МОНИТОРИНГ СОЦИАЛЬНЫХ СЕТЕЙ С ИСПОЛЬЗОВАНИЕМ ОБЛАЧНЫХ ВЫЧИСЛЕНИЙ

А.В. Якушев
СПбГУ ИТМО
E-mail: yaja30@gmail.com

Последнее время набирают все большую популярность различные социальные сети, которые служат отпечатком явлений, протекающих в реальном мире [1,2]. Поэтому анализ и мониторинг информации, содержащейся в социальных сетях, позволит на качественно новом уровне обеспечить исследования в области социодинамики – раздела социологии, посвященного количественным методам моделирования взаимоотношений между индивидами или группами.

Мониторинг представляет собой систематический сбор и обработку информации о состоянии объекта. Краулинг (crawling) сети можно рассматривать как процесс мониторинга явлений, протекающих в сети. Краулинг может осуществляться разово, для оценки текущего состояния сети, или периодически, в случае если необходимо выявить изменения в сети или необходимо принять решение по определенному вопросу. Для осуществления периодического краулинга необходим полный анализ параметров сети, полученных на текущей итерации, и их сравнение с параметрами, полученными на предыдущих итерациях.

Последний факт означает, что от системы, осуществляющей краулинг, требуются большие мощности, как вычислительные, так и мощности интернет-канала, которые должны быть доступны по запросу клиента. Для периодического мониторинга характерным является неравномерность запросов к ресурсам, поэтому для технической реализации такой системы хорошо подходит модель облачных вычислений.

В работе рассматривается система мониторинга социальных сетей на основе многофункциональной инструментально-технологической платформы CLAVIRE (CLoud Applications VIRtual Environment). Платформа обеспечивает доступ к сервисам и композитным приложениям в среде облачных вычислений, в которую краулер и программы для анализа собранных данных о сети встраиваются как сторонние сервисы. На рис. 1 представлена схема функционирования системы мониторинга, связанная с предоставлением сервисов доступа к данным и приложениям в области социодинамики и краулинга социальных сетей.

Реализуемые в системе сервисы сбора данных в социальных сетях используют различные модели краулинга (обход в глубину, в ширину) с оценкой общности по различным факторам, включая семантический профиль узлов сети. В распределенной среде эффективным является распараллеливание сетевого канала в рамках модели облачных вычислений, когда запросы к базе отправляются одновременно с разных целевых систем. На каждой целевой системе функционирует рабочий агент краулера. Он получает задание на просмотр определенного множества узлов сети. После выполнения задания он передает данные в централизованное хранилище. Действия отдельных агентов не синхронизируются. Функции управляющего узла (мастера) заключаются в том, что он определяет порядок, в котором будут обходиться пользователи сети, тем самым он реализует политику обхода краулера. Архитектура краулера с централизованным управлением позволяет динамически добавлять и удалять агентов, обеспечивая масштабируе-

мость системы в целом. Помимо классических политик обхода (обход в ширину и в глубину) таким образом может быть дополнительно реализована политика обхода по степени влияния: сначала посещаются те узлы, на которые ведет самое большое число ссылок. Эта эвристика позволяет обходить сеть по топологическим сообществам – множествам тесно связанных друг с другом вершин.

Для эффективного сбора информации в социальных сетях важно обеспечить высокую производительность краулера, что достигается за счет баланса операций по просмотру и записи данных в социальной сети и операций по их передаче в Интернет. Например, в социальной сети Live Journal (ЖЖ) за один день функционирования краулер обрабатывает данные около 700 тысяч пользователей сети со средней скоростью работы 490 пользователей в минуту. При этом выполняется около 270 итераций (которые соответствуют заданиям отдельным агентам). Анализ структуры временных затрат показал, что наиболее ресурсоемкими являются операции работы с базой данных (около 70%), в частности, сохранение связей между пользователями (18.6%) и списков интересов пользователей (39.4%). Временные затраты на работу с сетью не превышают 27%, что указывает на необходимость оптимизации доступа к базе данных.

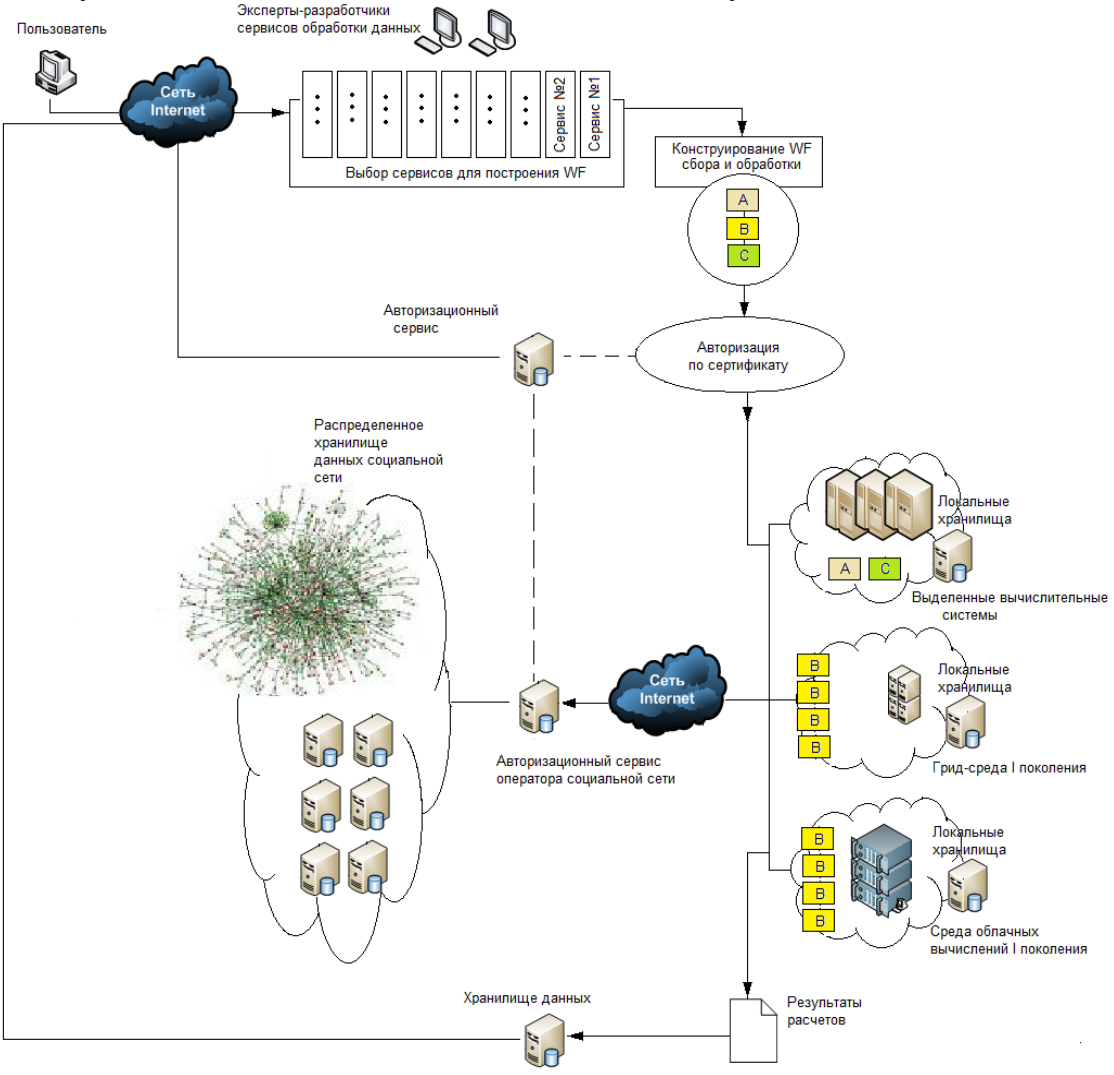

Рис. 1. Схема функционирования производственно-исследовательского web-центра в области социодинамики

Построение линейной регрессионной модели работы краулера позволило оценить его производительность для сетей с другими параметрами (например, с другим законом

распределения степеней вершин), а так же позволило выделить итерации, в которых краулер работал аномально долго, и понять причины возникновения таких ситуаций. Анализ регрессионных остатков показал, что их значения уменьшаются с ростом номера итерации, то есть время работы краулера уже не так хорошо аппроксимируется прямой линией. Это объясняется тем, что с увеличением размера базы, время работы с ней увеличивается.

Дальнейшее развитие мы видим в улучшении производительности краулера и расширении его функциональности, позволяющем помимо базовой информации сохранять посты пользователей. Анализ этих постов, посредством методов текстового анализа, позволит лучше понимать законы распространения информации в сети.

Литература

1. Mika P. Social Networks and the Semantic Web (Semantic Web and Beyond), Springer, 2007. P. 234. ISBN: 9780387710006.
2. Hu D., Kaza S., Chen H. Identifying Significant Facilitators of DarkNetwork Evolution, J. of the American Society for Inf. Science and Technology 60(4), 2009. P. 655–665.