Веб-ориентированный центр в области социодинамики: концепция и принципиальная архитектура

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

Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики
sergei.v.ivanov@gmail.com, katerina.bolgova@gmail.com, kashirin.victor@gmail.com, yaja30@gmail.com, chugunov@egov-center.ru, avb_mail@mail.ru

Аннотация

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

1. Введение

Развитие информационных технологий стимулирует появление новых методов и направлений исследований в различных предметных областях. В частности, бурный рост числа пользователей социальных сетей в Интернет 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 - возможностями наблюдения (измерения) соответствующих процессов в обществе, поскольку измерения и анализ парных или групповых взаимодействии индивидов гораздо сложнее, чем, например, их индивидуальных характеристик в рамках выборочного подхода. Однако в глобальных социальных сетях подобные взаимоотношения виртуализируются, формируя, таким образом, слепок общественной структуры в пространстве Интернет [1,2]. Несмотря на то, что поведение и характеристики пользователей социальных сетей в ряде аспектов могут отличаться от реальных, во многом эти факторы носят систематический характер, что дает возможность их учета (в форме 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 данной статье рассматривается концепция и принципиальная архитектура специализированного web-ориентированного производственно-исследовательского центра, который предоставляет возможности сбора, анализа и использования в моделях социометрических данных социальных сетей в Интернет, на основе технологий облачных вычислений, в соответствии с бизнес-моделью SaaS (Software as a Service).

2. Концепция web-ориентированного центра в области социодинамики

Специфика выполнения количественных исследований на социальных сетях связана с рядом особенностей, ограничивающих доступность таких данных для широкого круга исследователей. В частности:

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 и принципы их обхода. Как следствие, проведение измерений на различных сетях требует разработки специализированных средств сбора данных, что весьма трудоемко для отдельных исследователей.

Сбор данных в социальных сетях является достаточно ресурсоемкой операций. Как следствие, это требует от пользователя соответствующих выделенных вычислительных ресурсов. Дополнительно, регулярное выполнение таких операций различными пользователями увеличивает нагрузку на инфраструктуру оператора сети, что является нежелательным фактором.

Алгоритмы обработки и анализа данных социальных сетей во многих случаях имеют нелинейную сложность, поскольку описывают взаимоотношения «каждый с каждым». Потому исследование сетей достаточно большого объема

Труды XIV Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2011), Санкт-Петербург, Россия, 2011.

требует применения соответствующих вычислительных ресурсов и программного обеспечения, допускающего эффективное распараллеливание.

Визуализация результатов исследований на социальных сетях связана с применением достаточно изощренных когнитивных алгоритмов, позволяющих наглядно представить различные процессы на непланарных структурах данных большого объема. Как следствие, это приводит к необходимости использования специализированного программного обеспечения.

В совокупности перечисленные проблемы препятствуют развитию методов и технологий современной социометрии на основе социальных сетей в Интернет, что усугубляется тем, что их потенциальные пользователи (специалисты в науках об обществе) в большинстве случаев не обладают специализированными навыками для их самостоятельного решения. Как следствие, 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.

В состав проблемно-ориентированной среды входят следующие элементы:

1. **Управляющая оболочка** (ядро платформы), которая состоит из набора взаимодействующих системных web-сервисов, осуществляющих функции работы с низкоуровневой вычислительной структурой, поддержки образа облачной среды, мониторинга и управления ресурсами, конструирования и исполнения сценариев, а также поддержки пользовательских интерфейсов. Управляющая оболочка реализуется на основе концепции iPSE (Intelligent Problem Solving Environment) [4]. Фактически, веб-центр будет представлять 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.

2. **Набор прикладных сервисов вычислений и доступа к данным.** В их число входят как различные прикладные пакеты в области 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 [5]. Все сервисы строятся на основе предметно-ориентированных описаний пакетов на языке EasyPackage, регистрируемых в базе пакетов управляющей оболочки.

3. **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 online с использованием графических и текстовых средств общения. Кроме того, предусматриваются сервисы интерактивной консультации экспертов, возможность сохранения результатов выполненных задач для последующего использования другими членами сообщества или для совместного обсуждения и исправления. В качестве отдельной задачи рассматривается 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 ним, равно как и оптимизация распределения вычислительной нагрузки, осуществляется средствами управляющей оболочки без привлечения пользователя.

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

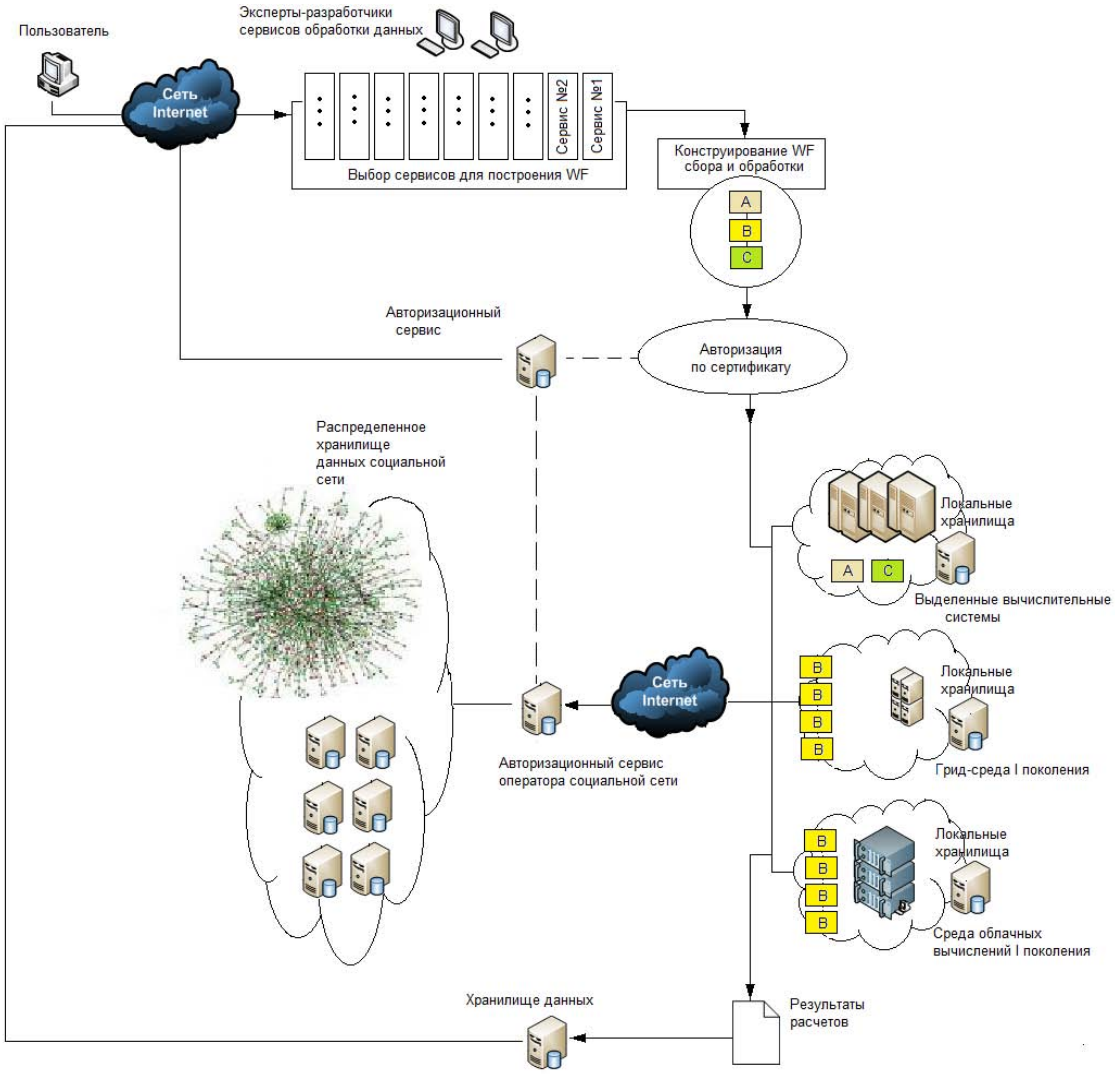


Рис. 1. Схема функционирования производственно-исследовательского web-центра в области социодинамики

Для подготовленного описания композитного приложения пользователь конфигурирует условия вычислений: определяет требуемые ему параметры WF, редактирует (при необходимости) его описание, готовит и загружает в хранилище среды входные данные для расчетов. В ряде случаев такие данные могут предоставляться провайдером web-центра (например, учебные базы данных фрагментов социальных сетей).

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, определяемые использованием разных социальных сетей и привлечением ресурсов сред распределенных вычислений. Окончательно, пользователь запускает задачу на выполнение 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.

3. Прикладные сервисы в составе web-ориентированного центра

Прикладные сервисы можно условно разделить на четыре функциональные группы: сервисы сбора данных в социальных сетях, сервисы

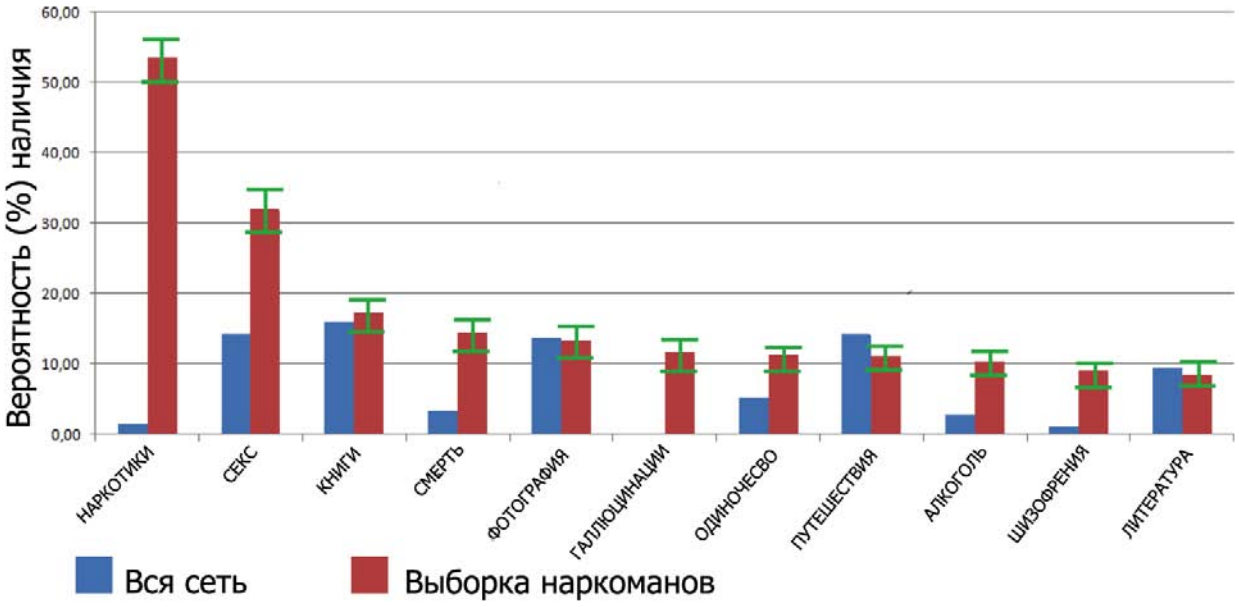


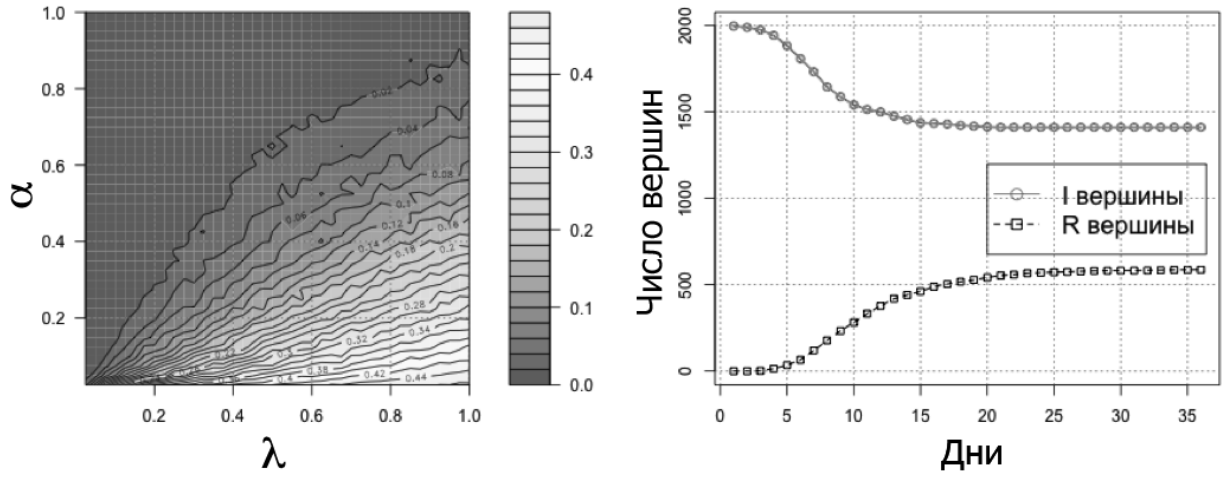
Рис. 2. Предпочтения интересов пользователей социальной сети ЖЖ

статистической обработки и анализа данных, сервисы моделирования сценариев, сервисы 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 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. Динамическая архитектура краулера с централизованным управлением позволяет динамически добавлять и удалять агентов, обеспечивая масштабируемость системы в целом. Помимо классических политик обхода (обход в ширину и в глубину) таким образом может быть дополнительно реализована политика обхода по степени влияния, согласно которой 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.

Сервисы статистической обработки и анализа данных используют общий подход к описанию многомерных комплексных сетей – стохастических графов с многомерными характеристиками вершин. Под комплексной сетью [6] понимается граф с достаточно большим числом узлов различной природы (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 мерности (обобщение метода главных компонент для графов), методов дискриминантного анализа для выявления характерных структур в сети, а также методов ординации (шкалирования) для учета



(а) (б)

Рис. 3. Моделирование процесса распространения слухов в социальной сети

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 использовать совокупно статистическую информацию о топологической структуре сети и наборе характеристик каждого индивида. В качестве иллюстрации на рис. 2 представлен пример анализа предпочтений сообщества индивидов из социальной сети ЖЖ, в профиле интересов которого присутствуют упоминания о наркотиках, по сравнению со всей сетью. Из рис. 2 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 включают в себя различные вероятностные модели социодинамики в терминах микро- и (или) макро-параметров 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 дает возможность провести исследование чувствительности 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 моделирования распространения слухов в социальной сети на основе модели Далея-Кендалла [7]. Он оперирует тремя множествами вершин: неинформированные (I), распространяющие (S) и неактивные (R). На каждом шаге, при взаимодействии на вершины типа I с вершиной типа S, она с заданной вероятностью 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 так же 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.

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 во множествах I и R на определенном шаге работы алгоритма: видно, что с течением времени 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).

4. Заключение

Несмотря на то, что перечисленные сервисы сами по себе достаточны для выполнения разного рода расчетов 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. Описания типовых задач (в форме соответствующих WF) могут быть представлены в репозитории для общего использования. К ним, в частности, относятся:

- 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 офф-лайн 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 смогут 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.

Литература

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

[2] Hu, D. Identifying Significant Facilitators of DarkNetwork Evolution / Hu D., Kaza S., Chen H. // J. of the American Society for Inf. Science and Technology. 2009. Vol. 60(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. On-line textbook available at http://faculty.ucr.edu/~hanneman/nettext/.

[4] Бухановский, А.В. Интеллектуальные высокопроизводительные программные комплексы моделирования сложных систем: концепция, архитектура и примеры реализации / Бухановский А.В., Ковальчук С.В., Марьин С.В. // Известия высших учебных заведений. Приборостроение. 2009. Т. 52. №10. С. 5-24.

[5] Chau, D.H. Parallel crawling for online social networks / Chau D.H., Pandit S., Wang S., Faloutsos C. // Proceedings of the 16th international conference on World Wide Web - WWW ’07, 1283. New York, New York, USA: ACM Press. 2007.

[6] Newman, M. E. J. The Structure and Function of Complex Networks // Society for Industrial and Applied Mathematics. 2003. Vol. 45, №. 2. P. 167–256.

[7] Daley, D. Epidemics and Rumours / Daley D., Kendall D. // Nature. 1964. Vol. 204. № 4963. P. 1118.

Web-oriented center in sociodynamics: concept and architecture principles

S.V. Ivanov, E.A. Bolgova, V.V. Kashirin, A.V. Yakushev, A.V. Chugunov, A.V. Boukhanosky, P.M.A. Sloot

Concept and architecture principles of cloud-computing problem-based environment for production and research center in sociodynamics and its applications in the frame of Web 2.0 paradigm is considered.