**КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ**
**И МОДЕЛИРОВАНИЕ 2015 Т. 7 № 5 С. 1127−1139**

Ки&М

**МОДЕЛИ ЭКОНОМИЧЕСКИХ И СОЦИАЛЬНЫХ СИСТЕМ**

УДК: 004.048

# Подходы к кластеризации групп социальной сети

**Е. П. Охапкина[1,a], В. П. Охапкин[2]**

[1] ФГБОУ ВПО «Российский государственный гуманитарный университет»,
факультет информационных систем и безопасности,
Кафедра информационных систем и моделирования,
Россия, 117534, г. Москва, ул. Кировоградская, д. 25, корп. 2

[2] ФГБОУ ВПО «Вятский государственный университет»,
Факультет экономики и менеджмента,
к афедра математического моделирования в экономике,
Россия, 610000, г. Киров, ул. Московская, д. 36

E-mail: [a] lenaokhapkina@mail.ru

*Получено 6 марта 2015 г.,*
*после доработки 24 июня 2015 г.*

Исследование посвящено проблеме использования социальных сетей в качестве инструмента в противозаконной деятельности и источника информации, способного нести опасность обществу. В статье приводится структура мультиагентной системы, под управлением которой может осуществляться кластеризация групп социальной сети по критериям, однозначно определяющим группу в качестве деструктивной. Приведен алгоритм, который используют агенты системы для кластеризации.

Ключевые слова: социальные сети, мультиагентная система, кластерный анализ, безопасность

# Approaches to a social network groups clustering

**E. P. Okhapkina**, **V. P. Okhapkin**[2]

*[1] Russian State University for the Humanities, Faculty of Information Systems and Security, 25 Kirovogradskaya st., block 2, Moscow, 117534, Russia*
*[2] Vyatka State University, Faculty of Economics and Management, 36 Moskovskaya st., Kirov, 610000, Russia*

**Abstract.** — The research is devoted to the problem of the use of social networks as a tool of the illegal activity and as a source of information that could be dangerous to society. The article presents the structure of the multi-agent system with which a social network groups could be clustered according to the criteria uniquely defines a group as a destructive. The agents' of the system clustering algorithm is described.

Keywords: multi-agent system, cluster analysis, system interaction, security

Citation: *Computer Research and Modeling*, 2015, vol. 7, no. 5, pp. 1127–1139 (Russian).

## Введение

Социальные сети уверенно интегрировались в жизнь современного человека. На сегодняшний день они предоставляют возможность для совершения покупок, геопозиционирования (геолокации[1]), разработки собственных приложений. Если же рассматривать основную идею, с которой создавались социальные сети, — способ общения, то его частота в социальной сети, обмен информацией личного характера не уступает связи людей посредством телефона. По результатам исследования, проведенного ВЦИОМ в 2011 году, на вопрос «Пользуетесь ли вы интернетом? Если да, то посещаете ли вы социальные сети или блоги («Одноклассники», «ВКонтакте», «Мой круг», «Мой мир» и т. п.)?» 26 % респондентов ответили, что посещают социальные сети, и 4 % — что посещают блоги. На этот же вопрос, но годом ранее 23 % опрошенных ответили, что посещают социальные сети (ответ «Посещаю блоги», в опросе 2010 года не предлагался). Популярность социальных сетей возрастает. Причем из 23 % использующих социальные сети в опросе 2010 года 57 % — это люди в возрасте от 18 до 24 лет, а в опросе 2011 года из 26 процентов 58 % респондентов находятся в таком же возрастном диапазоне [ВЦИОМ, 2014]. Рост интереса к использованию социальных сетей может быть объяснен их коммуникационными возможностями. В действительности в отличие от телефонии социальные интернет-сервисы предоставляют возможность сообщать информацию в различных форматах: текстовом, графическом, видео, аудио. Однако наиболее важным моментом в возрастающем интересе к использованию социальных сетей и блогов является возрастная категория, которая проявляет этот интерес. С социальной точки зрения возрасту до 30 лет свойственна наибольшая социальная активность и не полностью сформировавшиеся взгляды на протекающие в обществе, экономике, политике процессы, что нередко становится основой протестных настроений. А значит, с использованием социальных интернет-сервисов появляются значительные возможности для объединения, манипулирования и координации людей со стороны заинтересованных, преследующих определенные цели лиц, принимая во внимание, что личные страницы в социальной сети нередко снабжены достаточным объемом первичной информации о пользователе. Однако глава ВЦИОМ Валерий Федоров отмечает: «Само по себе распространение социальных сетей не усиливает протестных настроений, не революционизирует общество. Здесь другая взаимосвязь: если в конкретном городе или стране такие настроения возникают, облегчается координация участников». Именно социальные сети активно использовались организаторами массовых митингов против фальсификации выборов. Так, на митинг «За честные выборы» (прошел 24 декабря 2011 года на проспекте Сахарова) в специальной группе на Facebook согласились пойти более 54 тыс. человек и еще около 100 тыс. получили приглашения. На приглашение пойти на шествие 4 февраля на этом же ресурсе откликнулись почти 29 тыс. человек. При этом общее число участников акций, по данным организаторов, оказалось больше» [Иванов, 2014]. Уже сегодня новые разработки в области коммуникационного программного обеспечения позволяют обходиться без операторов сотовой связи и интернет-соединения. В частности, участники протестов, начавшихся в Гонконге в октябре 2014 года, использовали мессенджер FireChat. Программа обладает возможностью мгновенного обмена сообщениями посредством Bluetooth в радиусе 60 метров, при этом не используя интернет-соединение. Каждое устройство с FireChat фактически становится ретранслятором для других устройств с этой программой. С помощью этого клиента можно создавать собственные сети, в которых каждое устройство имеет равные права. Разработчики программы анонсируют приложение как средство для мгновенного обмена сообщениями на той территории, где может возникнуть проблема со связью. Принимая во внимание технологическое развитие и коммуникационный потенциал средств

[1] Термин, возникший в английской печати, посвященной IT-технологиям (англ. *geolocation*). В информатике означает определение географического местоположения интернет-пользователя. Однако прямое использование английского слова «геолокация» без соответствующего перевода может ввести в заблуждение отечественного читателя, поскольку термин «геолокация» также используется при геологоразведке и означает «неразрушающее обнаружение и исследование объектов грунтовых сред методом радиолокационного зондирования».

связи и социальных сетей, можно заключить, что социальные сети могут выступать мощным инструментом в незаконной деятельности, нарушающей общественный порядок и безопасность. С увеличением доступности широкополосного доступа в сеть Интернет и развитием мобильного интернета показатели использования социальных сетей будут только возрастать.

В России число пользователей, имеющих доступ к сети Интернет, растет из года в год (преимущественно это происходит в сегменте частных пользователей). Так, по данным государственной статистики, в России за 2011 год число абонентов фиксированного широкополосного доступа к сети Интернет составило 17.42 млн пользователей, среди них 16.41 млн пользователей из числа физических лиц, а объем информации, переданной от/к абонентам сети при доступе в Интернет составил 8492.2 петабайт. Уже в 2012 году эти показатели продемонстрировали значительный рост и, соответственно, составили 20.70 млн пользователей фиксированного широкополосного доступа, из которых физических лиц — 19.54 млн пользователей, а объем переданной информации в сети Интернет — 9923.7 петабайт. Причем число абонентов мобильного широкополосного доступа в сеть Интернет в 2011 году составило 68.39 млн пользователей, а в 2012 году этот показатель увеличился до 91.21 млн пользователей [Росстат, 2014].

С точки зрения соблюдения правопорядка относительной сложностью в многомиллионной аудитории социальной сети является выявление большого числа групп, содержание которых способно нанести вред как индивиду в частности, так и обществу в целом. Цель исследования заключена в следующем: выявить сообщества (группы)[2] в социальных сетях, представляющие опасность и несущие прямые угрозы. В связи с этим необходимо разрешить две актуальные задачи. Во-первых, необходимо по определенному критерию идентифицировать массив групп в качестве деструктивных. Во-вторых, провести семантический анализ текстовых сообщений внутри группы социальной сети.

## Архитектура мультиагентной системы

Поставленная цель многогранна, и для ее достижения необходимо использование технологий, позволяющих производить комплексный анализ разнородных по составу и качеству данных, которые генерирует социальная сеть.

В условиях большого объема информации, представленной в социальных сетях, возникает одна из нетривиальных задач: семантический анализ текста. Эффективным решением этой проблемы является создание интеллектуальной системы, которая включила бы предварительную кластеризацию сообществ сети в однородные группы, и последующий семантический анализ внутригрупповых сообщений. Анализ современного состояния в области разработки систем кластеризации и анализ данных показали, что в качестве решения возможно использование технологий мультиагентных систем.

Использование этой технологии нашло широкое применение в тех областях и сферах практики, где для достижения поставленной цели (или целей) требуется параллельное решение разнородных по составу и качеству задач. Например, отрасли производства, системы управления городской средой, поисковые системы Интернета. Одной из таких многоаспектных задач является модерирование и выявление деструктивных сообществ.

В рамках поставленной цели и задач архитектура мультиагентной системы анализа пользовательских групп социальной сети может выглядеть следующим образом [Филатов, 2002].

1. Интерфейс пользователя. Позволяет пользователю назначать задания.
2. Блок управления выполнением задач, расположенный в подсистеме администрирования мультиагентной системы, осуществляет передачу полученного списка заданий в блок «Постановщик задач». Отметим, что поставленные администратором задачи могут различаться

[2] Далее по тексту слова «группа» и «сообщество» социальной сети будут употребляться в качестве синонимов, обозначая виртуальную площадку для общения людей.

по качеству. Например, семантический анализ, анализ изображений, определение геопозиции. «Постановщик задач» является центром формирования заданий для агентов. Кроме того, подсистема управления выполнением задач содержит мониторинговые блоки: база данных агентов и их состояний для контроля функционирования агентов и база данных (далее — БД) результатов выполненных заданий.

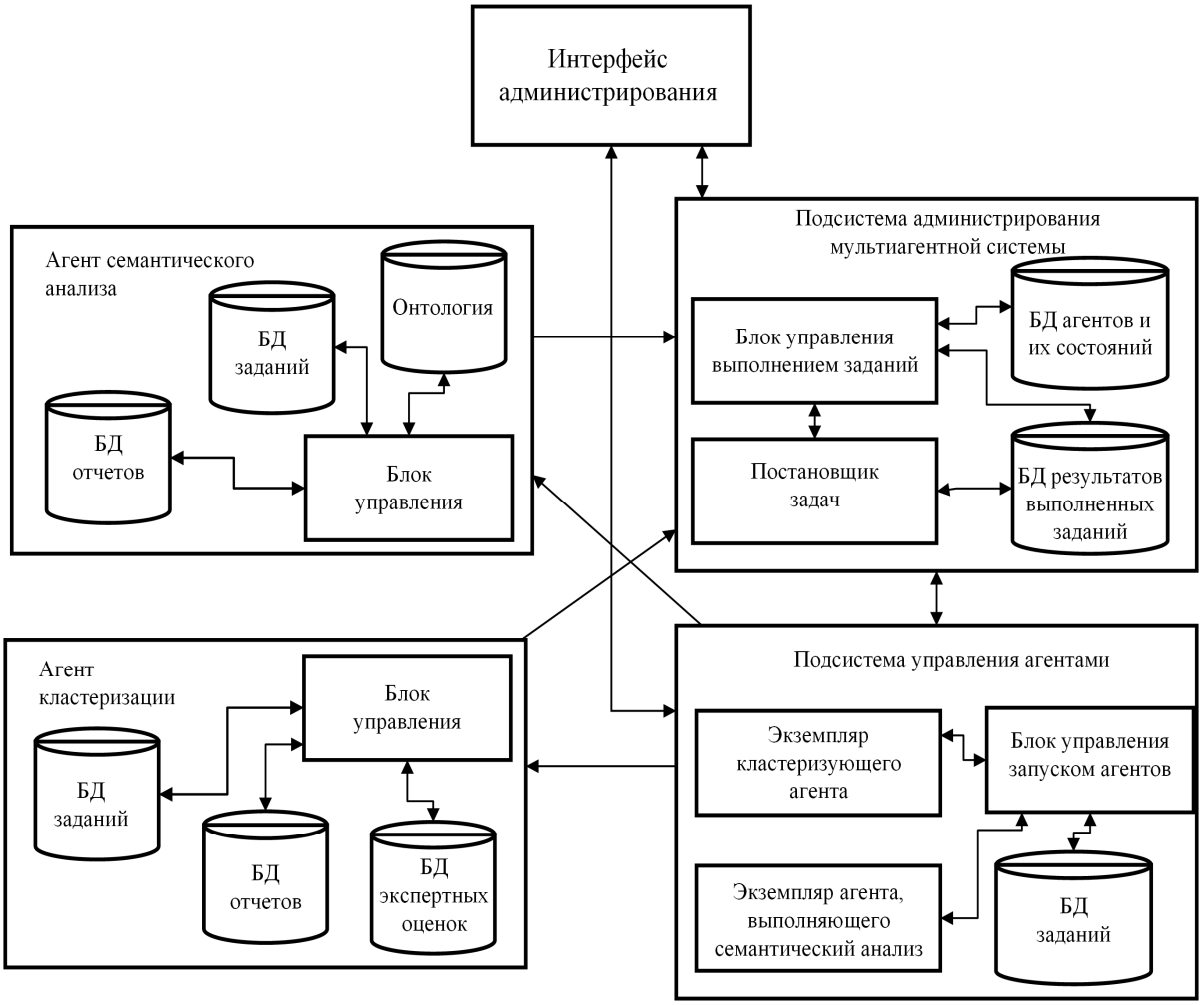



Рис. 1. Мультиагентная система для модерирования групп социальной сети

3. Подсистема управления агентами содержит блок управления запуском агентов, в который поступает информация из подсистемы администрирования мультиагентной системы о необходимости запуска агентов определенного типа, их количестве и поставленных заданиях, происходит формирование БД заданий, выполняется формирование и запуск агентов соответствующих типов.
4. В блок управления агентом семантического анализа направляется информация из подсистемы управления агентами о необходимости выполнить задание. Поскольку объектом семантического анализа является словесное выражение, то для адекватного заключения о высказывании необходимо понятие, способное отразить в структурированном виде значение этого высказывания. Например, при анализе текста из предметной области о распространении запрещенных веществ, суициде, призывах к террористической деятельности. С этой целью в подсистеме управления агентом предусмотрен блок онтологий. «Онтология — это подробная спецификация структуры определенной проблемной области, включающей: словарь (список) логических констант и предикатных символов для описания предметной области и набор логических высказываний, формулирующих существующие в данной проблемной

области ограничения и определяющих интерпретацию словаря» [Болотова, 2012]. Подсистема управления выполнением задач и подсистема управления агентом содержат блок мониторинга: список задач, поставленных перед агентом (БД заданий); отчеты о результатах выполнения заданий (БД отчетов).

5. Агент кластеризации содержит блок управления агентом, в который поступает информация о необходимости выполнить задание из подсистемы управления агентами, все задания формируются в БД заданий, также подсистема содержит БД экспертных оценок, содержащую значения эталонных кластеров, близость к которым определяется в процессе кластеризации; также агент кластеризации содержит БД отчетов, в которую поступают отчеты о результатах выполнения заданий по окончании работы агента.

Разработанная архитектура содержит базовые модули, необходимые для выполнения кластеризации групп социальной сети и семантического анализа содержащихся в них сообщений. Предложенная архитектура обладает гибкостью и может быть расширена за счет добавления новых задач, таких как анализ изображений, определение геопозиции, а также предоставляет возможность параллельного решения этих задач. Еще одним достоинством предложенной архитектуры является возможность внесения изменений в отдельные подсистемы без необходимости изменять всю архитектуру в целом.

## Кластерный анализ в классификации групп социальной сети

В описанной архитектуре предусмотрена подсистема управления агентами, осуществляющими поиск, распознавание и классификацию групп в социальной сети. Известно, что процедура классификации выполняется на основе критериев, заложенных в коде агента.

Всё множество сообществ некоторой социальной сети обозначим через $\Omega$. Тогда для того, чтобы классифицировать $\Omega$ и сформировать однородные по признаку кластеры, воспользуемся методом *k*-средних (англ. *k-means*).

В качестве расстояния между наблюдениями используем следующую метрику: евклидово расстояние

$$\rho_{ik} = \sqrt{\sum_{i=1}^{m} \left( z_{ji} - z_{jk} \right)^2}, \tag{1}$$

где $\rho_{ik}$ — это расстояние между *i* и *k* наблюдениями. Отметим, что исходные данные прежде стандартизуются по соотношению

$$z_{ij} = \frac{t_{ij} - \overline{t}_j}{\sigma_j}, \tag{2}$$

где $z_{ij}$ образует матрицу стандартизованных данных. Такое нормирование данных вызвано недостатком евклидовой и других метрик, которой заключается в том, что «оценка сходства сильно зависит от различий в сдвигах данных» [Факторный, дискриминантный…, 1989]. Иначе говоря, объекты, имеющие большие абсолютные значения и стандартные отклонения, могут оказывать подавляющее влияние на объекты с меньшими абсолютными значениями и стандартными отклонениями. Более того, метрические расстояния изменяются под воздействием преобразований шкалы измерения переменных, при которых не сохраняется ранжирование по евклидову расстоянию [Факторный, дискриминантный…, 1989].

Метод *k*-средних относится к группе итеративных методов эталонного типа. «Для начала процедуры классификации задаются *k* объектов, которые будут служить эталонами, или центрами, кластеров. На 1-м шаге из оставшихся $n - k$ объектов извлекается наблюдение и с использованием одной из метрик определяется, к какому кластеру оно ближе. Эталон заменяется новым, пересчитанным с учетом присоединенного наблюдения, при этом вес его (количество

наблюдений в кластере) увеличивается на единицу. После обработки всех наблюдений процесс продолжается уже для всех $n$ наблюдений до тех пор, пока разбиение не будет «устойчивым», т. е. новое разбиение, полученное после просмотра $n$ наблюдений, не будет отличаться от предыдущего» [Заречнев, 2004]. В качестве центров кластеров могут выступать группы, по категориям которых производится кластеризация.

Выполним кластеризацию, используя приближенные к реальности данные о группах. Пусть множество $\Omega$ содержит 10 различных групп, включая группы экстремистского и призывающего к насилию характера, пропагандирующие суицид. Любую группу будем характеризовать по следующим признакам:

1) название группы в социальной сети;
2) количество участников;
3) наличие специфической лексики.

Тогда необходимо среди имеющегося множества групп выделить три кластера, первый из которых содержит группы экстремистского характера, второй — группы, пропагандирующие суицид, а третий — нейтральные по оцениваемым признакам.

Сделаем допущение о том, что информация обо всех вышеперечисленных признаках доступна для оценки и измерения. Таким образом, разрабатываемое программное обеспечение должно быть интегрировано в архитектуру социальной сети и обладать правами администратора. Для того чтобы оценить признак, воспользуемся экспертными оценками и введем шкалу оценок, распределенную на промежутке от 0 до 1, где нижняя граница означает нейтральный характер признака, а верхняя граница — ярко выраженный деструктивный. Тогда множество $\Omega$ можно представить в матричном виде:

$$\Omega = \begin{pmatrix} 0.85 & 0.45 & 0.32 & 0.18 & 0.22 & 0.91 & 0.82 & 0.78 & 0.80 & 0.09 \\ 0.65 & 0.17 & 0.62 & 0.78 & 0.99 & 0.74 & 0.11 & 0.36 & 0.58 & 0.74 \\ 0.49 & 0.37 & 0.87 & 0.21 & 0.46 & 0.78 & 0.89 & 0.01 & 0.05 & 0.23 \end{pmatrix}, \qquad (3)$$

где в первой строке находятся оценки по первому признаку, во второй — оценки количества участников группы и в третьей — оценки, указывающие на наличие специфической лексики. В свою очередь, отдельный столбец матрицы — это одна группа.

Пользуясь (2), выполним стандартизацию матрицы (3). Тогда, получим

$$\Omega = \begin{pmatrix} 0.96 & -0.29 & -0.69 & -1.13 & -1.00 & 1.15 & 0.87 & 0.74 & 0.80 & -1.41 \\ 0.27 & -1.45 & 0.16 & 0.74 & 1.49 & 0.59 & -1.66 & -0.77 & 0.02 & 0.59 \\ 0.17 & -0.20 & 1.34 & -0.70 & 0.07 & 1.06 & 1.40 & -1.31 & -1.19 & -0.64 \end{pmatrix}. \qquad (4)$$

В качестве эталонных кластеров выберем первые три группы (1–3 столбцы матрицы $\Omega$). Выполним первую итерацию и определим расстояние четвертой группы до каждого из трех эталонных кластеров. Пользуясь (1), вычислим расстояние от четвертой группы до первого эталонного кластера:

$$\rho_{41} = \sqrt{\sum_{i=1}^{3} \left(z_{1i} - z_{i4}\right)^2} = \sqrt{\left(0.96 + 1.13\right)^2 + \left(0.27 - 0.74\right)^2 + \left(0.17 + 0.70\right)^2} = 2.306291.$$

Найдем расстояния от четвертой группы до второго и третьего эталонных кластеров, а полученные результаты занесем в таблицу 1.

Следуя алгоритму $k$-средних, выполним пересчет координат третьего кластера по следующей формуле:

$$\alpha_i = \frac{v_i z_{ij} + z_{kj}}{v_i + 1}, \qquad (5)$$

где $\alpha_i$ — это пересчитанные координаты $i$-го кластера (кластера, с которым выполнено объединение), $v_i$ — вес пересчитываемого кластера, $z_{ij}$ — координаты (элементы) пересчитываемого эталонного кластера, а $z_{kj}$ — координаты группы, которая объединяется с эталонным кластером.

Таблица 1. Расстояние четвертой группы до эталонных кластеров

| Рассматриваемый кластер | Расстояние | Вывод* |
|---|---|---|
| $4 \to 1$ | 2.306291 | |
| $4 \to 2$ | 2.390827 | |
| $4 \to 3$ | 2.160588 | Наименьшее расстояние от рассматриваемой группы до данного эталонного кластера. Четвертую группу можно включить в третий кластер |

* Вывод о принадлежности группы к тому или иному эталонному кластеру можно сделать только в отношении одного кластера.

Выполним пересчет координат третьего кластера по формуле (5) и увеличим его вес на единицу, тогда получим

$$\alpha_3 = \begin{pmatrix} -0.91 \\ 0.45 \\ 0.32 \end{pmatrix}, \quad v_3 = 2.$$

Вес третьего эталонного кластера после включения четвертой группы равен 2. Рассчитаем расстояния до эталонных кластеров для всех групп и сделаем вывод о включении группы в тот или иной кластер. Полученные расстояния запишем в таблицу 2.

Таблица 2. Расстояния групп социальной сети до эталонных кластеров

| Рассматриваемый кластер | Расстояние | Вывод** | Рассматриваемый кластер | Расстояние | Вывод** |
|---|---|---|---|---|---|
| $5 \to 1$ | 2.310828 | | $8 \to 1$ | 1.821972 | |
| $5 \to 2$ | 3.033260 | | $8 \to 2$ | 1.659418 | Объединить |
| $5 \to 3$ | 1.857854 | Объединить | $8 \to 3$ | 3.156779 | |
| $6 \to 1$ | 0.969490 | Объединить | $9 \to 1$ | 1.389789 | Объединить |
| $6 \to 2$ | 2.795550 | | $9 \to 2$ | 2.077683 | |
| $6 \to 3$ | 1.907653 | | $9 \to 3$ | 2.943175 | |
| $7 \to 1$ | 2.295020 | | $10 \to 1$ | 2.520349 | |
| $7 \to 2$ | 1.987622 | Объединить | $10 \to 2$ | 2.367411 | |
| $7 \to 3$ | 2.400033 | | $10 \to 3$ | 2.144788 | Объединить |

** До стрелки указан номер группы, которую необходимо включить в эталонный кластер с номером, указанным после стрелки.

После объединения и пересчета координаты эталонных кластеров примут следующие значения:

$$\alpha_1 = \begin{pmatrix} 0.97 \\ 0.30 \\ 0.01 \end{pmatrix}, \quad \alpha_2 = \begin{pmatrix} 0.44 \\ -1.29 \\ -0.04 \end{pmatrix}, \quad \alpha_3 = \begin{pmatrix} -1.06 \\ 0.74 \\ 0.02 \end{pmatrix}.$$

Вес каждого кластера после кластеризации будет соответственно равен $v_1 = 3$, $v_2 = 3$, $v_3 = 4$. Таким образом, после использования процедуры *k*-средних получены результаты кластеризации представлены в таблице 3.

Таблица 3. Результат кластеризации

| Кластер 1 | Кластер 2 | Кластер 3 |
|---|---|---|
| 1, 6 и 9 группы | 2, 7 и 8 группы | 3, 4, 5 и 10 группы |

Выше отмечалось, что аудитория социальной сети значительна и насчитывает десятки тысяч сообществ. В связи с этим представляется разумной автоматизация процесса кластеризации. Вариантом автоматизации может быть создание программируемого агента в структуре мультиагентной системы. Работа этого агент является предварительным этапом перед выполнением семантического анализа (вторая задача в поставленной цели).

## Об устойчивости полученного решения

Недостатком кластерных методов является то, что в них объекты распределяются по кластерам лишь за один проход, поэтому плохое начальное разбиение множества данных не может быть изменено на последующих шагах процесса кластеризации [Gower, 1967]. Кроме того, неустойчивость некоторых методов кластеризации (например, иерархических агломеративных методов) проявляется в том, что в результате переупорядочивания элементов в матрице сходства могут быть получены разные решения. Новые варианты кластеризации можно наблюдать и в тех случаях, когда некоторые объекты исключаются из рассмотрения. Устойчивость — это важное свойство любой классификации, так как устойчивые группы с большим правдоподобием представляют собой «естественные» группировки по сравнению с теми группами, которые исчезают, если некоторые объекты переупорядочены или исключены из анализа [Факторный, дискриминантный…, 1989]. Вопрос об устойчивости становится особенно существенным, когда мы имеем дело с малыми выборками объектов [Jardine, Sibson, 1971].

Для оценки устойчивости решения, полученного в таблице 3, выполним кластеризацию методом Уорда (англ. *Ward's method*). Существующий метод относится к классу иерархических агломеративных методов и основан на минимизации дисперсии внутри отдельного кластера. Целевая функция внутригрупповой суммы квадратов (СКО) имеет вид

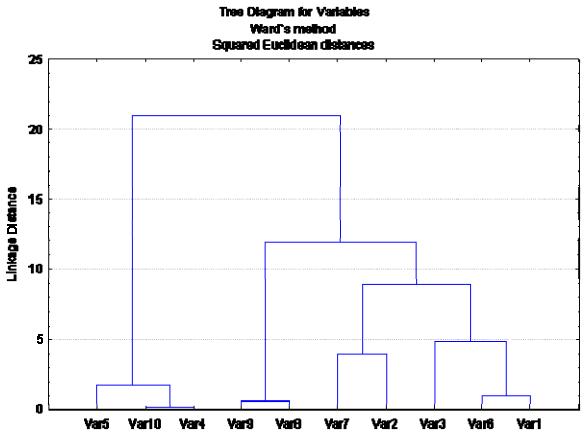
$$СКО = x_j^2 - \frac{1}{n}\left(\sum x_j\right)^2, \tag{6}$$

где $x_j$ — значение признака *j*-го объекта.

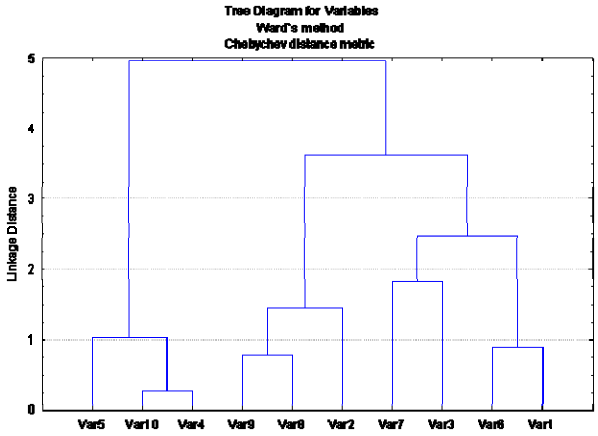
На начальном этапе, когда каждый кластер насчитывает один объект, внутригрупповая сумма квадратов равна 0. В соответствии с методом Уорда необходимо объединять в кластер те объекты, для которых СКО образует минимальное приращение. Метод имеет тенденцию к нахождению (или созданию) кластеров приблизительно равных размеров и имеющих гиперсферическую форму [Факторный, дискриминантный…, 1989, с. 174–175]. В контексте проводимого исследования важно отметить, что «метод Уорда широко используется во многих социальных науках» [Blashfield, 1980].

Выполним кластеризацию в пакете статистического анализа Statistica v 6.0. Для репрезентативности полученных результатов кластеризации воспользуемся различными видами метрик: квадратичным евклидовым расстоянием (рис. 2а), метрикой Чебышева (рис. 2б), манхэттонским расстоянием (рис. 2в) и евклидовой метрикой (рис. 2г).

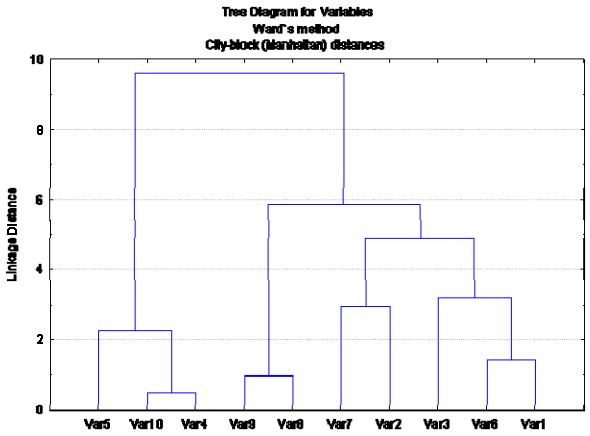
Заметим, что на рисунках 2а, 2б и 2в при различном количестве итераций наблюдается устойчивое относительно количества и состава кластеров разбиение. Оценивая все использованные метрики, отметим, что дерево, порожденное методом Уорда, ясно указывает на три кластера



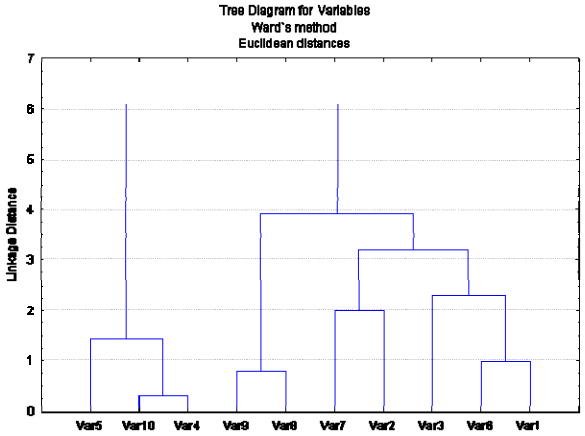

(а) Квадрат евклидова расстояния




(б) Метрика Чебышева




(в) Манхэттонское расстояние




(г) Евклидова метрика

Рис. 2

в полученном решении. Говоря об устойчивости используемых в исследовании алгоритмов, скажем, что иерархический агломеративный метод Уорда позволяет получить за $n$-ое количество итераций набор кластеров, не меняющих свой состав, что крайне важно при дальнейшем качественном исследовании этих кластеров. К такому исследованию относится семантический анализ открытых пользовательских сообщений внутри групп отдельного кластера. Сравнивая итеративный метод $k$-средних и иерархический агломеративный метод Уорда, нужно отметить, что несомненным достоинством первых является работа с первичными данными, что позволяет с их помощью обрабатывать значительные объемы данных. В условиях обширного информационного поля социальных сетей это преимущество. Более того, итеративные методы делают несколько просмотров данных и могут компенсировать последствия плохого исходного разбиения[3], тем самым устраняя самый главный недостаток иерархических агломеративных методов. Эти методы порождают кластеры одного ранга, которые не являются вложенными и поэтому не могут быть частью иерархии. Большинство итеративных методов не допускают перекрытия кластеров [Факторный, дискриминантный…, 1989].

[3] В 2007 году Дэвидом Артуром и Сергеем Вассильвитским была предложена улучшенная версия алгоритма $k$-средних, которая получила название k-means++. Идея метода осталась прежней, но на начальном этапе авторами предлагается выбирать центроиды кластеров, основываясь на вероятностях. Алгоритм реализован на языке Java и включен в популярную библиотеку Apache.

## Программная реализация кластер-процедуры

Код агента кластеризации реализован на платформе Embarcadero RAD Studio XE3, язык реализации — C++.

Программирование кода агента начинается с процедуры загрузки экспертных оценок. Для удобства хранения и заполнения экспертных оценок использован MS Excel. Форма, в которой хранятся оценки экспертов, представлена таблицей, где по строкам расположены признаки, а по столбцам — группы социальной сети. Для взаимодействия агента и офисного приложения используется OLE-объект, через который происходит подключение к электронной таблице MS Excel и загрузка экспертных оценок в операционный массив агента кластеризации. В листинге в строках 010–028 определены команды, выделяющие память для загрузки экспертных оценок (010, 015–19), хранения стандартизованных оценок (011, 015–019), вспомогательные массивы для временного хранения эталонных групп (012, 020–025) и оставшихся групп социальной сети (013, 020–025). Перед стандартизацией оценок необходимо предварительно вычислить среднее значение по признаку и его среднеквадратическое отклонение. В строках с 026 по 028 выделяется пространство памяти для сохранения этих параметров, а в 030–042 реализована процедура вычисления стандартизованных значений по соотношению (2). Участок кода с 057 по 080 строку представляет собой алгоритм, который сочетает поиск наименьшего расстояния от рассматриваемого кластера до эталонных кластеров, включение групп социальной сети в кластеры и пересчет координат эталонных кластеров. Строки в диапазоне 082–088 выполняют очистку памяти, выделенной под операции агента.

Листинг. Программный код агента, реализующего кластеризацию сообществ социальной сети методом *k*-средних

```
001  int amtOfParam = 0;
002  double tempAver = 0, tempDiver = 0;
003  double dist = 0;
004  int weight[3];
005  double **mas, **aver, **standMas, **tempEtalon;
006  double **tempStandMas, **DisMatrix;
007
008  amtOfParam = StrToInt(txtAmountOfClasters->Text);
009
010   mas = (double**) calloc(amtOfParam, sizeof(double));
011   standMas = (double**) calloc(amtOfParam, sizeof(double));
012   tempEtalon = (double**) calloc(amtOfParam, sizeof(double));
013   tempStandMas = (double**) calloc(amtOfParam, sizeof(double));
014
015    for (int i = 0; i < amtOfParam; i++){
016      mas[i] = (double*) calloc(10, sizeof(double));
017      standMas[i] = (double*) calloc(10, sizeof(double));
018      weight[i] = 1;
019    }
020     for (int i = 0; i < amtOfParam; i++){
021      tempEtalon[i] = (double*) calloc(amtOfParam,
022  sizeof(double));
023      tempStandMas[i] = (double*) calloc(10 - amtOfParam,
024  sizeof(double));
025     }
026    aver = (double**) calloc(amtOfParam, sizeof(double));
027   for (int i = 0; i < amtOfParam; i++)
```

```
028     aver[i] = (double*) calloc(2, sizeof(double));
029
030    for (int i = 0; i < amtOfParam; i++){
031      tempAver = 0;
032      tempDiver = 0;
033     for (int j = 0; j < 10; j++)
034       tempAver += mas[i][j];
035        aver[0][i] = tempAver/10;
036       for (int k = 0; k < 10; k++)
037         tempDiver += pow(mas[i][k] - aver[0][i], 2);
038     aver[1][i] = pow(tempDiver/9, 0.5);
039    }
040    for (int i = 0; i < amtOfParam; i++)
041     for (int j = 0; j < 10; j++)
042       standMas[i][j] = (mas[i][j] - aver[0][i])/aver[1][i];
043
044    for (int i = 0; i < amtOfParam; i++)
045     for (int j = 0; j < amtOfParam; j++)
046       tempEtalon[i][j] = standMas[i][j];
047
048    for (int i = 0; i < amtOfParam; i++)
049     for (int j = 0; j < 10 - amtOfParam; j++)
050       tempStandMas[i][j] = standMas[i][j + 3];
051
052     DisMatrix = (double**) calloc(amtOfParam, sizeof(double));
053   for (int i = 0; i < amtOfParam; i++)
054     DisMatrix[i] = (double*) calloc(10 - amtOfParam,
055  sizeof(double));
056
057    for (int i = 0; i < 10 - amtOfParam; i++)
058     for (int j = 0; j < amtOfParam; j++){
059         dist = 0;
060      for (int k = 0; k < amtOfParam; k++)
061        dist += pow(tempEtalon[k][j] - tempStandMas[k][i], 2);
062        DisMatrix[j][i] = pow(dist, 0.5);
063     }
064
065     double min;
066    for (int index = 0, i = 0; i < 10 - amtOfParam; i++){
067      min = DisMatrix[0][i];
068      index = 0;
070     for (int j = 0; j < amtOfParam; j++){
071       if (DisMatrix[j][i] < min){
072         min = DisMatrix[j][i];
073         index = j;
074       }
075     }
076      weight[index] += 1;
077    for (int k = 0; k < amtOfParam; k++)
078       standMas[k][index] = ((weight[index] -
079  1)*standMas[k][index] + tempStandMas[k][i])/(weight[index]);
080    }
```

```
081
082  free(aver);
083  free(standMas);
084  free(weight);
085  free(tempStandMas);
086  free(tempEtalon);
087  free(DisMatrix);
088  free(mas);
```

Отметим, что разработанный алгоритм требует унификации на пространство всех групп социальной сети. В частности, в циклах резервирования памяти (016, 017, 023) и обработки стандартизованных данных (033, 036, 041, 049, 057, 066) указана длина цикла, равная 10, что соответствует количеству групп множества $\Omega$ в экспериментальной задаче кластеризации. Однако модификация алгоритма в листинге не является сложной задачей: изменению подвергнется длина цикла, которая в унифицированной задаче будет представлена переменной со значением, равным количеству групп (или исследуемой совокупности групп) социальной сети.

## Перспективы развития

Решаемая задача анализа данных — это комплексная задача, и кластеризация, являясь важным базовым этапом, всё же лишь часть этой задачи. Представленная выше архитектура мультиагентной системы реализует следующую иерархию действий. Через интерфейс администрирования поступает набор задач в систему, после чего выполняется загрузка экспертных оценок и кластеризация. Для каждого кластера в зависимости от типа информации в нем системой создается агент с соответствующей программой действий: семантический анализ. Причем кластеризация и семантический анализ — это ветвь в решении объемной задачи анализа данных, находящихся и распространяющихся в социальной сети. Следующий этап — выделить анализ данных, представленных изображениями и аудиозаписями. Подходами и методами решения для такого вида информации могут выступать спектральный анализ изображения и частотный анализ звуковой записи.

## Заключение

В последнее десятилетие социальные сети со всей очевидностью демонстрируют свой коммуникационный потенциал в организации и управлении массовыми собраниями, в продвижении идей. Революционные события в Гонконге, странах Ближнего Востока и СНГ, беспорядки во время проведения шествий на Болотной площади в Москве, организация терактов в Волгограде показали, что этот потенциал может быть использован в негативном ключе и направлен на разрушение и поддержку общественные волнения.

Развитыми в сфере IT-технологий странами активно ведется работа по интеллектуальному анализу данных, генерируемых социальными сетями, сервисами мгновенного обмена сообщениями и др. В частности, в США построен дата-центр (англ. *data-center*) в штате Юта как центр всеобъемлющей национальной кибербезопасности и обработки разведывательных данных.

Система в автоматическом режиме, способная выявлять и классифицировать противоправные действия, которые еще только обсуждаются в виртуальном пространстве, является мощным инструментом в сохранении общественного порядка и безопасности.

Бесспорно, многие участники социальных сетей, систем коротких сообщений обладают технической возможностью контроля за содержанием информации. Однако обработка жалоб пользователей в ручном режиме отнимает большое количество времени и требует многочисленного штата сотрудников. Кроме того, не вся информация деструктивного характера отмеча-

ется пользователями и попадает на анализ модератору сети. Таким образом, разработка мультиагентной системы для анализа открытых данных социальных сетей является сверхактуальной задачей в современных мировых условиях. На этапе разработки такой системы необходимо решить важную задачу кластеризации данных как способа оптимизации работы МАС.

## Список литературы

*База данных исследований ВЦИОМ* // Всероссийский центр исследования общественного мнения. — URL: http://wciom.ru/data-base/ (дата обращения 10.11.2014).

*Болотова Л. С.* Системы искусственного интеллекта: модели и технологии, основанные на знаниях: Учебник / ФГБОУ ВПО «РГУИТП»; ФГАУ «ГНИИ ИТТ «Информатика». — М. : Финансы и статистика, 2012. — 664 с.

*Заречнев В. А.* Статистическое моделирование. Методы, алгоритмы, реализация. — Киров : Изд-во ВятГГУ, 2004. — 160 с.

*Иванов М.* Социальные сети лидируют на выборах // КоммерсантЪ. 2012. — URL: http://www.kommersant.ru/doc-y/1872265 (дата обращения 02.11.2014).

*Ким Дж.-О., Мьюллер Ч. У., Клекка У. Р., Олдендерфер М. С., Блэшфилд Р. К.* Факторный, дискриминантный и кластерный анализ / Под ред. И. С. Енюкова. — М. : Финансы и статистика, 1989. — 215 с.

*Регионы России. Социально-экономические показатели* // Федеральная служба государственной статистики. — URL: http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/publications/catalog/doc_1138623506156 (дата обращения: 15.11.2014).

*Филатов В. А., Цыбульник Е. Е., Чалая Л. Э.* Модель мультиагентной системы автономного администрирования информационных систем и распределенных баз данных // Искусственный интеллект. — 2002. — № 1. — С. 620–627.

*Blashfield R. K.* The growth of cluster analysis: Tryon, Ward and Johnson // Multivariate Behavioral Research. — 1980. — Vol. 15. — P. 439–458

*Gower J. C.* A comparison of some methods of cluster analysis // Biometrics. — 1967. — Vol. 23. — P. 623–637.

*Jardine N., Sibson R.* Mathematical taxonomy. — NY. : John Wiley & Sons Ltd., 1971. — 304 p.