О проблеме выбора зерновых ресурсов в задаче автоматического пополнения каталога веб-ресурсов на основе выявления компонент сильной связности с последующей контентной фильтрацией♣

Сычев А.В.

Воронежский государственный университет
sav@cs.vsu.ru

Баженов М.М

Воронежский государственный университет
bazhenov@vsu.ru

Аннотация

В статье представлены результаты исследования задачи автоматического пополнения каталога веб-ресурсов на основе нахождения компонент сильной связности в веб-графе, построенном исходя из зерновых ресурсов, взятых из уже существующих рубрик каталога.
Исследование основано на результатах экспериментов, проводившихся с веб-графами, реконструируемыми из реальной сети WWW. В качестве зерновых были использованы ресурсы из веб-каталога Яндекс.

1 Введение

Если рассматривать существующие ныне способы поиска релевантных документов в сети веб-ресурсов, то наилучшие результаты с точки зрения точности поиска достигаются при поиске по веб-каталогам, формируемым усилиями людей-экспертов (при достаточном размере каталога и его правильной структуре). При этом, однако, каталоги обладают рядом серьезных недостатков, такими как недостаточная полнота и актуализированность результатов поиска. Попытки автоматизации процедуры пополнения каталогов путем использования формальных методик оценки содержимого документов не могут обеспечить такого же уровня точности как в экспертном варианте, либо недостаточно эффективны с точки зрения затрат времени и вычислительных ресурсов.

Между тем “экспертная” оценка имеющихся в сети Веб ресурсов косвенно выполняется самими пользователями данной сети и опосредованно выражается в виде формируемых ими вручную гиперссылок. Однако, “язык” гиперссылок в HTML не обладает необходимой степенью выразительности для того, чтобы адекватно выразить интенцию их создателя. Поэтому подходы, на основе графа гиперссылок имеют серьезные недостатки и уязвимости, используемые поисковыми спамерами.

Тем не менее, подходы на основе идентификации веб-сообществ предоставляют фактически механизм интеграции распределенных в сети экспертных знаний, хотя и не являются при этом самодостаточными. Крайне необходимым дополнением является корректировка полученных таким способом результатов с помощью содержательного анализа выявленных веб-ресурсов.

Большинство работ по автоматической рубрикации веб-страниц для пополнения веб-каталога, основаны на методах автоматической обработки текста страниц. Некоторые из методов рассматриваются в работах [1-2].

Однако, как указывается в [3-4], привлечение информации о связи документов из каталога с документами, находящимися вне каталога, позволяет существенно улучшить результаты классификации. В [4] приведена идея комбинированного (учет гиперссылок + анализ контента) метода классификации веб-документов и результаты проведенного на его основе эксперимента.

С другой стороны, имеются подходы, основанные исключительно на анализе топологии гипертекстового графа, в частности, на основе решения задачи идентификации веб-сообществ. В работе [5] описана суть таких подходов, и подробно описан эксперимент по иерархической кластеризации веб-графа, построенного на основе ресурсов, взятых из каталога DMOZ. Результаты эксперимента показали высокую степень перекрытия полученных кластеров с рубриками DMOZ. Дополнительные модификации базового алгоритма поиска веб-сообществ рассмотрены в [6].

Авторами данной статьи было предложено в

[7] использовать уже имеющиеся в веб-каталоге ресурсы в качестве зерновых для процедуры идентификации тематических веб-сообществ. Сама процедура основывается на выявлении в сформированном на основе зерновых ресурсов локальном веб-графе компоненты сильной связности с использованием алгоритма Тарьяна. Список выявленных веб-ресурсов затем подвергается фильтрации на основе сравнения их содержимого с содержимым зерновых ресурсов. Для этого была разработана комплексная методика идентификации веб-сообществ на основе учёта как информации о гиперссылочной структуре Сети, так и о содержимом её ресурсов.

Как показали эксперименты, выбор зерновых ресурсов, необходимых для формирования веб-графа, существенным образом влияет на полученное в конечном итоге веб-сообщество и его качество. Поэтому была поставлена задача выяснить характер данной зависимости.

2 Методика и исходные данные

2.1 Описание методики

Идея предложенного подхода, заключается в последовательном выполнении нескольких этапов обработки информации, включающих: построение локального веб-графа на основе online-процедуры (с выполнения запроса с оператором [link=url] через систему Yandex.XML) и выбранных “зерновых” ресурсов; извлечение связанных компонент и укрупнение компоненты сильной связанности (КСС) до доменных узлов (что позволяет избавиться от гиперсвязей, формирующих навигационную структуру веб-сайтов и отрицательно влияющих на тематические связи); оценка укрупнённой компоненты на основе автоматической численной оценки с дальнейшим ранжированием узлов по результатам оценки. На рисунке 1 приведены основные составляющие комплексной методики (частные методики и алгоритмы), реализующие предлагаемый подход.

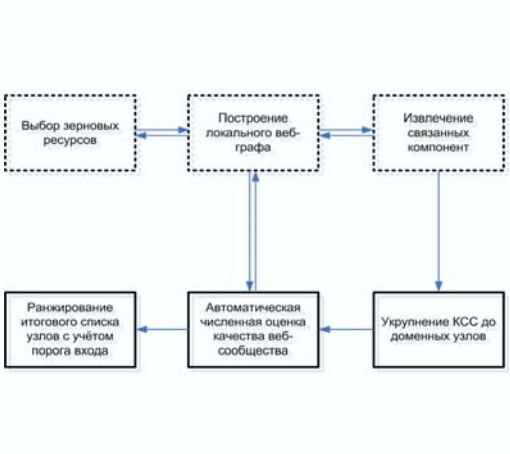

Рис. 1. Комплексная методика обнаружения и оценки веб-сообществ

Методика автоматической численной оценки качества веб-сообществ позволяет на основе “зерновых” ресурсов сформировать список ключевых слов тематики веб-сообщества в целом, после чего, идентифицированные члены веб-сообщества оцениваются на качество по соответствию списка своих ключевых слов и ключевых слов, характеризующих тематику веб-сообщества.

В итоге каждому члену из веб-сообщества ставится в соответствие определенный ранг, соответствующий оценке качества, полученной этим ресурсом.

2.2 Оценка качества КСС

Алгоритм численной оценки качества веб-сообщества (фактически членов КСС) без участия экспертов, основан на сравнении частотных характеристик текста веб-страниц с соответствующими частотными характеристиками зерновых ресурсов в сообществе. Диаграмма последовательности действий для этого алгоритма приведена на рисунке 2.

Общая идея алгоритма заключается в том, что из зерновых ресурсов выделяются ключевые слова, характеризующие их тематику, затем происходит объединение этих слов в список, представляющий собой тематику веб-сообщества в целом. Для зерновых ресурсов изначально определено, что они должны соответствовать тематике, что и позволяет применять такой подход. Далее, ключевые слова каждого члена веб-сообщества сравниваются с ключевыми словами тематики, и по степени их соответствия делается вывод о принадлежности к тематике в действительности.

Значение параметра t в алгоритме (своего рода топ-рейтинг ключевых слов) стоит выбирать, руководствуясь законами Ципфа. Численная оценка принимает значения из диапазона неотрицательных вещественных чисел [0;1].

Принятие решения о смысловом соответствии словоформ при сравнении ключевых слов может быть реализовано различными способами – от простого посимвольного сравнения, до усовершенствованных вариантов метрики TF*IDF, что возможно благодаря простоте получения насыщенного словаря в масштабах КСС.

2.3 Исходные данные

Экперимент проводился на примере каталога Яндекс (*http://yaca.yandex.ru*). В качестве зерновых ресурсов использовались сайты раздела каталога “*Учеба/Науки/Технические науки*” и раздела “*Учеба/Науки/Гуманитарные науки/История/История России*”. Характеристики подразделов приведены в таблицах 1 и 2. Объем исходных данных и выбор разделов каталога был обусловлен ограниченными техническими возможностями на момент проведения эксперимента.

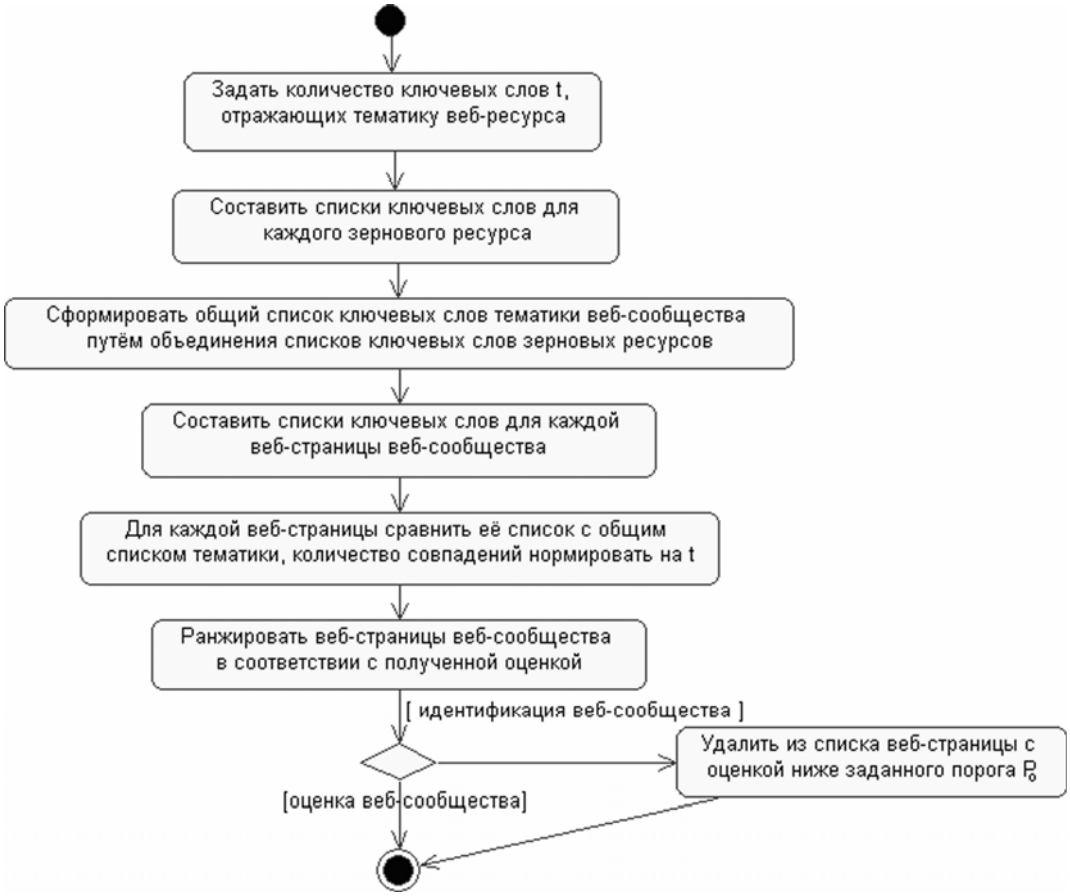

Рис. 2. Диаграмма последовательности действий алгоритма численной оценки качества веб-сообществ

Для хранения скачанных документов и последующих экспериментов с ними на жестком диске был организован кэш, содержащий папки (соответствующие именам доменов) и сами документы (внутри папок-доменов). Характеристики кэша для обеих рубрик приведены в таблице 3.

Таблица 1. Характеристики рубрик подрубрики каталога "*Учеба/Науки/Технические науки*".
Обозначения: A = *Авиация и космонавтика*, HT = *Высокие технологии*, CHE = *Вычислительная техника и электроника*, CSIS=Информатика, информационные системы, O = *Прочее*, U = *Универсальное*

№	Рубрика	Кол-во зерен,	Вх. ссылок, R	Доменов вх.ссылок, DR	R / DR
1	A	81	23422	2959	7,92
2	HT	25	2315	642	3,61
3	CHE	47	7228	1317	5,49
4	CSIS	102	13598	3022	4,50
5	O	49	10998	1869	5,88
6	U	18	1910	790	2,42

Таблица 2. Характеристики рубрик подрубрики каталога "*Учеба/Науки/Гуманитарные науки/История/История России* ".

Обозначения: Arc = *Археология*, WH = *Военная история*, G = *Генеалогия*, Anc = *Древний мир*, N = *Новая и новейшая история*, O = *Прочее*, M = *Средние века*, U = *Универсальное*, E = *Этнография и история народов*

№	Рубрика	Кол-во зерен	Вх. ссылок, R	Доменов вх.ссылок, DR	R / DR
1	Arc	4	250	104	2,40
2	WH	43	8250	1980	4,17
3	G	4	1402	177	7,92
4	Anc	2	12	9	1,33
5	N	23	3797	1385	2,74
6	O	11	1224	419	2,92
7	M	13	1729	573	3,02
8	U	20	2036	787	2,59
9	E	30	5035	858	5,87

Таблица 3. Размер кэша, сформированного модулем *WebCrawler*

Подрубрика веб-каталога	Общий размер, *Мбайт*	Папок (доменов), *тысяч*	Файлов, *тысяч*
История России	4696	30	200
Технические науки	13468	118	704

3 Результаты экспериментов

Машинный эксперимент проводился по схеме, представленной на рисунке 3.

3.1 Характеристики веб-графов

В таблицах 4a - 5 приведены характеристики построенных с помощью краулера веб-графов.

В таблицах 6-7 приведены данные по пересечению рубрик по входящим гиперссылок (для зерен) и по узлам в построенном веб-графе.

Как показали эксперименты, существенную часть построенного модулем *WebCrawler* веб-графа G(V,E) составляют узлы, используемые для навигации внутри домена и другие малорелевантные для тематики ресурсы. И хотя значительная их часть удаляется после работы модуля *DomainGraph* (который выполняет укрупнение гиперссылок в графе до уровня доменов – веб-граф DG), приходится затрачивать существенную часть машинных ресурсов на их обработку на этапе построения веб-графа G модулем *WebCrawler*. В связи с этим возникла идея сокращения числа ресурсов из общего домена, обрабатываемых краулером еще на этапе построения веб-графа G. С этой целью была модифицирована работа модуля *WebCrawler* таким образом, чтобы можно была задавать в качестве параметра долю ресурсов из общего домена, используемых при формировании веб-графа.

В результате при несущественном сокращении числа узлов в КСС в конечном итоге происходит существенное сокращение размера веб-графа G (и соответственно затрат машинных ресурсов, времени – до 10 раз). При этом размер доменного графа DG уменьшается незначительно (именно на его основе в дальнейшем происходит выявление веб-сообществ).

Сравнение таблиц 6 и 7, показывает, что расширение веб-графа на 1 уровень приводит к существенному увеличению степени пересечения между рубриками (в основном за счет каталогов ресурсов, счетчиков и других малорелевантных ресурсов). Данная статистика может быть использована для оптимизации тематической структуры рубрик веб-каталогов. Например, высокая степень пересечения существующих рубрик даже на уровне входящих гиперссылок приводит к мысли о необходимости их слияния или переструктурирования. Высокая степень пересечения между рубриками оправдывает необходимость использования техники кэширования закачиваемых ресурсов для сокращения затрат времени на скачивание (особенно при использовании каналов с низкой пропускной способностью).

Таблица 4a. Характеристики веб-графа для рубрик подрубрики каталога "*Учеба/Науки/Технические науки*" (без использования прореживания гиперссылок).

№	Рубрика	Размер графа *V*, узлов	Доменов *VD*	*V* / *VD*
1	A	н/д	н/д	н/д
2	HT	26993	7343	3,68
3	CHE	48743	13039	3,74
4	CSIS	187112	31613	5,92
5	O	н/д	н/д	н/д
6	U	44540	12291	3,62

Таблица 4b. Характеристики веб-графа для рубрик подрубрики каталога "*Учеба/Науки/Технические науки*" (с параметром прореживания, равным 2).

№	Рубрика	Размер графа *V*, узлов	Доменов *VD*	*V* / *VD*
1	A	57066	27025	2,11
2	HT	20166	6672	3,02
3	CHE	43587	21071	2,07
4	CSIS	114244	27430	4,16
5	O	31657	16365	1,93
6	U	32255	9915	3,25

Таблица 5. Характеристики веб-графа для рубрик подрубрики каталога "*Учеба/Науки/Гуманитарные науки/История/История России* " (с параметром прореживания, равным 2).

№	Рубрика	Размер графа *V*, узлов	Доменов *VD*	*V* / *VD*
1	Arc	1762	543	3,24
2	WH	34828	16465	2,12
3	G	6908	1896	3,64
4	Anc	432	90	4,80
5	N	47928	8388	5,71
6	O	9192	5248	1,75
7	M	40817	7774	5,25
8	U	16372	8574	1,91
9	E	38128	8906	4,28

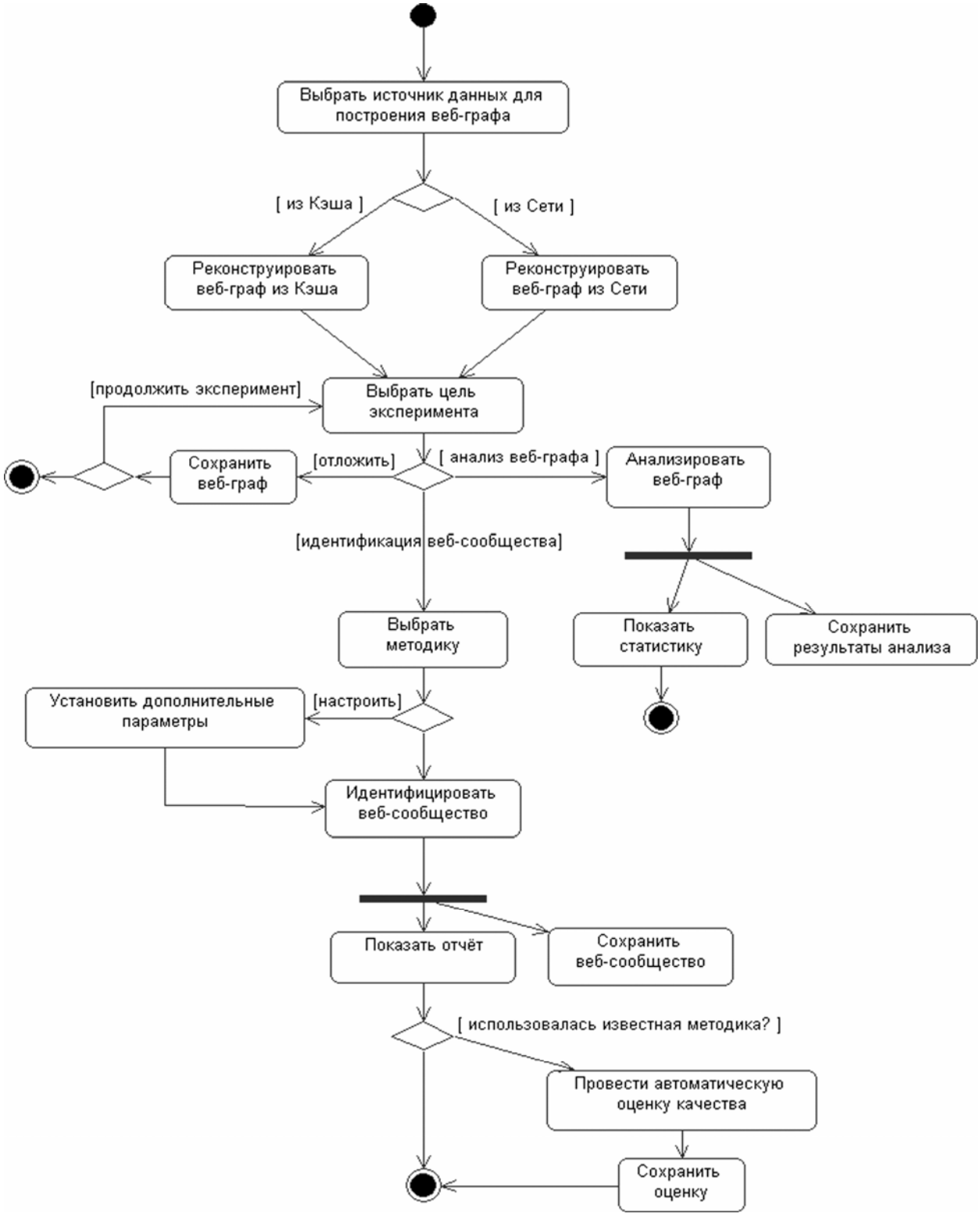

Рис. 3. Схема проведения машинного эксперимента

Таблица 6. Пересечения между рубриками подрубрики каталога "*Учеба/Науки/Технические науки*" по входящим гиперссылкам (для зерен).

	A	HT	CHE	CSIS	O	U	*Всего*	*%*	*Max*	*%*
A	23422	73	132	298	199	44	746	3,2	298	1,3

HT		2315	42	174	35	75	399	17,2	174	7,5
CHE			7228	275	87	78	614	8,5	275	3,8
CSIS				13598	193	103	1043	7,7	298	2,2
O					10998	75	589	5,4	199	1,8
U						1910	375	19,6	103	5,4

Таблица 7. Пересечения между рубриками подрубрики каталога "*Учеба/Науки/Технические науки*" по узлам веб-графа G (с параметром прореживания, равным 2).

	A	**HT**	**CHE**	**CSIS**	**O**	**U**	**Всего**	**%**	**Max**	**%**
A	57066	2270	5148	7299	3697	3711	22125	39	7299	13
HT		20166	3087	4873	2150	2221	14601	72	4873	24
CHE			43587	39884	3719	3103	54941	126	39884	92
CSIS				114244	5064	6564	63684	56	39884	35
O					31657	3012	17642	56	5064	16
U						32255	18611	58	6564	20

Таблица 8. Сводные результаты по выявлению и оценке веб сообществ для подрубрики каталога "*Учеба/Науки/Технические науки*" по узлам веб-графа (* - означает, что граф построен со значением параметра прореживания, равным 2).

Рубрика	**размер КСС**	**Среднее качество КСС**	**Распределение оценки качества по узлам КСС**											
			1	**0,8**	**0,75**	**0,67**	**0,6**	**0,5**	**0,4**	**0,33**	**0,25**	**0.2**	**0**	**Зерновых**
A*	1965	0,241	17	61	0	1	206	4	430	4	3	499	712	28
	%		0,88	3,15	0,00	0,05	10,64	0,21	22,20	0,21	0,15	25,76	36,76	1,42
HT	335	0,189	2	6	0	0	23	0	53	0	0	101	143	7
	%		0,61	1,83	0,00	0,00	7,01	0,00	16,16	0,00	0,00	30,79	43,60	2,09
CHE	126	0,281	1	8	0	1	15	0	27	0	0	28	39	7
	%		0,84	6,72	0,00	0,84	12,61	0,00	22,69	0,00	0,00	23,53	32,77	5,56
CSIS	6452	0,279	66	338	8	2	920	10	1410	11	6	1576	2060	45
	%		1,03	5,28	0,12	0,03	14,36	0,16	22,01	0,17	0,09	24,60	32,15	0,70
O*	447	0,227	1	5	0	1	40	4	95	0	2	145	144	10
	%		0,23	1,14	0,00	0,23	9,15	0,92	21,74	0,00	0,46	33,18	32,95	2,24
U	1070	0,137	2	2	0	0	32	2	133	3	1	338	549	8
	%		0,19	0,19	0,00	0,00	3,01	0,19	12,52	0,28	0,09	31,83	51,69	0,75

3.2 Результаты выявления и оценки веб-сообществ

В таблицах 7-8 приведены результаты выявления и оценки веб-сообществ для подрубрик каталога "*Учеба/Науки/Технические науки*").

3.3 Выбор зерновых ресурсов

На рисунке 4 представлена зависимость количества входящих гиперссылок *InL* на зерна из рубрик от ранга зерен *r*. Ранг зерна формировался в результате сортировки всех зерен рубрики по *InL* в порядке убывания.

Данный график показывает, что не все зерна в рубрике являются равнозначимыми, хотя параметр *InL* сам по себе не является прямым индикатором ценности зерна. Было проведено 2 исследования, в которых рассматривалась зависимость результата идентификации веб-сообществ (с последующей оценкой качества) от :

1) выбора единственного зернового ресурса из рубрики (веб-граф G строился на основе одноэлементного зернового множества) – схема *Singles*;
2) размера зернового множества (наращиваение множества происходило за счет инкрементного добавления зерен в порядке убывания их ранга *r*) - схема *Reduced*.

В таблице 9 приведен результат эксперимента по схеме *Singles*. Заштрихованы зерна, которые оказались недоступными, либо для них размер веб-сообщества оказался равным нулю.

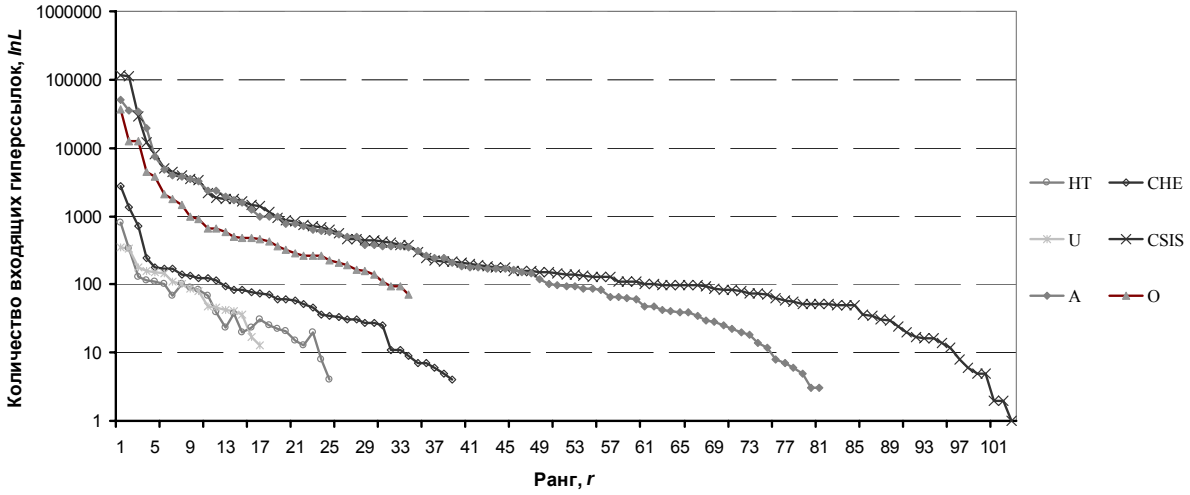

Рис.4. Распределение зерен в рубриках по количеству входящих гиперссылок (для подрубрики каталога "*Учеба/Науки/Технические науки*").

Таблица 9. Результат эксперимента по схеме *Singles* для рубрики *Universal* из подрубрики каталога "*Учеба/Науки/Технические науки*").

№	Зерно	InL	N(G), узлов	N(DG), узлов	E(G), ребер	E(DG), ребер	Размер КСС	MQ	W	W * MQ
1	http://www.cta.ru	354	6046	1800	98356	4087	21	0,21	0,10	0,020
2	http://www.tesis.com.ru	331	6134	1428	91672	2829	17	0,19	0,08	0,015
3	http://www.vniitfa.ru	177	2174	423	40754	794	0	-	0,00	-
4	http://www.ntpo.com	158	3928	1433	54797	3801	11	0,07	0,05	0,003
5	http://www.nd.ru/dk	148	4204	1524	50664	2578	8	0,07	0,04	0,002
6	http://www.virste.ru	143	3405	894	62259	2717	67	0,15	0,30	0,046
7	http://www.laboratory.ru	109	2980	1107	41386	4675	2	0,20	0,01	0,002
8	http://www.techbook.ru	100	2354	934	33016	1794	32	0,20	0,15	0,030
9	http://listlib.narod.ru	88	2877	1019	37171	4308	2	1,00	0,01	0,009
10	http://www.polarcom.ru	79	2452	802	35411	1292	4	0,07	0,02	0,001
11	http://www.k2kapital.com	47	1147	374	22342	550	2	0,20	0,01	0,002
12	http://www.bsuproduct.by	45	1682	430	33847	3174	46	0,14	0,21	0,029
13	http://aimatrix.nm.ru	42	880	374	7005	537	0	-	0,00	-
14	http://zntu.edu.ua/RIC	40	557	89	5446	127	0	-	0,00	-
15	http://www.inteltec.ru	36	1273	406	24612	3017	2	0,20	0,01	0,002
16	http://www.extech.ru	17	616	165	11816	2606	6	0,08	0,03	0,002
17	http://eko.org.ua	13	452	185	11738	2616	0	-	0,00	-
	Всего:						**220**		**1,00**	

В таблице приведены оценки по следующим параметрам:

- *InL* - ранг зерна;
- *N(G)* - количество узлов в построенном модулем *WebCrawler* веб-графе *G;*
- *N(DG)* - количество узлов веб-графе после доменного укрупнения (*DG*);
- *E(G)* - количество ребер в веб-графе *G;*
- *E(DG)* - количество ребер в веб-графе *DG;*
- *MQ* - среднее качество членов КСС (исключая зерновые);
- *W* - удельный размер КСС (вес).

Как видно из таблицы, наибольшую оценку получили зерна http://www.virste.ru (НТП ВИРАЖ-ЦЕНТР - издатель научно-технических журналов), http://www.techbook.ru (Горячая линия-Телеком) и http://www.bsuproduct.by (БГУ - Научно-

Таблица 10. Результат эксперимента по схеме *Reduced* для рубрики *Universal* из подрубрики каталога "*Учеба/Науки/Технические науки*"). В столбце "Всего сообществ" указано количество выявлнных сообществ, содержащих более 2 узлов.

Кол-во зерен	Размер графа G		Размер дом. графа DG		Всего сообществ	Размер КСС, узлов	Средняя оценка качества КСС	Прирост кол-ва вход.ссылок *InL*	Общее кол-во вход ссылок
	Узлов	Ребер	Узлов	Ребер					
2	11597	172240	3029	6888	63	38	0,181	331	685
3	13305	208177	3267	7502	69	38	0,181	177	862
4	17090	259565	4637	11613	100	89	0,136	158	1020
5	19748	318812	5243	13910	110	160	0,161	148	1168
6	23230	361954	6550	16842	128	200	0,155	143	1311
7	25731	396416	7354	21509	135	207	0,165	109	1420
8	27534	423543	8007	23390	115	740	0,175	100	1520
9	29797	457235	8704	25711	119	877	0,174	88	1608
10	31126	484625	9083	26764	119	940	0,175	79	1687
11	32658	514230	9417	27569	123	998	0,173	47	1734
12	33401	520681	9691	28205	123	1042	0,174	45	1779
13	33916	525146	9863	28635	123	1053	0,174	42	1821
14	34633	531389	10126	29407	129	1064	0,173	40	1861
15	34997	536011	10150	29603	130	1068	0,173	36	1897
16	35084	537453	10174	29693	131	1070	0,169	17	1914

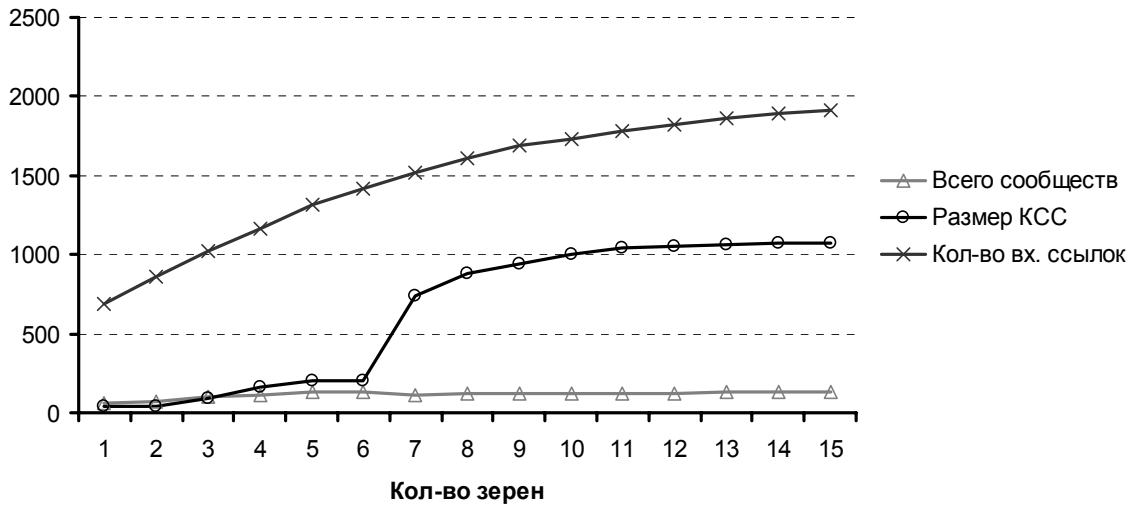

Рис.5. Зависимость результата выявления веб-сообществ из веб-графа от количества выбранных зерен (для рубрики каталога "*Учеба/Науки/Технические науки/Универсальное*")

Техническая Продукция), имеющие средний рейтинг по количеству входящих гиперссылок, между тем, как более рейтинговые зерна получили гораздо более низкую итоговую оценку.
В таблице 10 приведены реультаты эксперимента уже по схеме *Reduced*. Средняя оценка качества для КСС вычислялась после исключения зерновых узлов из компоненты, которые априори имели оценку качества, равную 1.0.

На рисунке 5 приведен график, показывающий влияние количества выбранных зерен на результат выявления веб-сообществ (по схеме *Reduced*).

3.4 Экспертная оценка качества членов сообщества

В таблице 11 представлены веб-страницы из выделенной (на основе 47 зерновых веб-сайтов из рубрики "*Каталог/Учеба/Науки/Технические науки/Вычислительная техника и электронка*") компоненты сильной связности, получившие оценку принадлежности превышающую значение 0.4. Общее количество входящих в эту компоненту узлов – 128.

Таблица 11. Выделенная КСС (для рубрики *Вычислительная техника и электронка* из подрубрики каталога "*Учеба/Науки/Технические науки*"). В первом столбце приведен результат автоматической оценки элементов КСС. Зерновые элементы выделены жирным шрифтом.

Автом. оценка	URL	Заголовок
1,00	**http://www.rntores.ru**	РНТОРЭС им. А.С. Попова - Титульная страница
1,00	**http://www.shema.ru**	Shema.ru - Анатомия электроники
1,00	**http://www.chipnews.ru**	Электронные компоненты
1,00	**http://www.dian.ru**	СХЕМОТЕХНИКА, схемы электрические, принципиальные
1,00	**http://dozen.mephi.ru**	Кафедра "Компьютерные системы и технологии" МИФИ
1,00	**http://www.microsystems.ru**	Журнал "Нано- и микросистемная техника"
1,00	**http://eldep.mephi.ru**	Кафедра электроники МИФИ
1,00	http://chipnews.gaw.ru	Новости микроэлектроники - счетчик электроэнергии, lcm, aduc, atmega
0,80	http://www.rlocman.ru	РадиоЛоцман - электроника, схемы, электронные компоненты, диаграммы, datasheet, новости, обзоры
0,80	http://www.radionet.pp.ru	RadioNet - информационно-поисковый портал по электронике
0,80	http://radionet.pp.ru	RadioNet - информационно-поисковый портал по электронике
0,80	http://www.radionet.com.ru	RadioNet - информационно-поисковый портал по электронике
0,80	http://www.compitech.ru	Статьи по электронным компонентам - attiny, промэлектроника, интерфейс usb, rs485
0,80	http://www.rlocman.com.ru	РадиоЛоцман - электроника, схемы, электронные компоненты, диаграммы, datasheet, новости, обзоры
0,80	http://www.kaf26.mephi.ru	МИФИ : Факультет "А" : Кафедра 26 : Общая справка о кафедре
0,80	http://www.hit.nsk.ru	Издательство Инфоэлектрон -- Главная
0,70	http://www.electronics.ru	Журнал Электроника НТБ
0,60	**http://www.soel.ru**	Современная Электроника : ЖУРНАЛ :
0,60	http://www.raai.org	Российская ассоциация искусственного интеллекта
0,60	http://display.chipexpo.ru	DISPLAY - специализированная выставка
0,60	http://www.chip-news.ru	Официальный сайт журнала Chip News
0,60	http://kaf29.mephi.ru	Кафедра 29 МИФИ
0,60	http://www.novtex.ru	Издательство Новые технологии
0,60	http://radionet.com.ru	RadioNet - информационно-поисковый портал по электронике
0,60	http://www.finestreet.ru	Издательский дом Finestreet
0,60	http://www.remserv.ru	Ремонт&Сервис/Новости
0,60	http://www.interself.ru	Поиск электронных компонентов | Поставка электронных компонентов в Россию :: Interself.ru
0,60	http://www.topplan.ru	Информационно-справочные системы TopPlan
0,60	http://www.kaf13.mephi.ru	Кафедра Теплофизики МИФИ
0,60	http://www.radioliga.com	Журнал РАДИОЛЮБИТЕЛЬ. Официальный сайт
0,60	http://www.saco.ru	SACO CONTROLS ::: главная страница
0,60	http://www.rudshel.ru	ЗАО "Руднев-Шиляев" - измерительные приборы платы сбора данных АЦП …..

На рисунке 6 представлен график зависимости показателей полноты и точности от порога фильтра $П_ф$. Расчет показателей основывался на экспертной оценке релевантности элементов данной КСС по отношению к тематике рубрики. Использовалась шкала: 1 – "да", 0.66 – "скорее да", 0.33 - "скорее нет", 0 - "нет". Порог релевантности $П_р$ был равен 0.5.

Рис.6. Зависимость показателей точности и полноты от величины порога фильтра П (для рубрики каталога "*Учеба/Науки/Технические науки/Вычислительная техника и электроника*")

3. Выводы

Проведенное исследование в целом подтвердило жизнеспособность предложенной авторами статьи методики автоматического пополнения каталога веб-ресурсов. При этом, однако выбор зерновых ресурсов из рубрик каталога имеет принципиально важное значение. Эксперименты по схемам *Singles* и *Reduced* показали неравноценность веб-ресурсов, размещаемых в рубриках каталогов, причем количество входящих на ресурс гиперссылок является достаточно приблизительным косвенным показателем его ценности.

К сожалению, существенно ограниченные технические возможности не позволили авторам статьи провести исследование достаточно большого количества рубрик каталога.

Разработанный для данного исследования программный комплекс может быть использован также для целей диагностики рубрик каталогов и оптимизации распределения веб-ресурсов между ними.

Для оценки качества членов КСС была использована предельно простая методика, которая позволила получить неплохие результаты. Собственно контентный анализ документов не являлся предметом исследования. Для последующего повышения показателей точности и полноты возможно использование уже хорошо проработанных методов, например на основе TF-IDF метрики.

Литература

[1] Дунаев Е.В., Шелестов А.А. Автоматическая рубрикация web-страниц в интернет-каталоге с иерархической структурой//Сборник работ стипендиатов по гранту компании Яндекс [Электрон. ресурс] -2005. – Режим доступа: http://company.yandex.ru/grant/2005/08_Shelestov_103119.pdf.

[2] Киселев М.В. Оптимизация процедуры автоматического пополнения веб-каталога //Сборник работ стипендиатов по гранту компании Яндекс. [Электрон. ресурс] -2005. – Режим доступа: http://company.yandex.ru/grant/2005/08_Kiselev_102710.pdf

[3] Козлов Д.Д., Белова А.А. Исследование эффективности применения методов совместного анализа текстов и гиперссылок для поиска тематических сообществ //Сборник работ стипендиатов по гранту компании Яндекс [Электрон. ресурс] -2005. – Режим доступа: http://company.yandex.ru/grant/2005/06_Kozlov_102805.pdf.

[4] Calado P., Cristo M., Moura E. et al. Combining link-based and content-based methods for web document classification. – CIKM'03, Nov. 3-8. – [Electronic resource].- Mode of access: homepages.dcc.ufmg.br/~nivio/papers/cikm03.ps

[5] Flake G., Tarjan R., Tsioutsiouliklis K. Graph clustering and minimum cut trees. Internet Mathematics Vol. I, No. 4, 2004: 385-408. [Electronic resource].- Mode of access: www.internetmathematics.org/volumes/1/4/Flake.pdf.

[6] Ino H., Kudo M., Nakamura A. Partitioning of web graphs by community topology. WWW 2005, May 10-14, 2005, Chiba, Japan. [Electronic resource].- Mode of access: www2005.org/cdrom/docs/p661.pdf

[7] Баженов М.М., Сычёв А.В. Идентификация веб-сообществ на основе сильно связанных компонент и контентного анализа // Вестн. ВГУ. Серия Системный анализ и информационные технологии. 2006 .— N1 .— С.13-19.

The Problem of The Seeds Selection for an Automatic Web-directory Resource Discovery Based on Strongly Connected Components Identification Followed by Content Filtering

Alexander V. Sytchev, Michael M. Bazhenov

The paper presents results of experimental research of an approach proposed by authors for an automatic web-directory resource discovery. Using data sets gathered from Yandex web-directory (http://yaca.yandex.ru) 2 approaches for seeds selection from web-directory rubrics were examined. It was demonstrated that seeds selected from from web-directory have rather different importance for the automatic resource discovery. The number of inlinks of seeds may be considered approximately like indirect indicator of its importance.

♣ Данная работа поддержана исследовательским грантом компании Яндекс.