Welcome to the class of
Web and Information Retrieval !

Min Zhang

z-m@tsinghua.edu.cn

Coffee Time
The Sixth Sense

By 费伦

Web Search Technologies (II)

Min Zhang

z-m@tsinghua.edu.cn

III. Key Techniques of Web IR

- Using web specific page features
- **Link-based analysis**
- User click analysis

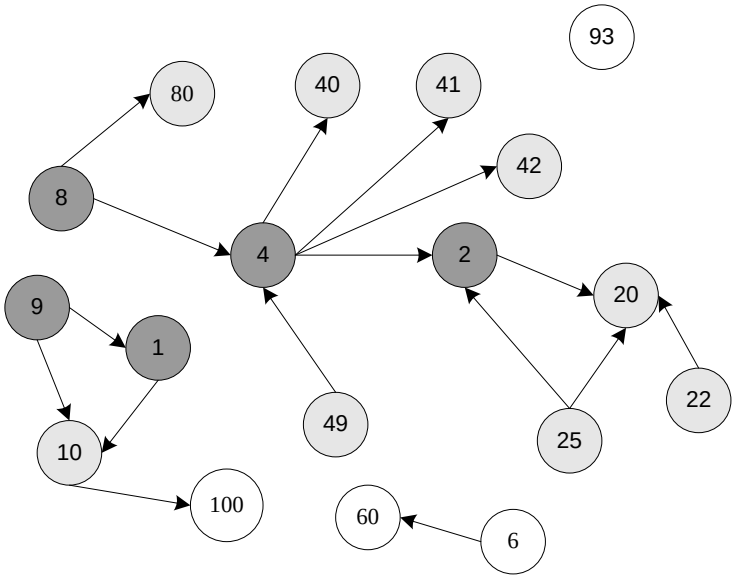
Web IR Techniques: Link-based IR

- Counting node degree
- PageRank
- HITS (authorities and hubs)
- TrustRank
- Spreading activation
- Anchor text

Web IR Techniques: Link-based IR

- Counting node degree
- PageRank
- HITS (authorities and hubs)
- TrustRank
- Spreading activation
- Anchor text

Counting node degree



- Index node:

  Whose out-degree is significantly larger than the average out-degree.

- Reference node:

  Whose in-degree is significantly larger than the average in-degree

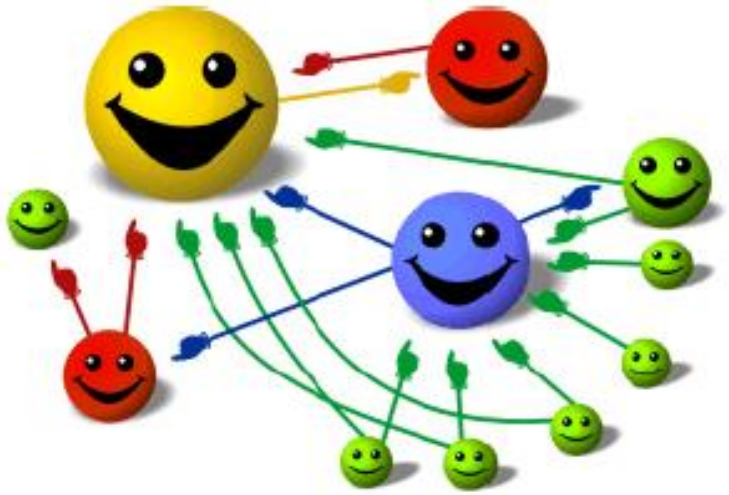
- Propose measures of centrality
  - Based on node-to-node distances in the link structure graph
  - Rank = in-degree + out-degree
- A "counting" (and mostly "directionless") notion of WWW link structure

Web IR Techniques: Link-based IR

- Counting node degree
- PageRank
- HITS (authorities and hubs)
- TrustRank
- Spreading activation
- Anchor text

PageRank -- background

- Measure the importance of the Web pages
- Recommendation assumption:
  - A has a link that points to B, then the author of A recommend page B.
  - The more recommended, the better.
  - Good recommender gives better recommendations.



PageRank – basic idea

- Sergey Brin and Larry Page 1998 (Google)
- A model of user behavior
- A "Random surfer" – random walk model:
  - Randomly given a webpage
  - Keep clicking
  - Never hitting "back"
  - Eventually get starts on another random page
- Simulate the user's navigation procedure with Markov chain
  - $t \rightarrow \infty$, the probability of each page that the user stays: PR

PageRank -- computation

- Computes a document's score based on the scores of documents that link to it.

$$PR(A) = \frac{d}{N} + (1-d)\left(\frac{PR(T_1)}{L(T_1)} + \cdots\cdots + \frac{PR(T_n)}{L(T_n)}\right)$$

T_i links to A, $L(T_i)$: out-degree of page T_i N: # of pages

d: probability of re-start (i.e. not following links), (0,1) ~ generally 0.15

- a probability distribution over web pages, Sum of all PR(A) = 1
- A simple iterative algorithm
- Corresponding to the principal eigenvector of the normalized link matrix of the web

L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998

PageRank – computation (cont.)

- Dealing with pages with NO out-link

In each loop:

- $I(A) = \frac{d}{N}$
- For each page T_i
  - if outdegree(T_i) > 0,
    - *For each A that* $T_i \rightarrow$ A , $I(A) = I(A) + (1-d)\frac{PR(T_i)}{L(T_i)}$
  - else (i.e. outdegree(T_i) = 0)
    - *For each A,* $I(A) = I(A) + (1-d)\frac{PR(T_i)}{N}$
- $PR(A) = I(A)$

Improvements on PageRank

- On speed-up of the computation
- On refinement and enrichment of the model
  - PageRank – nothing about content
    - Topic-sensitive PageRank, Query-dependent PageRank
  - Block-based PageRank, ……
- Others
  - Modifying the personalized vector
  - Introducing inter-domain and intra-domain link weights
    - HostRank, SiteRank, ……
  - TrustRank, anti-TrustRank

Web IR Techniques: Link-based IR

- Counting node degree
- PageRank
- HITS
- TrustRank
- Spreading activation
- Anchor text

HITS: Hypertext-Induced Topic Search

- John Kleinberg 1998 (Cornell)
  - *"Authoritative Sources in a Hyperlinked Environment"*
  - 6520 citations
- ONR Young Investigator Award
- A MacArthur Foundation Fellowship
- A Packard Foundation Fellowship
- A Sloan Foundation Fellowship
- Member of the National Academy of Engineering
- And the American Academy of Arts and Sciences

HITS: Hypertext-Induced Topic Search

- 3 types of queries
  - Specific queries
    - “Does Netscape support the JDK 1.1 code-signing API?"

far too large, need authoritative or definitive one

  - Broad-topic queries
    - “Find information about the Java programming language."
  - Similar-page queries
    - “Find pages ‘similar' to java.sun.com."

HITS: Hypertext-Induced Topic Search

- Authority page
  - Good sources of **content**
  - Large **in-degree**, e.g.
  - www.amazon.com, www.wikipedia.com , zhidao.baidu.com
- Hub pages
  - good sources of **links**
  - Pull together authorities on a given topic
  - Throw out unrelated pages of large in-degree
  - E.g. **dir.yahoo.com**
- Relationship of authorities and hubs
  - **Mutually reinforcing relationship**
  - A good hub points to many good authorities
  - A good authority is pointed by many good hubs

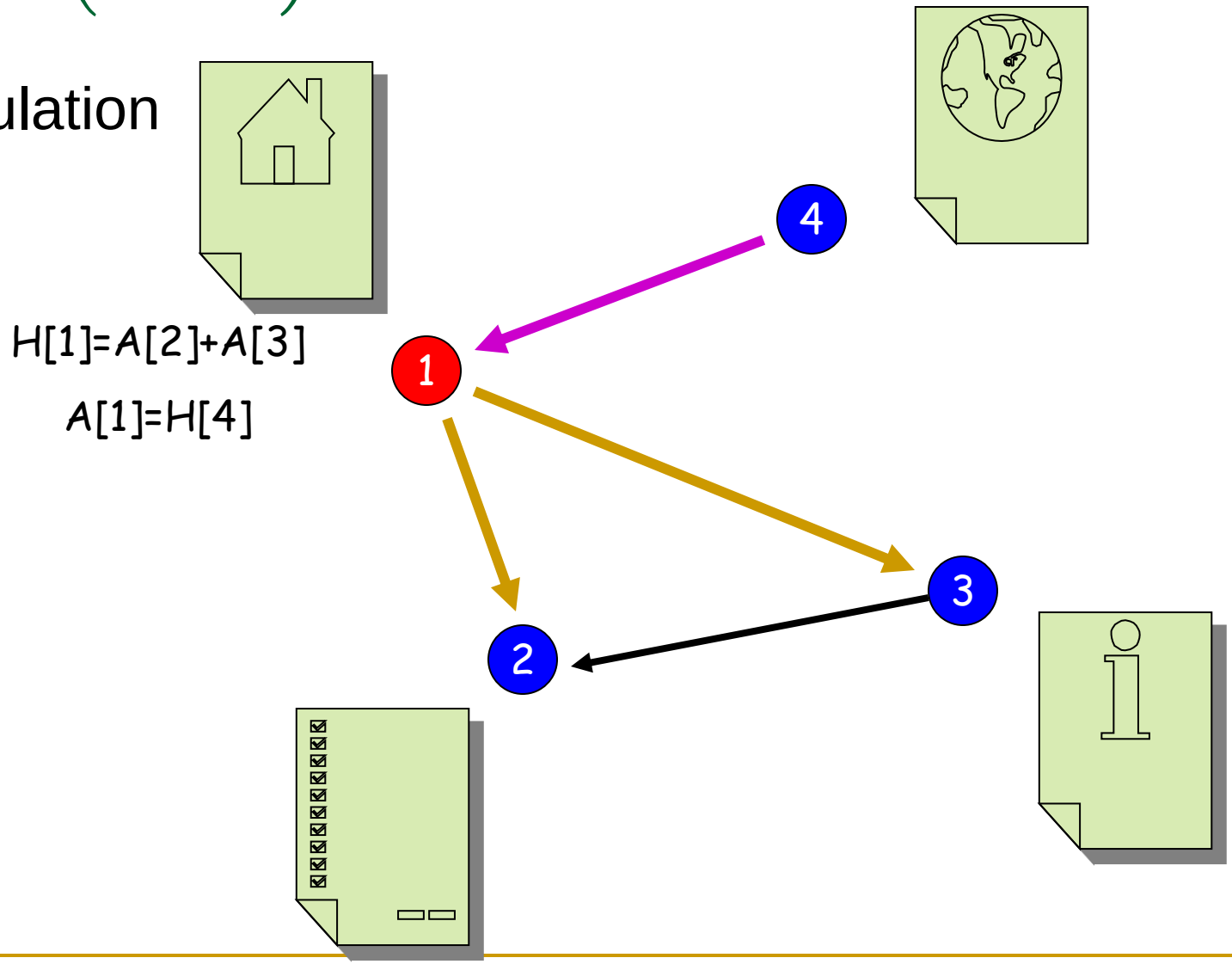
HITS

- A good *Base Set* R_σ is quite important
  - *a)* R_σ is relatively small.
  - *b)* R_σ is rich in relevant pages.
  - *c)* R_σ contains most (or many) of the strongest authorities.
- A root set
  - Collect the *t* highest-ranked pages for the query σ (e.g. t=200)
    - Good for a) and b), not enough for c)
- Expand root set to base set
  - Add all pages that the root set link to
  - Add 50 pages that link to the root set
  - Remove in-site links

Then its size is generally 1000-5000

HITS (cont.)

- Calculation



HITS (cont.)

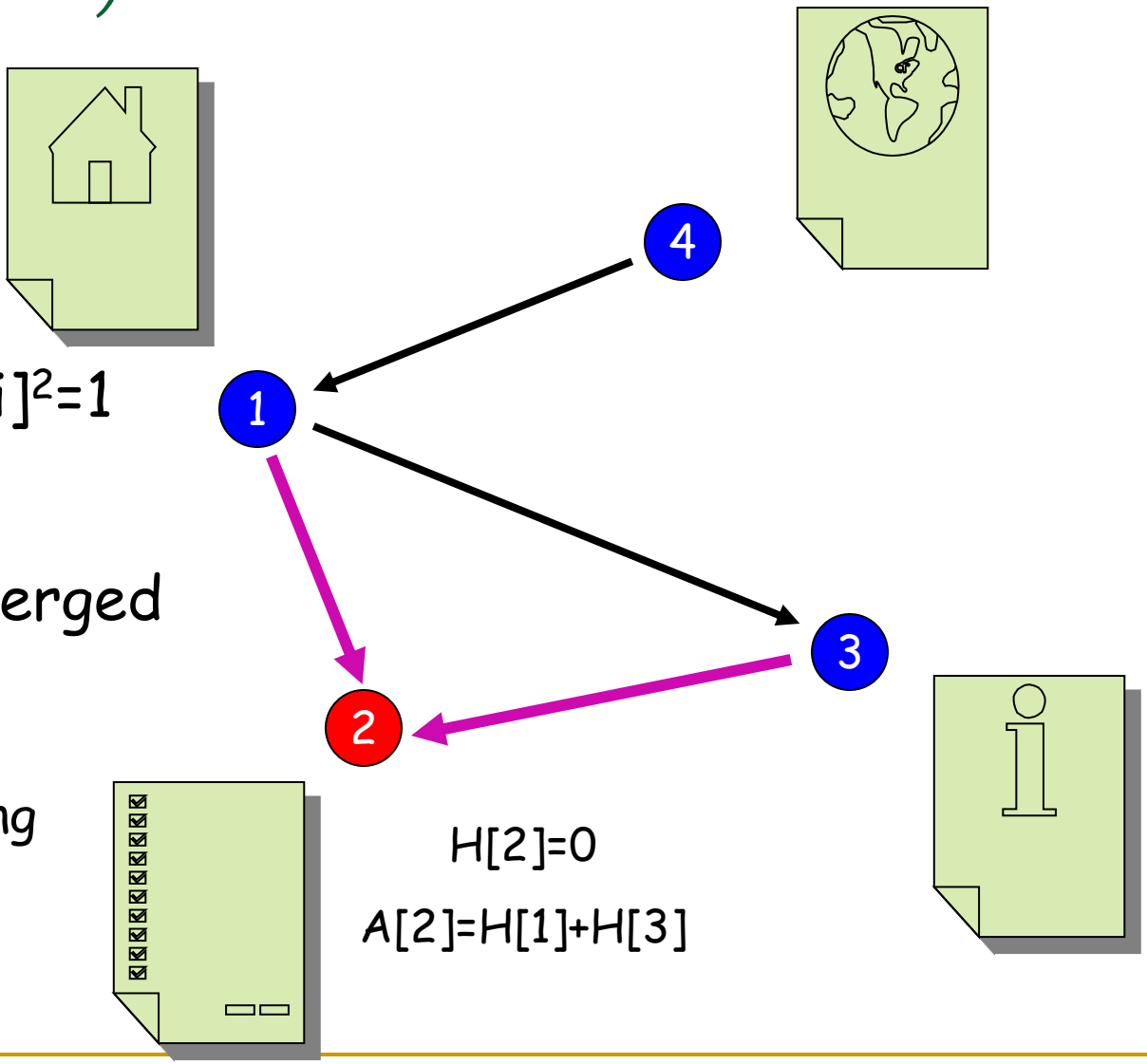
- Calculation

Normalization:

$\Sigma A[i]^2=1, \quad \Sigma H[i]^2=1$

Repeat until converged

Generally be used within the searching result documents -- topic specific / query dependent



Improvements on HITS

- Problems: can not work well in all cases
  - Mutually reinforcing relationships between hosts
  - Automatically generated links
  - Non-relevant nodes (topic drifting problem)

    e.g. "mango fruit" → "fruit"
- Some ideas
  - **Hosts problem** solution: A(n) → A(n)/k H(n) → H(n)/l
  - 2 basic approach to tackle **topic drift**
    - Elimination non-relevant nodes from the graph
    - Regulating the influence of a node based on its relevance

Web IR Techniques: Link-based IR

- Counting node degree
- PageRank
- HITS (authorities and hubs)
- TrustRank
- Spreading activation
- Anchor text

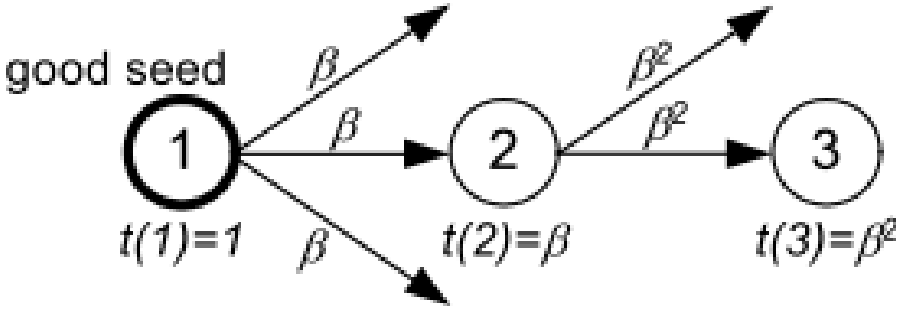
TrustRank – link-based spam page detection

- Trust Rank
  - Rationale: The approximate isolation of the good set
    - Good pages seldom point to bad ones.
  - Basic Idea: in-link based trust propagation
  - Select seed set
    - With high Invert PageRank (outlink-based PR)
      - To spread the trust score quickly
    - (And) high PageRank
    - Initial score t* = d
      - *d*: *normalize static score distribution, seed:1 others: 0*

Trust Attenuation

Type I:

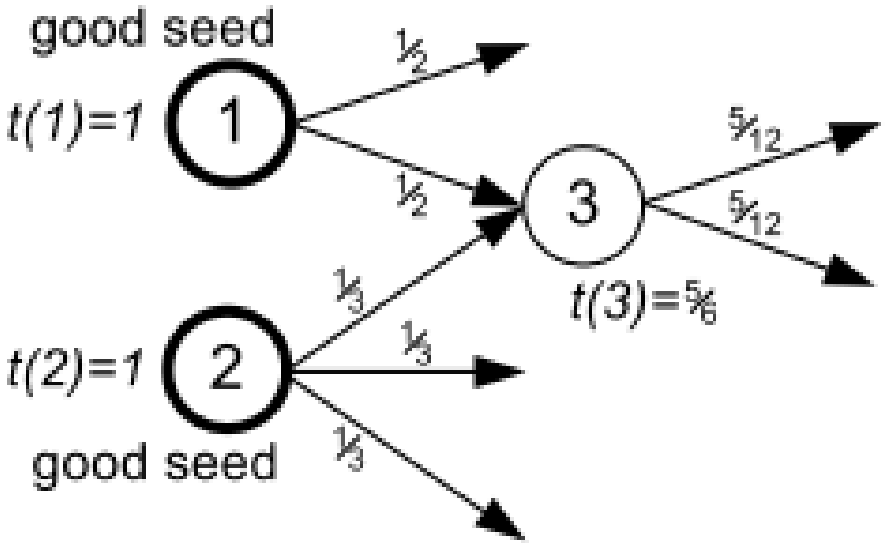
- Trust dampening



$0 < \beta < 1$

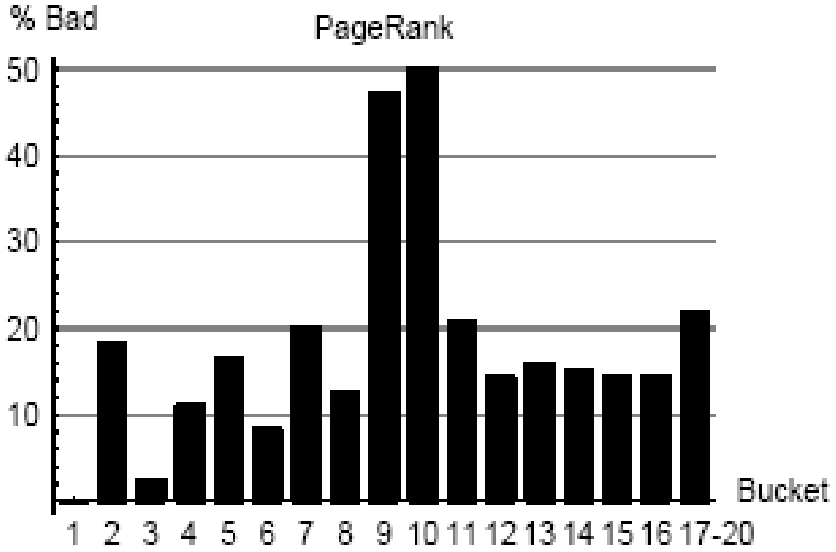
Type II:

- Trust splitting



Link Based Detection

- Trust Rank Result
  - Precision and recall, PageRank vs. TrustRank



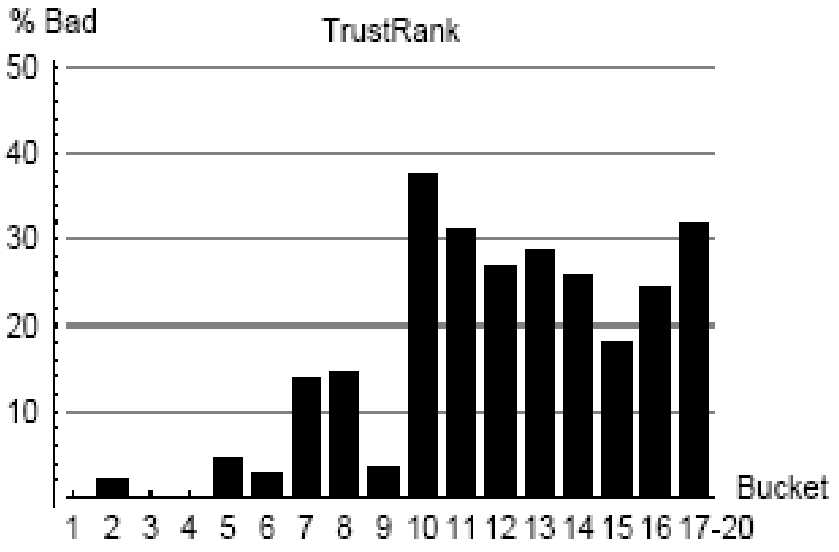


Figure 10: Bad sites in PageRank and TrustRank buckets.

Rank pages with scores. Rank 1: highest score.
Bucket setting: The sum of scores in each bucket is equal.

Z. Gyongyi, et al. Combating web spam with trustrank. In *VLDB '04, 576–587, 2004.*

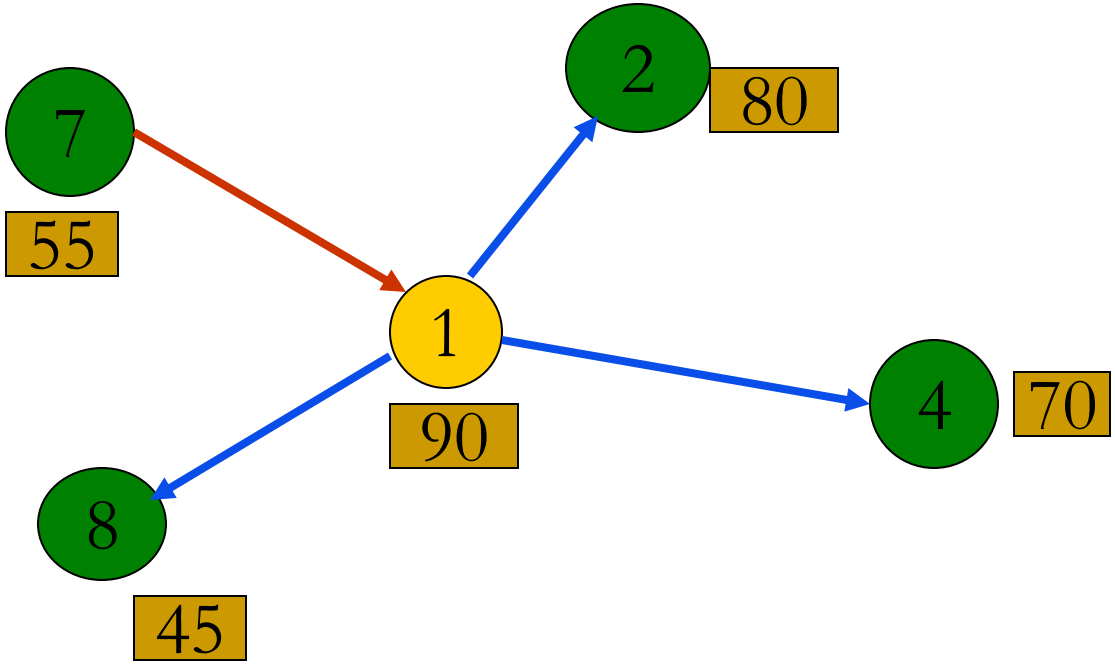
Web IR Techniques: Link-based IR

- Counting node degree
- PageRank
- HITS (authorities and hubs)
- TrustRank
- Spreading activation
- Anchor text

Spreading Activation

- For result re-ranking:

When the initial relevance scores have been generated for result pages, using SA to re-rank the results.



$$RSV(D_1) = 90 + \lambda_1 * (80 + 45 + 70) + \lambda_2 * 55$$

Web IR Techniques: Link-based IR

- Counting node degree
- PageRank
- HITS (authorities and hubs)
- TrustRank
- Spreading activation
- Anchor text


Google - Microsoft Internet Explorer
文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)
后退 搜索 收藏夹
地址(D) http://www.google.cn/
转到 链接
网页 图片 地图 资讯 视频 博客 更多 登录
Google
 Google 大全
Google 搜索 手气不错
使用偏好
语言工具
所有网页 中文网页 简体中文网页 中国的网页
广告计划 - Google 大全 - Google.com in English
©2008 Google - ICP证合字B2-20070004号
http://www.google.cn/intl/zh-CN/about.html
Internet

On the use of anchor text

- Three ways
  - Use anchor text as the complementary document information
    - $Document_i' = Document_i \cup Anchor_i$
      - Where $Anchor_i = \{anchor\_text_{ij} \mid \text{where doc } j \text{ has link to doc } i\}$
    - $Final\_score_i = ranking\ score\ on\ Document_i'$
  - Use anchor text search result to re-rank
    - $Final\_score = f(Doc\_search\_score, Anchor\_search\_score)$
  - Only use anchor text $Final\_score = Anchor\_search\_score$
- An observation
  - Only use anchor text is significantly better than use the original webpage on site finding tasks

    -- by Nick Craswell, David Hawking and Stephen Robertson, SIGIR01

Improvements on the use of anchor text

- Filtering anchor text noise
- Use anchor text for finding web synonyms
- For query expansion
- For abbreviations
- Anchor text weighting with click information
- ……

III. Key Techniques of Web IR

- Using web specific page features
- Link-based analysis
- User click analysis

To be continue…

References

- Broder, A. A taxonomy of Web Search. SIGIR Forum 36(2), 2002
- Zoltán Gyongyi and Hector Garcia-Molina. Web spam taxonomy. In First International Workshop on Adversarial Information Retrieval on the Web, 2005.
- "Structural analysis of hypertext: Identifying hierarchies and useful metrics", R. Botafogo, et al, ACM Trans. Inf. Sys. 10(1992), pp.142-180
- "WebQuery: searching and visualizing the Web through connectivity", J. Carriere, R. Kazman, www6, 1997
- "The Anatomy of a Large-Scale Hypertextual Web Search Engine", S. Brin, L. Page, www7, 1998
- "Authoritative Sources in a Hyperlinked Environment",Jon M. Kleinberg, Proc. of the 9th annual ACM-SIAM symposium on Discrete Algorithms, pp 668-677, 1997
- L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998

- “Searching for information in a hypertext medical handbook”, M. E. Frisse, Communications of the ACM, 31(7), pp880-886
- “Experiments in Topic Distillation”, S. Chakrabarti, B. Dom, D. Gibson, S.R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins, ACM SIGIR workshop on hypertext information retrieval on the web, 1998
- “Improved Algorithms for Topic Distillation in a Hyperlinked Environment”, K. Bharat, M. R. Henzinger, SIGIR 1998
- “Effective site finding using link anchor information”, N. Craswell, D. Hawking, S. E. Robertson, SIGIR 2001
- Does “Authority Mean Quality? Predicting Expert Quality Ratings of Web Documents”, B. Amento, L. Terveen, W. Hill, SIGIR 2000
- Yiqun Liu, *Min Zhang,* Liyun Ru, Shaoping Ma. Data Cleansing for Web Information Retrieval using Query Independent Features. Journal of the American Society for Information Science and Technology (JASIST), Volume 58, Issue 12, Pages 1884-1898, 2007
- 张敏，高剑峰，马少平，基于链接描述文本及其上下文的Web信息检索，计算机研究与发展，Vol. 41, No.1, pp221~226, 2004.

- Daniel E. Rose, Danny Levinson, Understanding User Goals in Web Search, www2004.
- Ricardo Baeza-Yates, et al, The Intention Behind Web Queries, SPIRE 2006.
- T. Haveliwala. Efficient computation of pageRank. Technical Report 1999-31, 1999.
- T. H. Haveliwala. Topic-sensitive pagerank. In *WWW ' 02,* Honolulu, Hawaii, May 2002.
- M. Richardson and P. Domingos. The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank. In *Advances in Neural Information Processing Systems 14. 2002.*
- A. N. Langville and C. D. Meyer. Deeper inside pagerank. *Internet Mathematics, 1(3):335–400, 2004.*
- Z. Gyongyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In *VLDB '04, pages 576–587, 2004*.
- F. McSherry. A uniform approach to accelerated pagerank computation. In *WWW '05, pages 575–582,* USA, 2005.