Сравнительное изучение атрибутов профилей сообществ LiveJournal в 2010 и 2011 годах

© А.В. Сычёв

Воронежский государственный университет
sav@cs.vsu.ru

Аннотация

Для атрибутов, указанных в профилях сообществ блог-хостинга LiveJournal была исследована временная динамика журналов русскоязычных сообществ путем сравнения данных по атрибутам профилей сообществ в 2010 и 2011 годах. Были рассмотрены изменения: общего числа уникальных интересов по всему множеству сообществ, частот употребления интересов, кластеров интересов, кластеров сообществ и пользователей.

1. Введение

По статистике ВЦИОМ (февраль 2011), в социальных сетях зарегистрированы сегодня более половины пользователей Интернета в России (52%), сетевыми журналами (блогами) пользуются 7% из них [1]. Всего же в марте 2010 года компании Яндекс было известно более 17 миллионов русскоязычных блогов, что оказалось в 2.5 раза больше чем в 2009 году. При этом количество активных блогов составило всего лишь 6% от их общего числа, увеличившись за год всего лишь на 12%. Что касается сервиса микроблогов Твиттер, то за один год (2009-2010) число блогов в нем выросло в 26 раз, увеличившись практически с нуля до более миллиона.

Как видно из приведенной статистики, блогосфера привлекает к себе внимание огромного количества пользователей Интернета и развивается очень динамично. Имеется немало публикаций, посвященных изучению динамики блогосферы и ее моделированию, например [7],[8]. В России научных публикаций по системным исследованиям социальных сетей на данный момент практически нет, имеются в основном отдельные отчеты коммерческих компаний, например Яндекс, содержащие, как правило, только статистические данные. В этих отчетах результаты аналитической обработки, например с привлечением методов Data Mining, часто отсутствуют.

В данном докладе автором приведены результаты исследования, продолжающего проводившееся ранее. В исследовании [3] на основе информации, представленной в профилях сообществ наиболее популярного из блог-хостингов LiveJournal, решались такие задачи как реконструкция хронологии создания сообществ, вычисление корреляции между атрибутами профилей; был изучен характер распределения интересов в профилях как путем расчета простейших числовых характеристик, так и с помощью процедуры кластеризации; изучены особенности регионального распределения интересов сообществ.

В рамках текущего исследования предметом изучения были изменения значений атрибутов профилей сообществ блог-хостинга LiveJournal за 2010 год и за первую половину 2011 года.

2. Исходные данные

Для решения задачи были использованы общедоступные данные сервиса «Рейтинг сообществ» компании Яндекс (http://www.livejournal.com/ratings/community/) в 2011 году и сервиса «Поиск по блогам» компании Яндекс (http://blogs.yandex.ru) в 2010 году.

Исходный набор данных содержал профили русскоязычных сообществ блог-хостинга LiveJournal. Набор был сформирован путем скачивания профилей в соответствии со списком из рейтинга сообществ. Общее количество профилей составило порядка 126 тысяч в январе 2011 года и порядка 134 тысяч в июле, а объем коллекции на жестком диске составил более 5Гбайт. В результате обработки html-страниц профилей сообществ был сформирован набор нормализованных реляционных таблиц, содержащих описание атрибутов самих сообществ, интересов и пользователей. Некоторые сообщества из рейтинга оказались «замороженными» либо удаленными, поэтому их профили были исключены из списка обработки. В таблице 1a приведены сравнительные данные по 2010-2011 гг. Также в таблице 1b сведены средние и медианы, а в таблице 1c - изменения величин ключевых атрибутов профилей сообществ за периоды наблюдения.

Труды 13[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2011, Воронеж, Россия, 2011.

Таблица 1a. Сводные данные по набору профилей сообществ LiveJournal.

Параметр	Год			Изменение			
				Янв 2011 - янв 2010		Июл 2011 - янв 2011	
	янв.10	янв.11	июл.11	D	%	D	%
Пользователей	948139	1143274	>1152066	195135	20,6	>8792	>0,8
Интересов	418107	432797	444734	14690	3,5	11937	2,8
Сообществ	109311	125843	134071	16532	15,1	8228	6,5
Сообществ (Доступных)	99163	125609	133980	26446	26,7	8371	6,7

Таблица 1b. Сводные данные по атрибутам профилей сообществ LiveJournal.

Номерами в заголовке таблицы обозначены следующие атрибуты: *1* - количество интересов сообщества; *4* - количество участников сообщества; *5* - количество читателей журнала сообщества; *9* - число записей в журнале; *11* - число полученных комментариев.

Период	Атрибут профиля	1	4	5	9	11
Янв 2010	*Среднее:*	13,40	81,30	72,2	135,80	859,70
	Медиана:	3	6	5	7	4
Янв 2011	*Среднее:*	11,10	81,87	62,7	140,55	720,41
	Медиана:	0	9	3	9	2
Июл 2011	*Среднее:*	10,77	81,89	61,9	139,62	710,44
	Медиана:	0	9	3	5	2

Таблица 1c. Сводные данные по изменениям значений атрибутов профилей сообществ LiveJournal.

Период	Атрибут профиля	1		4		5		9		11	
		D1	*%*	*D4*	*%*	*D5*	*%*	*D9*	*%*	*D11*	*%*
Янв 2010-Янв 2011	***Среднее:***	-2,30	**-17,14**	0,57	**0,70**	-9,52	**-13,18**	4,75	**3,50**	-139,29	**-16,20**
	Медиана:	-3	**-100,0**	3	**50,0**	-2	**-40,0**	2	**28,6**	-2	**-50,0**
Янв 2011-Июл 2011	***Среднее:***	-0,33	**-3,02**	0,02	**0,03**	-0,76	**-1,22**	-0,93	**-0,66**	-9,98	**-1,38**
	Медиана:	0	**0**	0	**0**	0	**0**	-4	**-2,9**	0	**0**

Из этих таблиц можно заметить, что если средние значения за два периода наблюдения изменились в пределах 18%, то медианы претерпели более значительные изменения. Так по интересам за год наблюдается практически 100% уменьшение, по числу участников 50% увеличение, по числу комментариев 50% уменьшение. Данная тенденция указывает на заметное увеличение «хвоста» в ранговом распределении сообществ за счет появления большого числа спам-сообществ, зачастую не имеющих списка интересов в профиле и содержащих заметное число «участников», созданных одним спамером. Также можно констатировать тенденцию к уменьшению числа комментариев от читателей в журналах сообществ.

По некоторым оценкам [4] доля спама в блогах варьируется от 56 до 75%. При этом на сайте популярного сервиса классификации блог-спама Akismet суммарная оценка доли спам-комментариев, проходящих через него, составляет около 87% [5]. Как указывается в отчете [2], летом 2008 года начался резкий рост регистраций спамовых блогов, и за год их количество выросло более чем в 30 раз. По состоянию на весну 2009 примерно треть всех записей в блогах определялась Яндексом как спам. Обнаружение спама на основе выявления аномалий в статистических распределениях было предложено в [6].

На рисунке 6 представлены гистограммы распределения четырех атрибутов профилей по сообществам. Характерно, что около половины всех сообществ вообще не имеют комментариев от читателей, причем эта тенденция усиливается во времени.

Наблюдаемые на графиках 6a) - 6c) скачки достигают величины 7-8%, что соизмеримо с общим приростом количества сообществ (16.5 тыс. за 2010 год и 8 тыс. за первое полугодие 2011 года). Такие «аномалии», скорее всего, имеют отношение к спам-блогам.

Наблюдается устойчивое по времени существенное доминирование сообществ,

содержащих всего одного читателя. В целом имеет место тенденция уменьшения числа читателей. Процент сообществ с нулевым числом записей в журнале остается довольно высоким.

3. Исследование изменения интересов

Несмотря на общее увеличение числа уникальных интересов (на 3.5%), среднее число интересов, указанных в профиле пользователя уменьшилось с 13,4 в январе 2010 года до 11,10 в январе 2011 года (10,8 в июле 2011 года).

Максимальное число интересов, которое можно указать в профиле сообщества – 150.

Распределение сообществ в зависимости от числа интересов в профиле показано на рисунке 1.

Видно, что за год характер распределения не изменился. Наибольший прирост сообществ наблюдается при числе интересов в профиле сообщества, близком к нулю.

Если рассмотреть распределение сообществ в зависимости от изменения размера списка интересов в профиле, то можно заметить, что у подавляющего большинства (~97%) сообществ, изначально имевших непустой список интересов, размер списка не изменился (рисунок 2).

Для сообществ, имевших пустой список интересов в профиле, было получено

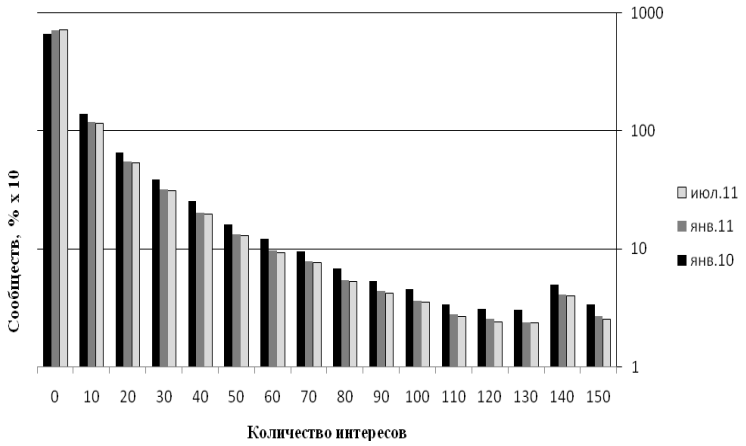

Рис.1. Распределение сообществ в зависимости от количества интересов, указанных в профиле.

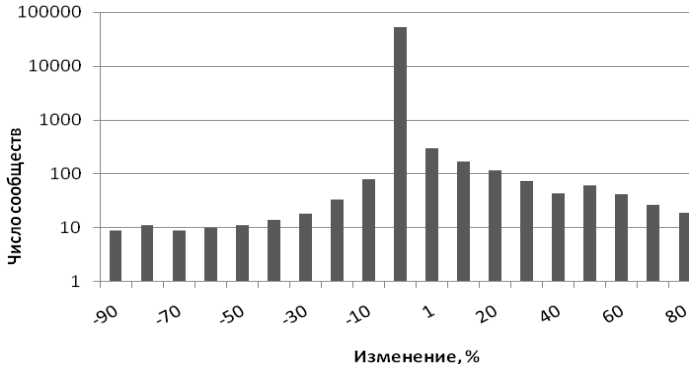

Рис.2. Распределение сообществ в зависимости от изменения количества интересов (с ненулевым значением

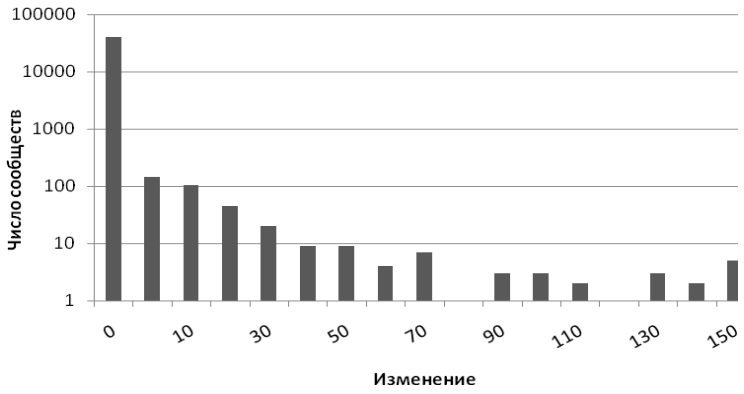

Рис.3. Распределение сообществ в зависимости от изменения количества интересов (с нулевым значением в 2010 г.).

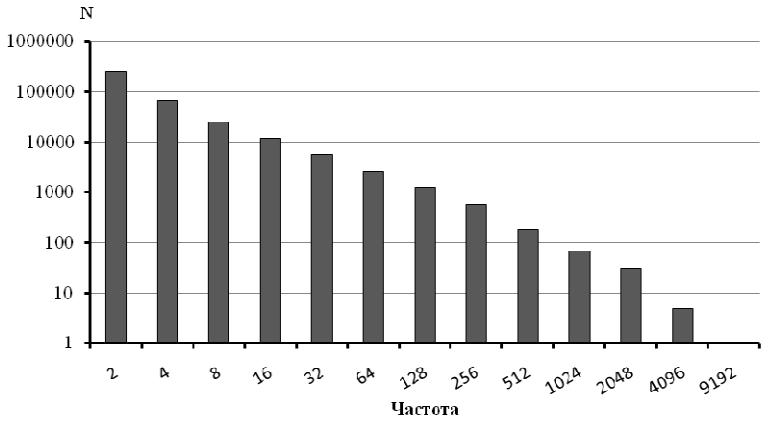

Рис.4. Распределение количества интересов в зависимости от их частоты

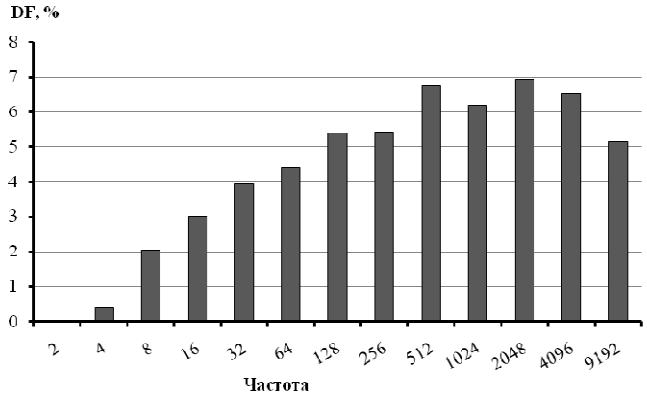

Рис.5. Зависимость процента изменения частоты (DF) интереса от значения его частоты

распределение, представленное на рисунке 3. У порядка 99% таких сообществ список интересов так и остался пустым спустя год.

По-сути, содержательные изменения списка интересов в профиле охватывают относительно небольшое число сообществ.

В рамках проведенного исследования также было изучено изменение частот упоминания интересов в профилях сообществ, а также кластеров интересов, построенных на основе схожести распределения этих интересов по профилям сообществ.

Из общего числа интересов в январе 2010 года (порядка 420 тысяч) около 40 тысяч не оказалось в профилях сообществ в январе 2011 года, кроме того, появилось порядка 50 тысяч новых интересов.

На рисунке 4 представлено распределение количества интересов в зависимости от их частоты упоминания в профиле. Для интересов, которые присутствовали в профилях в 2010 и 2011 годах, частота увеличилась в среднем на 0.43%. При этом наибольшие изменения наблюдались для высоко- и среднечастотных интересов (рисунок 5).

Для исследования структурных изменений на основе матрицы сопряженности типа “интерес-интерес” была проведена иерархическая агломеративная кластеризация интересов по методу Ланса-Уильямса. Первичное расстояние между интересам рассчитывалось по формуле:

$$\rho(\boldsymbol{i}_k,\boldsymbol{i}_l)=\frac{\left|\boldsymbol{i}_k\cap\boldsymbol{i}_l\right|}{\sqrt{\left|\boldsymbol{i}_k\right|}\cdot\sqrt{\left|\boldsymbol{i}_l\right|}}$$

Интерес i_k рассматривался как множество сообществ, в которых упоминался данный интерес. При проведении процедуры кластеризации расстояние между кластерами рассчитывалось по формуле среднего расстояния.

В качестве исходных данных для процедуры кластеризации интересов были использованы значения из матрицы “интерес-сообщество”, построенной для примерно 18 тысяч интересов, у которых частота встречаемости в профилях сообществ была не ниже 10 (сообществ). Из нее затем была сформирована матрица “интерес-интерес”, по которой и была проведена кластеризация.

На рисунке 7 показаны размеры кластеров в зависимости от выбранного значения порога *Th*. Такое распределение демонстрирует в целом стабильность средних размеров кластеров интересов во времени.

Размер кластера можно рассматривать в качестве индикатора неспецифичности интересов. Чаще всего в одном кластере будут оказываться интересы, имеющие похожее распределение по сообществам.

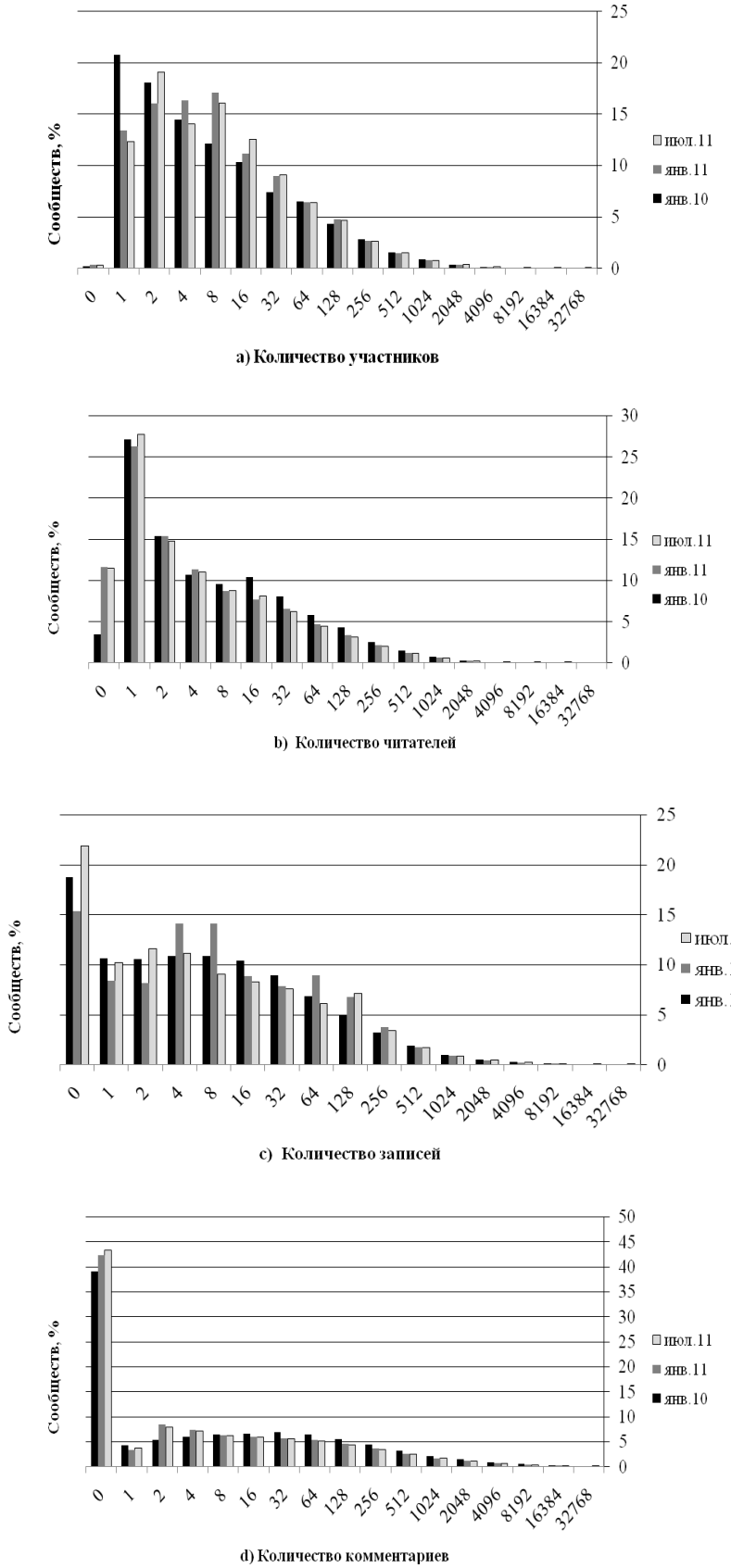

Рис.6. Распределение атрибутов профилей по сообществам: a) количество участников, b) количество читателей, c) количество записей, d) количество комментариев.

Наибольшие изменения за прошедший год затронули кластеры, полученные при значениях порога $Th < 0.125$.

Таким образом, интерес можно отнести к группе атрибутов профилей сетевого журнала сообщества, которые изменяются в наименьшей степени во времени. В большей степени изменяются средне- и высокочастотные интересы, при этом их частотные ранги достаточно устойчивы. Результаты кластеризации (если рассматривать изменения средних величин и медиан) свидетельствуют о том, что списки интересов, указанные в профилях сообществ достаточно стабильны.

4. Изменение кластеров сообществ

В качестве исходных данных для процедуры кластеризации сообществ по общим интересам были использованы значения из матрицы “сообщество-интерес”, построенной для примерно 34-37 тысяч сообществ, содержавших в профилях не менее 10 интересов. Из нее затем была сформирована матрица “сообщество-сообщество”, по которой и была проведена кластеризация. Результаты представлены на рисунке 8, подтверждая устойчивость кластеров сообществ, построенных на основе учета списка интересов из профиля. Средние размеры кластеров сообществ также демонстрируют стабильность, заметный на гистограмме прирост происходит за счет увеличения общего числа сообществ. Также как и в случае с интересами размер кластера сообществ можно рассматривать в качестве индикатора неспецифичности сообществ.

В качестве исходных данных для процедуры кластеризации сообществ по общим участникам были использованы значения из матрицы “сообщество-участник”, построенной для примерно 37 тысяч сообществ, содержавших не менее 27 участников в июле 2011 года (не менее 25 участников в январе 2011 г. и не менее 13 участников в 2010 году). Из нее затем была сформирована матрица “сообщество-сообщество”, по которой в дальнейшем и проводилась кластеризация. Результаты представлены на рисунке 9. На гистограмме можно увидеть, что характер изменения средних размеров кластеров

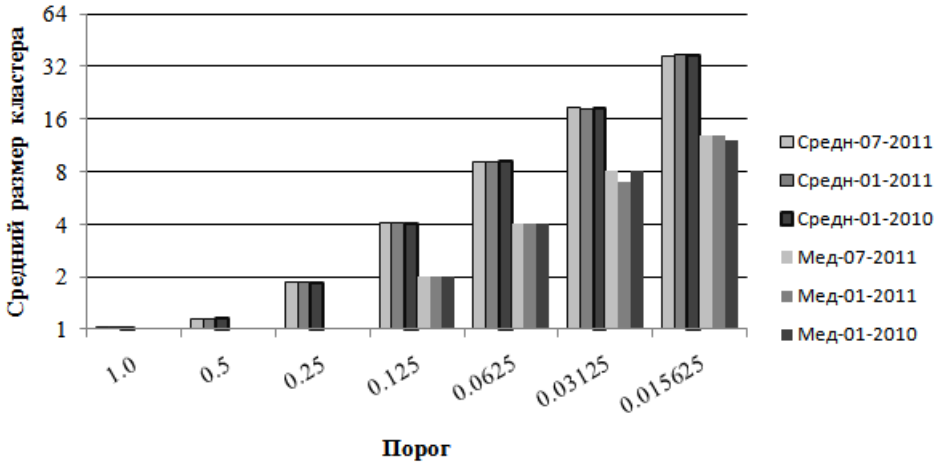

Рис 7. Оценки размера кластера интересов для различных значений порога *Th*.

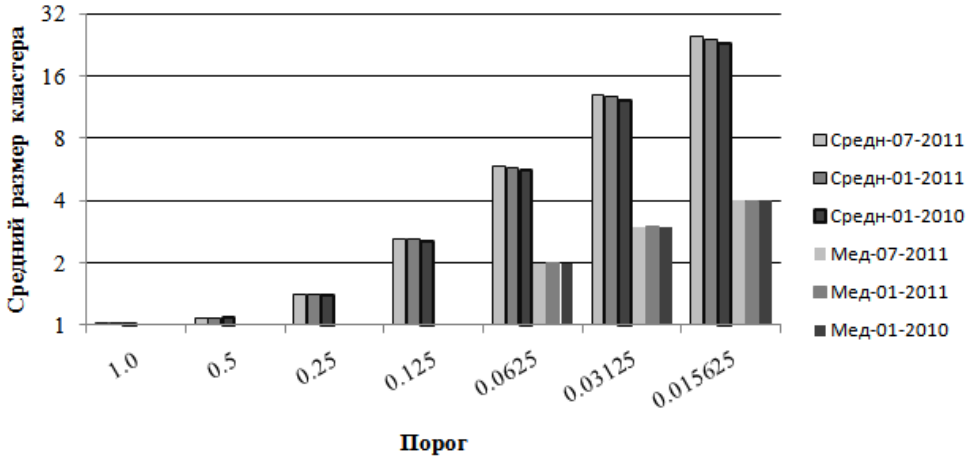

Рис 8. Оценки размера кластера сообществ (на основе их общих интересов) для различных значений порога *Th*. Для кластеризации были использованы порядка 34 тыс. сообществ в январе 2010 г., 36 тыс. – в январе 2011 г. и 37 тыс. – в июле 2011 г.

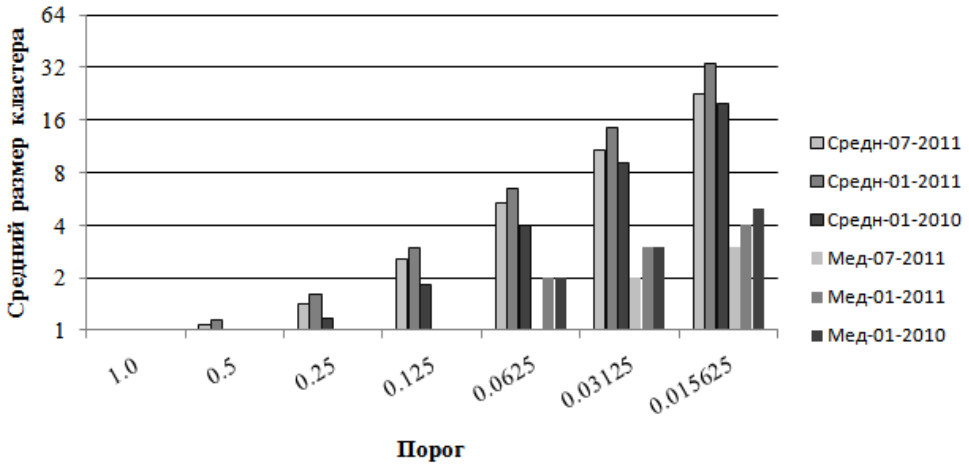

Рис.9 Изменение размера кластера сообществ (на основе их общих участников) для различных значений порога *Th*.

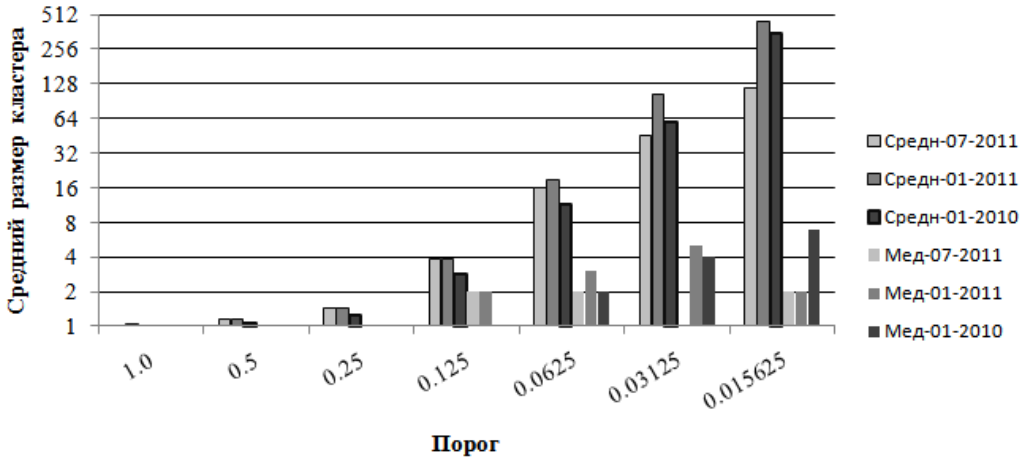

Рис.10 Изменение размера кластера пользователей (на основе их общих сообществ) для различных значений порога *Th*.

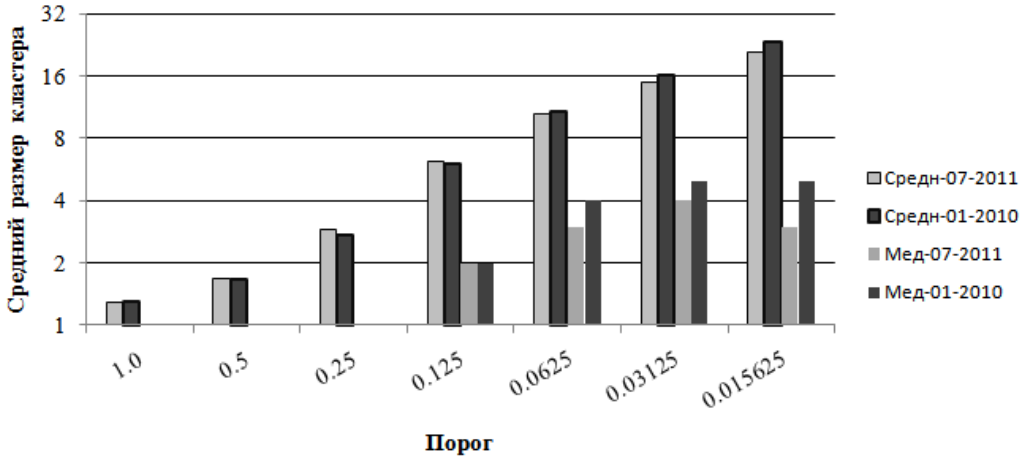

Рис.11 Изменение размера кластера пользователей со случайной выборкой (на основе их общих сообществ) для различных значений порога *Th*.

сообществ очень похож при различных значениях порога кластеризации *Th*. Учитывая высокий процент инфильтрации блогосферы спамом, можно предположить, что заметные колебания среднего размера как в положительную так и в отрицательную сторону помимо естественного увеличения числа сообществ отражают также активность спам-блогов и результаты противодействия им со стороны пользователей и администраторов блог-хостинга. При этом инфильтрация сообществ спам-

блогами носит скорее неизбирательный характер, что может объяснять подобие динамики размера кластера при различных значениях порога Th.

5. Изменение кластеров пользователей (участников)

В качестве исходных данных для процедуры кластеризации пользователей-участников сообществ по общим сообществам были использованы значения из матрицы "*сообщество-участник*", построенной для примерно 19 тысяч пользователей, участвующих не менее чем в 62 сообществах в 2011 году (не менее чем в 65 сообществах в 2010 году). Из нее затем была сформирована матрица "*участник-участник*", по которой и была проведена кластеризация (рисунок 10).

Средний размер кластера пользователей-участников сообществ можно рассматривать как индикатор неизбирательности пользователей. Тогда увеличение активности спам-блогов и степени инфильтрации сообществ спамом должно проявляться в заметном увеличении среднего размера кластера пользователей, поскольку эффективность инфильтрации спамом достигается за счет автоматических методов, ориентированных на уязвимости блогов, а не на их содержание.

Как видно из рисунков 9 и 10, наибольшие изменения произошли для низких значений порога *Th*. При этом средний размер кластера при тех же значениях порога быстрее растет для пользователей, нежели для сообществ.

Для сравнения на рисунке 11 приведен результат кластеризации пользователей со случайной выборкой пользователей. Тенденция похожа на ту, что приведена на рисунке 10, однако изменения имеют более плавный характер.

6. Заключение

Сравнительное изучение атрибутов профилей сообществ LiveJournal в 2010 и 2011 годах показало, что содержательные изменения списка интересов в профиле охватывают относительно небольшое число сообществ, причем изменяются в большей степени характеристики средне- и высокочастотных интересов.

Заметными и устойчивыми тенденциями являются: доминирование сообществ с нулевым числом читателей и с нулевым числом комментариев (при общем уменьшении обеих величин со временем у остальных сообществ), устойчивое процентное преобладание сообществ с нулевым числом записей в журнале.

Кластеры на основе интересов и кластеры на основе пользователей демонстрируют различную динамику изменений во времени. Первые - более консервативны, и для них характерна «эволюционная» модель изменений, тогда как во втором случае динамика больше тяготеет к «шумоподобной» модели, отражая в существенной степени влияние привходящих извне факторов.

Очень заметные изменения по кластерам для пользователей-участников сообществ при малых значениях порога *Th* могут отражать активность спам-блогов и мер противодействия спаму со стороны пользователей и администраторов блог-хостинга.

Средний размер кластера может рассматриваться как индикатор специфичности (для интересов и сообществ) или избирательности (для пользователей).

Конечно, исследование проводилось в целом по всей коллекции (или ее большому подмножеству) профилей сообществ, что позволило увидеть только самые общие тенденции. Применение отработанных на всей коллекции подходов к отобранным специальным образом подмножествам сообществ могло бы дать более конкретные и содержательные результаты.

Для систем мониторинга информационных процессов в глобальных и корпоративных социальных сетях необходима разработка методики комплексного изучения социально-сетевой структуры, в том числе образованной множеством профилей сообществ и пользователей. Как представляется, данная методика должна включать в себя анализ изменений частот, рангов и характеристик распределений атрибутов профилей сообществ или пользователей.

Литература

[1] «О чем врут пользователи социальных сетей?». Пресс-выпуск №1691// Всероссийский центр изучения общественного мнения. [Электрон. ресурс] - Режим доступа: (http://wciom.ru/index.php?id=459&uid=111364&ashmanov&ashmanov)

[2] Состояние блогосферы российского интернета. По данным поиска по блогам Яндекса. Весна 2009 г. [Электрон. ресурс] – Режим доступа: (http://download.yandex.ru/company/yandex_on_blogosphere_spring_2009.pdf)

[3] Сычев А.В., Гадебский И.А. Изучение характеристик сообществ русскоязычной блогосферы // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: труды X Всероссийской научной конференции «RCDL'2008», 7-11 октября 2008 г., Дубна – Дубна: ОИЯИ, 2008. – С.200-209.

[4] A.Thomason "Blog Spam: A Review". [Электрон. ресурс] – Режим доступа: (http://www.ceas.cc/2007/papers/paper-85.pdf).

[5] Akismet.com – The automatic spam killer (http://akismet.com/stats/)

[6] D.Fetterly, M.Manasse, M.Najork "Spam,damn spam, and statistics. Using statistical analysis to locate spam web pages" // 7-th

International Workshop on the Web and Databases, June17-18, 2004, Paris, France.
[7] M. Goetz, J. Leskovec, M. McGlohon, C. Faloutsos. Modeling blog dynamics // Proceedings of the International Conference on Weblogs and Social Media (May 2009) [Электрон. ресурс] - Режим доступа: (www.cs.cornell.edu/~goetz/pdf/2009ICWSM.pdf)
[8] M. Z.Shafiq, A.X. Liu. A Random Walk Approach to Modeling the Dynamics of the Blogosphere//Proceedings of the International Conferences on Networking (Networking), Valencia, Spain, May 2011. [Электрон. ресурс] - Режим доступа: (https://www.msu.edu/~shafiqmu/files/Networking 2011-BlogModeling.pdf).

A Comparative Study of Profile Attributes for LiveJournal Communities in 2010-2011 Years

© A.V. Sychev

Comparative study (2011 year vis 2010 year) of attributes presented in profiles of the Russian language communities hosted on LiveJournal servers was conducted. Results of the study demonstrate dynamics of the total number of unique interests among all communities, interests frequencies, clusters of interests, communities and community members.