ПРОВЕРКА ГИПОТЕЗ О ПОВЕДЕНИИ ПОЛЬЗОВАТЕЛЕЙ СОЦИАЛЬНОЙ СЕТИ С ИСПОЛЬЗОВАНИЕМ ROLAP-MINING СИСТЕМЫ

Петров А.И.

Петров Алексей Иванович – бакалавр,
кафедра инструментального и прикладного программного обеспечения,
Институт информационных технологий
Московский Технологический Университет, г. Москва

Аннотация: *в статье рассматривается процесс работы с ROLAP-Mining системой для проверки гипотез о поведении пользователей социальной сети «ВКонтакте» на основе открытых данных. В частности, рассматривается процедура проектирования хранилища данных, процедура его наполнения и, непосредственно, проверка гипотез.*

Ключевые слова: *интеллектуальный анализ данных, проверка гипотез, анализ социальных сетей, ROLAP, R.*

УДК 004.4

Последние несколько лет активно развивается такое направление в информационных технологиях, как интеллектуальный анализ данных (data mining). Оно подразумевает анализ больших объемов данных разными методами и разными инструментами, как например, аналитический анализ данных с использованием OLAP-системы или статистический анализ с помощью языка программирования R.

Одна из актуальных на сегодняшний день тем для анализа данных является анализ открытых данных социальных сетей[3]. Цели такого анализа: выявление разного рода неявных закономерностей в данных, что может поспособствовать в описании поведений как отдельных пользователей, так и целых сообществ.

В данной статье рассматривается процесс проектирования хранилища данных для ROLAP-Mining системы[1], сбор открытых данных из социальной сети «ВКонтакте»[2], наполнение этими данными хранилища и, наконец, пример проверки гипотез.

ROLAP-Mining система работает с хранилищем данных, которое обычное имеет специфичную структуру и не используется в качестве основной базы данных проекта. Данные в такие хранилища выгружаются из обычных баз данных, используемых на конечных программных системах. Поэтому сперва была спроектирована и создана база данных, в которую был загружен большой набор данных из социальной сети. Стоит отметить, что перед загрузкой данных, был произведен анализ информационных моделей социальной сети, в результате которого и была составлена схема базы данных системы с учетом необходимых для дальнейшего анализа атрибутов моделей и их оптимального использования. Схема базы данных представлена на рисунке 1.

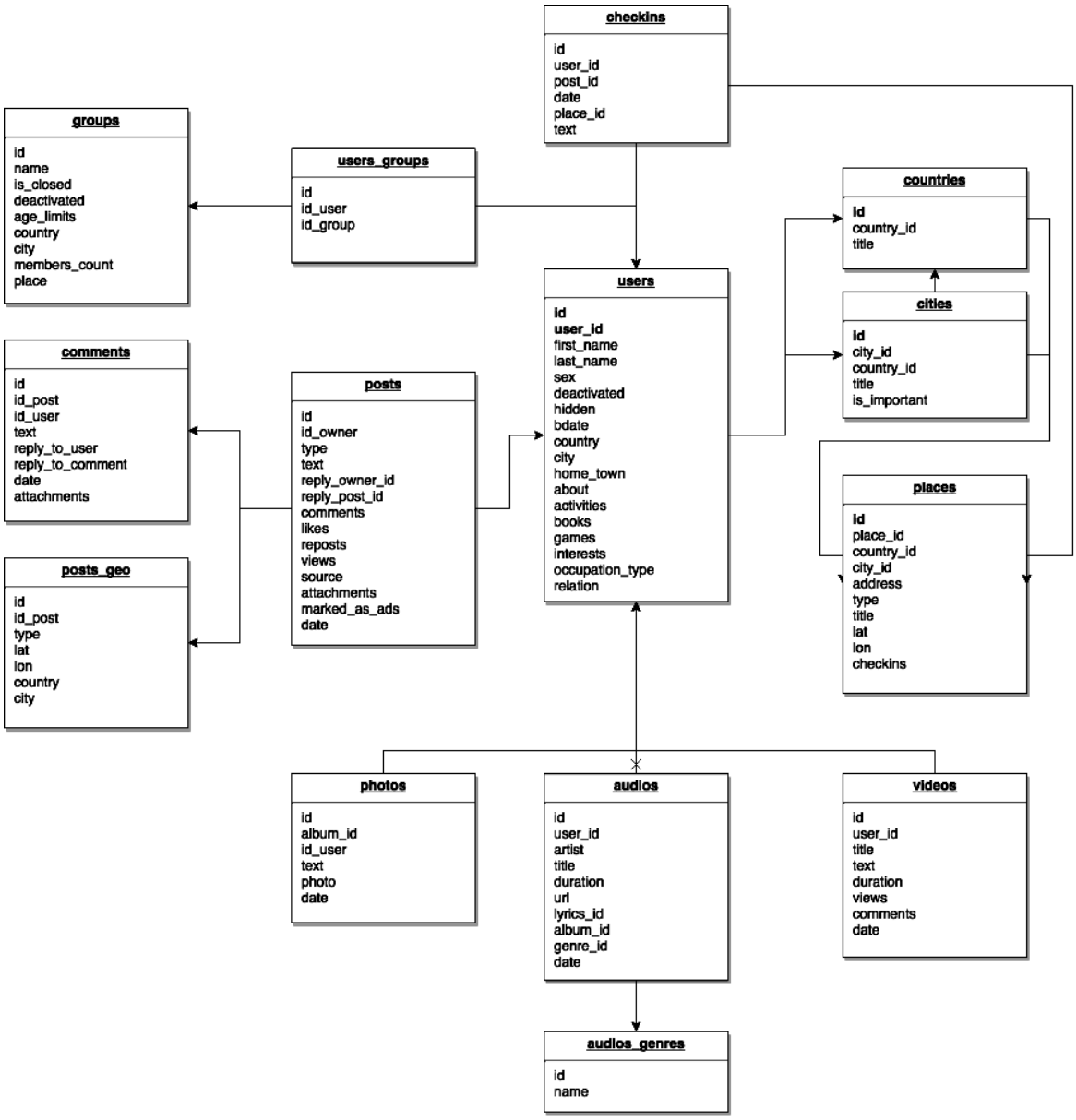

Рис. 1. Схема базы данных

Затем имеющаяся структура базы была проанализирована и были выработаны несколько вариантов возможных схем хранилища данных для дальнейшего его использования.

Конечная схема хранилища данных, выбранная для использования, представлена на рисунке 2.

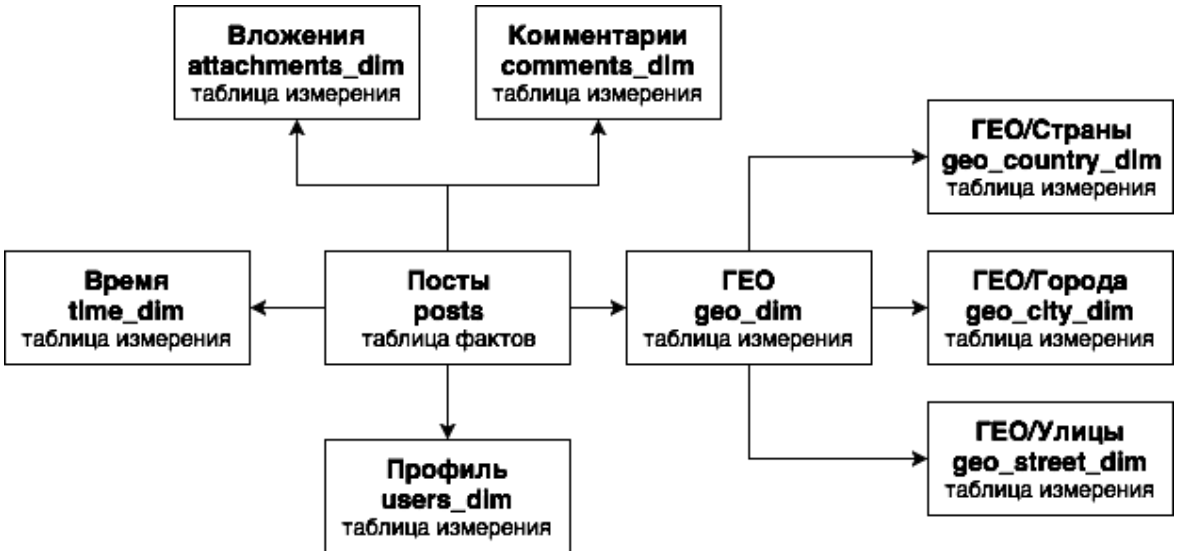

Рис. 2. Схема хранилища данных

После проектирования и создания хранилища данных, используя язык Python и библиотеку vk_requests, для работы с API ВКонтакте, были разработаны скрипты для загрузки данных из социальной сети в локальную базу данных.

Общее время работы скриптов по загрузке данных составило около 1 дня. За это время было загружено:

– 59 городов из России, Украины, Белоруссии и Казахстана;

– 12 007 самых известных мест (по данным соц.сети) по каждому и городов;

– около 140 тысяч публикаций и профилей пользователей.

Параллельно с загрузкой данных разрабатывался скрипт для переноса данных из БД в хранилище данных.

Наконец, после переноса данных в хранилище данных, была запущена ROLAP-Mining система и соответствующим образом настроена для работы с хранилищем данных.

Следующим этапом является составление гипотез и их проверка с целью поиска неявных закономерностей в собранных данных.

В качестве гипотез для проверки, были рассмотрены следующие:

1) В выходные дни большинство пользователей социальных сетей отдыхают в том числе от использования социальных сетей, по этой причине в понедельник наблюдается наибольшая активность пользователей;

2) Есть ли зависимость между месяцем года и тем, прикрепляют ли пользователи к своим записям файлы.

Для проверки данной гипотезы достаточно произвести выборку из хранилища данных в виде куба:

- количество отметок «мне нравится» или количество просмотров записей в качестве меры;
- дни недели и города как измерения куба.

Результат запроса представлен на рисунках 3-4.

	0	1	2	3	4	5	6
City	Views count	Views count	Views count	Views count	Views count	Views count	Views count
Волгоград	16464	13782	9436	16509	11410	13204	17625
Воронеж	54630	26816	24189	34114	24168	32840	57391
Екатеринбург	72456	48262	46108	74914	87830	66080	57422
Казань	64984	35130	39501	35819	48273	56085	55966
Краснодар	42309	25869	22362	27816	21156	29479	25176
Красноярск	33788	22849	12605	22199	20171	20325	29655
Москва	211135	142780	161088	149426	148490	196763	177815
Новосибирск	46650	28289	27285	24356	34867	41313	40065
Ростов-на-Дону	37880	29995	24243	29429	22493	36252	30806
Самара	26627	14383	7091	14092	15643	14777	20543
Санкт-Петербург	206490	124416	123046	111024	130767	138545	192934
Челябинск	46240	25267	23840	27974	23080	28893	37513
Днепропетровск (Днепр)	24688	15831	15276	21278	19549	22784	31424
Донецк	17554	6570	12398	10735	9421	12996	21549
Киев	98396	54324	64871	76466	54713	48711	65592

Рис. 3. Результат запроса для проверки 1-й гипотезы

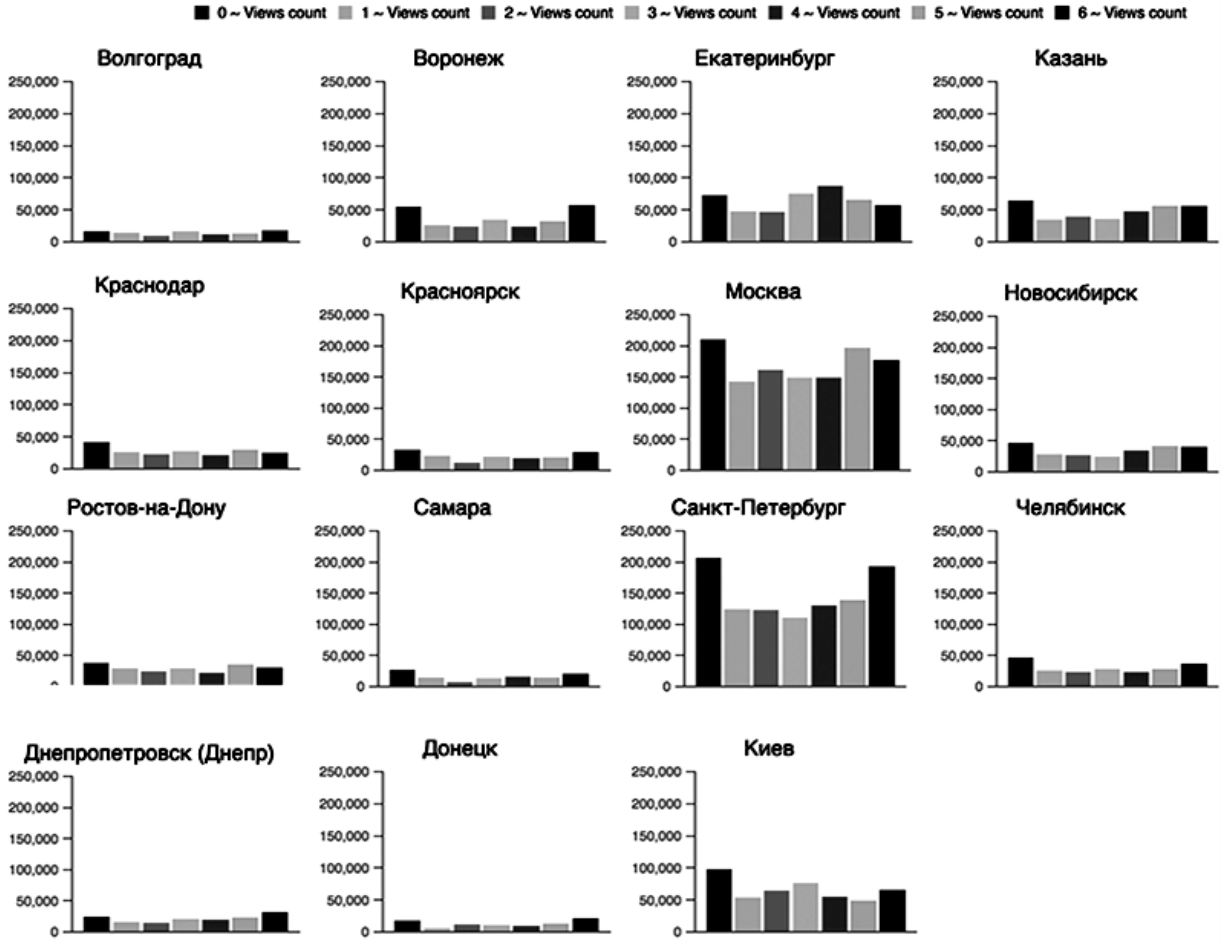

Рис. 4. Результат запроса проверки 1-й гипотезы

Из полученных результатов видно, что количество просмотров записей в городах Киев, Санкт-Петербург, Москва, Казань, Екатеринбург и Воронеж максимально в понедельник, немного меньшее количество просмотров в субботу и воскресенье.

Таким образом, можно утверждать, что гипотеза подтверждена.

Для проверки данной гипотезы необходимо произвести выборку из хранилища данных в виде куба:

– количество записей с вложениями в качестве меры;

– месяцы и города как измерения куба.

Результат запроса представлен на рисунках 5-6.

City	янв	фев	март	апр	май	июнь	июль	авг	сен	окт	ноя	дек
Волгоград	200	189	271	259	198	142	142	218	255	179	126	171
Воронеж	351	261	349	301	149	87	121	203	192	226	200	286
Екатеринбург	595	513	639	799	486	359	404	418	382	465	480	533
Казань	582	406	499	671	435	308	441	595	374	429	360	380
Краснодар	351	325	401	495	296	190	228	246	258	314	296	267
Красноярск	204	190	280	385	141	141	130	133	166	183	202	161
Москва	1420	1116	1776	1622	667	352	405	502	545	941	768	1239
Новосибирск	362	328	385	452	159	194	162	167	222	242	251	270
Ростов-на-Дону	298	289	376	453	270	169	177	242	237	246	256	258
Санкт-Петербург	1347	1146	1496	1446	724	466	525	642	555	686	843	904
Челябинск	299	252	376	345	207	172	156	199	281	254	221	292
Днепропетровск (Днепр)	264	248	272	283	176	117	144	147	181	236	198	226
Донецк	178	121	188	188	98	71	80	95	87	132	91	126
Киев	655	537	622	967	440	272	308	333	376	423	383	626

Рис. 5. Результат проверки 2-й гипотезы

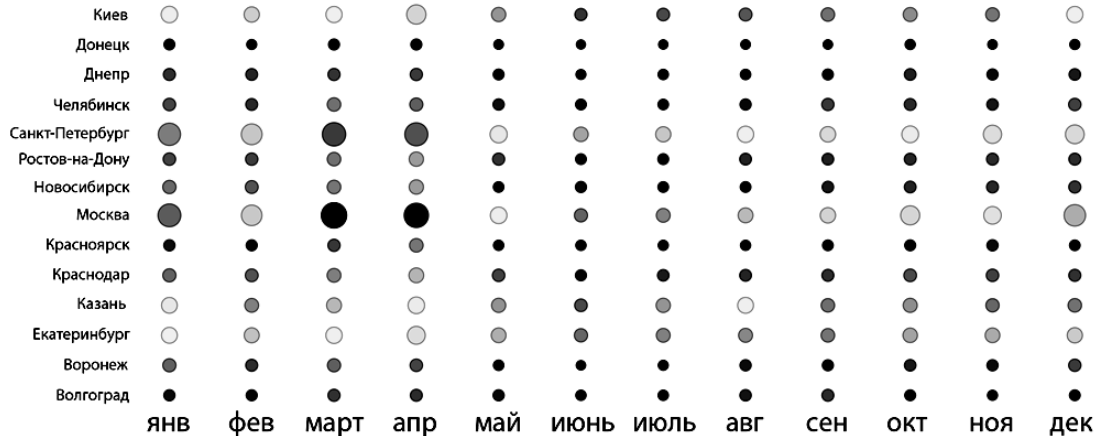

Рис. 6. Результат проверки 2-й гипотезы

Из полученных результатов видно, что в некоторых городах есть зависимость между месяцем года и тем, прикреплены ли файлы к записям. При этом видно, что максимальное количество файлов прикрепляются к записям в марте и апреле. Также можно заметить, что в июле и августе количество файлов прикрепленных к записям для городов Санкт-Петербург и Казань превышает по значению Москву, хотя в остальные месяцы показатели в Москве выше.

Таким образом, можно утверждать, что для некоторых городов действительно существует зависимость между месяцем года и количеством записей с прикрепленными к ним файлами.

В заключении можно сделать вывод о том, что рассмотренная ROLAP-Mining система уже готова для проведения научных исследований и проверки гипотез, что было продемонстрировано на конкретных примерах. В дальнейшем планируется автоматизировать некоторые из этапов работы с данной системой.

Список литературы

1. *Петров А.И., Чумак Б.Б.* ROLAP-Mining система на основе свободно-распространяемого программного обеспечения // Научный альманах 2017 N 5-3 (31).
2. Документация / Разработчикам. [Электронный ресурс]. Режим доступа: https://vk.com/dev/manuals/ (дата обращения: 02.10.2017).
3. *Смирнова О.С., Петров А.И., Бабийчук Г.А.* Основные методы анализа, используемые при исследовании социальных сетей // Современные информационных технологии и ИТ-образование. Т. 12 (№ 3). Часть 1, 2016. С. 151–158.