ПОСТРОЕНИЕ ХРАНИЛИЩА ДАННЫХ ДЛЯ ЗАДАЧ ОБРАБОТКИ ДАННЫХ ПОЛЬЗОВАТЕЛЕЙ СОЦИАЛЬНЫХ СЕТЕЙ

Мысливец О. Р., Рудикова Л. В.
Кафедра программного обеспечения информационных технологий, Белорусский государственный университет информатики и радиоэлектороники
Минск, Республика Беларусь
E-mail: myslivec.oleg@yandex.ru, rudikowa@gmail.com

В задачах многомерного анализа данных важными моментами являются правильная разработка и построение хранилища данных. В данной статье описывается процесс построения хранилища данных для задач обработки данных пользователей социальных сетей для последующего выявления закономерностей в образовании социальных групп в сети Интернет.

Введение

Обработка данных пользователей социальных сетей является в настоящий момент довольно перспективным направлением в анализе данных. Анализ поведения пользователей, сбор и накопление данных, анализ характерных признаков пользователей, анализ поведения пользователей, развитие систем рекомендации на основе предпочтений пользователей и прогнозирование связей в социальных группах - наиболее распространенные приложения для анализа данных социальных сетей и интернет-ресурсов. В связи с этим возрастает необходимость в развитии концепций сбора, хранения и обработки информации подобного рода. Для решения задач подобного рода разумно использовать многомерный анализ данных [1]. Несмотря на то, что огромное количество информации накапливается во всем мире, часто подобные данные имеют весьма разрозненный вид и далеки от предоставления общей картины жизни общества.

I. Данные социальных сетей

Данные социальных сетей, которые будут в последствии поступать в хранилище данных собираются с помощью API соответствующих социальных сетей. Несмотря на то, что часть данных многих пользователей может быть скрыта, имеется возможность получить такую основную информацию как [2]:

- базовая информация о пользователе (имя, фамилия, год рождения, город и т.д);
- информация о месте учебы (школа, университет, институт);
- информация о подписках пользователей;
- информация о постах пользователей.

Информация о постах пользователей является наиболее ценной для анализа, так как, как правило, несет в себе краткую характеристику пользователя, что позволяет охарактеризовать его в некой мере и выделить группы, к которым пользователь может принадлежать. Также каждый пост пользователя может включать геолокацию, хеш-теги и некоторые полезные сведения (текст, видео и аудио файлы). Исходя из всех полученных сведений с помощью создаваемого хранилища станет возможным анализировать пользователей исходя из их предпочтений и выявлять скрытые зависимости и тенденции интересов пользователей.

II. Проектирование хранилища данных

Основным аспектом в концепции проектирования системы на уровне хранения и работы с данными является подход с использованием хранилища данных (Data - Warehouse) – предметно-ориентированной информационной базы данных, построенной на основе схемы «созвездие фактов», специально разработанной и предназначенной для подготовки отчётов и бизнес-анализа с целью поддержки принятия решений. Данные, которые поступают в хранилище данных, доступны, в основном, только для чтения и поступают из внешних источников [3].

Предполагается, что данные будут поступать из различных социальных сетей, а это значит, что для получения и обработки данных необходимо использовать ETL-процесс. Данные социальных сетей являются слабоструктурированными и тот факт, что каждая социальная сеть хранит информацию в своих собственных структурах и предоставляет доступ к ней различными методами, делает процесс получения и очистки данных важным этапом при проектировании всей системы сбора и анализа данных.

Данные из различных социальных сетей собираются путем использования соответствующих API-сервисов. Полученная информация проходит процесс очистки - убираются все неинформативные данные. Далее производится структурирования данных - информация из различных социальных сетей объединяются и записываются в промежуточную базу данных.

Концептуальная модель базы данных для промежуточного хранения информации о пользователях представлена на рисунке 1.

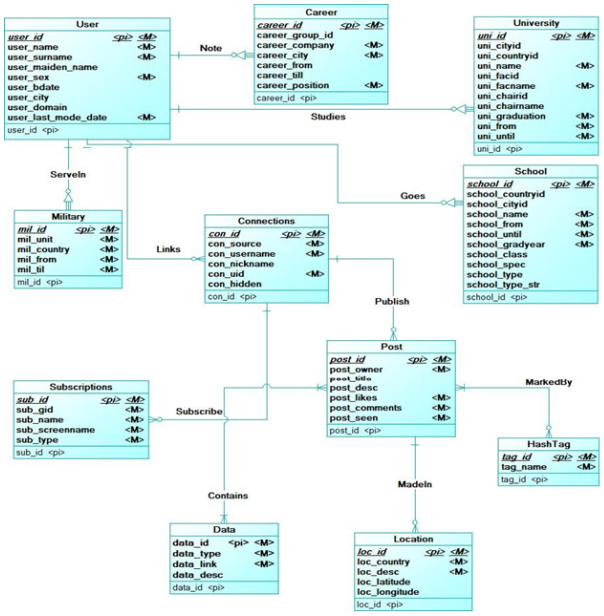

Рис. 1 – Концептуальная модель базы данных для промежуточного хранения информации о пользователях

Так как данные в хранилище будут поступать из различных социальных сетей, важно выявить и реализовать соответсвия между неструктурированными данными пользователя и данными, которое будут записываться в хранилище. Основные данные о пользователе, которые возможно получить через API социальных сетей представлены в сущности user (см. рис 2):

User			
user_id	<pi>	Integer	<M>
user_name		Text	<M>
user_surname		Text	<M>
user_maiden_name		Text	
user_sex		Byte	<M>
user_bdate		Date	
user_city		Text	
user_domain		Text	
user_last_mode_date		Date & Time	<M>
user_id <pi>			

Рис. 2 – Получаемая информация о пользователе

Действия пользователей в социальных сетях,в основном, связаны с написанием постов на своих страницах и репостом сообщений из групп или страниц других пользователей. Данная информация и представляет наиболее ценный интерес. Можно получить следующие данные о записях пользователя:

- данные геолокации;
- хеш-теги;
- группа или сообщество, откуда пользователь взял информацию;
- тип информации.

Основными объектами анализа являются географическое расположение пользователя, тип информации и набор хеш-тегов. Несмотря на тот факт, что из социальных сетей можно собрать гораздо больше дополнительной информации, вышеперечисленные данные будут являтся основой для построения OLAP-куба. Куб (см. рис. 3) реализован на основе схемы звезда и позволит анализировать действия пользователей на основе записей на их странице. Как видно из модели хранилища данных, часть информации не будет использована при анализе пользователей.

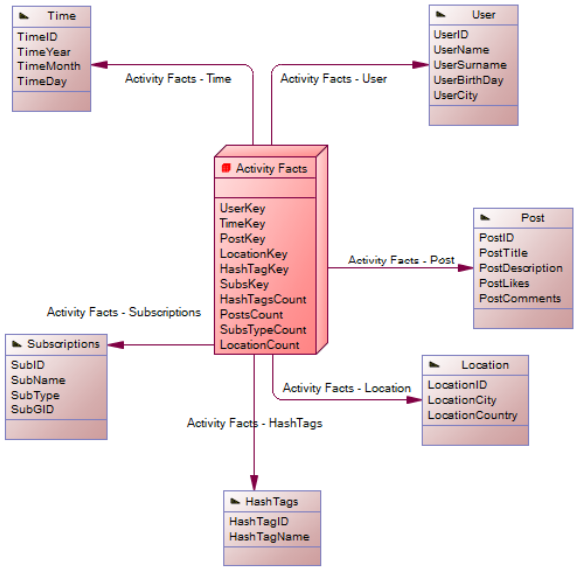

Рис. 3 – Модель OLAP-куба

III. Заключение

В данной статье представлена информация о построении хранилища данных для проверки простых гипотез и выявления закономерностей в образовании пользователями социальных групп на основе их предпочтений и записей на из страницах в социальных сетях. Представлена модель OLAP-куба для послебующего анализа информации о пользователях. В последствии планируется развитие модели для возможности проверки более сложных гипотез и поиска скрытых предпочтений пользователей на основе имеющейся информации.

IV. Список литературы

1. Анализ данных социальных сетей: методы и приложения [Электронный ресурс] / ИСП РАН Режим доступа: http://www.ispras.ru/ – Дата доступа: 18.09.2018.
2. Open API [Электронный ресурс] / Знакомство с API ВКонтакте Режим доступа: https://vk.com/dev/openapi – Дата доступа:18.09.2018.
3. Рудикова, Л.В. Об общей архитектуре универсальной системы хранения и обработки данных практико-ориентированной направленности // Л.В. Рудикова / Системный анализ и прикладная информатика. – Мн.: БНТУ, 2017. – №2. – С. 12-19.