Анализ ссылок

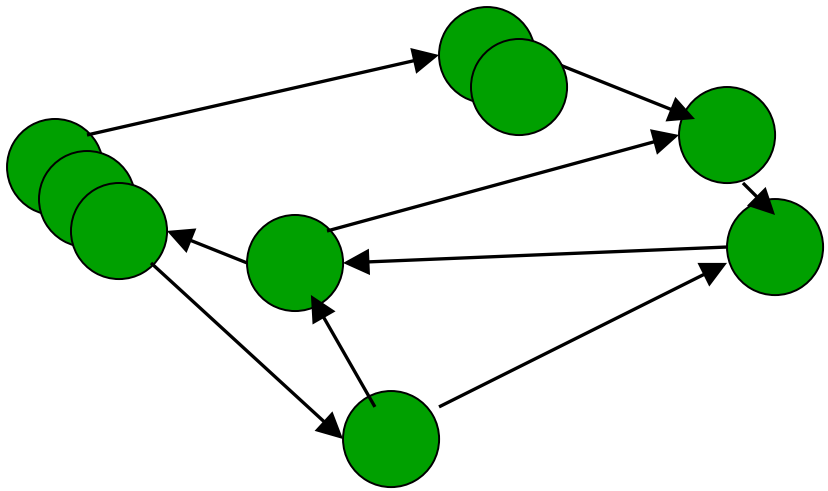
Тест

$$\cos(\vec{q},\vec{d}) = \frac{\vec{q}\cdot\vec{d}}{|\vec{q}||\vec{d}|} = \frac{\vec{q}}{|\vec{q}|}\cdot\frac{\vec{d}}{|\vec{d}|} = \frac{\sum q_i d_i}{\sqrt{\sum q_i^2}\sqrt{\sum d_i^2}}$$

- Если при подсчете косинусной меры сходства запроса с документом забыть разделить на длину запроса, то
- a. Вперед продвинутся документы, в которых запрос больше совпадает с заголовком
- b. Порядок выдачи документов не изменится
- c. Вперед продвинутся более короткие документы
- d. Вперед продвинутся более длинные документы

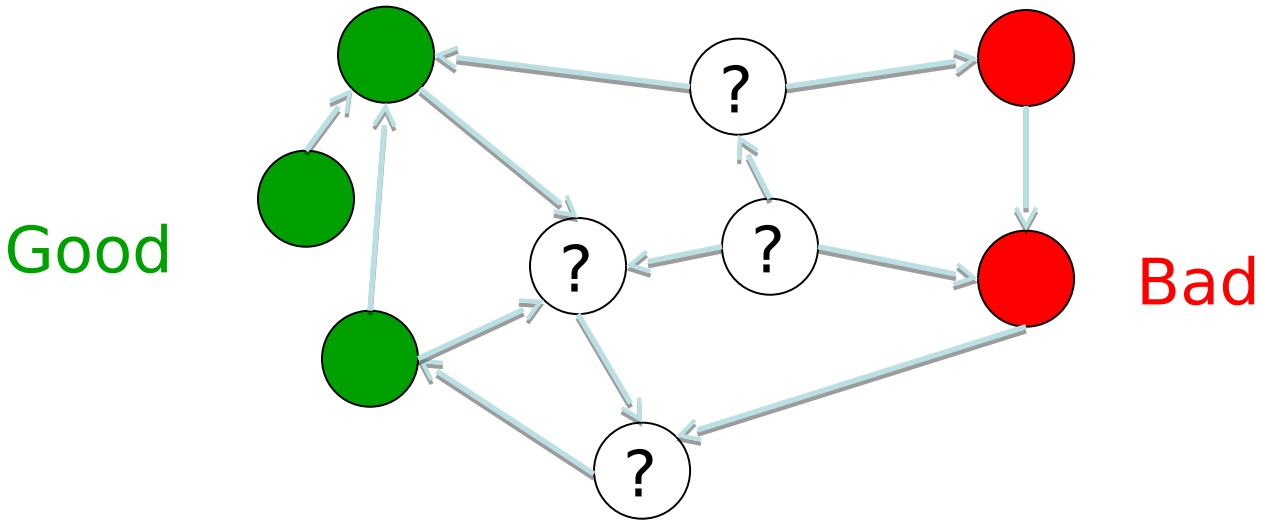
Гипертекст и ссылки

- Что есть кроме содержания документов
  - Гиперссылки между документами
- Вопросы
  - Могут ли ссылки продемонстрировать авторитетность страниц? Полезны ли они для ранжирования?
- Применение
  - Интернет
  - Email
  - Социальные сети



Ссылки – везде!

- Мощные источники информации об авторитетности
  - Почтовый спам – какие почтовые адреса – спамерские?
  - Качество хостов – какие хосты – «плохие»?
  - Логи телефонных звонков
- Good, Bad и Unknown – алгоритмы label propagation



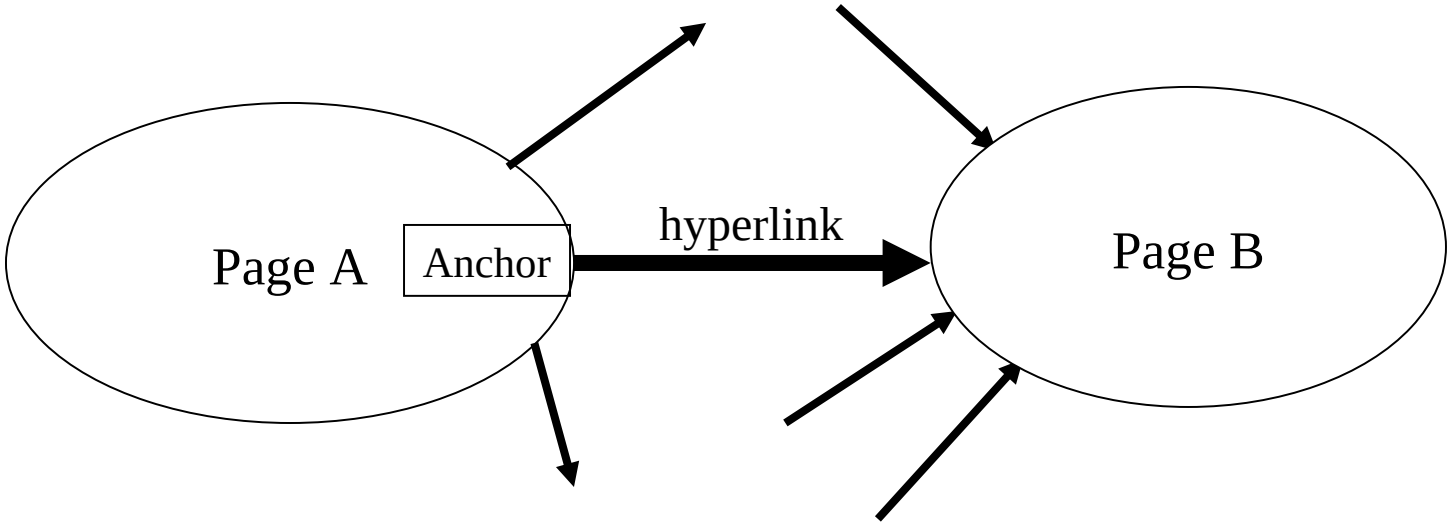
Много других видов учета ссылок

- Социальные сети
- «Сообщества шопоголиков» (Goel+Goldstein 2010)
  - Потребители, друзья которых много покупают, также много покупают
  - Выделение групп и лидеров групп в социальных сетях

Ссылки в информационном поиске

- Анализ ссылок
  - Вычисление веса и ранжирование
  - Кластеризация, основанная на ссылках
  - Ссылки как признаки в классификации
    - Ссылки друг на друга
    - Ссылки на другие документы
  - Краулинг

Веб как направленный граф

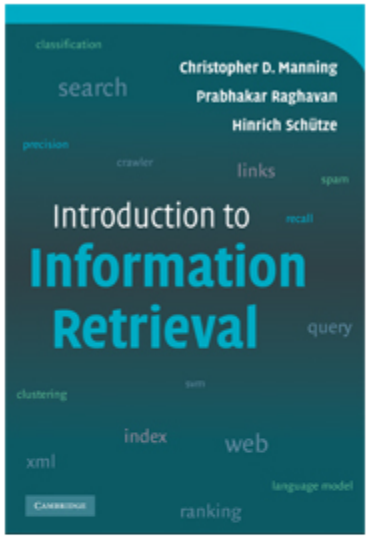


Гипотеза 1: Гиперссылка между страницами дает сигнал авторитетности

Гипотеза 2: Текст якоря в гиперссылке на странице A описывает целевую страницу B

Гипотезы: авторитетные сайты и текст якоря

Introduction to Information Retrieval



This is the companion website for the following book.

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *Introduction to Informat*

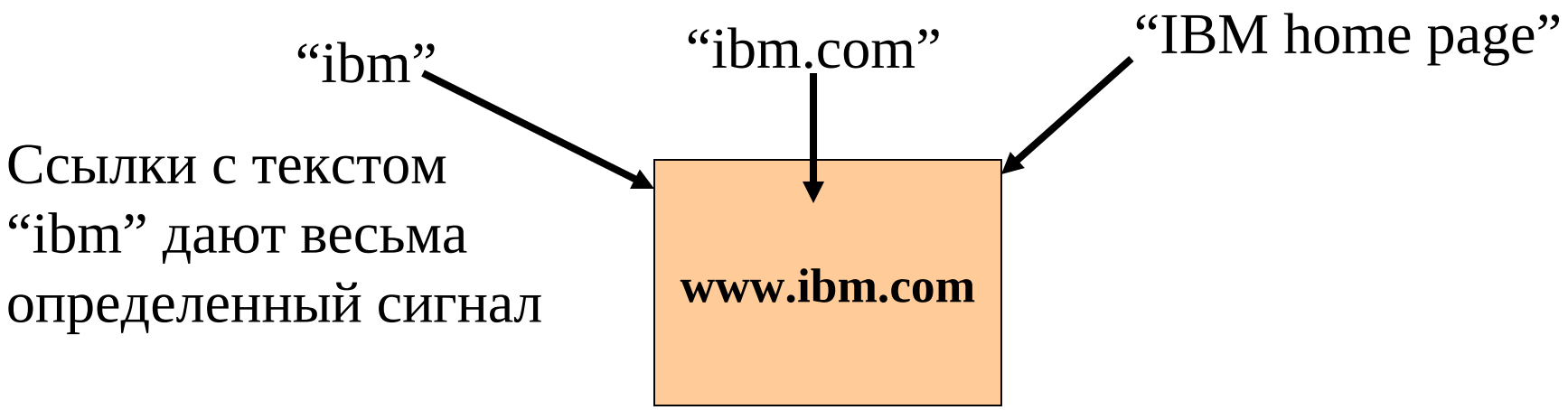
You can order this book at CUP, at your local bookstore or on the internet. The best search

The book aims to provide a modern approach to information retrieval from a computer scie
University and at the University of Stuttgart.

We'd be pleased to get feedback about how this book works out as a textbook, what is m
comments to: informationretrieval (at) yahoogroups (dot) com

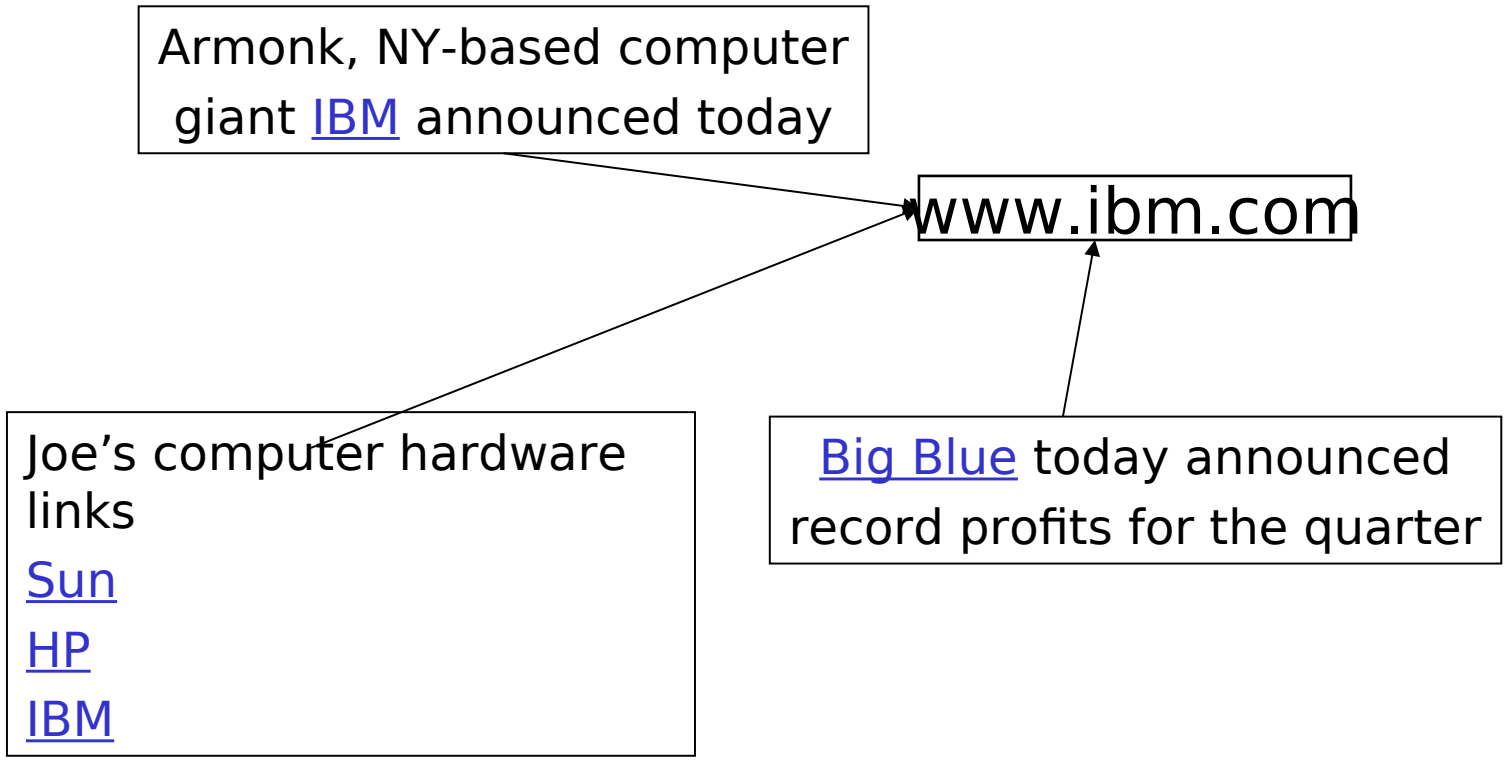
Текст ссылки (анкер)

- Для ***ibm***, как различить между:
  - IBM’s home page
  - IBM’s copyright page
  - Rival’s spam page



Индексирование анкеров

- При индексировании документа D нужно включить (с некоторым весом) текст ссылки, указывающий на *D*.



Индексирование анкеров

- Может привести к неожиданным эффектам (спам)

- Можно присваивать вес анкеру в зависимости от авторитетности сайта, которому принадлежит страница со ссылкой
  - Например, если мы доверяем сайтам cnn.com илиr yahoo.com, то можно присваивать более высокие веса ссылкам из них
  - Увеличиваем веса ссылкам, идущим извне сайта (non-nepotistic scoring)

Анализ ссылок. PageRank

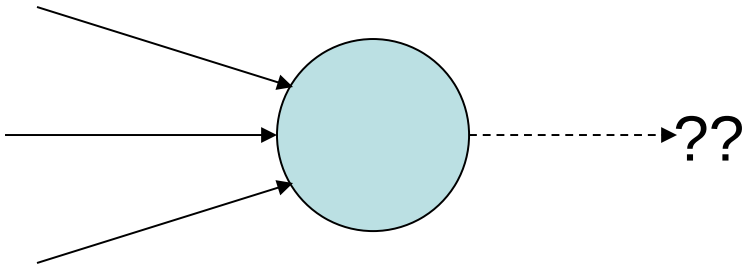
Вес Pagerank

- Представим, что пользователь случайно бродит по страницам:
  - Начинает на случайной странице
  - На каждом шагу переходит на следующую по ссылку с равной вероятностью

- В пределе каждая страница получит рейтинг посещений – можно использовать как вес страницы

Этого недостаточно

- Поскольку есть много тупиковых страниц.
  - В которых остановится случайное блуждание.
  - Это обессмысливает рассуждения о рейтинге посещений.



«Телепортация»

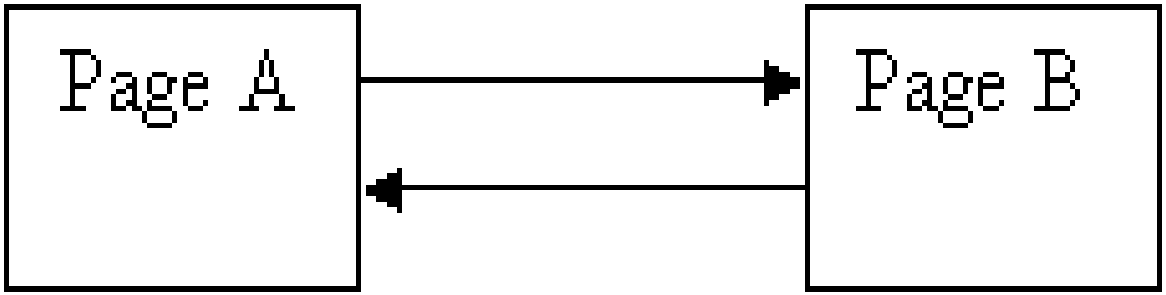
- В тупиковой странице – переход на случайную страницу
- Для любой нетупиковой страницы - с вероятностью 0.1, переходим на случайную страницу
  - С оставшейся вероятностью (0.9) – переход по одной из исходящих ссылок (с равной вероятностью)
  - 0.1 - параметр
- Теперь можно говорить о посещаемости страницы как о ее рейтинге
- Как можно посчитать такой рейтинг?

Формула PageRank

- PR(A) = d + (1-d)(PR(T1)/C(T1) +…+PR(Tn)/C(Tn))

- **PR(Tn)** – исходная значимость страницы
- **C(Tn)** – количество исходящих ссылок со страницы
- **PR(Tn)/C(Tn)** – значимость страницы равномерно распределяется по исходящим ссылкам и переносится в значимость страницы A по входящим в нее ссылкам
- **d** – например, 0.15 значимость страницы, без входящих ссылок (коэффициент телепортации)
- **(1-d)(…)** – 0.85

Простой пример



- Каждая страница имеет исходящую ссылку. Это означает, что C(A) = 1 и C(B) = 1.
- Мы не знаем с чего начать
- -начнем с первой страницы
- И предположим, что исходная значимость у всех = 1

Итак, PR(..)=1

- d (коэффициент телепортации) = 0.15

- PR(A)= d + (1-d) PR(B)/1)
- PR(B)= d + (1-d) (PR(A)/1)

т.е.

- PR(A)= 0.15 + 0.85 * 1= 1

- PR(B)= 0.15 + 0.85 * 1= 1

Возьмем другое число, например, PR(..)=0

- PR(A)= 0.15 + 0.85 * 0 = 0.15
  PR(B)= 0.15 + 0.85 * 0.15 = 0.2775

- Продолжаем
- PR(A)= 0.15 + 0.85 * 0.2775 = 0.385875
- PR(B)= 0.15 + 0.85 * 0.385875 = 0.47799375

И далее…

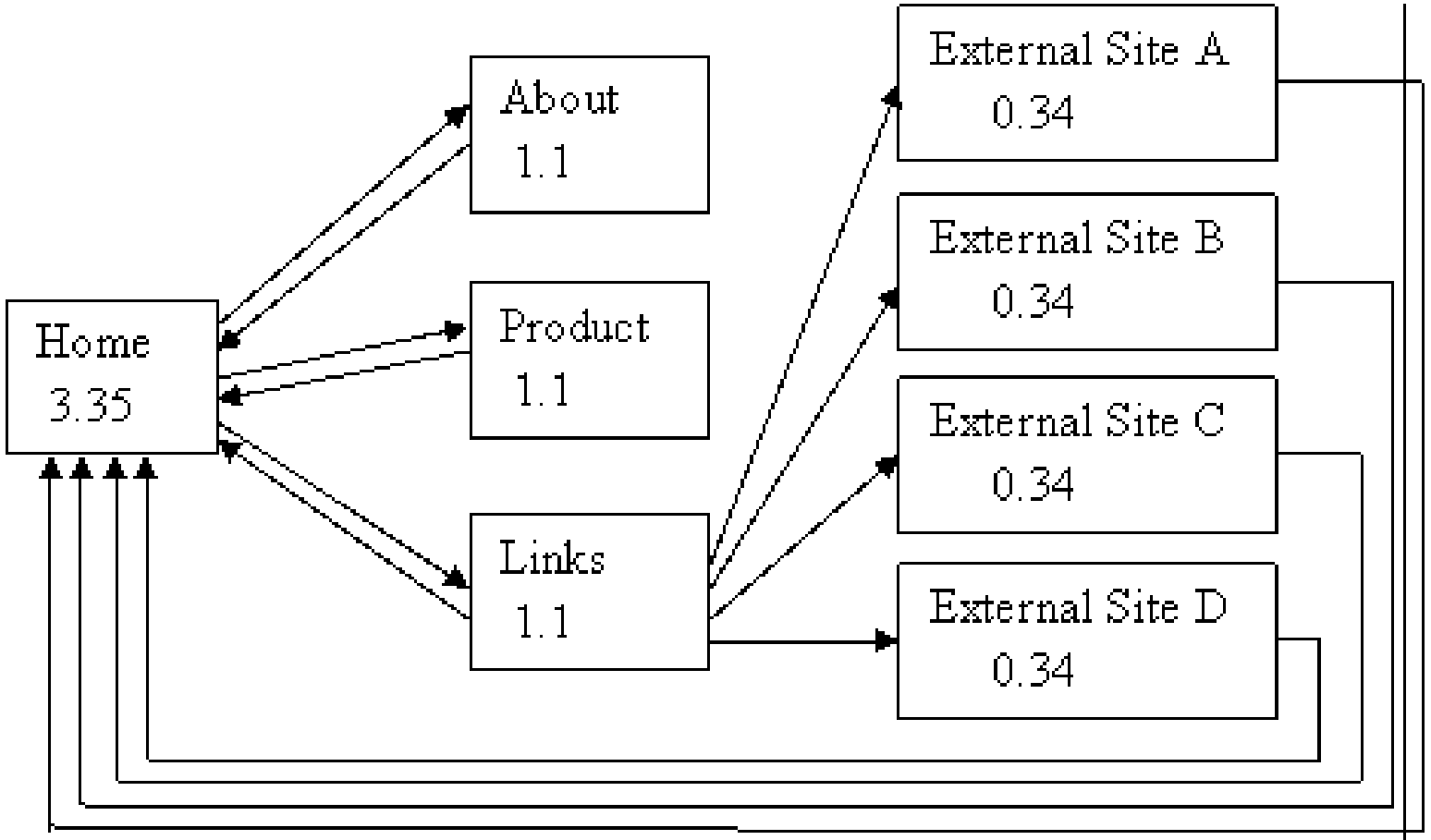
- PR(A)= 0.15 + 0.85 * 0.47799375 = 0.5562946875
- PR(B)= 0.15 + 0.85 * 0.5562946875 = 0.622850484375

Теперь пусть PR()=40

- PR(A)= 0.15 + 0.85 * 40
  = 34.25
- PR(B)= 0.15 + 0.85 * 34.25
  = 29.1775

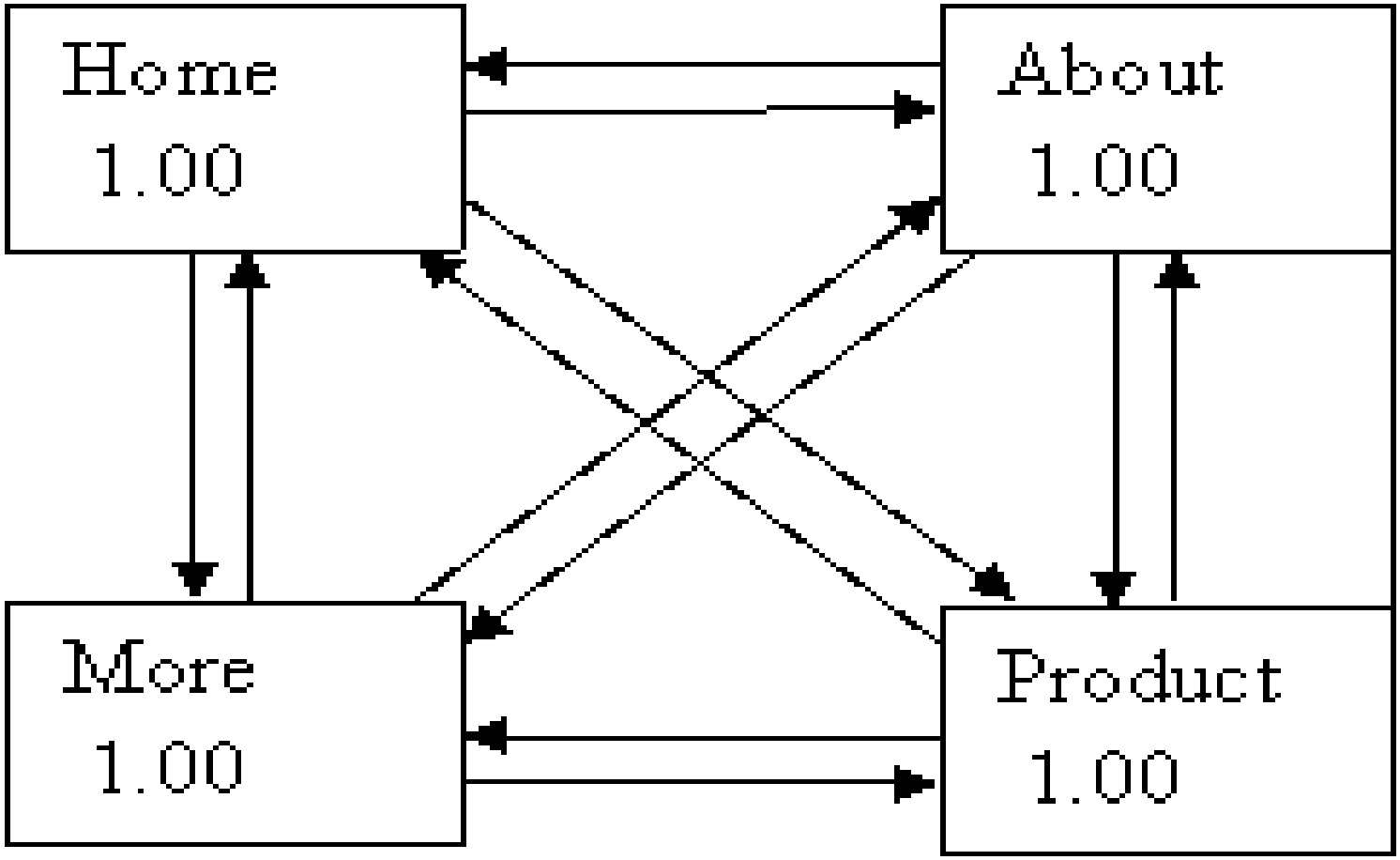
- И далее:
- PR(A)= 0.15 + 0.85 * 29.1775
  = 24.950875
- PR(B)= 0.15 + 0.85 * 24.950875
  = 21.35824375
- Процесс начинает сходиться к 1

Иерархия с возвратом ссылок

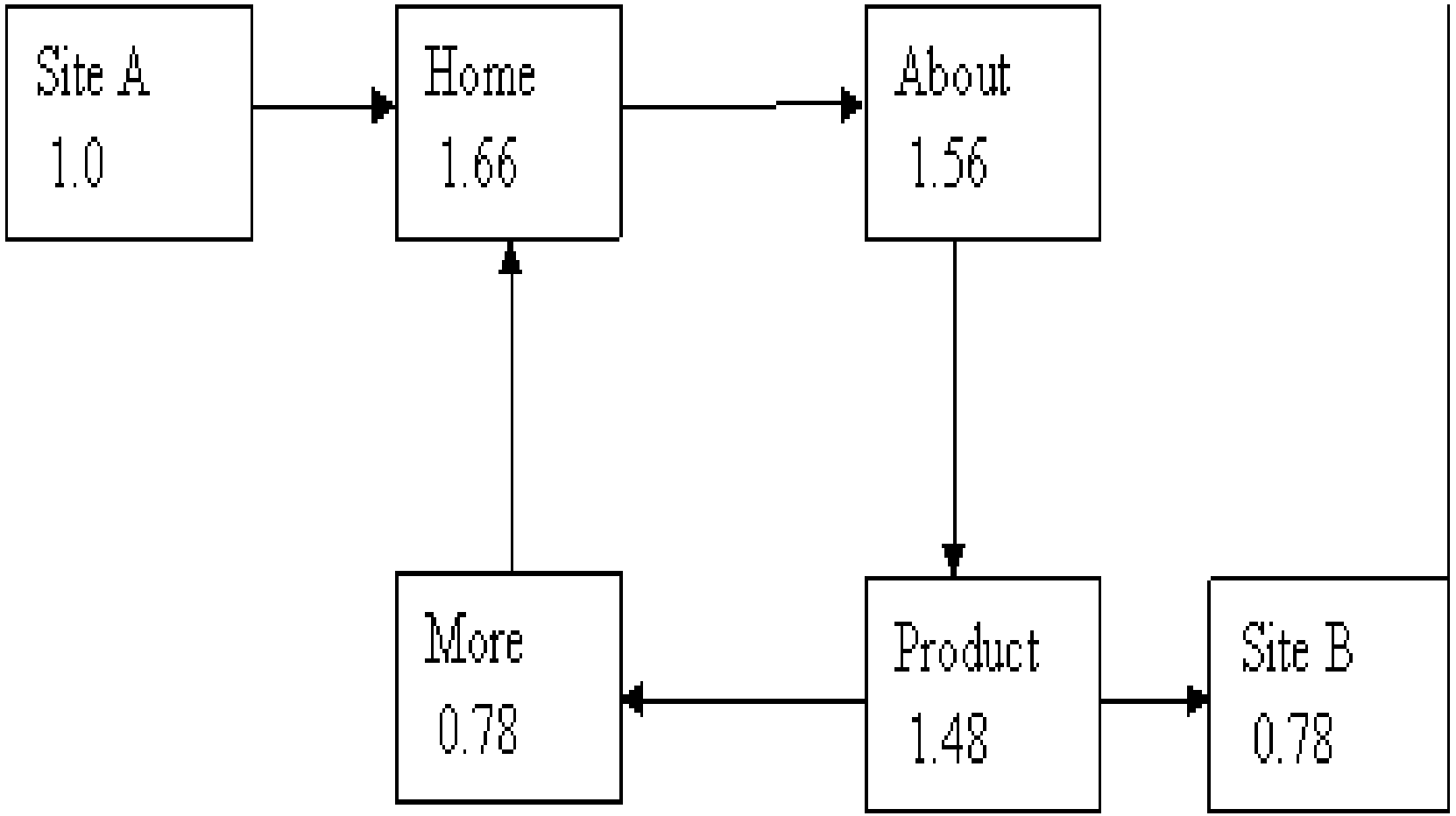


Average PR: 1.000

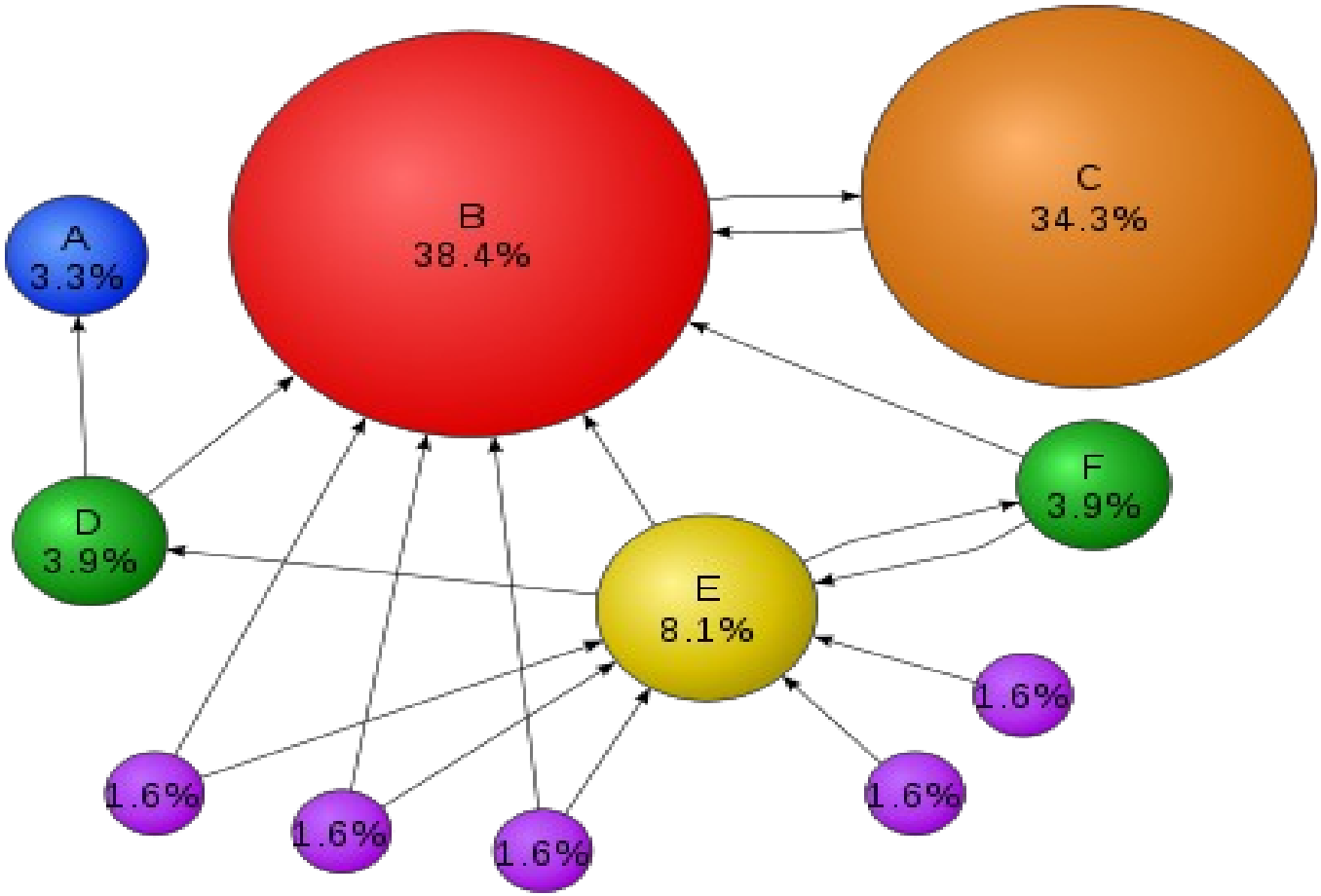
Все взаимосвязано



Еще пример



PageRank страницы



Уточнение по формуле

- Была дана формула вида (как в исходной статье (Brin, Page, 1998) и соотв. примеры

$$PR(A) = 1 - d + d\left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \cdots\right).$$

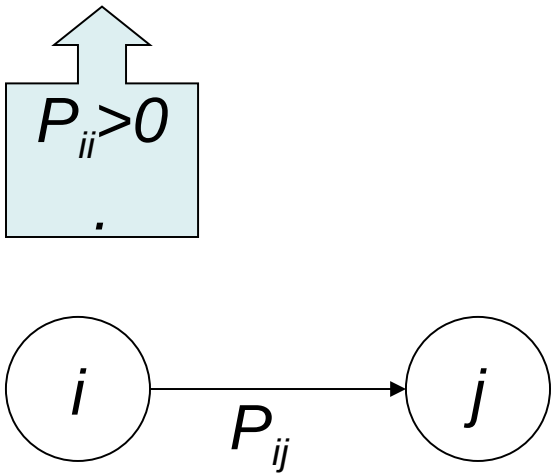
- Считается, что правильнее использовать такой вариант (сумма pagerank страниц в сети будет равна 1)

$$PR(A) = \frac{1 - d}{N} + d\left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \cdots\right).$$

Математическая основа PageRank

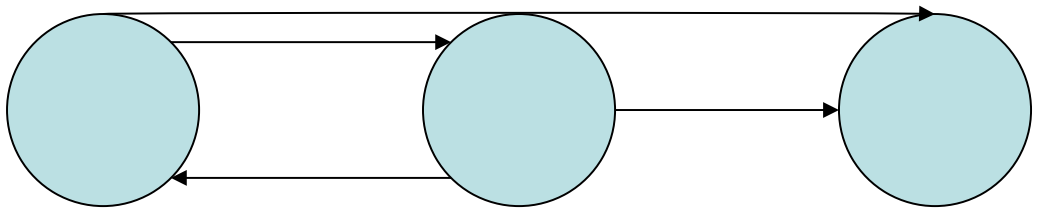
Марковские цепи

- Цепь маркова состоит из *n состояний*, плюс $n \times n$ матрица вероятностей переходов **P**.
- На каждом шаге, мы в одном из состояний
- Для $1 \leq i,j \leq n$, *элемент матрицы* P_{ij} означает вероятность перехода в *j (следующее состояние), при условии, что сейчас состояние - i*.



Марковские цепи-2

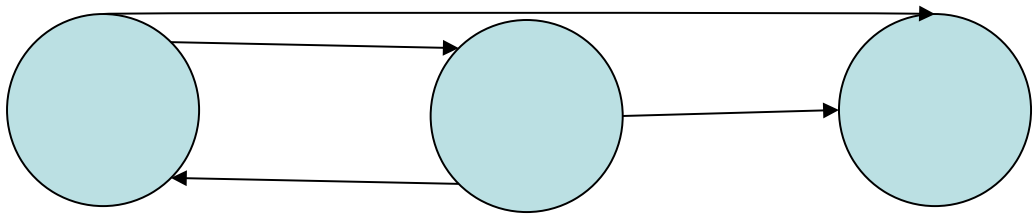
- Очевидно $\sum_{j=1}^{n} P_{ij} = 1.$
-
- *Задача*: представить случайное блуждание с «телепортацией» как Марковскую цепь



Марковские цепи-2

$$\sum_{j=1}^{n} P_{ij} = 1.$$

- Коэффициент телепортации – 0.1
- | | 1 | 2 | 3 |
- | 1 | 1/30 | 1/30+0.9*1/2 | 0.9*1/2+1/30 |
- | 2 | 1/30+0.9*1/2 | 1/30 | 0.9*1/2+1/30 |
- | 3 | 1/3 | 1/3 | 1/3 |



Алгоритм составления матрицы переходов

- Составить матрицу NxN по количеству страниц
- Для каждой страницы (строчки) исходящие ссылки обозначить 1 и поставить в соответствующих столбцах
- Нормализовать в каждой строке единицы, поделив на количество единиц
- Единицы умножить на коэффициент сглаживания (= (1-d), где d - коэффициент телепортации)
- Ко всем элементам строчки добавить коэффициент телепортации, поделенный на N (т.е. d/N)
- Если со страницы не было ссылок, то по всем столбцам ставим 1/N

Эргодическая марковская цепь

- Марковская цепь называется эргодической если:
  - существует положительное число T_0 такое, что для любой пары состояний *i, j* марковской цепи, если она начинается во время 0 в состоянии *i,* тогда для всех $t>T_0$, $P(i, j) >0$
    - *Требуются ненулевые вероятности перехода из одного состояния в другое*
    - *Множество состояний не разбивается на изолированные подмножества*
    - *Т.е. процесс перехода по сети с телепортациями – это эргодическая марковская цепь*

Теорема об эргодической цепи

- Для любой эргодической Марковской цепи, которая задана матрицей переходов P, существует единственный вектор вероятностей π, который представляет собой левый собственный вектор P такой, что если $\eta\ (i,\ t)$ – это число посещений узла i за t шагов, то

$$\lim_{t\to\infty}\frac{\eta(i,t)}{t}=\pi(i)$$

  - Это означает, что любая эргодическая марковская цепь имеет периодичность захода в узлы, т.е. наблюдается как бы «рейтинг посещений».
- Процесс сходится, и не важно, откуда начать и с каким исходным распределением

Изменения в векторе вероятности

- Пусть вектор вероятностей нахождения в узлах сети $\mathbf{x} = (x_1, \ldots x_n)$ *на данном шаге*, какой он будет на следующем шаге?

- Ряд *i* матрицы вероятностей переходов **P** говорит нам, куда мы идем из состояния*i*.

- Таким образом, следующее состояние будет **xP**
  - Далее $\mathbf{xP}^2$, $\mathbf{xP}^3$, и др.
  - Как посчитать, куда это сходится

Аналитический подсчет вектора стационарного состояния

- Пусть **a** = $(a_1, \ldots a_n)$ обозначает вектор стационарных вероятностей.
- Наша текущая позиция описывается как **a**, тогда и следующая позиция будет **aP**.
- Но **a** – это постоянное состояние, поэтому **a**=**aP**.
- Получается матричное уравнение, из которого можно вычислить **a**.
  - Т.е. **a** – это левый собственный вектор **P**.

Итеративный метод подсчета PageRank

- X - случайный вектор начальных состояний
- Подсчитываем матрицу P (матрица переходов с телепортациями)
- Нужно выполнить умножение xP
- затем xP^2 …
- Достаточно быстро достигается сходимость

Из истории интернет-поисковиков

- До 1998: поиск по ключевым словам
  - Altavista, Excite, Infoseek, Inktomi, Lycos
- 1998+: Google запускает ранжирование с учетом ссылок - PageRank
  - Вытеснил с рынка практически все ранние поисковые машины

Другая запись матрицы переходов PageRank

- Есть граф G с N вершинами: N1…Nn
- di – количество исходящих ссылок Ni,
- M: Матрица NxN описывает исходящие ссылки
- В каждой строке Ni указывается, на какие другие страницы есть переходы
- Mij=1/di – если переход из i в j существует
  - = 0 если нет
  - Если страница тупиковая, то значения 1/N

  Учет телепортации: изменение матрицы

  Pr=(1-c)M+c v, где v – матрица с элементами 1/N

  c – коэффициент телепортации (например, c=0.1)

Персонализированный PageRank

- Ранее равновероятная матрица случайных переходов
- Теперь!: Специально заданная матрица случайных переходов V
- Например, все случайные переходы возвращаются в вершину n_i (телепортации)
- Ранг n_i – высокий
- Ранг всех вершин около n_i – высокий
- Используется для вычисления тематической значимости страниц в информационном поиске

Анализ ссылок: HITS

Hyperlink-Induced Topic Search (HITS)

- В ответ на запрос вместо упорядоченного множества страниц найдем два множества взаимосвязанных страниц:
  - ***Hub pages*** *– хорошие списки ссылок по теме*.
    - “Bob’s list of cancer-related links.”
  - ***Authority pages*** **–** часто упоминаются на страницах хабов

- Хорошо работает на широких тематических запросах

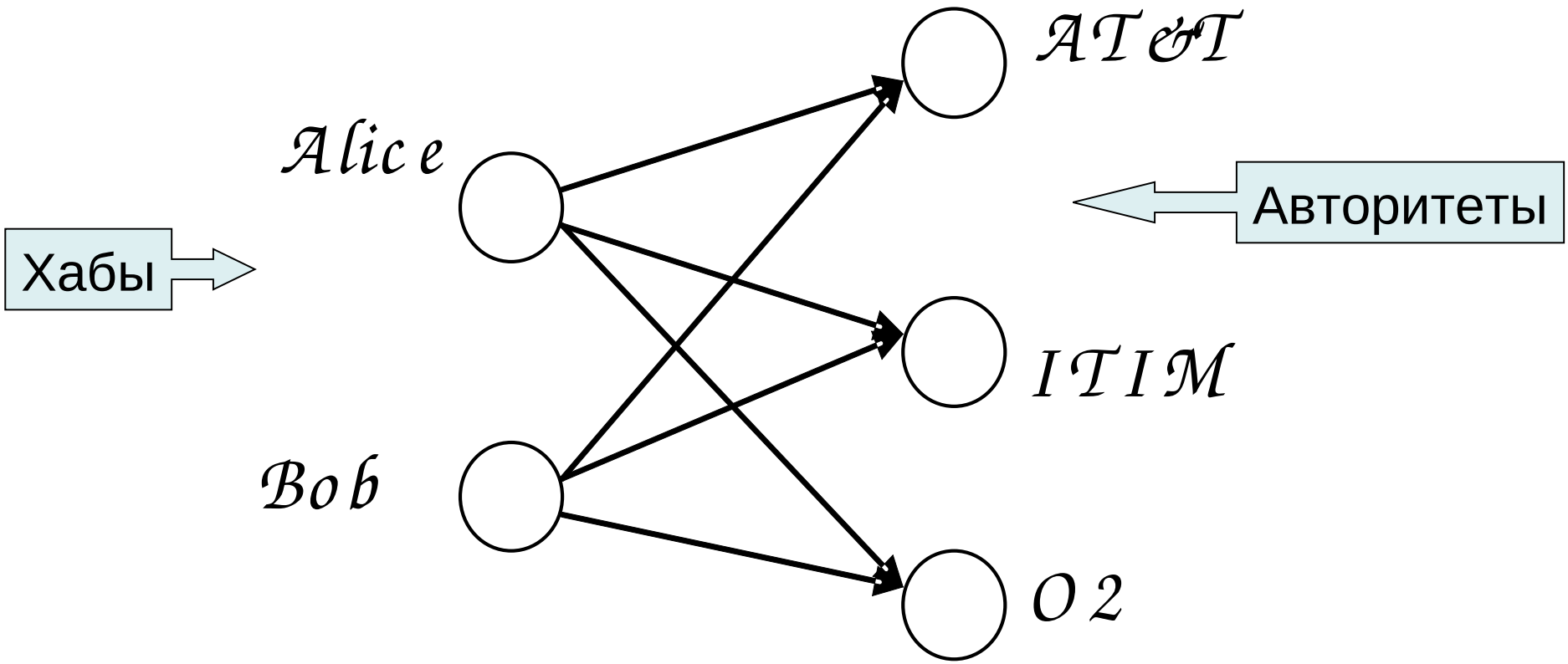
Хабы и Авторитеты

- Хорошая хаб-страница (посредник) для какой-то темы указывает на многие авторитетные страницы для этой темы

- Хорошая авторитетная страницы по теме указывается большим количеством хороших хабом по этой теме

- Для каждой страницы рекурсивно вычисляется ее значимость как посредника и как авторитета (автора)

Предположение



Мобильные компании

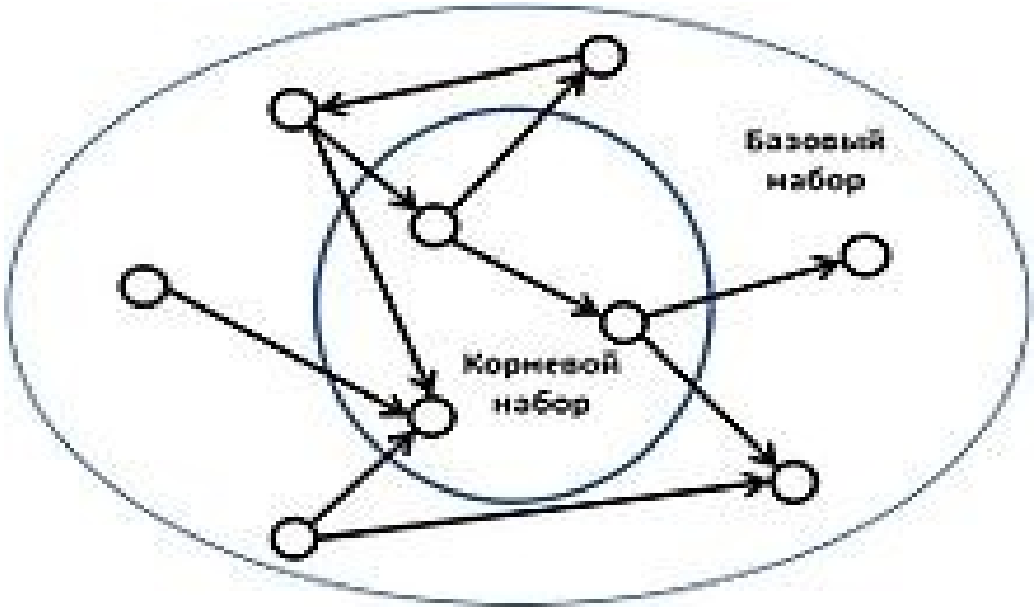
Основная схема

- Извлечь исходное множество (базовый набор) потенциально хороших хабов или авторитетов
- Из них формируем небольшой топ-лист хабов или авторитетов
  - Итеративный алгоритм

Базовый набор страниц

- Дан текстовый запрос (например, ***браузер***), получаем страницы, содержащие слово ***браузер.***
  - Это **корневой набор** страниц.
- Добавляем любую страницу, которая
  - указывает на страницу из корневого набора или
  - На которую есть ссылка со страницы корневого множества.
- Это **базовый** набор

Корневой и базовый набор страниц



Разделение хабов и авторитетов

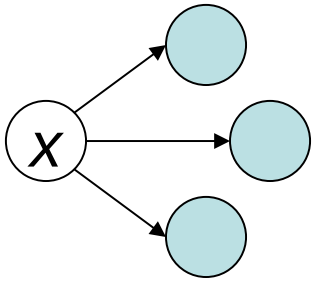
- Вычисляем для каждой страницы в базовом наборе hub score *h(x)* и authority score *a(x).*
- Инициализация: для всех *x, h(x)←1; a(x) ←1*;
- Итеративно пересчитываем *h(x), a(x)*;

- В результате итераций
  - Выдать страницы с наивысшими *h()* как топ-хабы
  - С наивысшими *a()* scores как топ-авторитеты

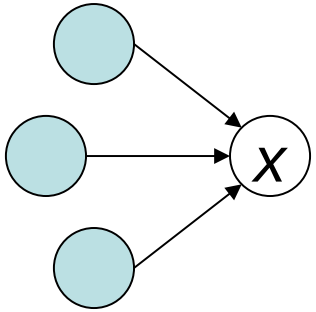
Итеративный пересчет

- Повторяем следующий пересчет для всех x:

$$h(x) \leftarrow \sum_{x \mapsto y} a(y)$$



$$a(x) \leftarrow \sum_{y \mapsto x} h(y)$$



Итеративный пересчет

- Таким образом,
  - оценка авторитетности страницы вычисляется как сумма значений оценок посреднических страниц, которые указывают на эту страницу.
  - посредническая оценка страницы вычисляется как сумма значений оценок авторитетности страниц, на которых она ссылается.
- Рост значений авторитетности и посредника – необходима нормализация.
- Значения, полученные в результате этого процесса, в конечном итоге сходятся.
- Обычно требуется около 5 итераций

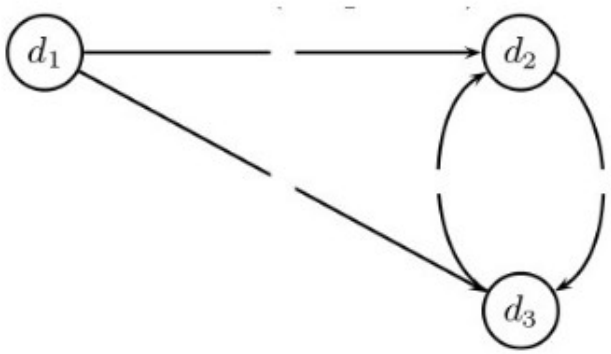
HITS vs. PageRank

- Алгоритм HITS вычисляет не только ранг каждого узла, но также дает посредническую оценку.
- Алгоритм PageRank содержит свободный параметр α, который обычно не включен в алгоритм HITS.
- Приоритетом, в результате работы алгоритма PageRank, пользуются, как правило, более старые ресурсы, в то время как HITS алгоритм имеет меньший уклон в этом отношении.
- Алгоритм PageRank может находить единственное уникальное решение

Недостатки HITS

- Сдвиг темы (Topic drift)
  - Нерелевантные документы могут вызвать сдвиг темы
- Нерелеватные страницы на первых позициях выдачи
  - приводят к ошибочным результатам
- Взаимное усиление страниц, ссылающихся друг на друга
- Поисковая оптимизация SEO: создание искусственного множества ссылок

Задание 6. Задача 1



Дан веб-граф, составить матрицу переходов.
Коэффициент телепортации = 0.1.
Составить матрицу переходов и вычислить
pagerank для узлов сети
Начальный вектор состояний можно
взять с равными вероятностями для каждого состояния

Задание 6. Задача 2. Задание такое же