

## Анализ гиперссылок в сети Веб: модели, подходы и алгоритмы

### Лекция 3. Поиск ресурсов в WWW

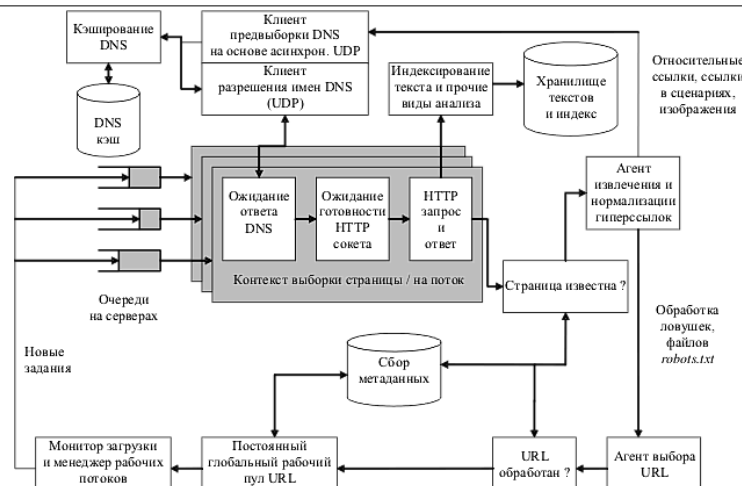
RUSSIR-2007. Сычев А.В.

## Принципы работы сетевого робота

- Обход или исследование веб-графа – процесса поиска узлов и ребер графа, начиная корневого подмножества узлов.
- Данная процедура реализуется с помощью компоненты, которая называется сетевым роботом-”пауком” или просто *пауком*.

RUSSIR-2007. Сычев А.В.

## Типовая структура “Паука”



RUSSIR-2007. Сычев А.В.

## Стратегии обхода веб-графа

- Приоритет *в ширину*.
- Приоритет *в глубину*.
- Эвристические методы (приоритет для более *качественных* ресурсов).

RUSSIR-2007. Сычев А.В.

## Релевантно-приоритетный сбор ресурсов из Веб. Идеи.

- При выдаче ответа на запрос релевантными оказывается лишь небольшая доля из всех выданных и следовательно ещё меньшая доля из скачанных пауком из Веб.
- Возможно ли организовать приоритетный сбор ресурсов из Веб?
- Следует руководствоваться следующими целями:
  - Предпочтение должно отдаваться наиболее релевантным ресурсам
  - Часто изменяемые документы могут быть предпочтительными для некоторых приложений, например новостных порталов
  - Для вертикальных порталов, обслуживающие запросы относящиеся к одной или нескольким обширным темам, требуется строить индексы из документов, относящихся только к этим темам.

RUSSIR-2007. Сычев А.В.

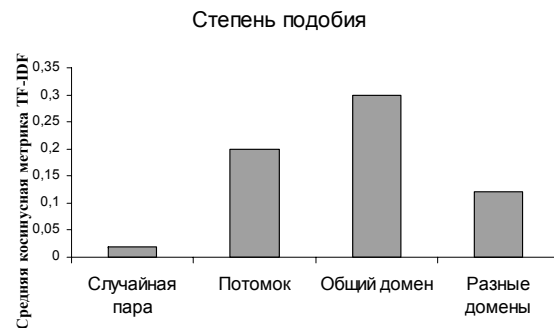
## Тематическая локализация

### Эксперимент

- Случайным образом выбираются два документа, вычисляется косинусная метрика подобия  $TF-IDF$ .
- Случайным образом выбираются два документа, связанных гиперссылкой с общим для них документом, выбранным случайным образом из коллекции; также оценивается их подобие.
- Для случайной страницы  $u$  выбираются документы, связанные с ней гиперссылками, но имеющие общий хост.
- Для случайной страницы  $u$  выбираются документы, связанные с ней гиперссылками, но принадлежащие различным хостам.

RUSSIR-2007. Сычев А.В.

## Тематическая локализация



В эксперименте использовалась выборка размером в 100000 документов из хранилища исследовательской ИПС *Disco Web*.

RUSSIR-2007. Сычев А.В.

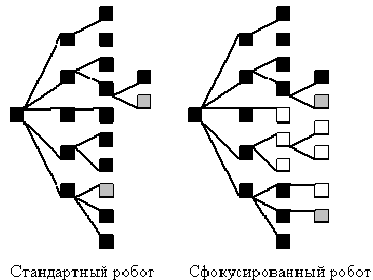
## Тематическая локализация

### Интерпретация результатов эксперимента:

- Экспериментально подтверждается *тематическая кластеризация* в сети Веб.
- Локальность типа “Разные домены” может быть использована для формирования стратегии расширения веб-графа для сетевого робота.
- Однако, проявление данного типа локальности быстро убывает с расстоянием в графе.
- По этой причине “агрессивное” расширение *графа соседства* исключено.
- В тоже время в графе существуют относительно длинные пути, сохраняющие тематическую релевантность

RUSSIR-2007. Сычев А.В.

## Стандартный vs. сфокусированный сетевой робот



- Стандартный сетевой робот, использующий стратегию “*приоритет в ширину*” будет последовательно уровень за уровнем скачивать страницы независимо от их релевантности, прежде чем достигнет целевого документа.
- Сфокусированный робот игнорирует нерелевантные документы. И таким образом, скачает лишь небольшое подмножество документов по пути к целевому документу.

RUSSIR-2007. Сычев А.В.

## Радиус-1 гипотеза

Если страница  $u$  соответствует тематике, и с ней связана гиперссылкой страница  $v$ , тогда вероятность того, что  $v$  соответствует тематике выше чем соответствующая вероятность у случайно выбранной из страницы.

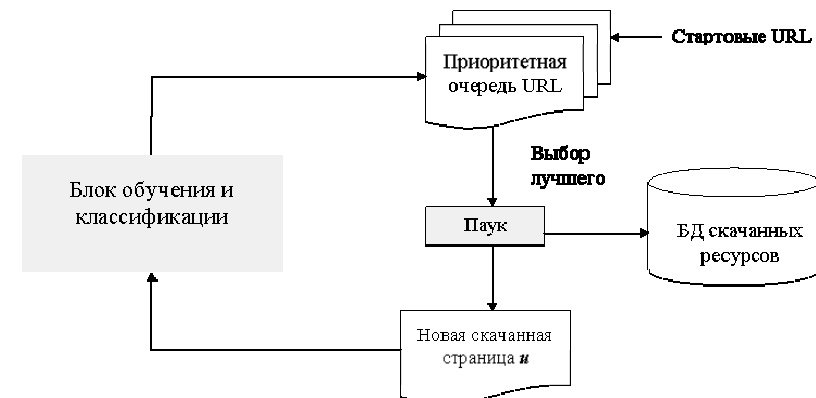
RUSSIR-2007. Сычев А.В.

## Идея сфокусированного поиска

Обучаемому классификатору предъявляется каждый вновь закаченный роботом документ  $u$ . Если классификатор по нему принимает положительное решение, то исходящие из него ссылки добавляются в очередь сетевого робота, иначе — игнорируются.

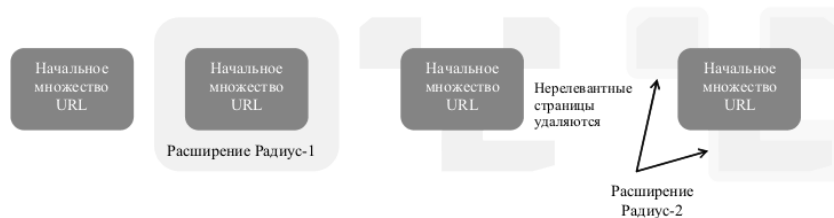
RUSSIR-2007. Сычев А.В.

## Блок-схема сфокусированного робота



RUSSIR-2007. Сычев А.В.

## Операционное представление сфокусированного поиска



RUSSIR-2007. Сычев А.В.

## Сфокусированный поиск Варианты реализации

- “Жесткий”:  
– Байесовский классификатор выносит бинарное пороговое решение по документу  $u$  (и соответственно по документам, на которые ссылается  $u$ ).
- “Мягкий”:  
– Оценка вероятности определяет приоритет документов в очереди, ссылки на которые, извлекаются из документа  $u$ .

RUSSIR-2007. Сычев А.В.

## Сфокусированный поиск

- Откуда брать стартовую обучающую выборку?
- Решение: иерархический тематический каталог (*Yahoo!*, *Open Directory* и т.п.)

RUSSIR-2007. Сычев А.В.

## Формализация задачи сфокусированного поиска

- Имеются:
  - ориентированный гипертекстовый граф  $G$
  - иерархический тематический каталог  $C$
  - каждому узлу  $c \in C$  соответствуют некоторые документы из  $G$ , т.е.  $D(c)$
  - потребность пользователя задается подмножеством  $C' \subseteq C$
  - байесовский классификатор на основе *TF-IDF-косинусной метрики* для каждого поступающего документа  $q$  формирует оценку релевантности  $R_{C'}(q)$
  - стартовое множество документов задается как  $D(C^*)$
- Целью является поиск на графе множества вершин  $V \supset D(C')$  достижимых из  $D(C^*)$ , таких, что максимизируется величина:

$$(\sum_{v \in V} R(v)) / |V|$$

RUSSIR-2007. Сычев А.В.

## Сфокусированный поиск Тестирование

Как показал эксперимент, усредненное значение релевантности  $(\sum_{v \in V} R(v))/|V|$  для закачанных сфокусированным роботом документов сохраняется на стабильно высоком уровне на протяжении закачки до тысяч документов, в отличие от обычного робота (для него это значение быстро убывает в пределах первой сотни скачанных роботом документов)

RUSSIR-2007. Сычев А.В.

## Поиск авторитетов и концентраторов сфокусированным роботом

- Хорошие авторитеты находятся сфокусированным роботом на расстоянии 10-12 гиперссылок от стартовых узлов.
- Концентраторы плохо выявляются “*радиус-1*” сфокусированным роботом, поскольку содержат относительно мало описательной текстовой информации, а также имеют ссылки по нескольким темам.

RUSSIR-2007. Сычев А.В.

## Радиус-2 гипотеза

Если страница *u* ссылается на много страниц *v* имеющих высокий показатель *R(v)*, тогда *u* является *хорошим концентратором*; соседние по графу страницы *w*, на которые ссылается *u*, вероятнее всего имеют более высокую релевантность нежели страницы случайно выбранные из веб.

RUSSIR-2007. Сычев А.В.

## Сфокусированный поиск

Когда показатель результативности сфокусированного робота “*радиус-1*” резко сокращается можно:

- Переключиться на документы, на которые указывают качественные концентраторы (используя один из вариантов HITS) – **сетевой робот “радиус-2”**
- Использовать обратные ссылки

RUSSIR-2007. Сычев А.В.

## Контекстно-сфокусированный поиск

Пример: если необходимо найти статьи по определенной ИТ-тематике, то можно начать поиск с веб-страниц факультетов ИТ профиля, содержащих ссылки на домашние страницы преподавателей и сотрудников, некоторые из которых могут содержать в свою очередь ссылки на статьи по интересующей тематике.

RUSSIR-2007. Сычев А.В.

## Контекстно-сфокусированный поиск

Необходимо построить обучающую структуру для нахождения путей в веб-графе, приводящих к релевантным документам:

- *объектами*, подлежащими классификации, являются документы (узлы графа);
- в качестве *атрибутов* документов рассматриваются термины, содержащиеся в них;
- по тексту документа на входе необходимо *предсказать* - сколько переходов по ссылкам необходимо выполнить до достижения релевантного документа;
- оценка указанного расстояния используется для отнесения документа к определенному *уровню*, определяющего порядок очередности обработки документов.

RUSSIR-2007. Сычев А.В.

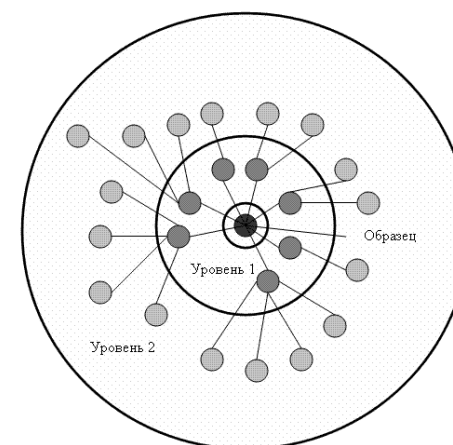
## Контекстный граф

Таким образом, строится следующий *контекстный граф*:

- Стартовый документ образует *0-й* уровень.
- Документы, ссылающиеся на документ *0-го* уровня попадают на *1-й* уровень и т.д.

RUSSIR-2007. Сычев А.В.

## Контекстный граф



RUSSIR-2007. Сычев А.В.

## Контекстный граф

- Построенный *контекстный граф* показывает:
  - темы, прямо или косвенно связанные с целевой темой;
  - пути связывающие документы в графе с целевыми документами.
- Далее, контекстные графы для всех стартовых документов объединяются, образуя *объединенный контекстный граф*.

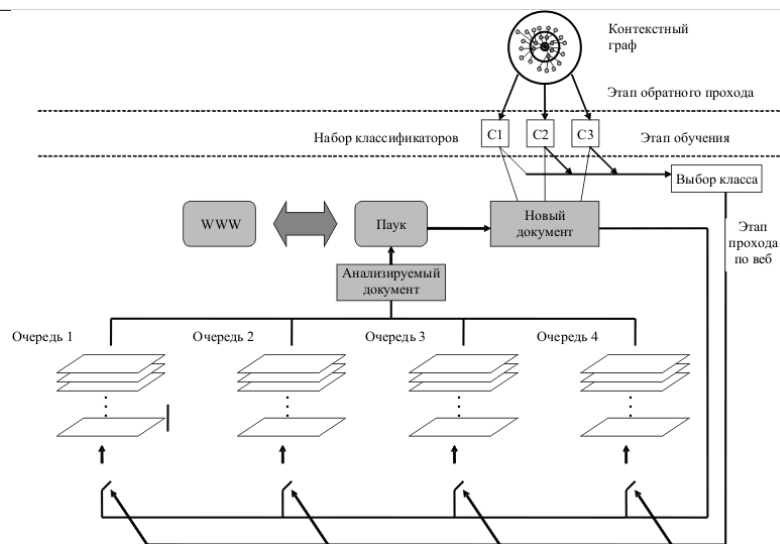
RUSSIR-2007. Сычев А.В.

## Контекстно-сфокусированный поиск

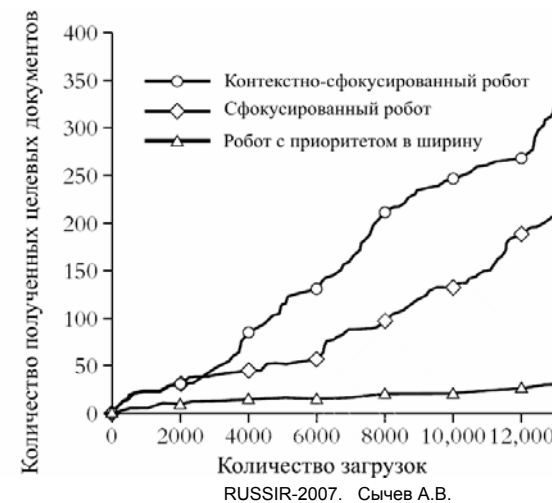
*Объединенный контекстный граф* используется байесовскими классификаторами для распределения вновь поступающих от сетевого робота документов по уровням контекстного графа и фактически для построения приоритетных очередей обработки документов. Наиболее приоритетными являются очереди документов, находящихся на нижележащих уровнях контекстного графа.

RUSSIR-2007. Сычев А.В.

## Структура контекстно-сфокусированного робота-паука



## Эксперимент



## Литература

- S. Chakrabarti “*Mining the Web. Discovering Knowledge from Hypertext Data*”. Morgan Kaufmann Publishers, 2003.