always converges to the dominant eigenvectors of $L^T L$ and LL^T, respectively.
- An issue: different choice of initial vector $x^{(0)}$ and $y^{(0)}$ may result in different limiting vectors. For example:

$$x^{(0)} = (1/4, 1/4, 1/4, 1/4)^T \longrightarrow x^{(\infty)} = (1/3, 1/3, 1/3, 1/3)$$
$$x^{(0)} = (1/4, 1/8, 1/8, 1/2)^T \longrightarrow x^{(\infty)} = (1/2, 1/4, 1/4, 0)$$

- The iterative algorithm can be viewed as an application of using power method to compute the dominant eigenvectors.

Strengths and weaknesses of HITS

Strengths

- Provide dual rankings(authority rankings and hub rankings) for each webpage.
- Cast the overall web information retrieval problem as a small problem.

weaknesses

- Query dependence: The hub score and authority score are derived from a local neighborhood graph for each query.
- Susceptible to spamming.
- Topic drift problem.

PageRank

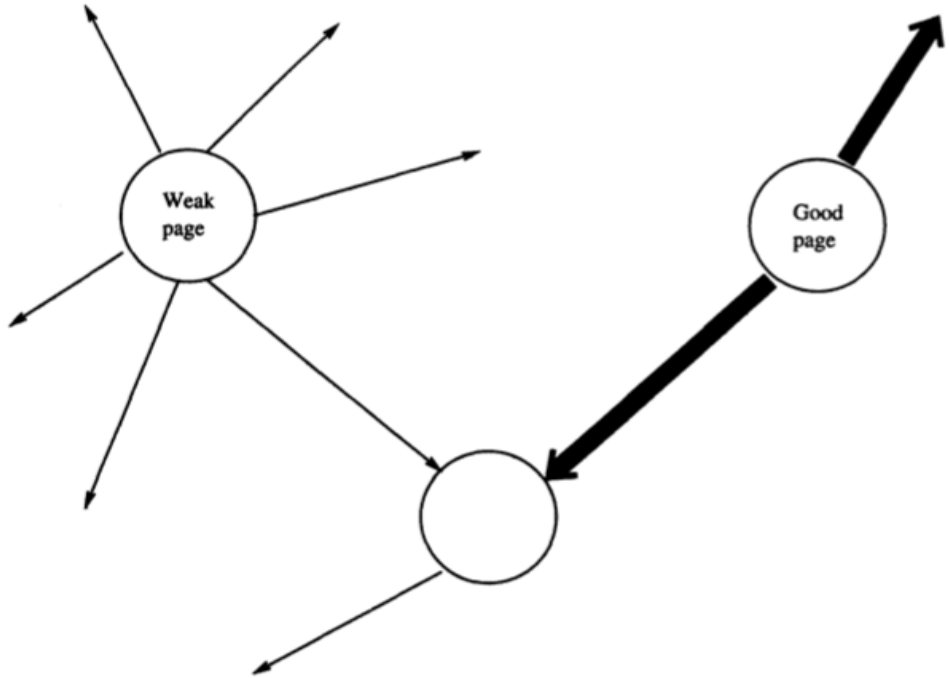
Background

- Created by google founders Larry Page and Sergey Brin.
- Provided basis for google search tools.

Idea

- Provide an importance score(ranking) for each page on the web.
- The importance score is determined by "votes" from other pages.

PageRank



- Votes from important sites should carry more weight than votes from less important sites.
- The significance of a vote from any source should be scaled by the number of sites the source is voting to.

PageRank

- PageRank importance score of a given page p is defined as:

$$r(P) = \sum_{Q \in B_p} \frac{r(Q)}{|Q|}$$

where $B_p = \{\text{all pages pointing to p}\}$, $|Q|$ = number of out links from Q

- If we have n webpages $P_1, P_2, \cdots, P_n$, initialize $r^{(0)}(P_i) = \frac{1}{n}$ and iteratively compute $r(P_i)$ by:

$$r^{(k)}(P) = \sum_{Q \in B_{p_i}} \frac{r^{(k-1)}(Q)}{|Q|}$$

PageRank

Set

$$\pi^{(k)^T} = (r^{(k)}(P_1), r^{(k)}(P_2), \cdots, r^{(k)}(P_n)),$$

then solving PageRank importance score is to iteratively compute:

$$\pi^{(k)^T} = \pi^{(k-1)^T} P \tag{1}$$

where

$$(P)_{ij} = \begin{cases} 1/|P_i| & if\ P_i\ links\ to\ P_j, \\ 0 & otherwise \end{cases}$$

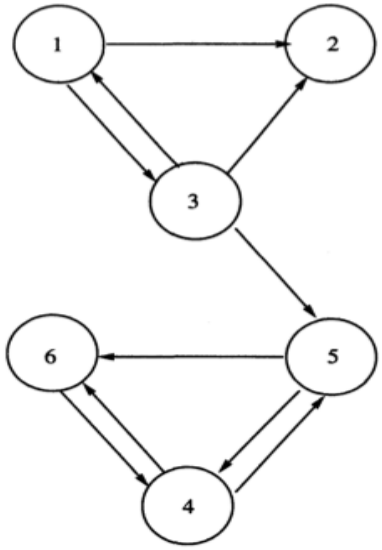
PageRank

- Under some weak assumptions, the PageRank iteration (1) represents a random walk on the graph.
- $\pi^{(k)}$ converges to the stationary distribution of the random walk.
- Google intuitively characterizes the PageRank value of each site as the long-run proportion of time spent at that site by a Web surfer eternally clicking on links at random.

Adjustment of P in practice

- Replace zero row with $\mathbb{1}^T/n$ to make P a stochastic matrix.
- Adjust P to be irreducible to assure the unique convergence the random walk.

A PageRank Example



$$\Longrightarrow \quad P = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

$$\bar{P} = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \quad \bar{\bar{P}} = \begin{bmatrix} \frac{1}{60} & \frac{7}{15} & \frac{7}{15} & \frac{1}{60} & \frac{1}{60} & \frac{1}{60} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{19}{60} & \frac{19}{60} & \frac{1}{60} & \frac{1}{60} & \frac{19}{60} & \frac{1}{60} \\ \frac{1}{60} & \frac{1}{60} & \frac{1}{60} & \frac{1}{60} & \frac{7}{15} & \frac{7}{15} \\ \frac{1}{60} & \frac{1}{60} & \frac{1}{60} & \frac{7}{15} & \frac{1}{60} & \frac{7}{15} \\ \frac{1}{60} & \frac{1}{60} & \frac{1}{60} & \frac{11}{12} & \frac{1}{60} & \frac{1}{60} \end{bmatrix}$$

A PageRank Example

The stationary vector(PageRank vector) is:

$$\pi^T = (0.03721, 0.05396, 0.04151, 0.3751, 0.206, 0.2862)$$

- A query: term1 and term2

  term1 $\rightarrow$ doc 1, doc 4, doc 6

  term2 $\rightarrow$ doc 1, doc 3
- The relevancy set for this query on term1 and term2 is {1,3,4,6}
- Document 4 is most important, followed by document 6, 3, and 1.

Strengths and Weaknesses of PageRank

Strengths

- Query Independent. PageRank is a global measure.
- Imperviousness to spamming.

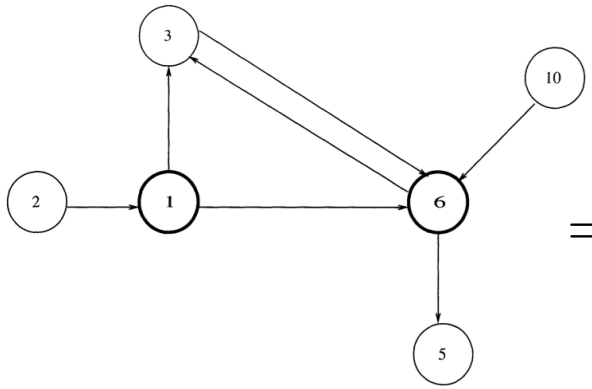
Weaknesses

- Topic drift problem.

The Stochastic Approach for Link Structure Analysis(SALSA)

- A combination of ideas from HITS and PageRank
- Both hub and authority scores are created(like HITS).
- Scores are created through Markov chains(like PageRank).
- A hub Markov Chain with transition probability matrix H and an authority Markov Chain with transition probability matrix A.

An Example of SALSA



$$\Longrightarrow \quad L = \begin{bmatrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$L_r = \begin{bmatrix} 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \quad L_c = \begin{bmatrix} 0 & 0 & \frac{1}{2} & 0 & \frac{1}{3} & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 0 \end{bmatrix}$$

An Example of SALSA

$$L_r L_c^T = \begin{bmatrix} \frac{5}{12} & 0 & \frac{2}{12} & 0 & \frac{3}{12} & \frac{2}{12} \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} \end{bmatrix} \quad L_c^T L_r = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{6} & 0 & \frac{5}{6} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Take the nonzero columns and rows of $L_r L_c^T$ to form H. Take the nonzero columns and rows of $L_c L_r^T$ to form A

$$\Longrightarrow \quad H = \begin{bmatrix} \frac{5}{12} & 0 & \frac{2}{12} & \frac{3}{12} & \frac{2}{12} \\ 0 & 1 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ \frac{1}{4} & 0 & 0 & \frac{3}{4} & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{3} \end{bmatrix} \quad A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{6} & 0 & \frac{5}{6} \end{bmatrix}$$

Strengths and Weaknesses for SALSA

Strengths

- Provide dual rankings(Hub scores and authority scores)
- Not affected by topic drift problem.
- Less susceptible to spamming compared to HITS.

weaknesses

- Convergence issue.
- Query dependence.