

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

### Лекция 4. Веб-сообщества

RUSSIR-2007. Сычев А.В.

## Социальные сети

- Большинство социальных сетей демонстрируют структуру типа “сообщество”, т.е. содержат группы вершин, имеющих высокую плотность концентрации ребер внутри группы и низкую – между группами.

RUSSIR-2007. Сычев А.В.

## Веб-сообщества

- Неформально *веб-сообщество* определяется как подграф веб-графа, в котором плотность внутренних связей превышает плотность внешних связей.
- Формальное определение: **Веб-сообщество** есть подмножество вершин  $C \subset V$ , таких, что для всех вершин  $v \in C$ ,  $v$  имеет множество ребер, соединяющих её с вершинами в  $C$  и практически не имеет ребер, соединяющих с вершинами в  $(V \setminus C)$ .
- Данная задача является NP-полной.

RUSSIR-2007. Сычев А.В.

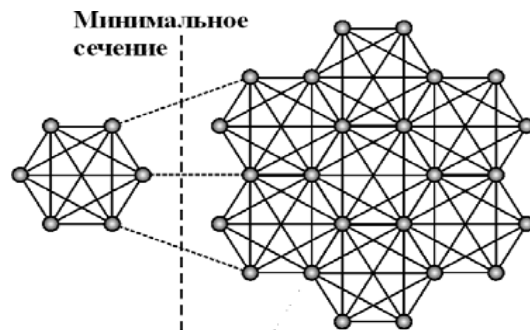
## “Зерновые” веб-ресурсы

- Тем не менее, если исходить из факта существования одного или более “зерновых” веб-ресурсов и использовать систематические закономерности в структуре веб-графа, задача может быть сформулирована в виде, который позволяет эффективно идентифицировать веб-сообщества.
- Под “зерновым” понимают веб-ресурс (веб-страницу), который является признанным авторитетом в тематической области идентифицируемого веб-сообщества и однозначно ему принадлежит.

RUSSIR-2007. Сычев А.В.

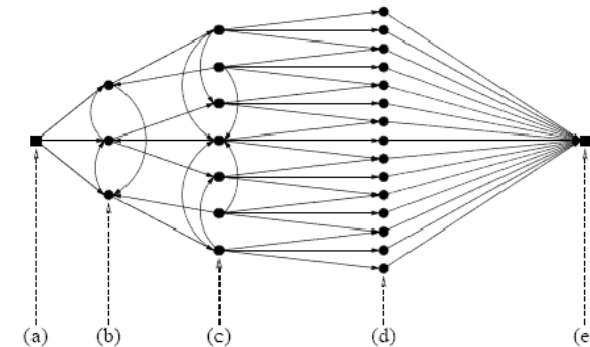
## Поиск веб-сообществ

Решение задачи о поиске веб-сообщества сводится к задаче поиска минимального сечения для потока в сети.



RUSSIR-2007. Сычев А.В.

## Направленное извлечение сообщества и построение графа



(a) виртуальный источник; (b) вершины зерновых веб-сайтов; (c) вершины веб-сайтов на расстоянии одной ссылки в глубину от любого зернового сайта; (d) ссылки на сайты не из (b) или (c); (e) вершина виртуального стока.

RUSSIR-2007. Сычев А.В.

## Направленное извлечение сообщества и построение графа

- Начиная с зерновых веб-страниц (b), находятся все страницы, которые ссылаются или на которые ссылается зерновое подмножество страниц.
- Исходящие ссылки извлекаются при анализе HTML-кода страницы.
- Входящие ссылки находятся путём запроса к поисковому сервису, который поддерживает модификатор “link”.

RUSSIR-2007. Сычев А.В.

## Направленное извлечение сообщества и построение графа

Как только URL из множества (c) идентифицированы, их HTML скачиваются и все исходящие ссылки запоминаются. Некоторые из этих исходящих ссылок могут ссылаться на страницы уже посещённые (такие как ссылки из (c) на (c) и (c) на (b)); тем не менее, большинство исходящих ссылок из (c) ведут на ещё не скачанные страницы (из множества (d)). Страницы, составляющие множество (d) фактически являются эффективно очищенной составной вершиной стока, т.к. каждая из них ссылается на вершину виртуального стока.

RUSSIR-2007. Сычев А.В.

## Алгоритм для выделения веб-сообществ FLG (Flake-Lawrence-Giles )

```
procedure EXACT-FLOW-COMMUNITY
input : graph:  $G = (V; E)$ ; set :  $S \subset V$ ; integer :  $k$ .
// Создаёт искусственные вершины,  $s$  и  $t$  и
// добавляет их в  $V$ .
for all  $v \in S$  do
Add ( $s; v$ ) to  $E$  with  $c(s; v) \equiv \infty$ .
end for
for all ( $u; v$ )  $\in E$  do
Set  $c(u; v) \equiv k$ .
if ( $v; u$ )  $\notin E$  then add ( $v; u$ ) to  $E$  with  $c(v; u) \equiv k$ .
end for
for all  $v \in V; v \notin S \cup \{s; t\}$  do
Add ( $v; t$ ) to  $E$  with  $c(v; t) \equiv 1$ .
end for
call : MAX-FLOW ( $G, s, t$ ).
output : all  $v \in V$  всё ещё соединённых с  $s$ .
end procedure
```

```
procedure APPROXIMATE-FLOW-COMMUNITY
input : set :  $S$ .
while число итераций меньше желаемого do
Построить  $G = (V; E)$  путём просмотра сети на
фиксированную глубину, начиная с  $S$ .
Set  $k$  to  $|S|$ .
call :  $C = \text{EXACT-FLOW-COMMUNITY}(G; S; k)$ .
Посчитать ранг для всех  $v \in C$  по числу рёбер в  $C$ .
Добавить не зерновые вершины с высоким рангом в
 $S$ .
end while
output : all  $v \in V$  всё ещё соединённых с  $s$ .
end procedure
```

RUSSIR-2007. Сычев А.В.

## Альтернативные подходы к поиску веб-сообществ

- На основе классического алгоритма *HITS*
- На основе *HITS* с использованием неглавных собственных векторов
- На основе комбинированного *HITS* и латентно-семантического анализа
- На основе комбинирования анализа гиперссылок с помощью *SALSA* и анализа текста с помощью *tf-idf* метрики.

RUSSIR-2007. Сычев А.В.

## Блогосфера

На протяжении нескольких последних лет в глобальной сети WWW происходят серьезные трансформации:

- формирование и быстрый рост коммуникативной среды пользователей сети WWW – блогосферы. Если в 90-х годах и в самом начале нынешнего века основным средством размещения контента в сети WWW было создание и обновление веб-сайтов, что являлось прерогативой технически подготовленных пользователей, то в наши дни программное обеспечение, поддерживающее работу с сетевыми дневниками (веб-блогами) существенно расширяет круг пользователей, способных создавать веб-контент и делать его доступным для больших и малых групп пользователей.
- простота создания контента приводит к постепенному перемещению в WWW некоторых функций, традиционно относившихся к электронной почте и системам передачи мгновенных сообщений.

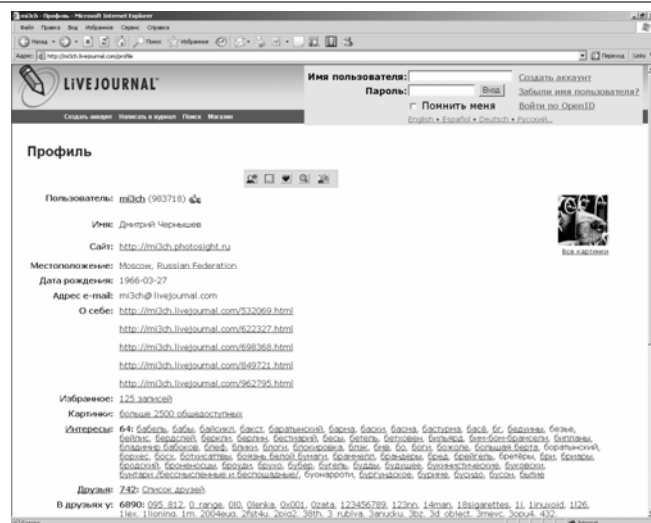
RUSSIR-2007. Сычев А.В.

## Блог

- Веб-дневник (блог):
  - представляет собой постоянно обновляемую страницу, на которой автор размещает свои записи (сообщения), расположенные в обратном хронологическом порядке.
  - другие пользователи могут просматривать эти сообщения и оставлять на странице свои комментарии.
  - аудитория блога разделяется на две большие группы пользователей: авторы и читатели.
  - имеет ссылки на страницы, содержащие информацию об авторе блога, его интересах, членстве в сообществах по интересам, друзьях (пользователях, дневники которых он читает) и другую информацию.

RUSSIR-2007. Сычев А.В.

## Блог



RUSSIR-2007. Сычев А.В.

## Блогосфера

- Совокупность блогов, связанных между собой ссылками, образуют коммуникационную группу. Техническая поддержка таких дневников реализуется с помощью общедоступного программного обеспечения.
- Существование таких блогов существенно изменяет модель поведения пользователя сети WWW и приводит к формированию веб-сообществ, вся совокупность которых называется *блогосферой*.

RUSSIR-2007. Сычев А.В.

## Статистика

- По данным одного из ведущих англоязычных блог-поисковиков *Technorati.com* на сентябрь 2006 года в мире было более чем 54 миллиона блогов.
- В русскоязычной блогосфере, по данным службы Яндекса «Поиск по блогам» (*Blogs.yandex.ru*), сегодня более чем 1 150 000 блогов, а записей — более 80 миллионов.
- Каждую секунду появляется в среднем три новых записи.

RUSSIR-2007. Сычев А.В.

## Исследование блогосферы

Систематическое исследование блогосферы представляет интерес по ряду причин.

- Блогосфера структурно отличается от сети традиционных веб-страниц.
- Традиционные методы изучения сети WWW основаны на формировании веб-графа с помощью специальных программных агентов — роботов, скачивающих из сети веб-страницы, следуя по гиперссылкам. Полученный граф представляет собой статический “снимок” сети. Что вынуждает исследователей периодически делать такие “снимки” для изучения динамики сети. При этом сложно понять когда именно и какая часть страницы или ссылка изменилась. Блогосфера предлагает уже готовый инструмент для наблюдения эволюции во времени, поскольку все записи в блогах имеют встроенную временную метку.

RUSSIR-2007. Сычев А.В.

## Исследование блогосферы

- Сообщества в блогосфере представляют собой гигантский механизм совместной фильтрации информации, построенный на основе неформальной сети доверия между членами сообщества.
- Исследование причин и механизмов резонанса в ответ на события реального мира в сети подобной блогосфере интересно с точки зрения того, как из хаоса спонтанно происходит самоорганизация.
- В то время как социальные сети сами по себе зачастую с трудом поддаются прямому изучению, on-line системы могут предоставить достаточно много информации относительно социальных сетей.
- Члены блог-сообщества явным образом выражают свою принадлежность к конкретной группе или сообществу. Тем самым отпадает необходимость в решении ряда трудоемких вычислительных задач на графах.
- Анализ блогосферы может проводиться в двух измерениях: *временном* (динамика или эволюция во времени) и *пространственном* (структурном).

RUSSIR-2007. Сычев А.В.

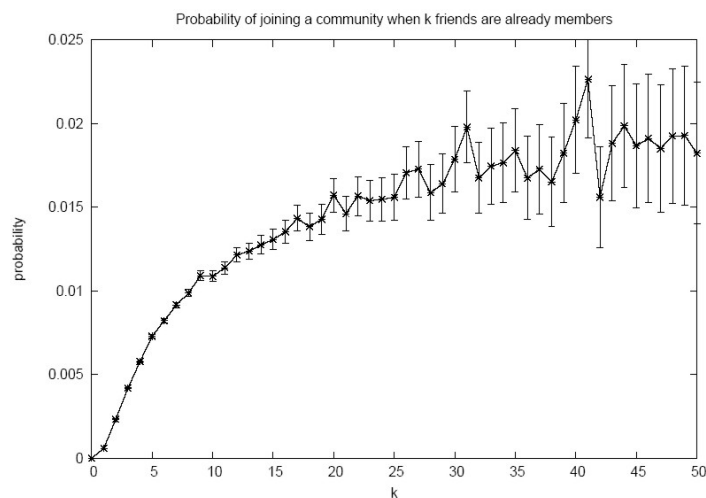
## Исследование блогосферы Вопросы

В настоящее время исследователи блогосферы пытаются найти ответы на следующие вопросы:

- чем определяется структура сообщества?
- какова динамика развития сообщества, основные этапы его развития?
- каковы признаки, позволяющие относить данный элемент к тому или иному сообществу?

RUSSIR-2007. Сычев А.В.

## Исследование блогосферы



RUSSIR-2007. Сычев А.В.

## Литература

1. J. Kleinberg, S. Lawrence “*The structure of the Web*” // Science, vol 294, November 2001. pp. 1849-185.
2. Э. Майника “*Алгоритмы оптимизации на сетях и графах*”. – М.: «Мир», 1981. – 323 с.
3. G. Flake, S. Lawrence, and C. L. Giles “*Efficient identification of web communities*”. In 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 150–160, 2000.
4. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. “*Crawling the Web for emerging cyber-communities*”. In Proceedings of the 8th International World Wide Web Conference, pp. 1481–1493, 1999.
5. Д.Д. Козлов, А.А. Белова. “*Исследование эффективности применения методов совместного анализа текстов и гиперссылок для поиска тематических сообществ*” ([http://company.yandex.ru/grant/2005/06\\_Kozlov\\_102805.pdf](http://company.yandex.ru/grant/2005/06_Kozlov_102805.pdf))

RUSSIR-2007. Сычев А.В.

# Литература

6. R. Kumar, J. Novak, P. Raghavan, A. Tomkins “*On the bursty evolution of blogspace*”// Proceedings of the 12th international conference on World Wide Web, May 20-24, 2003.
7. E. Adar, L. Zhang, L. A. Adamic, and R. M. Lukose “*Implicit Structure and the Dynamics of Blogspace*”// WWW2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics. 2004.
8. L. Backstrom, D. Huttenlocher, J. Kleinberg “*Group formation in large social networks: membership, growth, and evolution*” // KDD’06, August 20-23, 2006, Philadelphia, Pennsylvania, USA.
9. А.В. Сычёв, А.В. Кровопусков “*Исследование структурных зависимостей русскоязычных блог-сообществ в LiveJournal*”// Телематика' 2007 : тр. 14 Всерос. науч.-метод. конф., 18-21 июня 2007 г., Санкт-Петербург .— СПб, 2007 .— Т. 2. - С. 289-291.