Математические модели согласованного поведения малых Интернет-сообществ

Печников А.А.
Институт прикладных математических исследований Карельского научного центра Российской академии наук
pechnikov@krc.karelia.ru

Чуйко Ю.В.
Институт прикладных математических исследований Карельского научного центра Российской академии наук
julia@krc.karelia.ru

Аннотация

Одна из проблем поисковых алгоритмов, учитывающих наличие внешних ссылок на документ или сайт, заключается в возможности искусственного увеличения ссылочной популярности путем обмена ссылками. Основной целью данной работы является исследование согласованного поведения так называемых «малых профессиональных Интернет-сообществ» на основе двух предлагаемых математических моделей, оптимизация целевых функций которых трактуется как согласованное поведение.

Исследования, проведенные на базе предоставленного Яндексом хостграфа, позволили доработать данные модели, приблизив их к реальному описанию поведения сообществ в Интернете.

Исследования двадцати сообществ, отобранных экспертным способом, выявили ряд сообществ, с поведением, близким к согласованному.

1. Введение

1.1 Постановка задачи

Одна из проблем поисковых алгоритмов, учитывающих наличие внешних ссылок на документ или сайт, заключается в возможности искусственного увеличения ссылочной популярности путем обмена ссылками. Для решения этой проблемы обычно используют такие способы, как исключение сайтов из индекса, наложение фильтра на исходящие ссылки с сайтов и др. Однако, обмен ссылками, даже и договорной, не всегда является накруткой – многие владельцы сайтов обмениваются ссылками с действительно качественными ресурсами в своей тематике и не заслуживают штрафных санкций.

В качестве работ из этой области можно привести работу [7], посвященную расчетам ссылочной популярности, и [6], в которой рассматриваются вопросы идентификации веб-сообществ.

В данной работе в качестве объекта исследования выделены так называемые "малые профессиональные Интернет-сообщества", примерами которых могут служить интернет-ресурсы университетов, институтов и научных центров, предприятий одной сферы деятельности и т.п. (Термин «малые» обозначает небольшое количество участников и не имеет отношения к объемам интернет-ресурсов.)

Очевидно, что вследствие профессионального взаимодействия участники таких сообществ имеют потенциальную возможность согласовать свои действия по увеличению ссылочной популярности. При этом согласованность их действий может трактоваться как попытка увеличения ссылочной популярности «слабых» за счет «сильных», а поэтому математическая модель должна иметь соответствующую целевую функцию.

Целями данной работы являются:

- доработка и развитие математических моделей согласованного поведения малых сообществ с учетом реальных данных,
- апробация адаптированных математических моделей на фактическом материале,
- содержательный анализ согласованного поведения сообществ в Интернете в зависимости от таких факторов, как, например, принадлежности к определенной тематической области.

В авторских работах [3,4] рассматриваются математические модели согласованного поведения в Интернете, которые мы кратко опишем ниже.

Рассмотрим малое Интернет-сообщество (далее – сообщество) со следующими характеристиками:

n – количество участников сообщества,

c_i – значимость i-го участника, $c_i \ge 0, i = \overline{1,n}$,

m_i - количество прямых ссылок от i-го участника на других участников сообщества, $m_i > 0, \forall i = \overline{1,n}$.

Определим матрицу ссылок $X = \left(x_{ij}\right)$, $i, j = \overline{1,n}$, где x_{ij} = 1, если существует ссылка от i-го участника к j-му, и x_{ij} = 0, если ссылки не существует.

Система ограничений выглядит следующим образом:

$$x_{ii} = 0,\ i = \overline{1,n} \qquad (1)$$

$$x_{ij} = 0,1,\ i = \overline{1,n},\ j = \overline{1,n} \qquad (2)$$

$$\sum_{j=1}^{n} x_{ij} \le m_i,\ i = \overline{1,n} \qquad (3)$$

$$\sum_{j=1}^{n} x_{ij} \geq 1,\ i = \overline{1,n} \qquad (4)$$

Ограничения (1) учитывают, что ресурсу нельзя дать ссылку на самого себя, (2) - ссылки либо существуют, либо нет, (3) - количество исходящих ссылок ограничено некоторыми реальными соображениями (наверняка оно меньше *n*).

Ограничения (4) определяют один из принципов кооперативного подхода: участником сообщества может быть ресурс, который обязательно делится своей значимостью с другими участниками, поэтому от него должна быть хотя бы одна ссылка. Обратное, вообще говоря, не обязательно: если значимость участника изначально достаточно высока, то на него может и не быть ссылок в рамках сообщества.

Принципиальным моментом является определение функции приращения значимости. В современной литературе принят подход, основанный на линейном представлении функции приращения (или на решении системы линейных уравнений) [7,8]. Вкратце основные предположения, положенные в основу вычисления таких функций, заключаются в следующем:

- чем больше ссылок на ресурс, тем он становится «значимее»,
- чем больше значимость ресурса *i*, тем больше возрастает значимость ресурса *j*, если x_{ij} =1,
- чем больше исходящих ссылок от ресурса *i*, тем меньше приращение значимости каждого ресурса *j*, для которого x_{ij} =1.

Будем считать, что изменение значимости *j*-го участника сообщества может быть представлено следующей формулой:

$$\forall j = \overline{1,n}:\ \widetilde{c}_j = c_j + \sum_{i=1}^{n} x_{ij} \cdot c_i \cdot \alpha_i \qquad (5).$$

Здесь α_i – коэффициент, который содержательно означает, что при установлении ссылки от *i*-го участника на *j*-го (т.е. при x_{ij} =1), значимость *j*-го участника возрастает на некоторую часть значимости *i*-го участника, α_i >0. Уточнение вида α_i применительно к теме нашего исследования будет приведено в пункте 3.1.

Пусть *F(X)* - функция, характеризующая некоторый интегральный показатель значимости всех участников сообщества и зависящая от того, каким образом расставлены ссылки между ее участниками, т.е. от матрицы *X*. Тогда задача заключается в нахождении матрицы *X*, удовлетворяющей заданным ограничениям и доставляющей оптимальное значение целевой функции: $F(X) \underset{x_{ij}}{\rightarrow} opt$.

В работе [3] в качестве целевой функции взята функция среднеквадратичного отклонения по всем участникам. В этом случае оптимизационная задача имеет следующий вид:

$$F1(X) = \sum_{j=1}^{n} \left(\frac{\sum_{k=1}^{n} \widetilde{c}_k}{n} - \widetilde{c}_j \right)^2 \underset{x_{ij}}{\rightarrow} \min \qquad (6)$$

при ограничениях (1) – (4).

Договоренность участников сообщества можно сформулировать так: распределение ссылок среди участников сообщества должно привести к минимальному отклонению полученных значимостей каждого участника от нового среднего значения по всему сообществу.

В работе [4] в качестве целевой функции рассматривается линейная функция вида:

$$F2(X) = \sum_{j=1}^{n} \widetilde{c}_j \cdot \lambda_j \underset{x_{ij}}{\rightarrow} \max \qquad (7),$$

где коэффициенты $0 < \lambda_j \leq 1$ имеют следующий содержательный смысл, - чем больше начальное значение c_j, тем меньше значение λ_j.

В этом случае договоренность участников сообщества можно сформулировать так: распределение ссылок должно привести в первую очередь к увеличению значимости наименее значимых ресурсов, при этом суммарный прирост значимости по всем участникам системы должен быть максимальным.

Вид коэффициентов λ_j был определен как $\lambda_j = 1/c_j$, т.е. обратно пропорционально имеющейся значимости участника. В этом случае, при подстановке в (7) формул (5) и после ряда преобразований, целевая функция (7) приобретает вид:

$$F2(X) = \sum_{i=1}^{n} \sum_{j=1}^{n} x_{ij} \cdot \frac{c_i}{c_j} \cdot \alpha_i \underset{x_{ij}}{\rightarrow} \max \qquad (8).$$

Обозначим через *K* среднее значение получаемых значимостей, $K = \sum_{i=1}^{n} \widetilde{c}_i / n$. Тогда, если значимость некоторого участника системы изначально больше, чем получаемое в результате решения задачи *K*, то на него не нужно устанавливать ссылки, то есть

$$\sum_{i=1}^{n} x_{ij} = 0, \forall j : c_j \geq K \qquad (9).$$

1.2 Уточнение и конкретизация моделей для целей исследования

В случае нашего исследования ограничения (3) и (4) заменяются на ограничения вида:

$$\sum_{j=1}^{n} x_{ij} = m_i,\ i = \overline{1,n} \qquad (10).$$

Содержательная трактовка этой замены заключается в следующем: если поведение участников некоторого выбранного сообщества действительно является согласованным, то, в первую очередь, они должны договориться о количестве прямых ссылок $m_i > 0$, исходящих от каждого участника на других участников сообщества. Тогда количество исходящих ссылок на других участников сообщества является константой для каждого участника, что и фиксируется ограничениями (10).

Таким образом, получаем две модели согласованного поведения, адаптированные по отношению к целям нашего исследования:

Модель 1 с функцией приращения значимости (5), целевой функцией (6) и ограничениями (1), (2), (10), и

Модель 2 с функцией приращения значимости (5), целевой функцией (8) и ограничениями (1), (2), (9), (10).

Более подробное формальное описание ***Моделей 1 и 2*** дается в разделе 3.

2. Идея исследования

Как уже было отмечено ранее, считаем, что участники Интернет-сообществ имеют потенциальную возможность согласовать свои действия по увеличению ссылочной популярности вследствие профессионального взаимодействия. В качестве критериев, определяющих степень согласованности поведения сообществ, будем использовать отклонения значений функционалов *F1(X)* и *F2(X)*, вычисленных для реальных матриц сообществ, от их же значений на матрицах оптимальных решений.

Фактически речь идет о том, насколько велико отклонение функционалов *F1(X)* и *F2(X)*, вычисленных на реальной матрице $X^{real} = \left(x_{ij}^{real}\right)$, от оптимальных значений, вычисленных на матрицах X^{opt1}, X^{opt2} и отличающихся от X^{real} с точностью до перестановки единиц в строках.

Суть экспериментов по апробации обеих моделей по данным каждого Интернет-сообщества заключается в следующем:

1. экспертным путем (посредством перечисления интернет-ресурсов) выделяется некоторое сообщество,

2. «вручную», - то есть с помощью известных механизмов Яндекса [1], - определяются тематические индексы цитирования (тИЦ) участников сообщества, принимаемые в качестве значений $\widetilde{c}_i \geq 0, \forall i = \overline{1,n}$,

3. с использованием набора данных Яндекса "Хостграф" [2] определяются:

- наличие/отсутствие связей между участниками сообщества,

- характеристики констант n, m_i, $i = \overline{1,n}$,

- полное количество исходящих ссылок от каждого участника сообщества (далее оно будет обозначено как L_i, $i = \overline{1,n}$),

4. строится реальная матрица $X^{real} = \left(x_{ij}^{real}\right)$, $i, j = \overline{1,n}$,

5. вычисляются значения функционалов *F1(X)* и *F2(X)* на матрице X^{real},

6. перевычисляются исходные значения $c_i, i = \overline{1,n}$ на основе имеющихся значений $\widetilde{c}_i, i = \overline{1,n}$ (более подробно о механизме перевычисления – в пункте 3.5),

6. находятся оптимальные решения для Моделей 1 и 2, т.е. $F1(X^{opt1})$ и $F2(X^{opt2})$,

7. на основании отклонения реального значения целевой функции от оптимального, делается вывод о согласованном (или несогласованном) поведении участников.

3. Описание методов, алгоритмов и экспериментов

3.1 Анализ и уточнение математических моделей согласованного поведения

Продолжим рассмотрение ***Модели 1*** с функцией приращения значимости (5), целевой функцией (6) и ограничениями (1), (2), (10), и ***Модели 2*** с функцией приращения значимости (5), целевой функцией (8) и ограничениями (1), (2), (9), (10).

Возвращаясь к функции приращения значимости (5), будем полагать, что на фиксированном временно**м** интервале рассмотрения задачи согласованного поведения новые внешние ссылки на участников сообщества не появляются и имеющиеся не исчезают; таким образом, изменение функции значимости зависит только от расстановки ссылок между участниками сообщества друг на друга.

Введем следующие обозначения:

$L_i, i = \overline{1,n}$ - общее количество исходящих ссылок от *i*-го участника сообщества,

$\overline{L}_i, i = \overline{1,n}$ - количество ссылок, исходящих от *i*-го участника сообщества без учета ссылок, сделанных на других участников сообщества, то есть с учетом (10), $\overline{L}_i = L_i - m_i, i = \overline{1,n}$.

Возвращаясь к формуле (5), уточним вид α_i.

Поскольку, как было отмечено в п.1.1, увеличение количества исходящих ссылок уменьшает приращение значимости ресурса, выразим $\alpha_i = \beta / L_i$, считая, что β - параметр конкретного алгоритма вычисления значимости, существенно зависящий от поисковой машины, в нашем случае – от алгоритма вычисления

тематического индекса цитирования (тИЦ) Яндекса [1]. Далее мы будем считать β равным единице. (Содержательное обоснование такого выбора приводится в пункте 3.2). Тогда изменение значимости *j*-го участника будет выражаться следующей формулой:

$$\forall j = \overline{1,n} : \widetilde{c}_j = c_j + \sum_{i=1}^{n} \frac{x_{ij}}{L_i} \cdot c_i \quad (11).$$

Суммированием по *j* формул вида (11), с учетом того, что из (4) знаменатель никогда не равен нулю и замены, следующей из (10), несложно получить значение *K*, не зависящее от матрицы *X*, а именно:

$$K = \left(\sum_{j=1}^{n} c_j + \sum_{i=1}^{n} \frac{m_i \cdot c_i}{L_i} \right) / n \quad (12).$$

Данный результат существенно упрощает вычисление значений целевой функции в ***Модели 1***, а также работу алгоритма нахождения оптимального решения для ***Модели 2***.

Завершая тему формального определения и анализа используемых математических моделей, соберем их в одном месте.

Модель 1.

$$FI(X) = \sum_{j=1}^{n} \left(K - (c_j + \sum_{i=1}^{n} \frac{x_{ij}}{L_i} \cdot c_i) \right)^2 \to \min_{x_{ij}} \quad (13)$$

при ограничениях

$$x_{ii} = 0,\ i = \overline{1,n},$$

$$x_{ij} = 0,1\ ,\ i = \overline{1,n},\ j = \overline{1,n},$$

$$\sum_{j=1}^{n} x_{ij} = m_i,\ i = \overline{1,n}.$$

Модель 2.

$$F2(X) = \sum_{i=1}^{n} \sum_{j=1}^{n} x_{ij} \cdot \frac{c_i}{c_j} \cdot \frac{1}{L_i} \to \max_{x_{ij}} \quad (14)$$

при ограничениях

$$x_{ii} = 0,\ i = \overline{1,n},$$

$$x_{ij} = 0,1,\ i = \overline{1,n},\ j = \overline{1,n},$$

$$\sum_{j=1}^{n} x_{ij} = m_i,\ i = \overline{1,n},$$

$$\sum_{i=1}^{n} x_{ij} = 0,\ \forall j : c_j \geq K.$$

3.2 Некоторые соображения относительно значений исходных данных моделей

Прежде чем коснуться вопроса о поведении параметра β, хотелось бы остановиться на не менее важном и более общем вопросе о потенциальной возможности участников сообщества получить конкретные данные о значениях констант, входящих в модели согласованного поведения.

Вопрос о количестве исходящих ссылок в рамках самого сообщества $m_1,...,m_n$ – это вопрос регламента, который должен определяться самими участниками сообщества.

Значимость ресурсов определяется достаточно просто через тИЦ [1] и решение системы линейных уравнений (пункт 3.5). Единственное замечание по этому поводу заключается в том, что в случае сообщения Яндекса о том, что *«Индекс цитирования (тИЦ) ресурса меньше 10»*, остается возможность самостоятельного определения $\widetilde{c}_i$. Авторы в большинстве случаев принимали $\widetilde{c}_i = 0$. И лишь в силу особого интереса исключение составили сайты Карельского научного центра РАН, где в этом случае было принято решение взять минимальные $\widetilde{c}_i$ равными 5.

Количество ссылок, исходящих с *i*-го сайта ($L_i, i = \overline{1,n}$), может быть определено через поиск в HTML-тексте сайта гипертекстовых ссылок типа ***href.***

Ответ на вопрос о том, могут ли участники сообщества знать точное значение параметра β - категорическое нет. Первое объяснение такой категоричности заключается в том, что понять, как вычисляется параметр β, - это во многом понять механизм поведения поисковой машины, что является коммерческим секретом любой серьезной фирмы и уже поэтому достаточно хорошо защищен. Второе объяснение заключается в том, что механизмы ранжирования поисковых машин подвергаются постоянной корректировке, поэтому затраты на понимание этих механизмов на сегодняшний день вряд ли окупятся завтра. Эти объяснения подтверждаются публикацией [5].

Попытки авторов оценить значение β для ряда частных случаев приводят к достаточно большому разбросу значений в пределах от 0.3 до 2.83, что не слишком противоречит выбранному значению β=1.

3.3 Процедура отбора сообществ и их основные характеристики

Сообщества, рассмотренные в рамках проведенного исследования, можно классифицировать по следующим группам (в скобках – количество сообществ):

Университетские и научные ресурсы (4);
Сайты органов власти (2);
СМИ (2);
Баннерообменные сети и Web-ринги (4);
Участники ассоциаций (1);
Сайты из разделов Каталога Яндекса (5);
Сайты поисковых систем (1);
Выявленное сообщество на Narod.ru (1).

Первая группа была сформирована на основе данных, взятых на сайтах, которые являются в некотором смысле, «вышестоящими» по отношению к участникам сообществ. То же принцип отбора был принят и к сообществу сайтов

Комсомольской правды.

Сайты электронных СМИ Карелии отобраны на основе личных знаний авторов о них и проверены дополнительными поисками в Интернете. Сайты баннерообменных сетей и Web-рингов были включены по предложению экспертов Яндекса об исследовании поведения так называемых «спам-сообществ». Кроме того, включение сообщества сайтов участников Ассоциации организаций и предприятий целлюлозно-бумажной промышленности в некотором смысле «коррелирует» с одним из «спам-сообществ» - Целлюлозно-Бумажной Баннерной Сетью.

Сайты из разделов Каталога Яндекса выбирались по предпочтениям авторов, при этом в предполагаемое сообщество включались первые 20 сайтов соответствующего раздела Каталога (лишь в разделе «Интернет детям» сайтов оказалось меньше). Авторы не могли обойти рассмотрение вопроса о том, являются ли «сообществом» сами поисковые системы. Сообщество на Narod.ru было выявлено, можно сказать, случайно, посредством просмотра данных о ссылающихся друг на друга сайтах.

Всего вручную было отобрано 503 сайта, составивших 20 предполагаемых сообществ. Термин «предполагаемое сообщество» означает, что на этапе отбора сообществ в них были включены участники, которые не удовлетворяют ряду требований к участнику сообщества и по правилам, изложенным в пункте 3.6, впоследствии будут исключены из сообщества. Более подробная информация приведенf в сводной Таблице 1.

3.4 Автоматизация работы с хостграфом

При проведении экспериментов были использованы предоставленные компанией Яндекс наборы данных о структуре хостграфа русскоязычного Интернета, то есть о взаимных ссылках сайтов друг на друга, по состоянию на 7 декабря 2007 года.

Данные представляют собой два текстовых файла: первый (вспомогательный) содержит соответствия символьных имен хостов их идентификаторам в базе данных Яндекс, во втором хранится структура хостграфа в виде списка хостов, для каждого из которых приведен набор идентификаторов хостов, на содержимое сайтов которых данный сайт имеет исходящие ссылки.

В файлах содержится информация о 2,714,279 сайтах, занимая около 8 гигабайт дискового пространства. Большой размер текстовых файлов и необходимость их полного сканирования каждый раз при выполнении операций выборки данных затрудняют и существенно замедляют обработку информации.

Для ускорения процесса получения необходимых выборок исходные данные были помещены в базу данных **mysql**, что позволило использовать предоставляемые этой СУБД возможности индексирования и быстрого поиска по запросам.

Таблица 1. Список сообществ

№	Наименование сообщества	НЧ*	ОС	КС	КВ
1	Баннерообмен-ная сеть Medlinks	84	15	28	15043
2	Университеты РФ	60	38	223	27347
3	Кафедры факультета ВМиК МГУ	7	0	--	--
4	Электронные СМИ Карелии	12	4	9	205
5	Министерства Российской Федерации	17	7	18	1750
6	Сайты КарНЦ РАН	20	9	20	241
7	Комсомольская правда. Региональные сайты	33	12	34	4274
8	Сайты муниципалите-тов Саратовской области	29	4	6	181
9	Web-ресурсы ПетрГУ	43	18	35	2649
10	Российские поисковые системы	6	0	--	--
11	Кольцо сайтов "Законы, законодательст-во и право"	22	16	34	8269
12	Баннерная сеть Ket.Ru	39	7	12	630
13	Журналы для женщин	20	14	64	86509
14	Ассоциация ЦБП	25	7	15	11527
15	Целлюлозно-Бумажная Баннерная Сеть	12	5	10	91
16	Рок-музыка	20	16	63	35566
17	Религия. Православие	20	20	177	10101
18	Группа сайтов, выявленная на Narod.ru	7	4	12	30
19	Интернет детям	7	0	--	--
20	Каталоги	20	6	7	103498

* **Обозначения столбцов:**

НЧ – начальное количество участников,

ОС - оставшееся количество участников,

КС – общее количество ссылок исходящих ссылок от всех оставшихся участников собой,

КВ - общее количество исходящих ссылок от всех оставшихся участников.

Для того чтобы обеспечить пользователю удобный доступ к данным, на языке PHP был реализован web-интерфейс, позволяющий выполнять следующие операции:

- находить соответствующий хост по его идентификатору;
- находить идентификатор хоста;
- находить количество исходящих ссылок с указанного хоста;
- находить количество хостов, ссылающихся на указанный хост;
- делать выборку хостов, на которые ссылается указанный хост;
- делать выборку хостов, ссылающихся на указанный хост.

Операция выборки информации обо всех исходящих ссылках с указанного пользователем хоста на компьютере с характеристиками 2.20GHz/240Mb/40Gb в среднем выполняется за 7 минут.

3.5 Определение исходных данных для алгоритмов решения оптимизационных задач

Реализованный web-интерфейс для обработки хостграфа позволил определить списки всех хостов, ссылающихся на хосты сайтов, являющихся участниками предполагаемых сообществ. Далее, посредством «ручной» чистки для каждого предполагаемого сообщества были получены списки всех участников сообщества, ссылающихся на заданного участника, что явилось основой для формирования рабочих матриц сообществ (см. далее пункт 3.6).

Кроме того, с помощью web-интерфейса определялось общее количество исходящих ссылок с каждого хоста, входящего в одно из сообществ.

Как уже было сказано, в качестве значений $\widetilde{c}_i, i=\overline{1,n}$ были приняты значения тИЦ соответствующих сайтов. Однако, $\widetilde{c}_i, i=\overline{1,n}$ - это значения, которые получаются уже как результат согласованных действий участников сообщества (если гипотеза о согласованном поведении верна). Поэтому в качестве исходных значений $c_i, i=\overline{1,n}$ должны быть взяты тИЦ, которые находятся как решение системы линейных уравнений, получаемых из (11):

$$\begin{cases} c_1 + \sum_{i=1}^{n} \frac{x_{ij}}{L_i} \cdot c_i = \widetilde{c}_1 \\ \ldots\ldots\ldots\ldots \\ c_k + \sum_{i=1}^{n} \frac{x_{ij}}{L_i} \cdot c_i = \widetilde{c}_k \\ \ldots\ldots\ldots\ldots \\ c_n + \sum_{i=1}^{n} \frac{x_{ij}}{L_i} \cdot c_i = \widetilde{c}_n \end{cases}.$$

(Следует отметить, что для сообщества сайтов Карельского научного центра РАН вновь было сделано исключение, поскольку в результате решения системы были получены отрицательные значения для некоторых c_i, замененные на нули для ***Модели 1*** и на 1 для ***Модели 2.*** Это является также и лишним подтверждением более сложной модели для вычисления приращений значимости, и, в первую очередь, подтверждает рассуждения о сложности определения β.)

3.6 Рабочая матрица предполагаемого сообщества и правила исключения

Итак, для всех участников каждого предполагаемого Интернет-сообщества строилась начальная матрица *X*, фактически являющаяся матрицей смежности графа, вершинами которого выступают участники сообщества.

Уже на первом этапе из матрицы были исключены строки (и соответствующие столбцы), для которых $\widetilde{c}_i = 0$, поскольку вклад этих участников не изменяет значение целевой функции (13) и, более того, не допустим для целевой функции (14). (Об исключениях для сообщества сайтов Карельского научного центра РАН, уже было сказано ранее).

Затем, в соответствии с ограничениями (10) производилось исключение строк (и соответствующих столбцов), для которых $\sum_{j=1}^{n} x_{ij} = 0$. Процедура имеет итерационный характер, поскольку удаление строк и столбцов может вести к тому, что условие $\sum_{j=1}^{n} x_{ij} = 0$ выполняется вновь на уже на уменьшенной матрице.

Таким образом, были получены матрицы сообществ, которые явились малыми Интернет-сообществами как объектами для дальнейших исследований на предмет согласованного поведения.

Интересно отметить, что три предполагаемых сообщества на этом этапе были исключены из дальнейшего рассмотрения, как не имеющие ссылок между собой (Кафедры факультета ВМиК МГУ, Российские поисковые системы и Интернет детям). Если по поводу второго и третьего сообщества можно привести ряд аргументов в поддержку такого поведения, то по поводу кафедр факультета вычислительной математики и кибернетики нет ничего, кроме изумления.

3.7 Алгоритмы решения оптимизационных задач по Моделям 1 и 2

Исходными данными для алгоритмов решения задач являются:

n - количество сайтов сообщества;

вектор найденных «начальных» значений тИЦ $c_1,\ldots,c_n$;

вектор $m_1,\ldots,m_n$, каждый *i*-й элемент которого – регламентируемое количество исходящих ссылок со страниц *i*-го сайта на сайты-участники сообщества;

вектор $L_1,\ldots,L_n$ - общее число исходящих ссылок от каждого сайта.

Переменными являются элементы матрицы *X*. Алгоритмы для двух задач отличаются только целевыми функциями и наличием во второй модели ограничений, связанным со средним значением значимостей *K*.

Алгоритм решения задач является рекурсивным ограниченным перебором перестановок m_i единиц в каждой *i*-й строке матрицы *X*. На каждом шаге рекурсии, с учетом ограничений задачи, заполняется очередная ячейка матрицы *X* и вычисляется нижняя оценка значения целевой функции. Если она меньше минимума (для первой модели) или больше максимума (во второй модели) целевой функции из вычисленных ранее, то выполняется рекурсивный переход алгоритма вглубь к заполнению следующей ячейки. Для каждого случая полностью заполненной матрицы вычисляется значение целевой функции, и, если оно лучше имеющегося минимума из вычисленных ранее, то это значение и запоминается как минимум.

Описанный алгоритм был реализован в виде программы на языке Java.

3.8 Результаты

Результаты расчетов по ***Моделям 1 и 2*** приведены в Таблице 2. Для удобства сравнения с Таблицей 1 здесь сохранены строки 3, 10 и 19, хотя об исключении из дальнейшего рассмотрения этих трех сообществ уже было сказано ранее.

Значения функционалов *F1(X)* и *F2(X)* на матрице X^{real} вычислялись в MS Excel. Причем сразу же пришлось отказаться от ограничения $\sum_{i=1}^{n} x_{ij} = 0,\ \forall j : c_j \geq K$ в ***Модели 2***, поскольку в матрицах X^{real} всех 17 сообществ имелись ссылки от «слабого» к «сильному», нарушающие данное ограничение и делающие систему ограничений несовместной.

Таким образом, на этапе расчетов по построенным моделям произошло еще одно изменение ***Модели 2***. С учетом отбрасывания данных ограничений алгоритм нахождения оптимального решения для ***Модели 2*** упрощается и может быть посчитан в MS Excel.

Оптимальные значения $F1(X^{opt1})$ для каждого из 17 сообществ вычислялись с помощью программной реализации алгоритма, описанного в пункте 3.7.

Таблица 2. Результаты расчетов

№	Наименование сообщества	откл. по ***Модели 1***	откл. по ***Модели 2***
1	Баннерообменная сеть Medlinks	0.998	10.7
2	Университеты РФ	0.996	5.5
3	Кафедры факультета ВМиК МГУ	----	----
4	Электронные СМИ Карелии	0.992	1.1
5	Министерства Российской Федерации	0.925	2.2
6	Сайты КарНЦ РАН	0.905	1.3
7	Комсомольская правда. Региональные сайты	0.998	11.1
8	Сайты муниципалитетов Саратовской области	0.967	1.9
9	Web-ресурсы ПетрГУ	0.997	10.7
10	Российские поисковые системы	----	----
11	Кольцо сайтов "Законы, законодательство и право"	0.975	6.8
12	Баннерная сеть Ket.Ru	0.994	4.0
13	Журналы для женщин	0.997	1.2
14	Ассоциация ЦБП	0.893	112.9
15	Целлюлозно-Бумажная Баннерная Сеть	0.791	8.7
16	Рок-музыка	0.997	1.2
17	Религия. Православие	0.997	5.3
18	Группа сайтов, выявленная на Narod.ru	1.000	1.0
19	Интернет детям	----	----
20	Каталоги	0.999	1.3

В качестве отклонений по моделям приведено отношение $F1(X^{opt1}) / F1(X^{real})$ и $F2(X^{opt2}) / F2(X^{real})$ соответственно.

4. Заключение

Очевидного лидера – сообщество «Группа сайтов, выявленная на Narod.ru», - мы исключим из дальнейшего обсуждения, во-первых, по причине всего лишь 4 участников, а во-вторых, - полноты графа ссылок. Поведение этого сообщества согласовано с абсолютной очевидностью. Более того, соотношение ссылок между участниками сообщества и всех ссылок находится такой пропорции, как будто участники договорились не ставить «внешних необязательных ссылок».

Первый (и достаточно неожиданный для авторов) вывод заключается в том, что количество ссылок с сайта, проиндексированных Яндексом, столь велико. Например, сайты из сообщества «Журналы для женщин» вполне сравнимы по количеству исходящих ссылок с сообществом «Каталоги», а среднее количество исходящих ссылок, приходящееся на каждый из 503 исследованных сайтов равно 1500.

Вследствие этого, соотношение количества внутренних ссылок между участниками сообщества к общему количеству ссылок (даже без учета сообщества «Каталоги», где соотношение равно 14575) в среднем равно 200. Естественно, этот фактор существенно влияет на незначительный прирост значимости (см. формулу (11)).

Но именно это фактор легко объясняет столь идеальный результат отклонения $F1(X^{opt1})/F1(X^{real})$ для сообщества «Каталоги»: какими бы ни были согласованные действия, они дадут слишком малый реальный эффект.

Поэтому для адекватных выводов о согласованности поведения были отобраны только те сообщества, в которых, упрощенно говоря, было что распределять, то есть, во-первых, соотношение «внешних» ссылок к «внутренним» было не слишком большим, и, во-вторых, суммарный прирост тИЦ был не слишком маленьким.

Взяв указанное соотношение ссылок равным 100, минимальный суммарный прирост тИЦ равным 50 и отбросив сообщества, не попадающие в эти рамки, получим следующую пятерку сообществ, наиболее близких к согласованному поведению по ***Модели 1*** (в скобках – отклонение):

- Религия. Православие (0.997),
- Баннерная сеть Ket.Ru (0.994),
- Министерства РФ (0.925),
- Сайты КарНЦ РАН (0.905),
- Целлюлозно-Бумажная Баннерная Сеть (0.791).

В качестве вывода можно сказать, что поведение первых четырех сообществ является близким к согласованному.

Что касается ***Модели 2***, расстановка этих же сообществ по отклонению от оптимума дает следующий порядок:

- Сайты КарНЦ РАН (1.3),
- Министерства РФ (2.2),
- Баннерная сеть Ket.Ru (4.0),
- Религия. Православие (5.3),
- Целлюлозно-Бумажная Баннерная Сеть (8.7).

В данном случае, поведение, близкое к согласованному, фиксируется только у сообщества сайтов КарНЦ РАН.

5. Литература

[1] Индекс цитирования. [Электронный ресурс] – 2007. – Режим доступа: http://help.yandex.ru/catalogue/?id=873431.

[2] Наборы данных. [Электронный ресурс] – 2007. – Режим доступа: http://company.yandex.ru/grant/datasets_description.xml.

[3] Печников А.А. Задача рационального размещения ссылок в регламентируемой локализованной системе Интернет-ресурсов / А.А. Печников // Труды Института прикладных математических исследований КарНЦ РАН. – 2006. - Вып. 7. - С. 176-182.

[4] Печников А.А. Математические модели размещения ссылок в локализованной системе интернет-ресурсов / А.А.Печников // Системы управления и информационные технологии. – 2007. - №28. - С.92-96.

[5] Сегалович И. Мы умеем обходить, строить и отвечать на запросы примерно по 1 миллиарду документов. [Электронный ресурс] – 2006. – Режим доступа: http://webplanet.ru/news/interview/2006/2/6/ilya_segalovich.html.

[6] Сычев А.В Идентификация веб-сообществ в глобальной сети WAP-ресурсов ./ А.В. Сычев, М.М. Баженов // Информационные технологии. - 2006. - №6. - С. 38-44.

[7] Трофименко Е.А. Оптимизация расчета ссылочной популярности и учета ее при ранжировании результатов поиска / Е.А. Трофименко // Интернет-математика 2005. – 2005. - С.272-282.

[8] Brin S. The Anatomy of a Large-Scale Hypertextual Web Search Engine / S. Brin, L. Page // Computer Networks and ISDN Systems. – 1998. - №30. - P.107-117.

Mathematical models of coherent behavior for small web-communities

Andrey A. Pechikov, Julia V. Chuiko

One of the problems of the search algorithms that take into account the availability of external links to a document or site is the ability to artificially increase the link popularity by exchanging links. The main purpose of this work is to study the agreed conduct of the so-called "small professional web-communities" on the basis of two proposed mathematical models, optimization of target functions is treated as a coherent behavior.

Studies conducted on the basis provided by Yandex host-graph have completed the considered data models, bringing them to a real description of the Internet communities.

Studies twenty communities selected expert way, found a number of communities, to conduct close to the agreed.