УДК 681.3.069, 681.324

С. В. Иванов, Е. В. Болгова, В. В. Каширин, А. В. Якушев,
А. В. Чугунов, А. В. Бухановский

WEB-ОРИЕНТИРОВАННЫЙ ПРОИЗВОДСТВЕННО-ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР „СОЦИОДИНАМИКА“

Рассмотрены концепция и принципиальная архитектура проблемно-ориентированной среды облачных вычислений, обеспечивающей функционирование производственно-исследовательского центра, для исследования социодинамики и ее приложений, в рамках парадигмы web 2.0.

Ключевые слова: *социодинамика, социометрия, облачные вычисления, прикладные сервисы, социальная сеть, краулер, распространение слухов.*

Введение. Развитие информационных технологий стимулирует появление новых методов и направлений исследований в различных предметных областях. В частности, бурный рост числа пользователей социальных сетей в Интернете обеспечивает информационную базу, позволяющую на качественно новом уровне обеспечить исследования в области социодинамики — раздела социологии, посвященного количественным методам моделирования взаимоотношений между индивидами или группами. Традиционно развитие социодинамики ограничивалось социометрическим фактором — возможностью наблюдения (измерения) соответствующих процессов в обществе, поскольку измерения и анализ парных или групповых взаимодействий индивидов гораздо более сложны, чем их индивидуальных характеристик в рамках выборочного подхода. Однако в глобальных социальных сетях подобные взаимоотношения виртуализируются, формируя, таким образом, слепок общественной структуры в пространстве Интернета [1, 2]. Несмотря на то что поведение и характеристики пользователей социальных сетей в ряде аспектов могут существенно отличаться от реальных, во многом эти факторы имеют систематический характер, что позволяет их учитывать (путем смещения определенных параметров) при обработке и анализе соответствующих данных [3]. Таким образом, социальные сети в настоящее время, по-видимому, составляют основу социометрических исследований нового поколения.

В настоящей статье рассматриваются концепция и принципиальная архитектура специализированного web-ориентированного производственно-исследовательского центра, который предоставляет возможности сбора, анализа и использования в моделях социометрических данных социальных сетей в Интернете, на основе технологий облачных вычислений, в соответствии с бизнес-моделью SaaS (Software as a Service, программное обеспечение как услуга).

Концепция web-ориентированного центра. Специфика выполнения количественных исследований на основе социальных сетей связана с рядом особенностей, ограничивающих доступность таких данных для широкого круга исследователей; отметим некоторые из них.

— Доступ к данным глобальных социальных сетей регламентируется политикой оператора сети и соответствующим законодательством в области персональных данных. Для масштабного сбора и анализа соответствующих данных необходимо наличие предварительных соглашений с оператором.

— Социальные сети имеют технологически различные интерфейсы доступа к данным, существуют разные принципы обхода сетей. Как следствие, проведение измерений характеристик

различных сетей требует разработки специализированных средств сбора данных, для отдельных исследователей это весьма трудоемкий процесс.

— Сбор данных в социальных сетях является достаточно ресурсоемкой операцией и требует соответствующих выделенных вычислительных ресурсов. Регулярное выполнение таких операций различными пользователями увеличивает нагрузку на инфраструктуру оператора сети, что нежелательно.

— Алгоритмы обработки и анализа таких социальных сетей во многих случаях имеют нелинейную сложность, поскольку описывают взаимоотношения „каждого с каждым“. Поэтому исследование сетей достаточно большого объема требует применения соответствующих вычислительных ресурсов и программного обеспечения, допускающего эффективное распараллеливание.

— Визуализация результатов исследований на основе социальных сетей связана с применением достаточно сложных когнитивных алгоритмов, позволяющих наглядно представить различные процессы на непланарных структурах данных большого объема, это приводит к необходимости использования специализированного программного обеспечения.

В совокупности перечисленные проблемы препятствуют развитию методов и технологий современной социометрии на основе социальных сетей в Интернете, это усугубляется тем, что потенциальные пользователи сетей (специалисты в науках об обществе) в большинстве случаев не обладают специализированными навыками для их самостоятельного решения. Выходом из сложившейся ситуации является эксплуатация концепции облачных вычислений — создание проблемно-ориентированной среды, обеспечивающей доступ к соответствующим сервисам (ресурсам, данным и процедурам их обработки и моделирования) через web-интерфейс. В качестве основы для развития среды рассматривается многофункциональная инструментально-технологическая платформа CLAVIRE (CLoud Applications VIRtual Environment) обеспечения доступа к сервисам и композитным приложениям в среде облачных вычислений.

Проблемно-ориентированная среда включает следующие элементы:

— управляющая оболочка (ядро платформы), которая состоит из взаимодействующих системных web-сервисов, функционирующих с низкоуровневой вычислительной структурой, осуществляющих поддержку образа облачной среды, мониторинг и управление ресурсами, конструирование и исполнение сценариев, а также поддержку пользовательских интерфейсов. Управляющая оболочка реализуется на основе концепции iPSE (Intelligent Problem Solving Environment) [4]. Фактически web-центр представляет собой открытую интеллектуальную проблемно-ориентированную среду, объединяющую распределенные сервисы вычислений и доступа к данным и позволяющую эффективно управлять параллельными вычислительными процессами в распределенной иерархической среде на основе интеллектуальных технологий [5];

— набор прикладных сервисов вычислений и доступа к данным. В него входят как различные прикладные пакеты для обработки данных и социодинамического моделирования, доступные в рамках концепции облачных вычислений, так и специализированные инструменты для поиска и извлечения данных сбора данных из социальных сетей — краулеры [6]. Все сервисы строятся на основе предметно-ориентированных описаний пакетов на языке EasyPackage, регистрируемых в базе пакетов управляющей оболочки [7];

— дополнительные средства, обеспечивающие поддержку виртуального профессионального сообщества пользователей в рамках концепции web 2.0. Они включают в себя интерактивные средства общения, совместного выполнения проектов, поддержки единого рабочего пространства, а также средств, позволяющих вести дискуссии в режиме реального времени с использованием графических и текстовых средств общения. Кроме того, имеются сервисы интерактивной консультации экспертов, предусматривается возможность сохранения результатов выполненных задач для последующего использования другими членами сообщества

или для совместного обсуждения и исправления. В качестве отдельной задачи рассматриваются сбор, обработка и анализ текущей и ретроспективной информации о процессах в виртуальном профессиональном сообществе (включая ряд показателей индивидуальной и коллективной активности пользователей, характеристики востребованности сервисов и пр.).

Аппаратная составляющая поддержки web-ориентированного центра формируется в составе распределенной иерархической среды облачных вычислений, включающей в себя выделенные суперкомпьютеры, виртуальные машины в „облаке“ и целевые системы в составе Грид. Единообразный способ работы с центром, равно как и оптимизация распределения вычислительной нагрузки, осуществляется средствами управляющей оболочки без привлечения пользователя.

На рис. 1 представлена схема функционирования web-центра, демонстрирующая работу сервисов доступа к данным и приложениям в области социодинамики.

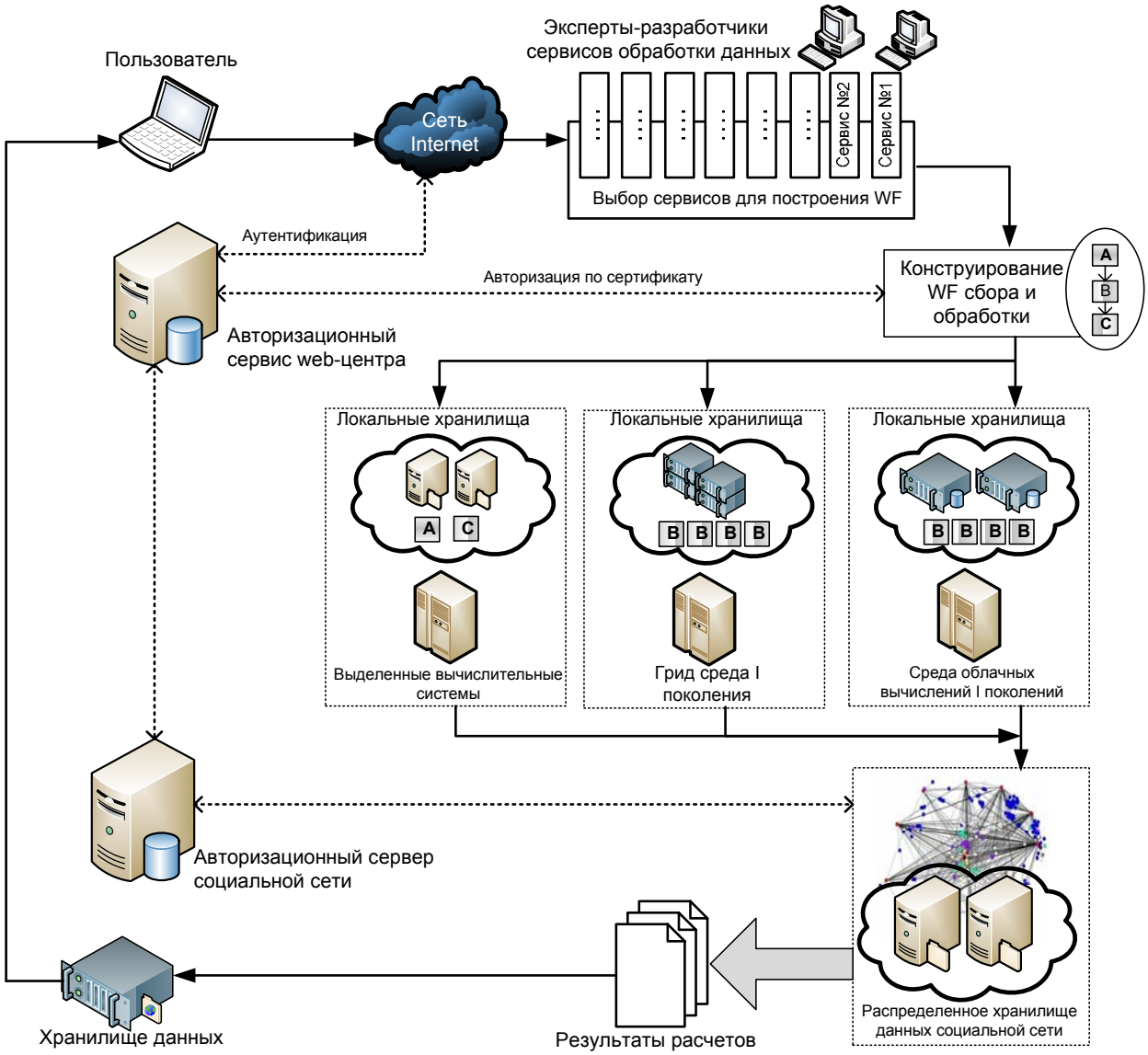

Рис. 1

Пользователь авторизуется в проблемно-ориентированной среде через портал провайдера. Через соответствующий web-интерфейс он может выбрать конкретные сервисы или шаблоны композитных приложений в форме потоков заданий (workflow, WF), а также получить (при необходимости) доступ к технической и эксплуатационной документации. Выбрав необходимые ему сервисы, пользователь средствами управляющей оболочки конструирует

соответствующее композитное приложение в форме WF, которое определяет правила сбора, обработки и анализа данных. При этом может использоваться как графическая, так и текстовая форма представления композитных приложений. Для подготовленного описания композитного приложения пользователь конфигурирует условия вычислений: определяет требуемые параметры WF, редактирует (при необходимости) его описание, готовит и загружает в хранилище среды входные данные для расчетов. В ряде случаев такие данные (например, учебные базы данных фрагментов социальных сетей) могут предоставляться провайдером web-центра.

Затем пользователь определяет режим исполнения задачи в среде (утверждает предлагаемые ему варианты) в соответствии с требованиями к временным характеристикам расчета и правилами доступа к различным источникам данных. При этом пользователю предлагаются различные тарифные варианты, связанные с использованием разных социальных сетей и привлечением ресурсов сред распределенных вычислений. Окончательно пользователь запускает задачу на исполнение в проблемно-ориентированной среде. Использование вычислительных ресурсов и сервисов сбора данных в социальных сетях производится с учетом единого сертификата, который обеспечивает права пользователя, делегированные провайдером. В процессе вычислений пользователь (при необходимости) осуществляет мониторинг процесса исполнения (в форме динамического отображения WF); при этом прогнозируется время завершения вычислений. Когда все расчеты завершены, результаты помещаются в хранилище данных web-центра; пользователю отправляется соответствующее уведомление (sms, e-mail и пр.). Пользователь может получить доступ к результатам расчетов через интерфейс проблемно-ориентированной среды.

Прикладные сервисы в составе web-ориентированного центра в соответствии с выполняемыми ими функциями можно условно разделить на четыре группы: сбора данных в социальных сетях, статистической обработки и анализа данных, моделирования сценариев, а также визуализации.

Сервисы сбора данных в социальных сетях реализуют различные модели краулинга (обход в глубину, в ширину) с оценкой общности по различным факторам, включая семантический профиль узлов сети. В рамках распределенной среды эффективно 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 передает данные в централизованное хранилище. Действия отдельных агентов не синхронизируются. Управляющий узел (мастер) 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 масштабируемость системы в целом. Помимо классических политик обхода (обход в ширину и глубину) таким образом может быть дополнительно реализована политика обхода по степени влияния, согласно которой сначала посещаются те узлы, на которые идет 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, а также операций по их передаче в Интернет. Например, в социальной сети Live Journal (Живой Журнал, ЖЖ) за один день краулер обрабатывает данные около 700 тысяч пользователей сети со средней скоростью 490 пользователей в минуту. При этом выполняется около 270 итераций (которые соответствуют заданиям для отдельных агентов). Анализ структуры временных затрат показал, что наиболее ресурсоемки операции 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.

Сервисы статистической обработки и анализа данных используют общий подход к описанию многомерных комплексных сетей — стохастических графов с многомерными характеристиками вершин. Под комплексной сетью [8] понимается граф с достаточно большим числом узлов, характеризуемых, в том числе, многомерным кортежем признаков, и динамически изменяющимися связями; распределение признаков узлов и характеристик связей может быть описано вероятностной моделью (многомерным распределением). Для их оценки используются методы многомерного статистического анализа, что обусловлено неопределенностью перехода к уравнениям относительно вероятностных характеристик сети в многомерном случае. Это требует совокупного применения формальных способов снижения размерности (обобщение метода главных компонент для графов), методов дискриминантного анализа для выявления характерных структур в сети, а также методов ординации (шкалирования) для учета многомерности признаков, описывающих узлы сети. При этом интерпретация результатов осложняется тем, что социальные сети могут включать в себя как формальные, так и неформальные сообщества. Формальным сообществом можно назвать группу индивидов, объединенных по какому-либо (формальному) признаку. Например, такие сообщества составляют взаимосвязанные пользователи, которые указали одним из своих интересов „книги“, „музыку“, „эзотерику“ и др. Напротив, неформальное сообщество включает индивидов с общими интересами, однозначно не отраженными в профиле. Поскольку пользователи социальных сетей зачастую указывают неполную или искаженную информацию в своих профилях, для исследования неформальных сообществ требуется совокупно использовать статистическую информацию о топологической структуре сети и наборе характеристик каждого индивида.

На рис. 2 представлен пример анализа предпочтений сообщества индивидов из социальной сети ЖЖ, в профиле интересов которых присутствуют упоминания о наркотиках; приведены интервальные оценки вероятности наличия интереса (P) для сети в целом и точечные — для выборки наркоманов.

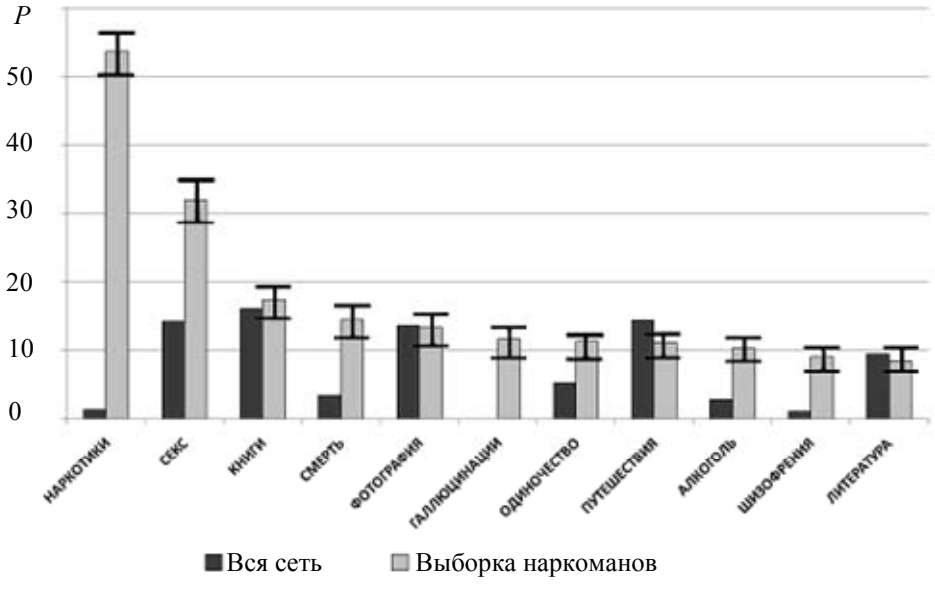

Рис. 2

Из рисунка следует, что проявление таких интересов у пользователей, как „музыка“, „книги“, „путешествия“, практически не связано с тем, что пользователь относит себя к формальному сообществу наркоманов. Наоборот, такие интересы, как „наркотики“, „смерть“, „одиночество“, характерны для пользователей из сообщества наркоманов и могут быть использованы в дальнейшем для выявления потенциальных потребителей наркотиков (группы риска).

Сервисы моделирования сценариев реализуют различные вероятностные модели социодинамики в терминах микро- и (или) макропараметров социальных сетей. Модель динамики комплексной сети задается в форме композиции стохастических эволюционных операторов над графом заданной структуры; каждый из операторов отражает определенный класс процессов в сети (присоединение новых пользователей, повышение рейтинга пользователя, расширение сферы научных интересов и пр.). Комплексная сеть характеризуется набором макропараметров (коэффициентов операторов), которые могут быть идентифицированы только по результатам измерений (посредством краулинга и обработки полученных данных). Посредством осреднения по ансамблю общее уравнение над графом сводится к системе обыкновенных дифференциальных уравнений, описывающих изменчивость отдельных вероятностных характеристик сети. Это позволяет исследовать чувствительность модели к изменению параметров на основе анализа фазовых портретов, в ходе чего могут быть выделены макропараметры, существенные для процесса мониторинга и управления.

В качестве иллюстрации на рис. 3 представлены результаты расчета распространения слухов в социальной сети на основе модели Далея—Кендалла [9]. В модели используются три множества вершин: неинформированные (*I*), распространяющие (*S*) и неактивные (*R*). На каждом шаге, при взаимодействии на вершины типа *I* с вершиной типа *S*, *I* с заданной вероятностью переходит во множество *S*, а при взаимодействии с вершиной из *S* или *R* — во множество *R*. Значения вероятностей можно задавать для всей сети, для каждого класса вершин и также для каждой пары классов, вершины которых вступают во взаимодействие. За шаг алгоритма можно взять сутки (что связано с цикличностью просмотра страниц пользователями социальных сетей), а взаимодействие между двумя вершинами определяется наличием связи между ними.

Сеть включает в себя два класса вершин. Распределение степеней первого класса характеризуется степенным законом с показателем степени 4, а второго — законом Пуассона с показателем 7. Вершины первого класса (80 %) определяют традиционную структуру социальной сети, тогда как второго (20 %) соответствуют сплоченному сообществу с большим количеством связей между участниками.

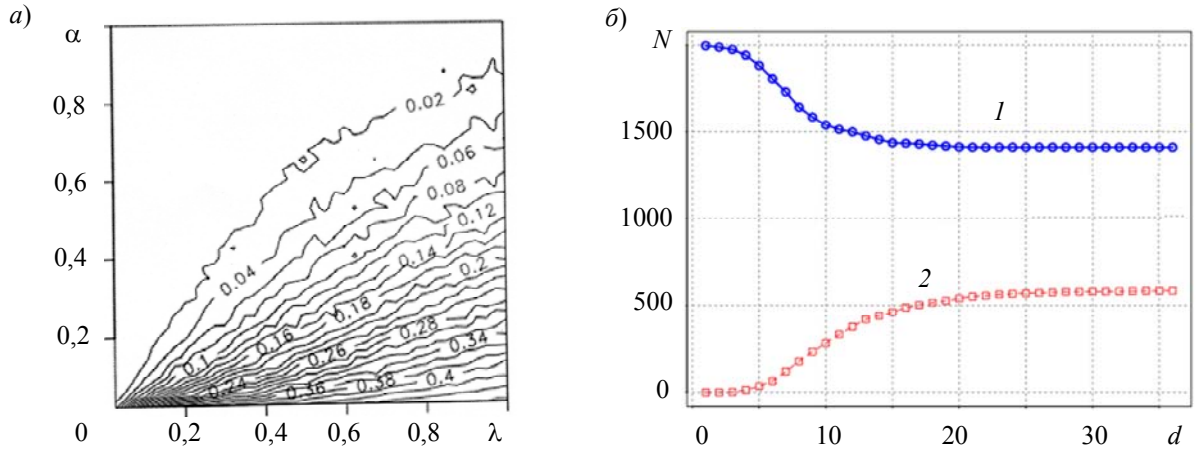

Рис. 3

На рис. 3, *а* приведен график, демонстрирующий степень покрытия вершин сети слухом в зависимости от параметров процесса распространения: вероятности передачи слуха (α) и его забывания (λ). Видно, что наилучшее покрытие обеспечивается при высокой вероятности передачи слуха и низкой вероятности его забывания. На рис. 3, *б* приведена динамика распространения слуха, выраженная через число вершин (*N*) во множествах *I* (*1*) и *R* (*2*) на определенном шаге работы алгоритма: видно, что с течением времени

(*d*) процесс выходит на определенную асимптоту — количество „осведомленных“ вершин не возрастает.

Сервисы визуализации ориентируются на применение моделей анализа и представления комплексных сетей, рассмотренных выше. В рамках web-центра используются пакеты Pajec, Enronic и JUNG, адаптированные к задачам серверной визуализации. В ходе выполнения расчетов визуализация выполняется на сервере, поддерживающем хранилище данных проблемно-ориентированной среды. Пользователь при этом имеет возможность просмотра статической картинки и (или) видеопотока средствами web-браузера (в зависимости от специфики решаемой задачи).

Заключение. Несмотря на то что, используя перечисленные сервисы, возможно выполнять разного рода расчеты в области социометрии и социодинамики, основным назначением web-центра является решение комплексных задач, требующих совокупного применения сервисов сбора, анализа, моделирования и визуализации. При этом сценарий решения задачи не является жестко заданным, а описывается пользователем в форме композитного приложения на языке EasyFlow [7]. Описания типовых задач (в форме соответствующих WF) могут быть представлены в репозитории для общего использования. К ним, в частности, относится:

— построение социограммы неформального сообщества, анализ скорости и каналов распространения информации;

— анализ и прогноз индексов общественных настроений;

— выявление групп влияний в социальной сети и определение „лидеров мнений“;

— мониторинг манипуляций мнениями;

— выявление призывов к общественным/экстремистским акциям;

— обнаружение источников умышленной дезинформации.

Перечисленный перечень задач не является полным: поскольку проблемно-ориентированная среда в рамках концепции iPSE обладает открытой архитектурой, пользователи могут добавлять в базу пакетов собственные сервисы и самостоятельно пополнять репозиторий собственными композитными приложениями. Таким образом, это обеспечивает дальнейшее развитие и востребованность web-центра в области социодинамики и ее приложений.

Работа выполнена в рамках реализации Постановлений № 218 и 220 Правительства Российской Федерации при поддержке ФЦП „Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007—2012 гг.“.

СПИСОК ЛИТЕРАТУРЫ

1. *Mika P.* Social Networks and the Semantic Web (Semantic Web and Beyond). Springer, 2007. 234 p.

2. *Hu D., Kaza S., Chen H.* Identifying Significant Facilitators of DarkNetwork Evolution // J. of the American Soc. for Inf. Sci. and Techn. 2009. Vol. 60, N 4. P. 655—665.

3. *Hanneman R. A. and Riddle M.* Introduction to social network methods. Department of Sociology at the University of California, Riverside [Электронный ресурс]: <textbook available at http://faculty.ucr.edu/~hanneman/nettext/>.

4. *Бухановский А. В., Ковальчук С. В., Марьин С. В.* Интеллектуальные высокопроизводительные программные комплексы моделирования сложных систем: концепция, архитектура и примеры реализации // Изв. вузов. Приборостроение. 2009. Т. 52, № 10. С. 5—24.

5. *Марьин С. В., Ковальчук С. В.* Сервисно-ориентированная платформа исполнения композитных приложений в распределенной среде // Там же. 2011. Т. 54, № 10. С. 21—29.

6. *Chau D. H., Pandit S., Wang S., Faloutsos C.* Parallel crawling for online social networks // Proc. of 16th Intern. Conf. on World Wide Web – WWW. 2007.

7. *Князьков К. В., Ларченко А. В.* Предметно-ориентированные технологии разработки приложений в распределенных средах // Изв. вузов. Приборостроение. 2011. Т. 54, № 10. С. 36—43.

8. *Newman M. E. J.* The Structure and Function of Complex Networks // Soc. for Industrial and Appl. Mathematics. 2003. Vol. 45, N 2. P. 167—256.

9. *Daley D., Kendall D.* Epidemics and rumours // Nature. 1964. Vol. 240, N 4963. P. 1118.

Сведения об авторах

Сергей Владимирович Иванов — канд. техн. наук; НИИ Наукоемких компьютерных технологий Санкт-Петербургского государственного университета информационных технологий, механики и оптики; старший научный сотрудник; E-mail: Sergey.v.ivanov@rambler.ru

Екатерина Владимировна Болгова — НИИ Наукоемких компьютерных технологий Санкт-Петербургского государственного университета информационных технологий, механики и оптики; младший научный сотрудник; E-mail: katerina.bolgova@gmail.com

Виктор Валерьевич Каширин — НИИ Наукоемких компьютерных технологий Санкт-Петербургского государственного университета информационных технологий, механики и оптики; младший научный сотрудник; E-mail: kashirin.victor@gmail.com

Андрей Владимирович Якушев — НИИ Наукоемких компьютерных технологий Санкт-Петербургского государственного университета информационных технологий, механики и оптики; младший научный сотрудник; E-mail: yaja30@gmail.com

Андрей Владимирович Чугунов — канд. политич. наук; Санкт-Петербургский государственный университет информационных технологий, механики и оптики, Центр технологий электронного правительства; директор; E-mail: chugunov@egov-center.ru

Александр Валерьевич Бухановский — д-р техн. наук, профессор; НИИ Наукоемких компьютерных технологий Санкт-Петербургского государственного университета информационных технологий, механики и оптики; директор; E-mail: avb_mail@mail.ru

Рекомендована НИИ НКТ

Поступила в редакцию
15.05.11 г.