# Модели претраге

# Шта је овде предмет разговора?

- ФОРМУЛАЦИЈА ПРОБЛЕМА
  - Претпостављамо да су нам дате колекције докумената или неке од јасних тема
    - пример:
      - еволуција, Београд ...
      - можда добијен слободан текст за претрагу
  - Да ли можемо да организујемо ове документе на неки начин?
    - PageRank нуди једну од солуција
    - HITS (Hypertext-Induced Topic Selection) је друга
- ТЕМА:
  - **PageRank**
  - **HITS**

# Рангирање

- **Циљ:** Одговори релевантни нашем упиту треба бити приказани у опадајућем поретку
  - **упит-независан:** Одредити суштинску вредност документа независно од актуелног упита
  - **упит-зависан:**Вредност је одређена само за поједине (одређене) упите
  - пример:
    - упит-независан: дужина, речник, публикација ...
    - упит-зависан: мера косинуса

# Неки критеријуми рангирања

- **Систем основних техника** (векторски терм модел, вероватносни модел...) – углавном упит-зависан
- **‘Ад–хок’ фактори** (локација/публикација података...) – углавном упит-независни
- **Људско бележење**
- **Повезивање база техникама**
  - упит-независно  PageRank
  - упит-зависно HITS

# Мотивација PageRank-а

- претпоставка:
  - линк од странице A до стране B, је препорука за страну B од аутора A (B је наследник A)
- Квалитет стране је повезан са улазном вреношћу
- рекурзија: Особина стране да је повезана и са:
  - улазним вредностима
  - и особинама страна које су линкови

# Дефиниција рангирања страна

- Замислимо бесконачно случајно сурфовање
  - увођење сурфовања је као случајна страна
  - следећи корак вожње обухвата
    - пробу случајног одабира web стране са вероватноћом d
    - пробу случајног одабира успешне конверзије стране са вероватноћом 1-d

# Рангирање страна

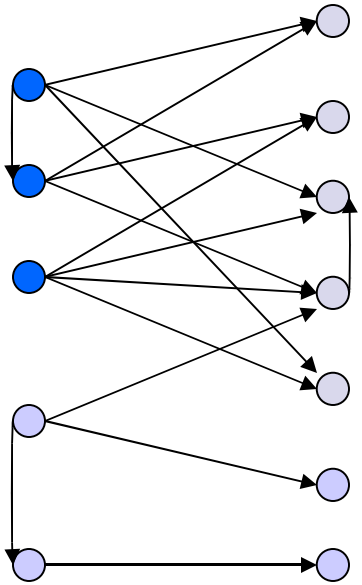
- Прелаз вероватносне матрице је:
  - dxU+(1-d)xA
  - где је U нормална расподела, а A је суседна матрица нормализације.
  - PageRank(u)=d/n+(1-d) ΣPageRank(v)/outdegree(v)
    - где је n број свих чворова у графу

# HITS - модел

- Рангирање докумената се врши у две класе
  1. **ауторитети** – странице које садрже корисне информације
  2. **везе** – странице које показују на ауторитете

# Преглед

Везе Ауторитети



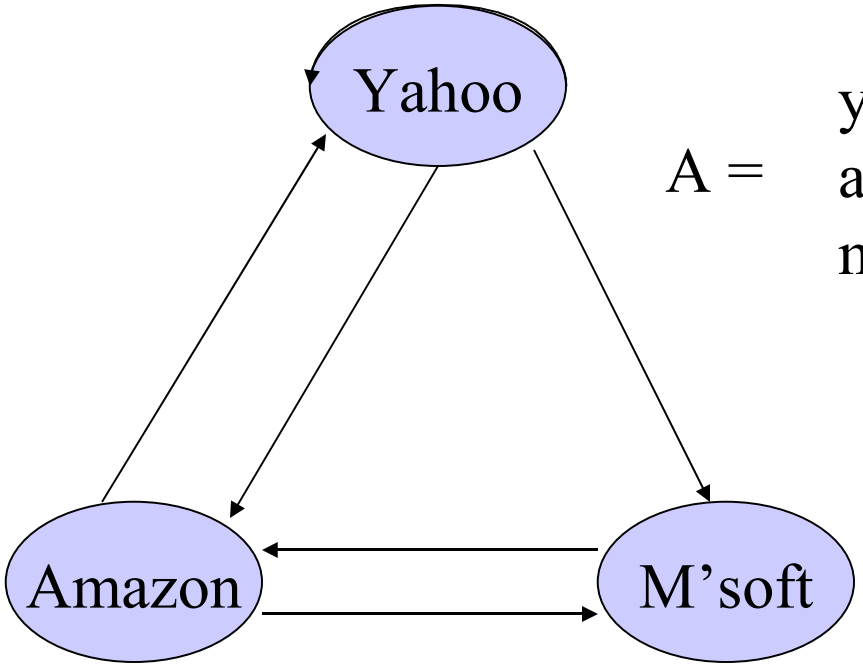
# Међусобна рекурзивна дефиниција

- добра веза показује на много добрих ауторитета
- добар ауторитет је показан од стране много добрих веза
- Модел користи 2 броја за сваки чвор
  - Веза број
  - Ауторитет број
- (приказујемо их као векторе h и a )

# Прелаз матрице A

- HITS користи матрицу A[i,j]=1 ако страна i показује на страну j, 0 иначе
- $A^T$ , транспонована матрица матрице A

# Пример




$$A = \begin{array}{c|ccc} & y & a & m \\ \hline y & 1 & 1 & 1 \\ a & 1 & 0 & 1 \\ m & 0 & 1 & 0 \end{array}$$

# Једначина веза и ауторитета

- Веза вредност стране P је пропорционална суми ауторитет броја странице које су њени линкови
  - $\mathbf{h} = \lambda A\mathbf{a}$
  - $\lambda$ је скалар
- Вредност ауторитета стране P је пропорционална суми вредности веза које показују на њу
  - $\mathbf{a} = \mu A^T\mathbf{h}$
  - $\mu$ је скалар

# Итеративни алгоритам

- Иницијализација **h** , **a** на јединице
- $\mathbf{h} = \mathbf{A}\mathbf{a}$
- $\mathbf{a} = \mathbf{A}^T\mathbf{h}$
- понављамо све док **h** , **a** конвергирају

# Пример

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \qquad A^T = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| a(yahoo) | = | 1 | 1 | 1 | 1 | . . . | 1 |
| a(amazon) | = | 1 | 1 | 4/5 | 0.75 | . . . | 0.732 |
| a(m’soft) | = | 1 | 1 | 1 | 1 | . . . | 1 |
| | | | | | | | |
| h(yahoo) | = | 1 | 1 | 1 | 1 | . . . | 1.000 |
| h(amazon) | = | 1 | 2/3 | 0.71 | 0.73 | . . . | 0.732 |
| h(m’soft) | = | 1 | 1/3 | 0.29 | 0.27 | . . . | 0.268 |

# Постојање и јединственост

$\mathbf{h} = \lambda A\mathbf{a}$

$\mathbf{a} = \mu A^T \mathbf{h}$

$\mathbf{h} = \lambda\mu A A^T \mathbf{h}$

$\mathbf{a} = \lambda\mu A^T A\ \mathbf{a}$

- Овај итеративни алгоритам конвергира векторима **h*** и **a*,** где су:
  - **h*** својствени вектор матрице $AA^T$
  - **a*** својствени вектор матрице $A^TA$

# Проблеми и решења

- Неки од путева су ‘погрешни’ тј. нису препоручљиви:
  - више путева од истог аутора
  - аутоматски генерисани
  - спамови...
  - **Решење:** дужина путева ограничава утицај
- Усмерење теме
  - Питање: +jaguar+car
  - резултат: странице о колима уопште
  - **Решење:** анализа контекста и додавање теме у резултат чворова

# Модификација HITS алгоритма

- Понављање све док $\overrightarrow{HUB}$ и $\overrightarrow{AUTH}$ конвергирају:
- Нормализација $\overrightarrow{HUB}$ и $\overrightarrow{AUTH}$
  - HUB[v] := Σ AUTH[ui] TopicScore[ui] weight[v,ui]
    - за све ui  са путем(v, ui)
  - AUTH[v] := Σ HUB[wi] TopicScore[wi] weight[wi,v]
    - за све wi са путем(wi, v)