Изучение характеристик сообществ русскоязычной блогосферы

© А.В. Сычев, И.А. Гадебский

Воронежский государственный университет
sav@cs.vsu.ru

Аннотация

В статье представлены результаты расчета характеристик для сообществ блог-хостингов LiveJournal и LiveInternet, полученных на основе атрибутов профилей сообществ. Также приведены графики и таблицы, полученные в результате кластеризации сообществ и их интересов.
Результаты могут быть использованы для решения ряда прикладных задач связанных с поиском сообществ и интересов.

1 Введение

Социальные сети в WWW — интересный феномен, который сформировался за буквально за последние 5 лет. Они объединяют в себе блоги (сетевые дневники), сети медиа-ресурсов, сети персональной информации (MySpace, LinkeIn, Facebook, Мой круг, Одноклассники и др.), системы закладок (del.icio.us), wiki-энциклопедии и другие. Количество пользователей в данных сетях увеличивается с беспрецедентной скоростью, и это вызывает вполне заслуженный интерес к данным сетям со стороны исследователей. В частности, во многих исследованиях рассматриваются вопросы о структуре и свойствах больших длительно развивающихся социальных групп и сообществ в этой электронной среде.

За рубежом проводятся исследования данного феномена по довольно широкому спектру направлений и имеются многочисленные научные публикации, например [1–8].

В [5] представлены результаты изучения изменений, происходящих в блогах в течение всего жизненного цикла, выделены отдельные фазы жизненного цикла сетевого журнала (блога) и соответствующие им индикторы. Применение методов кластеризации для решения таких задач как поиск «горячих» тем, аннотирование и классификация блогов рассматривается в работах [2, 3, 6]. Эффективный алгоритм поиска кластеров терминов из сообщений в блогах, существующих в рамках определенных временных интервалов предложен в [1]. Там же представлены результаты исследования временных характеристик таких кластеров. В работе [4] блог-сообщества рассматриваются в их структурной и временной динамике. Авторами предложен метод факторизации сообществ (на основе графа сообществ), позволяющий выявлять сообщества, не обнаруживаемые традиционными методами.

Задача поиска «скрытых» друзей по подобию в распределении тематик в их блогах рассмотрена в [7]. Исследованию он-лайн профилей пользователей посвящена работа [8]. В ней на основании экспериментальных данных показано, что пользователи реально используют лишь «тонкий срез» информации из профиля для того, чтобы составить представление о других пользователях. Там же рассмотрены такие характеристики атрибутов профилей как полезность для восприятия, предсказуемость и диагностируемость.

В России исследования по данной тематике носят эпизодический характер и являются чаще всего коммерчески ориентированными. Из числа наиболее известных исследований в России можно упомянуть исследования, регулярно проводимые компанией Яндекс [9] и проект «Пульс блогосферы» (http://blogs.yandex.ru/pulse/).

Целью исследования, результаты которого представлены в данной статье, является анализ структуры и связи между атрибутами профилей сообществ в наиболее популярных в России блог-хостингах LiveJournal и LiveInternet, а также поиск эффективных методик обработки данных из профилей сообществ, позволяющих получить дополнительную информацию о сообществах и их интересах.

2 Исходные данные

Для проведения исследования были использованы профили сообществ из блог-хостингов LiveJournal (http://www.livejournal.com) и в LiveInternet (http://www.liveinternet.ru). В качестве исходных данных был получен список русскоязычных сообществ LiveJournal. Список был извлечен путем скачивания и анализа веб-страниц, на которые указывают гиперссылки со страницы «Реестр русскоязычных сообществ “Живого журнала”» (http://lj.com.ru) и «Топ сообществ» LiveInternet (http://www.liveinternet.ru/top/community/). Для проведения исследования всего было скачано 2905 профи-

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL’2008, Дубна, Россия, 2008.

лей сообществ LiveJournal и 35984 профилей сообществ LiveInternet.

3 Хронология создания сообществ

На рис. 1 показана гистограмма, отражающая хронологию развития русскоязычной части блогхостинга LiveJournal с момента создания по настоящее время. На гистограмме хорошо прослеживаются сезонные циклы и основные этапы становления и развития данной среды.

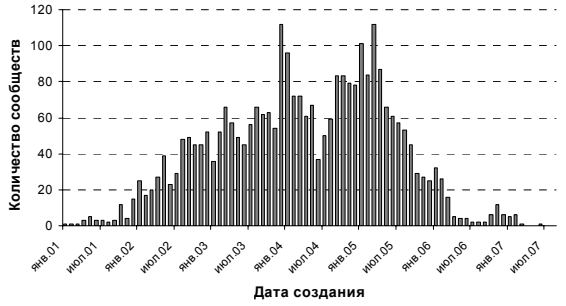

Рис. 1. **Распределение сообществ LiveJournal по дате создания**

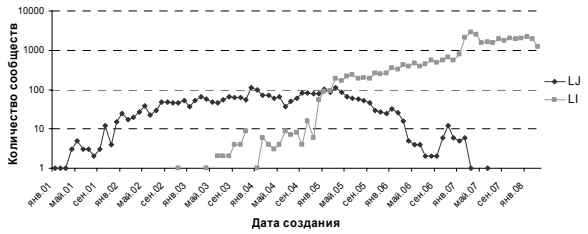

Рис. 2. **Распределение сообществ LiveJournal и LiveInternet по дате создания (в логарифмическом масштабе)**

Для сравнения на рис. 2 приведены в виде графика данные по хронологии создания сообществ одновременно для LiveJournal и LiveInternet.

На первый взгляд, исходя из этого графика можно сделать вывод об упадке блогхостинга LiveJournal (LJ) и его вытеснении LiveInternet (LI) и другими блогхостингами. Однако исследование, проведенное компанией Яндекс [1], показывает, несмотря на то, что LJ уступает LI по количество блогов, преимущество по количеству активных блогов все же на стороне LJ. Сравнение медианных значений для количества записей в таблицах 1 и 3 позволяет сделать вывод о том, что блогосфера в указанных блогхостингах находится просто на разных этапах развития.

4 Атрибуты профилей

4.1 LiveJournal

В таблице 1 приведены сводные данные по атрибутам профилей русскоязычных сообществ LiveJournal.

Для каждой пары атрибутов был рассчитан коэффициент корреляции, соответствующие значения корреляции между атрибутами профилей приведены в таблице 2

Из таблицы 2 следует, что наиболее тесно связанными между собой параметрами профиля являются количество членов (4), количество читателей (5), количество записей (11), количество пользователей с правом записи (14).

Таблица 1. **Усредненные количественные показатели профилей русскоязычных сообществ в LiveJournal**

№	Название поля профиля	Максимум	Среднее	Медиана
1	Количество интересов	150	25	11
2	Количество смотрителей	27	1,8	1,0
3	Количество модераторов	12	0,3	0,0
4	Количество членов	17653	599	195
5	Количество читателей	16191	551	186
6	Входит в сообщества (количество)	29	0,1	0,0
7	Тип аккаунта (0-беспл.,1-улучш., 2-платн.)	2	0,03	0,00
8	Дата создания	22.06.2007	01.04.2004	13.04.2004
9	Дата обновления	10.03.2008	13.10.2007	24.02.2008
10	Количество подарков	5	0,01	0,00
11	Количество записей	71271	1089	193
12	Написано	965	5	0
13	Получено	1335748	10302	627
14	Количество пользователей с правом записи	17650	578	171

Таблица 2. **Оценка корреляции между атрибутами профилей русскоязычных сообществ в *LiveJournal***

№	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0,11	0,09	0,10	0,11	0,01	0,05	0,16	0,19	0,03	0,04	-0,05	0,05	0,10
2		0,23	0,31	0,30	0,02	0,16	-0,02	0,13	0,07	0,30	0,01	0,30	0,30
3			0,26	0,29	-0,01	0,11	0,04	0,06	0,04	0,12	0,01	0,16	0,25
4				0,91	0,06	0,19	-0,21	0,17	0,10	0,73	0,08	0,62	0,99
5					0,05	0,15	-0,18	0,18	0,07	0,58	0,06	0,52	0,90
6						0,00	-0,11	0,02	0,00	0,05	0,05	0,03	0,04
7							0,07	0,02	0,00	0,05	0,05	0,03	0,04
8								0,12	-0,01	-0,20	-0,21	-0,15	-0,21
9									0,03	0,12	-0,05	0,08	0,18
10										0,13	0,09	0,14	0,10
11											0,15	0,74	0,73
12												0,13	0,08
13													0,62

Таблица 3. **Усредненные количественные показатели профилей русскоязычных сообществ в *LiveInternet***

№	Название поля	Максимум	Среднее	Медиана
1	Дата регистрации	23.03.2008	22.04.2007	04.06.2007
2	Записей в дневнике	20570	60	2
3	Комментариев в дневнике	1002402	256	3
4	Написано сообщений	990725	246	4

4.2 LiveInternet

В таблице 3 приведены сводные данные по атрибутам профилей сообществ LiveInternet.

Для каждой пары атрибутов был рассчитан коэффициент корреляции, результаты сведены в таблице 4.

Таблица 4. **Оценка корреляции между атрибутами профилей сообществ в *LiveInternet***

	2	3	4
1	-0,16	-0,05	-0,03
2		0,44	0,44
3			0,99

5 Распределение интересов в сообществах

Изучение интересов сообществ может предоставить исследователю вполне компактную и достаточно точную информацию, позволяющую судить о тематической направленности этих сообществ и предпочтениях пользователей – членов сообществ.

5.1 LiveJournal

В результате обработке профилей сообществ LiveJournal был сформирован список интересов, характеристики которого приведены в таблицах 5–6. Хотя бы 1 интерес был указан в профиле 2260 сообществ. Общее количество интересов получилось равным 43247.

В таблице 5 приведен список 30 наиболее упоминаемых интересов в профилях русскоязычных сообществ LiveJournal.

Таблица 5. **Список наиболее распространенных интересов в профилях русскоязычных сообществ *LiveJournal***

№	Интерес	Сообществ
1	музыка	167
2	любовь	120
3		115
4	литература	108
5	фотография	97
6	искусство	86
7	секс	86
8	книги	84
9	история	83
10	кино	83
11	фото	79
12	психология	77
13	красота	76
14	Россия	73
15	люди	68
16	женщины	67
17	философия	67
18	свобода	66
19	творчество	65
20	дети	63
21	путешествия	62
22	поэзия	62
23	music	61
24	дизайн	61
25	природа	60
26	юмор	59
27	жизнь	59
28	общение	59
29	реклама	55
30	интернет	52

В таблице 6 показаны характеристики распределения интересов в русскоязычных сообществах *LiveJournal*, усреднение проводилось по сообществам. Величина **ICW** рассчитывалась как сумма частот интересов, указанных в профиле сообщества. Вес интереса был равен частоте его встречаемости в профилях всех сообществ. Величина **CIC** рассчитывалась как количество интересов из профиля сообщества, указанных также в профиле хотя бы одного другого сообщества.

Таблица 6. **Характеристики распределения интересов в русскоязычных сообществах *LiveJournal***

	Взвешенное количество интересов (*ICW*)	**Простое количество интересов (*IC*)**	***ICW* / *IC***
Макс.	2486	150	167
Средн.	230	32	9,2
Мед.	120	18	5,7

Корреляция (ICW, IC)	0,63
Корреляция (IC, ICW / IC)	-0,16

	Совпадений по интересам (*CIC*)	**Всего интересов в сообществе (IC)**	***CIC* / *IC*, %**
Макс.	137	150	100
Сред.	18	32	59
Мед.	10	18	59

Корреляция (CIC, IC)	0,81
Корреляция (CIC, CIC/IC)	-0,21

5.2 LiveInternet

Характеристики распределения интересов в сообществах LiveInternet приведены в таблице 7.

Таблица 7. **Характеристики распределения интересов в русскоязычных сообществах *LiveInternet***

	Взвешенное количество интересов (*ICW*)	**Простое количество интересов (*IC*)**	***ICW* / *IC***
Макс.	59760	402	1598
Средн.	1778	18	131
Мед.	583	7	78

Корреляция (ICW, IC)	0,77
Корреляция (IC, ICW / IC)	-0,08

	Совпадений по интересам (*CIC*)	Всего интересов в сообществе (IC)	*CIC* / *IC*, %
Макс.	402	402	100
Сред.	15	18	84
Мед.	6	7	94

Корреляция (CIC, IC)	0.985
Корреляция (CIC, CIC/IC)	0.029

Как видно из сравнения таблиц 6 и 7, в LiveJournal существенно больше доля уникальных интересов, неповторяющихся в других сообществах.

6 Кластеризация сообществ по интересам, указанным в профиле

Для получения более содержательных результатов был проведен ряд экспериментов по кластеризации сообществ и интересов, указанных в профилях сообществ.

С этой целью был использован метод аггломеративной кластеризации Ланса-Уильямса. Первичное расстояние между сообществами рассчитывалось по формуле:

$$\rho(c_1, c_2) = \frac{|c_1 \cap c_2|}{\sqrt{|c_1|} \cdot \sqrt{|c_2|}}$$

Сообщество c_i рассматривалось как множество интересов, указанных в его профиле. При проведении процедуры кластеризации расстояние между кластерами рассчитывалось по формуле среднего расстояния.

При проведении кластеризации интересов расчет расстояния между интересами выполнялся по аналогичной формуле, при этом вместо размера сообщества подставлялся размер множества сообществ, в которых указан данный интерес.

В качестве исходных данных для процедуры кластеризации сообществ (интересов) была использована матрица «*сообщество–интерес*», на основе которой строилась матрица «*сообщество–сообщество*» («*интерес–интерес*»).

Характеристики матриц приведены в таблицах 8–10.

При проведении кластеризации по интересам в связи с ограничениями вычислительного характера учитывались только интересы, указывавшиеся в двух и более сообществах, т.е. фактически кластеризация выполнялась на прореженных матрицах.

Таблица 8. **Исходные данные для построения матрицы «*Сообщество–Интерес*»**

Блог-хостинг	Сообществ всего	Сообществ, в которых указаны интересы	Макс. кол-во интересов в профиле сообщества	Интересов (во всех сообществах)	Кол-во интересов, указанных более чем в 1 сообществе
LiveJournal	3083	2260	150	43245	9416
LiveInternet	35983	11200	402	43398	18542

Таблица 9. **Характеристики матриц «*Сообщество-Интерес*» и «*Сообщество-Сообщество*» (без прореживания)**

Блог-хостинг	Параметр	Матрица «Сообщество–Интерес»		Матрица «Сообщество–Сообщество»	
		Агрегирование по сообществам	Агрегирование по интересам	Число остальных сообществ, с которыми есть пересечение	Средняя степень пересечения сообщества с остальными сообществами, %
LiveJournal	Максимум	150 (интересов)	167 (сообществ)	638	80
	Среднее	31,7	1,7	66,0	6
	Медиана	18	1	33	5
	Сообществ (Интересов)	2260	43245	2065	
	Разреженность матрицы, %	99.93		96.81	
LiveInternet	Максимум	402 (интересов)	1598 (сообществ)	7602	100
	Среднее	17,6	4,2	465,5	12
	Медиана	7	1	188	10
	Сообществ (Интересов)	11200	47398	10488	
	Разреженность матрицы, %	99.96		95.56	

Таблица 10. **Характеристики матриц «*Сообщество-Интерес*» и «*Интерес–Интерес*» (с прореживанием)**

Блог-хостинг	Параметр	Матрица «Сообщество-Интерес» (прореженная)		Матрица «Интерес-Интерес» (прореженная)	
		Агрегирование по сообществам	Агрегирование по интересам	Число остальных интересов, с которыми есть пересечение	Средняя степень пересечения интереса с остальными интересами, %
LiveJournal	Максимум	139 (интересов)	167 (сообществ)	2657	100
	Среднее	17,7	4,02	68,08	49
	Медиана	10	2	24	50
	Сообществ (Интересов)	2142	9416	8867	
	Разреженность матрицы, %	99.81		99.23	
LiveInternet	Максимум	402 (интересов)	1598 (сообществ)	10374	100
	Среднее	15,6	9,09	228,07	41
	Медиана	6	3	38	40
	Сообществ (Интересов)	10829	18542	16600	
	Разреженность матрицы, %	99.92		98.63	

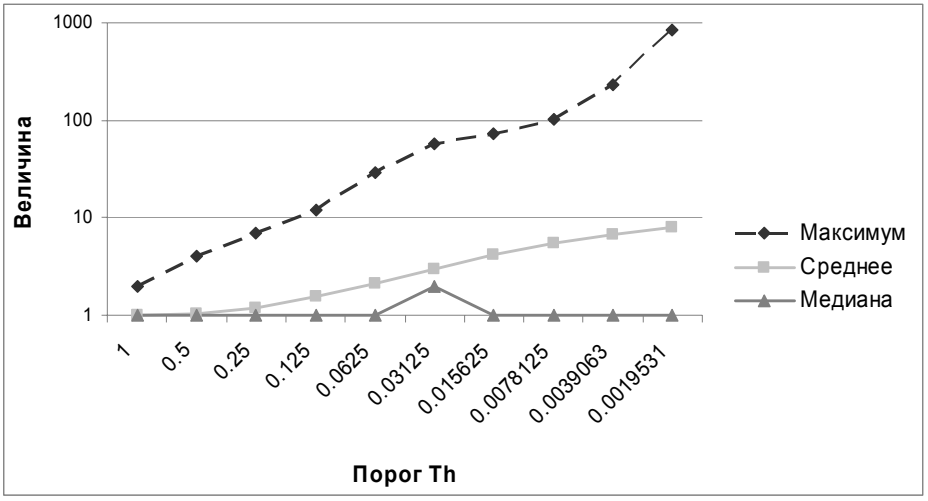

Рис. 3. **Характеристики кластеров сообществ *LiveJournal***

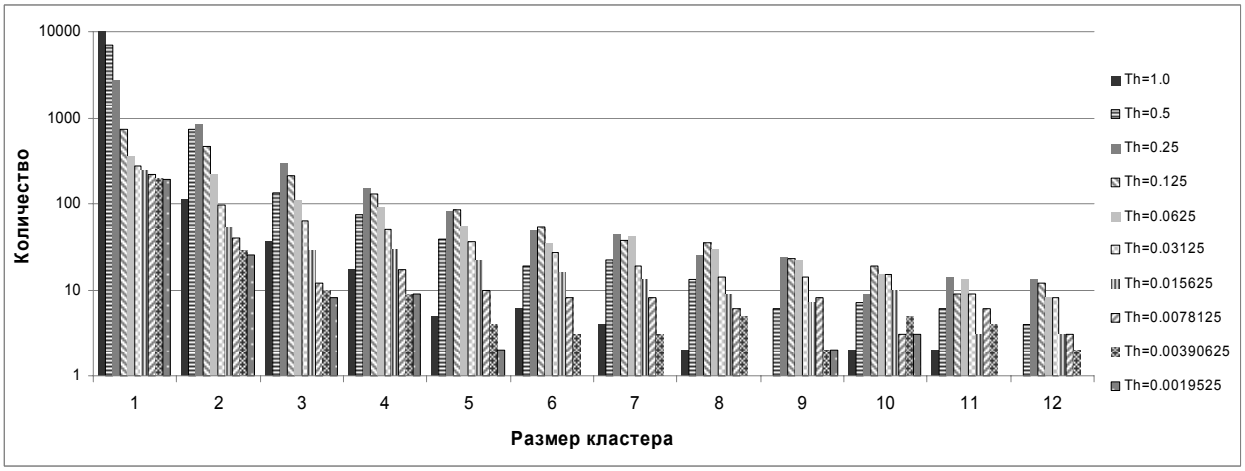

Рис. 4. **Распределение кластеров сообществ *LiveInternet* при различных значениях порога кластеризации Th**

6.1 Кластеризация сообществ

На основе матриц «Сообщество–Сообщество» были получены кластеры сообществ при различных значениях порога кластеризации *Th*. Характер зависимости размера кластера (максимум, среднее и медиана) от порога *Th* для LiveJournal представлен на рис. 3. Для LiveInternet была получена зависимость похожей формы, причем, в обоих случая наибольшее значение медианы достигается для значения порога в районе от 0.02 до 0.05. Распределение кластеров сообществ в зависимости от порога кластеризации для LiveInternet приведено на рис. 4. На гистограмме размер кластера ограничен 12 сообществами, поскольку самая существенная часть сообществ оказалась сконцентрированной в кластерах с размером в диапазоне от 1 до 12.

Таблица 11. **Сообщества *LiveInternet* из наибольшего кластера (для порога Th = 0.5)**

№	Название	№	Название	№	Название
1	___ДИЗАЙН___	21	HappyTreeFriends_4ever	41	вСе_ЛуЧшЕе_ЗдЕсЬ
2	-__Картинки__-	22	JannLissssy_Design	42	Всё_самое_лучшее_для_днева
3	_-Pictures-only-for-you-_	23	Lilac_style	43	Всевозможные_аватары
4	_ЗаКаЖи_АвАтАр_	24	-Making_avatars-	44	выполняем_заказики
5	_МеГо_КаРтИнКи_	25	mega_art_design	45	Делаем_аватары_на_заказ
6	_ТВОЁ_	26	MP_STUDIO	46	ДиЗайн
7	A_F_Y_D	27	Pics_and_Diz	…	……..
8	Anything_For_Your_Diary	28	PiCtUrEs_sTyLe	67	Просто_така
9	Avatarki_special_for_your	29	Pretty_Blogs	68	РисОваННые_ДизаЙны
10	AvatarLand_BG_	30	THE-BEST-PICTURES	69	Смени_ДиЗ
11	-Avatars-	31	-Voncuver-	70	собрание_картинок
12	Avatars_for_you_diary	32	we_make_designs	71	сообществО_х
13	Beauty_Diary_4_u	33	Your_orders	72	Фон_для_тебя
14	design_for_diarys	34	Аватары_для_Жизни	73	Фоны_для_дневников
15	Designes_By_-Rapsody-	35	авафонэпик	74	Фоны_и_эпиграфы_Ап
16	Dizzziki	36	Беру_на_Ли-ру	75	Фоны_Эпиги_Авики
17	Ducks	37	Всё__Лучшее_Для_Дневников	76	фсё_для_днеффничка
18	FASHOIN_PICTURES	38	Все_для_Вашего_днева-	77	хХСмотри_сукаХх
19	fine_arts_group	39	Все_для_Дневничк0в	78	Цветной_Клан
20	ForYourDiary	40	Все_картинки_здесь	79	Юзер_пиг_и_Ко

Таблица 12. **Интересы, указанные в профилях сообществ из таблицы 11**

№	Интерес	CF	ICF	PF	CF-ICF	№	Интерес	CF	ICF	PF	CF-ICF
1	фоны	73	218	364	0,821	20	lilac	1	1	1	0,031
2	эпиграфы	70	196	315	0,815	21	purple	1	1	1	0,031
3	аватары	70	393	699	0,633	22	violet	1	1	1	0,031
4	картинки	51	342	831	0,488	23	сиреневый	1	1	2	0,031
5	дизайны	21	116	186	0,285	24	фотошоп ...	1	1	1	0,031
6	заказы	9	30	59	0,168	25	схемы оформления	1	1	2	0,031
7	дизайн	11	205	463	0,126	26	велкомы	1	1	1	0,031
8	фотошоп	6	144	307	0,077	27	заголовки	1	1	1	0,031
9	готовые дизайны	3	7	11	0,072	28	фотографии.	1	1	3	0,031
10	фотографии	4	105	369	0,056	29	глиттер	1	1	1	0,031
11	подписи	2	4	6	0,052	…	……	…..	…..	…..	…..
12	поиск картинок	2	4	10	0,052	61	эмо	1	89	341	0,015
13	схемы	2	8	22	0,047	62	рисунки	1	91	224	0,014

14	оформление	2	8	27	0,047	63	гламур	1	97	288	0,014
15	анимации	2	12	29	0,044	64	интернет	1	99	335	0,014
16	дневники	3	138	288	0,039	65	критика	1	129	222	0,013
17	сообщества	2	31	95	0,037	66	стихи	1	147	507	0,013
18	создание аватарок	1	1	1	0,031	67	аниме	1	250	560	0,011
19	happy tree friends	1	1	9	0,031	68	любовь	1	267	1048	0,010

В таблице 11 приведен в качестве примера приведен наибольший по размеру кластер сообществ *LiveInternet***,** сформированный при значении порога *Th* = 0.5. Интересы, содержащиеся в профилях этих сообществ, представлены в таблице 12. Для каждого интереса также указаны: **CF** – как часто встречается интерес в профилях сообществ, образующих кластер, **ICF** – количество других кластеров, содержащих сообщества с этим интересом, **PF** – частота встречаемости интереса в профилях всех сообществ, **CF-ICF** – метрика, аналогичная TF-IDF, и показывающая специфичность интереса для данного кластера.

Расчет показал, что коэффициент корреляции между **CF-ICF** и **PF** равен 0.28, что указывает на возможность применения метрики **CF-ICF** для ранжирования кластерообразующих интересов**.** Процедура кластеризации может быть использована как для поиска латентных «суперсообществ», так и для выявления множества интересов, определяющих общую тематику таких «суперсообществ». Варьируя величину порога кластеризации *Th*, можно задавать соотношение между полнотой и точностью данного множества.

6.2 Кластеризация интересов

На основе матриц «Интерес-Интерес» были получены кластеры сообществ при различных значениях порога кластеризации *Th*.

Характеристики кластеров представлены на рис. 5–6. В таблицах 13–14 приведены интересы из наибольшего кластера и соответствующие им сообщества.

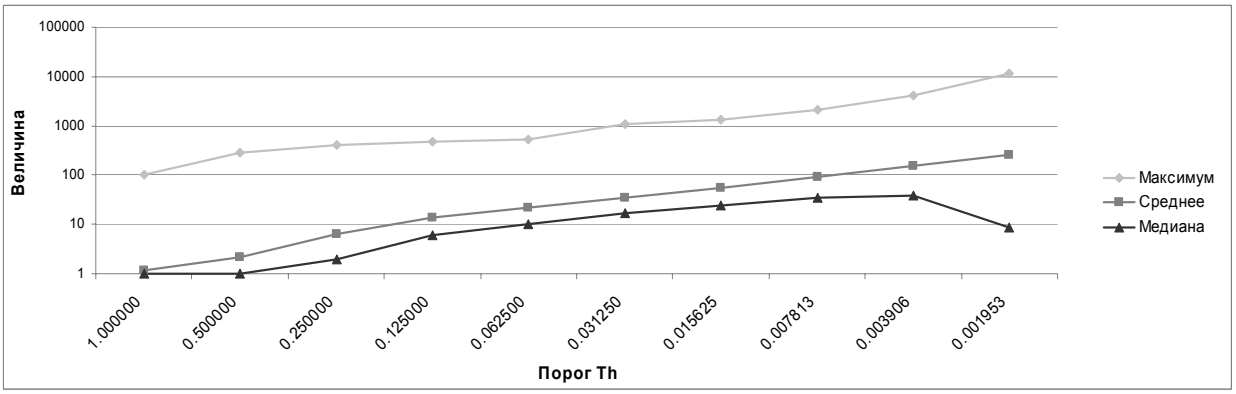

Рис. 5. **Характеристики кластеров интересов *LiveInternet***

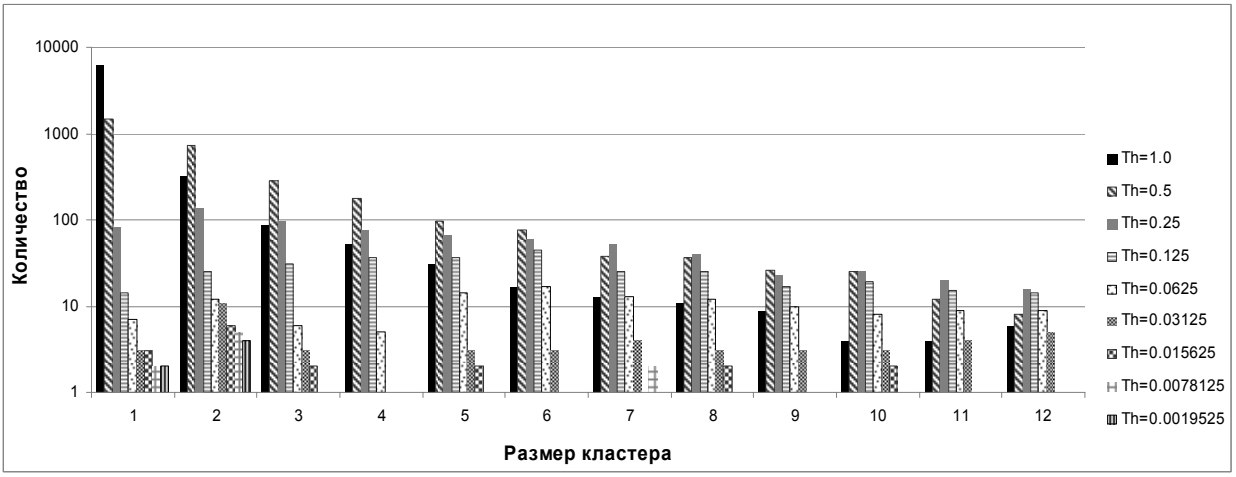

Рис. 6. **Распределение кластеров интересов *LiveJournal* при различных значениях порога кластеризации Th**

Таблица 13. **Интересы *LiveJournal* из наибольшего кластера (для порога Th = 0.5)**

№	title	№	title	№	title	№	title
1	alfa romeo	21	hyundai	41	renault	61	ВАЗ
2	audi	22	isuzu	42	rover	62	вождение
3	auto	23	jaguar	43	saab	63	ГАЗ
4	automobile	24	jeep	44	seat	64	ГАИ
5	bmw	25	kia	45	skoda	65	гаишники
6	buick	26	lancia	46	subaru	66	ГИБДД
7	cadillac	27	land rover	47	suzuki	67	давить на газ
8	car	28	lexus	48	toyota	68	дворники
9	cars	29	lincoln	49	volkswagen	69	джип
10	chevrolet	30	mazda	50	volvo	70	дизель
11	chrysler	31	mercedes	51	vw	…	……
12	citroen	32	mitsubishi	52	авария	110	схождение
13	daewoo	33	mustang	53	авто	111	сцепление
14	daihatsu	34	nissan	54	автозапчасти	112	теория
15	dodge	35	opel	55	автомобиль	113	топливо
16	fiat	36	peugeot	56	автопром	114	тормоз
17	ford	37	pontiac	57	автосервис	115	тормоза
18	geo	38	porsche	58	АЗС	116	тормозная жидкость
19	headlights	39	porshe	59	аккумуляторы	117	УАЗ
20	honda	40	range rover	60	бензин	118	фаркоп

Таблица 14. **Сообщества, указанные в профилях сообществ из таблицы 13**

№	nick	CF	№	nick	CF
1	avto_piter	120	11	ru_f1	3
2	spb_auto	112	12	ru_luxury	3
3	ru_auto	111	13	italian_cars	2
4	ru_auto_shop	43	14	mountains	2
5	ru_rightcars	9	15	ru_ford	2
6	avto_ru	8	16	ru_hyundai	2
7	_megafon_	4	…	…….	…
8	megafondv	4	44	strana_sovetov	1
9	moto_ru	3	45	su_tormoz	1
10	ru_4x4	3	46	surgut	1

7 Заключение

Изучение сообществ русскоязычной блогосферы – обширная и многогранная задача. В данной статье представлены некоторые результаты, касающиеся одного из аспектов общей задачи. На основе информации, представленной в профилях сообществ наиболее популярных блог-хостингов LiveJournal и LiveInternet в рамках проведенного исследования решались такие задачи как реконструкция хронологии создания сообществ, вычисление корреляции между атрибутами профилей. Наибольшее внимание в работе было уделено одному из самых информативных и значимых атрибутов – интересам сообщества. Был изучен характер распределения интересов в профилях как путем расчета простейших числовых характеристик, так и с помощью процедуры кластеризации. Последняя может быть эффективно использована для решения таких прикладных задач как поиск латентных «суперсообществ» и определения их тематики, автоматического структурирования пространства интересов, автоматической оценки тематической принадлежности и специфичности интересов, а также для ряда других интересных задач.

Литература

[1] N.Bansal, F.Chiang, N.Koudas, F W.Tompa. Seeking Stable Clusters in the Blogosphere. VLDB 2007, Vienna. pp. 806–817.

[2] C.H.Brooks, N.Montanez. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In: Proc. of WWW, 2006, pp. 625–632.

[3] A.Qamra, B.L.Tseng, E.Y.Chang. Mining blog stories using community-based and temporal clustering. In Proc. of the 15th CIKM Conference, 2006. pp. 58–67.

[4] Y.Chi, S.Zhu, X.Song, J.Tatemura, B. L. Tseng. Structural and temporal analysis of the blogosphere through community factorization. In: Proc. of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, 2007. pp. 163–172.

[5] D.Gurzick and W.G.Lutters. From the personal to the profound: understanding the blog life cycle. CHI 2006, April 22–27, 2006, Montreal, Quebec, Canada. ACM Press, New York: 2006. pp. 827–832.

[6] B.Li, S.Xu, J.Zhang. Enhancing Clustering Blog Documents by Utilizing Author/Reader Comments. In Proc. ACMSE'07, March 23-24, ACM.

[7] D.Shen,J.Sun,Q.Yang,Z.Chen. Latent Friend Mining from Blog Data. In: 6th International Conference on Data Mining, Hong Kong, China,2006. pp. 552–561.

[8] K.Stecher, S.Counts. Thin Slices of Online Profile Attributes. In Proc. of International Conference on Weblogs and Social Media – 2008. [Electronic resource]. – Mode of access: research.microsoft.com/~scottlt/pubs/ICWSM_ThinSlices.pdf.

[9] Состояние блогосферы российского интернета. По данным поиска по блогам Яндекса. Весна 2008 г. [Электрон. ресурс] – 2008. – Режим доступа: http://download.yandex.ru/company/yandex_on_blogosphere_spring_2008.pdf

Russian Blogspace Community Features Research

A.V. Sychev, I.A. Gadebskij

The paper is dedicated to the Russian blogspace research. Some numerical features related to community profiles (hosted in the most popular in Russia blogspaces www.livejournal.com and www.liveinternet.ru) are presented. The diagram of communities creation chronology is provided. The correlation factor between pairs of community profile attributes is estimated. The clustering of communities and its interests is illustrated by tables, histograms and graphics. This procedure could have important practical applications like detecting latent supercommunities and its topic distillation, interest space automatic stucturing and others.