Воронежский государственный университет
Факультет компьютерных наук
Кафедра информационных систем

Самоорганизация в сети Веб

Регулярность в распределении гиперссылок

Исследования показали, что гиперссылки в сети Веб не подчиняются модели независимой случайной генерации. В первом приближении вероятность появления новой ссылки у страницы подчиняется степенному закону:

$$\Pr(исх = k) \propto \frac{1}{k^{a_{исх}}} \qquad \Pr(вх = k) \propto \frac{1}{k^{a_{вх}}}$$

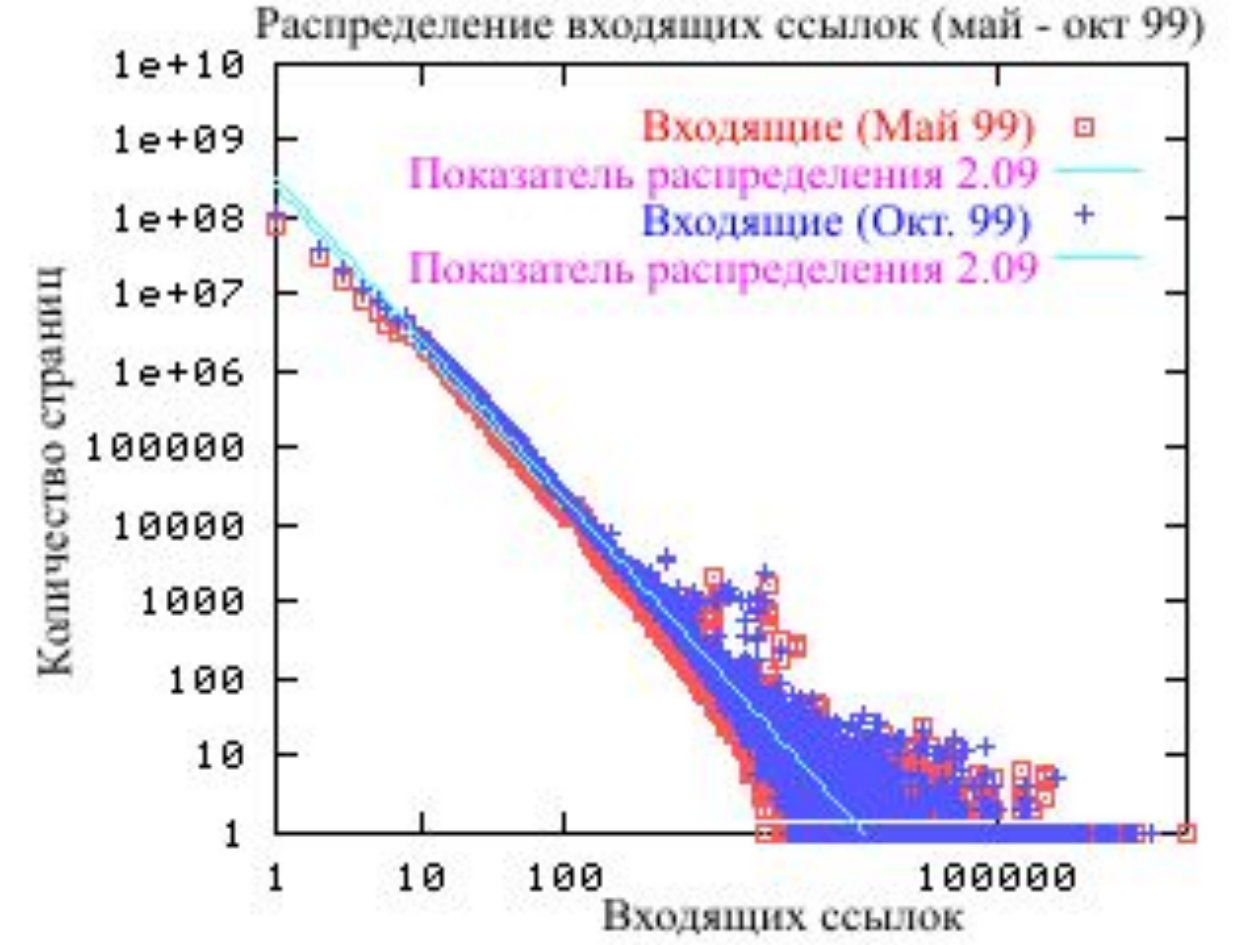
где ***k*** - количество исходящих или входящих гиперссылок,

$a_{исх} = 2.45$, $a_{вх} = 2.1$

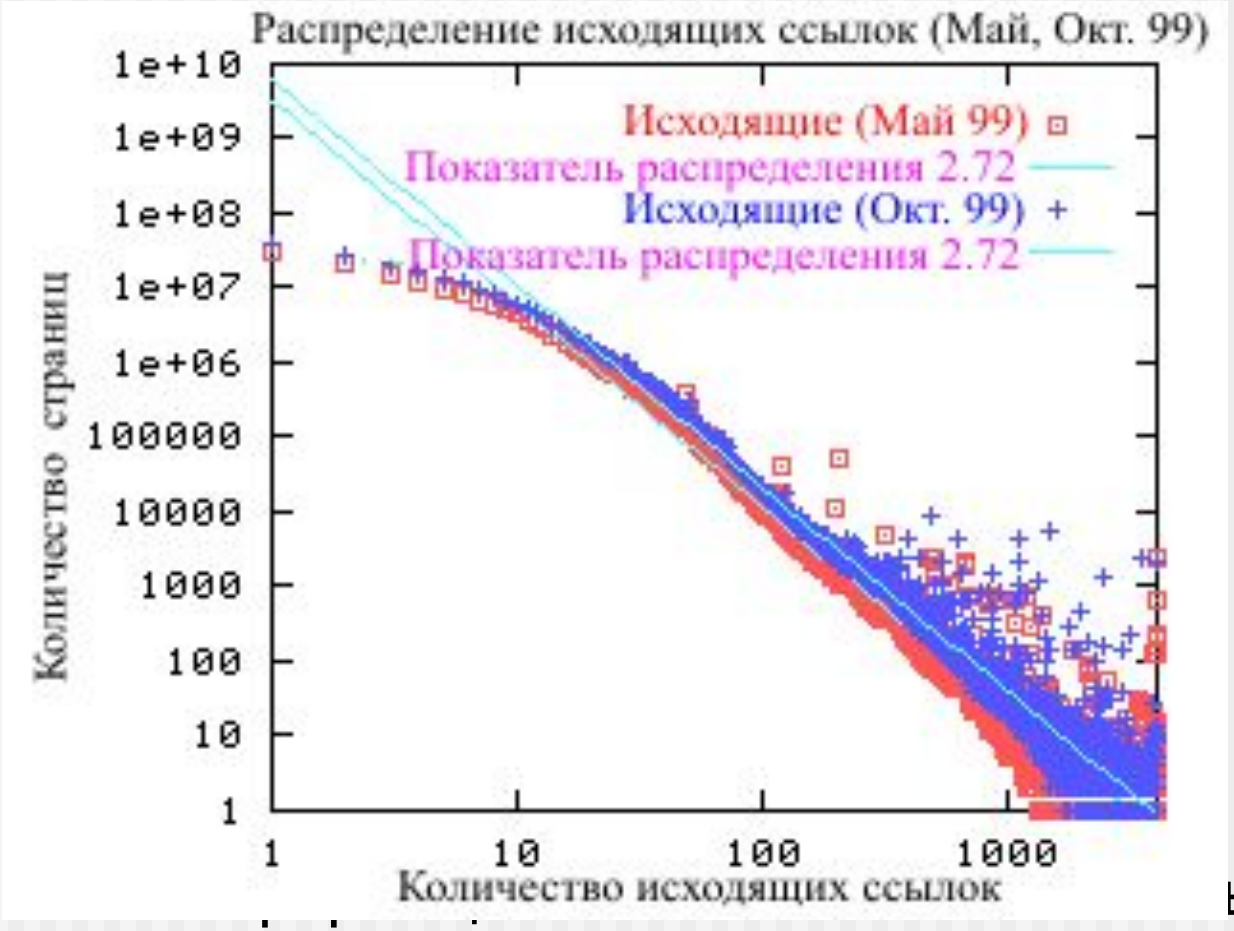
Модель предпочтительного прикрепления

- Вновь возникающий узел веб-графа устанавливает соединения с уже существующими узлами не равновероятно, но с большей вероятностью с узлами, имеющими большое количество связей.
- “Победителям достается все”.

Модель предпочтительного прикрепления

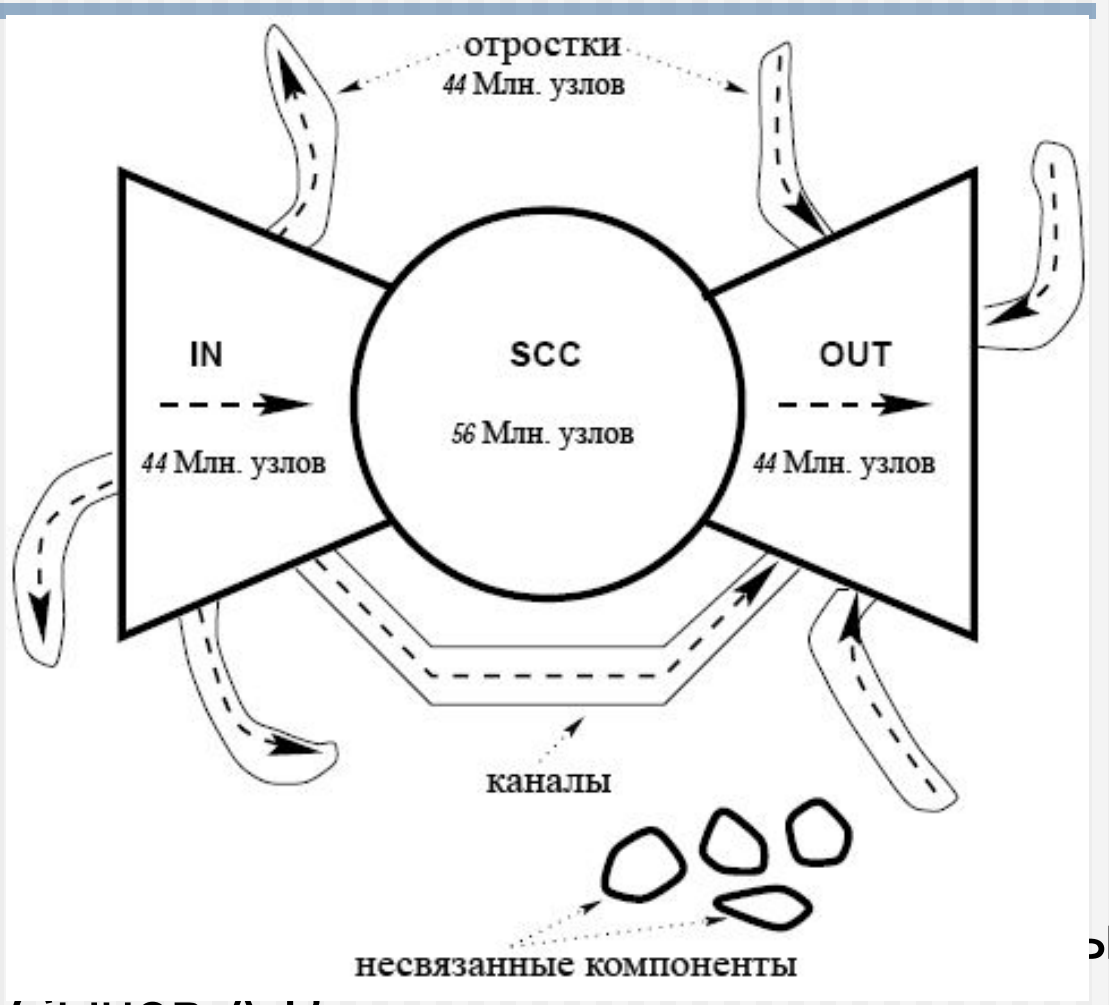


Модель предпочтительного прикрепления



ы.

Модель веб-графа “бабочка”



ы.

Модель “бабочка”

- В 1999 г. Было проведено исследование структуры веб-графа, содержащего около 200 млн. узлов. В результате исследования было обнаружено центральное сильной связное ядро (SCC), подграф, содержащий только направленные ссылки на ядро (IN), подграф, содержащий только направленные ссылки из ядра (OUT), относительно изолированные “отростки”, связанные с одной из трех крупных компонент, названных выше. Имелись также полностью изолированные компоненты, не имевшие связей с названными выше компонентами.

Веб-сообщества

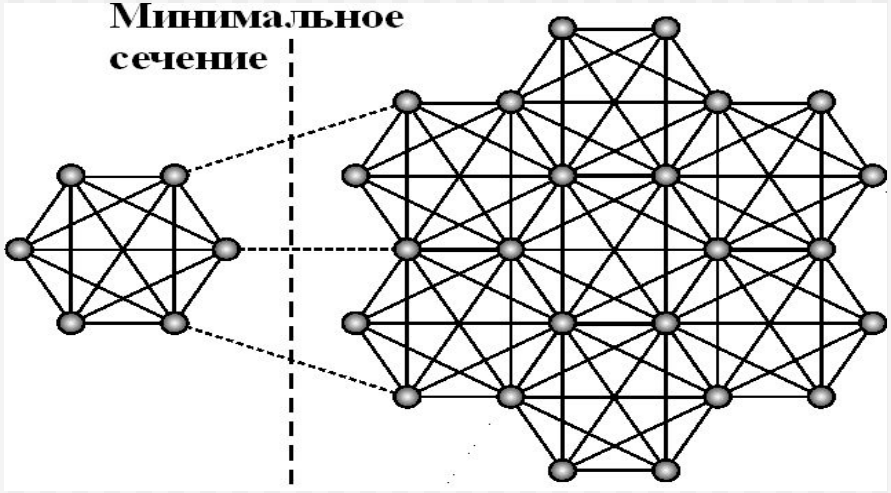
- Неформально *веб-сообщество* определяется как подграф веб-графа, в котором плотность внутренних связей превышает плотность внешних связей.
- Формальное определение: **Веб-сообщество** есть подмножество вершин $C \subset V$, таких, что для всех вершин $v \subset C$, v имеет множество рёбер, соединяющих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

“Зерновые” веб-ресурсы

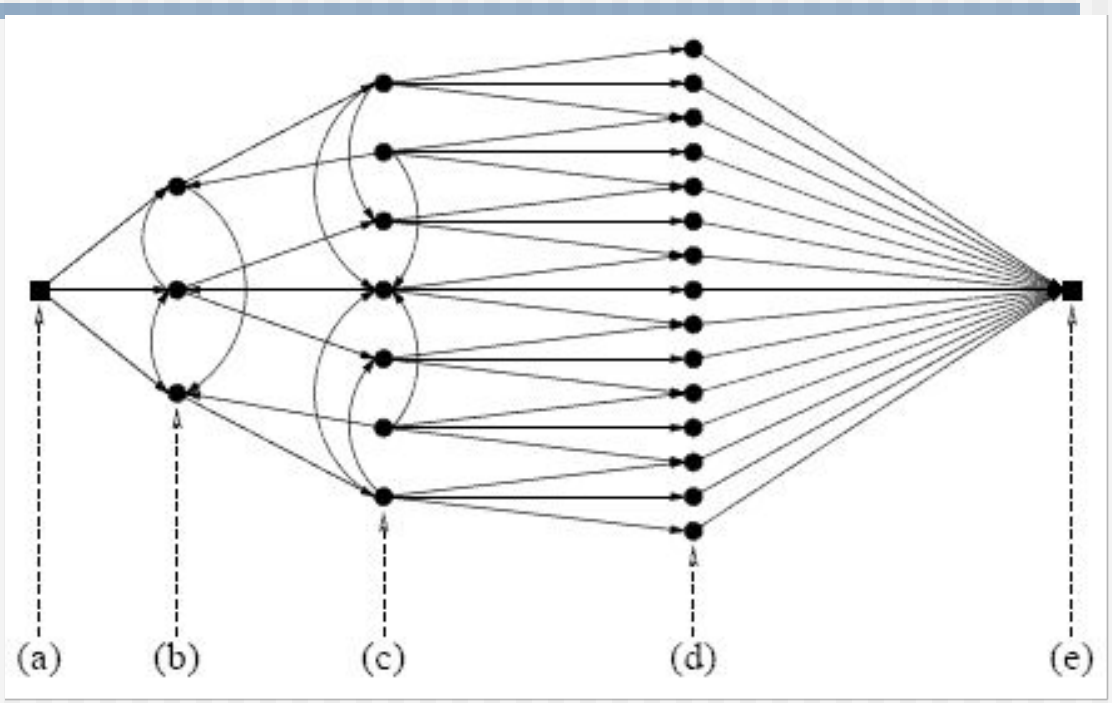
- Тем не менее, если исходить из факта существования одного или более “*зерновых*” веб-ресурсов и использовать систематические закономерности в структуре веб-графа, задача может быть сформулирована в виде, который позволяет эффективно идентифицировать веб-сообщества. Под “*зерновым*” понимают веб-ресурс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

Веб-сообщества

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



Направленное извлечение сообщества и построение графа



(a) виртуальный исток; (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.

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*”.

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 скаченные 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.

Алгоритм для выделения веб-сообществ (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 = 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

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.

Литература

- A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J.Wiener. "*Graph structure in the Web: Experiments and models*". In WWW9, pp. 309–320, Amsterdam, May 2000. Elsevier Science.
- S. Chakrabarti "*Mining the Web. Discovering Knowledge from Hypertext*" Data. Morgan Kaufmann Publishers, 2003.
- G. W. Flake, S. R. Lawrence, C. L. Giles, and F. M. Coetzee. *Self-Organization and Identification of Web Communities*. IEEE Computer, 35(3), 66–71, 2002
- N. Imafuji and M. Kitsuregawa, "*Finding a web community by maximum flow algorithm with hits score based capacity*." In 8th International Conference on Database Systems for Advanced Applications, pp. 101–106, 2003.

Литература

- J. Kleinberg, S. Lawrence. "*The structure of the Web*" // Science, vol 294, November 2001. pp. 1849-185.
- Майника Э. Алгоритмы оптимизации на сетях и графах. – М.: «Мир», 1981. – 323 с.
- 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.
- R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. "*Trawling the Web for emerging cyber-communities*". In Proceedings of the 8th International World Wide Web Conference, pp. 1481–1493, 1999.
- A.P. Dempster, N.M. Laird, and D.B. Rubin. "*Maximum likelihood from incomplete data via the EM algorithm*". J. R. Statist. Soc. B, 39:185-197, 1977.
- Д.Д. Козлов, А.А. Белова. "*Исследование эффективности применения методов совместного анализа текстов и гиперссылок для поиска тематических сообществ*".