Моделирование динамики онлайн-дискуссий в сети Интернет с использованием многоагентных систем[1]

Э.А. Бабкин
кандидат технических наук, PhD in Computer Science
профессор кафедры информационных систем и технологий
Национальный исследовательский университет «Высшая школа экономики»
Адрес: 603155, г. Нижний Новгород, ул. Большая Печерская, д. 25/20
E-mail: eababkin@hse.ru

Т.С. Бабкина
старший преподаватель кафедры информационных систем и технологий
Национальный исследовательский университет «Высшая школа экономики»
Адрес: 603155, г. Нижний Новгород, ул. Большая Печерская, д. 25/20
E-mail: tbabkina@hse.ru

Б.И. Улитин
старший преподаватель кафедры информационных систем и технологий
Национальный исследовательский университет «Высшая школа экономики»
Адрес: 603155, г. Нижний Новгород, ул. Большая Печерская, д. 25/20
E-mail: bulitin@hse.ru

Аннотация

Совместный анализ общей структуры онлайн-дискуссий в Интернете и различных психолингвистических характеристик отдельных сообщений является актуальной исследовательской задачей в фундаментальном и прикладном аспектах. Несмотря на успехи алгоритмических методов автоматического анализа сообщений с помощью методов машинного обучения, остаются нерешенными проблемы моделирования динамики структуры дискуссий и характеристик отдельных сообщений при наличии группы автономных авторов. Авторами предлагается использовать для решения этих проблем методы многоагентного имитационного моделирования. В данной работе представлены две многоагентные модели дискуссии, которые позволяют в полной мере учесть характеристики отдельных сообщений и наличие группы авторов с индивидуальными моделями поведения, сформированными на основе анализа реальных онлайн-дискуссий в сети Интернет. Одна из моделей является централизованной в том смысле, что поведение всех авторов идентично и описывается единым блоком управления, зависящим от нескольких параметров. В отличие от централизованной модели, вторая модель является распределенной и характеризуется индивидуализированным поведением каждого автора. Поведение автора в данном случае задается посредством иерархической марковской цепи особой формы. Такая структура модели позволяет не только максимально приблизить процесс ее работы к реальному процессу создания дискуссий, но и обеспечивает возможность сравнения результатов работы моделей с фактическими данными онлайн-дискуссий в Интернете. Важной особенностью предлагаемого подхода к моделированию является активное использование предобработанных фактических данных реальных дискуссий на различных Интернет-площадках. Предобработка данных включает как методы экспертной оценки психолингвистических характеристик (интент- и контент-анализа), так и методы математического статистического анализа. Поэтому в целом исследование является удачным примером междисциплинарного похода к изучению феноменов Интернет-коммуникаций.

[1] Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований, проект № 16-06-00184 А «Разработка и исследование моделей online-дискуссии на материале обсуждения политических новостей»

Ключевые слова: модели коммуникации, онлайн-дискуссии, многоагентные системы, имитационное моделирование.

Цитирование: Бабкин Э.А., Бабкина Т.С., Улитин Б.И. Моделирование динамики онлайн-дискуссий в сети Интернет с использованием многоагентных систем // Бизнес-информатика. 2018. № 2 (44). С. 17–29. DOI: 10.17323/1998-0663.2018.2.17.29.

Введение

Анализ общей структуры онлайн-дискуссий в сети Интернет и содержания отдельных сообщений является актуальной исследовательской задачей. В фундаментальном плане изучение феномена онлайн-дискуссий в различных формах (блоги, чаты и т.п.) позволяет углубить представление о феномене межличностных и сетевых коммуникаций. Прикладной аспект затрагивает проблемы создания интеллектуальных рекомендательных систем [1], повышения эффективности электронного маркетинга и электронного бизнеса [2], новых форм спонтанной и организованной политической деятельности [3, 4] и другие существенные проблемы современного общества и экономики.

Поставленная исследовательская задача имеет ярко выраженный междисциплинарный характер. Несмотря на успехи алгоритмических методов автоматического определения эмоциональной окраски сообщений с помощью методов машинного обучения, тщательный анализ сообщений в Интернет-дискуссии требует развития традиционных методов качественного и количественного анализа текста на основе многопараметрической экспертной оценки [5].

В данной работе предлагается две новых многоагентных имитационных модели динамики онлайн-дискуссии, основанные на результатах экспертной и статистической обработки большого объема фактического материала, полученного из Интернет-источников (сайты СМИ). В первой модели предполагается, что за поведение агентов-авторов отвечает централизованный блок генерации комментариев. Во второй модели поведение агентов-авторов является индивидуализированным и реализуется посредством многоуровневой иерархической марковской цепи.

По сравнению с другими подходами к моделированию дискуссий на основе машинного обучения [3, 6–9] предложенные решения дают исследователям новые возможности моделирования поведения отдельных участников дискуссии и прогнозирования ее развития в целом. Разработанная структура позволяет формировать не только такие традиционные характеристики дискурса, как интенции и референциальный объект отдельных сообщений [10, 11], но и генерировать структуру дискуссии по основным графовым метрикам, таким как центральная вершина/клика графа, количество ветвей дискуссии, среднее количество и дисперсия количества комментариев в ветвях дискуссии, совпадающей с реально существующими дискуссиями. Использование парадигмы многоагентных вычислений позволило осуществить разработку моделей генерации на индивидуальном уровне отдельных авторов.

В данной статье результаты представлены следующим образом. В разделе 1 проблема анализа и прогнозирования развития онлайн-дискуссий рассматривается в общем виде и ставится задача многоагентного имитационного моделирования. Раздел 2 содержит информацию о структуре и динамике предлагаемых многоагентных моделей. Основные сведения о программной реализации представлены в разделе 3. Раздел 4 посвящен описанию и анализу имитационных экспериментов, выполненных с помощью разработанного программного инструментария. В Заключении подводятся итоги исследования, выполняется сравнение с известными аналогами и определяются пути дальнейшего развития полученных результатов.

1. Ключевые направления исследований онлайн-дискуссий в сети Интернет

Значительная часть современных исследований онлайн-дискуссий в Интернете использует различные варианты статистических оценок появления слов или фраз на основе распределения Дирихле [12–14]. Среди исследований, непосредственно относящихся к теме нашей работы, можно отметить результаты Дж. Ванга и др. [8]. Оригинальный метод предсказания структуры также предлагается в работе Т. Яно, У. Когена и Н. Смита [3] и А. Риттера с коллегами [7]. В ряде важных аспектов интерес представляют результаты, полученные за послед-

ние годы в широкой исследовательской области «анализ тональности» [2, 9, 15–17]. Например, в работе [9] можно отметить эффективное применение техник машинного обучения с применением нейронных искусственных сетей.

Несмотря на важные теоретические и прикладные результаты, полученные в перечисленных работах, остаются нерешенными важные проблемы. В их число входит возможность моделирования отдельных комментариев дискуссии по нескольким параметрам (референциальный объект, интенция), а также моделирования последовательного формирования структуры самой дискуссии.

Решение перечисленных проблем может быть получено при использовании междисциплинарного подхода, в котором согласованным образом сочетаются методы качественной экспертной оценки текстов, свойственные психолингвистике, и методы имитационного моделирования на основе компьютерных моделей индивидуальных сущностей.

Качественная оценка текстовых сообщений реальных онлайн-дискуссий группой независимых экспертов позволяет сформировать необходимый для предварительной статистической обработки массив эмпирических данных. В рамках междисциплинарных научных исследований была организована психолингвистическая экспертная обработка 300 статей ведущих российских Интернет-СМИ с пользовательскими комментариями (от 24 до 200 комментариев на статью). В итоге были разработаны обобщенные типологии интенций, содержания (контент-коды) и референциальных объектов. На этой основе были сформированы необходимые категориальные матрицы и выполнена оценка статистической достоверности появления различных комбинаций интенций, контент-кодов и референциальных объектов в ходе дискуссий. Поскольку подробное описание этой части исследований выходит за рамки статьи, за более подробным описанием разработанных методики, типологий, результатов и анализа можно обратиться к работам [4, 5].

Имитационное моделирование на основе индивидуальных сущностей [18] естественным образом дополняет результаты качественной экспертной оценки текстов онлайн-дискуссий и позволяет построить класс повторно-воспроизводимых компьютерных моделей поведения отдельных авторов и структуры дискуссии в целом. Практическим инструментом реализации таких моделей является технология многоагентного программирования. Сегодня многоагентные системы активно используются для моделирования различных социальных феноменов [19, 20] или оптимизации [21, 22], однако исследования особенностей и возможностей многоагентных моделей онлайн-дискуссий только зарождаются.

2. Предлагаемые многоагентные модели дискуссии

Основой для предлагаемых многоагентных имитационных моделей служит формальная модель коммуникации и динамики дискуссии на основе графа в виде направленного дерева. В этом случае текст, послуживший основой дискуссии, является корнем дерева, а появляющиеся текстовые сообщения с комментариями становятся вершинами дерева. Каждая вершина графа содержит определенный набор параметров: набор интенций, код содержания (контент) и референциальный объект текстового сообщения.

Такая математическая структура используется в составе как централизованной, так и в распределенной имитационных моделей. Их отличает лишь механизм генерации новых вершин графа дискуссии.

2.1. Описание структуры модели в терминах многоагентной системы

В разработанных моделях присутствуют два вида агентов: активные (интеллектуальные, обладающие собственным поведением и принимающие решения) агенты-авторы и создаваемые ими пассивные (являющиеся продуктом решений активных агентов и не имеющие собственного поведения) агенты-комментарии.

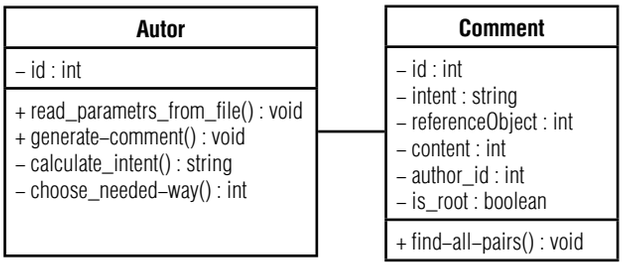

Рис. 1. UML–диаграмма агентов имитационной модели

Каждый агент-автор (*рисунок 1*) характеризуется своим уникальным номером и обладает способностью генерировать комментарии, основываясь на нескольких глобальных переменных окружения: ***ways***, хранящей все возможные пути от корня к листовым вершинам ветвей дискуссии, а также ***neededwayid*** и

cur_way, используемых для хранения ветви дискуссии, с которой проводится работа на текущем шаге работы модели.

Агенты-комментарии характеризуются только своими атрибутами: интенцией (***intent***), референциальным объектом (***referenceObject***), контентом (***content***), идентификатором принадлежности конкретному автору (***author_id***), а также вспомогательным атрибутом-флагом (***is_root***), позволяющим отличить оригинальную статью от комментариев к ней.

Также в модели есть ряд переменных окружения, необходимых для статистического анализа модели: ***branches_count***, хранящая количество ветвей дискуссии, ***comments_count_in_branches_M*** и ***comments_count_in_branches_D***, хранящие среднее количество и среднеквадратическое отклонение количества комментариев во всех ветвях дискуссии соответственно, и некоторые другие.

2.2. Описание правил поведения агентов

Поведение агентов-авторов в модели задается набором правил, применение которых определяется на каждом шаге работы модели значениями переменных окружения и состоянием дерева дискуссии.

Данные правила включают правило генерации комментария, правило определения характеристик комментария, правило расположения комментария по глубине и правило выбора ветви для комментария. Первые два правила одинаковы как для многоагентной, так и для централизованной моделей, остальные зависят от версии модели.

Правило генерации комментария выглядит следующим образом:

$$Igc = \begin{cases} 1, \text{ если } x < Cgc_i^t \\ 0, \text{ иначе} \end{cases},$$

где *Igc* – переменная-индикатор, отвечающая за необходимость генерировать комментарий;

x – значение равномерно распределенной случайной величины $X(0, 1)$;

$i = 1, 2, ..., I$ – индекс-идентификатор текущего агента-автора;

Cgc_i^t – пороговое значение, которое определяется моделью поведения агента-автора i в момент времени t.

В централизованной модели пороговое значение одинаково для всех авторов и не меняется с течением времени, то есть $Cgc_i^{t+1} = Cgc_i^t = Cgc_1^0, \forall i \geq 1 \; и \; t \geq 0$.

Правило определения характеристик генерируемого комментария создается на основании информации о парных правилах генерации комментариев, каждое из которых имеет вид:

$$i_1 r_1 \rightarrow i_2 r_2 N,$$

где $i_1 r_1$ и $i_2 r_2$ – интенция и референциальный объект комментария-родителя и комментария-потомка соответственно;

N – число раз, когда данная пара комментариев родителя-потомка встретилась в дереве дискуссии.

Группируя семейство таких правил по коммен-тарию-родителю, для каждого из них получаем следующий набор парных правил:

$$i_p r_p \rightarrow \{i_j r_j N_j\},$$

где $\{i_j r_j N_j\}, j = 1, 2, ..., J$ – семейство всех комментариев-потомков, связанных с текущим комментарием-родителем $i_p r_p$.

При этом:

$$j = \begin{cases} 1, \text{ если } 0 \leq x < N_1 \\ 2, \text{ если } N_1 \leq x < N_2 \\ \quad \dots \\ J, \text{ если } N_{J-1} \leq x < N_J \end{cases},$$

где x – значение равномерно распределенной случайной величины $X\left(0, \sum_j N_j\right)$.

В централизованной модели набор пороговых правил совпадает для всех агентов-авторов, участвующих в дискуссии, а в многоагентной – задается для каждого автора индивидуально, на основании модели его поведения в момент времени t.

Правило расположения комментария по глубине в случае централизованной модели принимает следующий вид:

$$Iclt = \begin{cases} 1, \text{ если } x < Clt \\ 0, \text{ иначе} \end{cases},$$

а в случае многоагентной модели обобщается до следующего:

$$Imlt = \begin{cases} 1, \text{если } x < Ct_i^t \\ 2, \text{если } Ct_i^t \leq x < Cnt_i^t \\ 3, \text{если } Cnt_i^t \leq x < Cm_i^t \\ 4, \text{если } Cm_i^t \leq x < Cnl_i^t \\ 5, \text{иначе} \end{cases},$$

где *Iclt* – переменная-индикатор, определяющая расположение генерируемого комментария по глубине в централизованной модели (1 – комментарий генерируется к листовым вершинам, в противном случае – к нелистовым);

Imlt – переменная-индикатор, определяющая расположение генерируемого комментария по глубине в распределенной модели (1 – комментарий генерируется к корневым вершинам, 2 – на глубину не более *M* от корневых вершин, 3 – на глубину более *M* от корневых но не более *M* от листовых вершин, 4 – на глубину не более *M* от листовых вершин, 5 – к листовым вершинам);

x – значение равномерно распределенной случайной величины *X*(0, 1);

i = 1, 2, ..., *I* – индекс-идентификатор текущего агента-автора,

Clt – пороговое значение, которое определяется значением переменной окружения ***leaf_top_connection_probability***;

M – критическое значение удаленности от корня/листа ветви дискуссии;

Ct_i^t, Cnt_i^t, Cm_i^t, Cnl_i^t – пороговые значения, которые определяются моделью поведения агента-автора *i* в момент времени *t* и описывают количественное распределение комментариев автора по глубине ветви дискуссии.

Правило выбора автором ветви для генерации комментария определяется на основании конфигурации дерева дискуссии в конкретный момент времени *t*. Пусть имеется *M* ветвей дискуссии, длина каждой из них $L_i\left(1 \le i \le M, i \in \mathbb{N}\right)$, в которых автор *A* оставил cc_i количество комментариев в *i*-й $\left(1 \le i \le M, i \in \mathbb{N}\right)$ ветви дискуссии. Тогда в централизованной модели правило выбора автором *i*-й ветви для генерации комментария можно сформулировать следующим образом:

$$i = \begin{cases} 1, \text{ если } 0 \le x < cc_1 \\ 2, \text{ если } cc_1 \le x < cc_2 \\ \cdots \\ M, \text{ если } cc_{M-1} \le x < cc_M \end{cases}.$$

В многоагентной модели данное правило обобщается до следующего:

$$i \in I = \begin{cases} Io, \text{ если } 0 \le x < Co_A^t \\ If, \text{ если } Co_i^t \le x < Cf_A^t \\ Imf, \text{ если } Cf_i^t \le x < Cmf_A^t \\ Im, \text{ иначе} \end{cases},$$

где $Io = \left\{i : cc_i = 0\right\}$, $If = \left\{i : \frac{cc_i}{L_i} < S\right\}$,

$Imf = \left\{i : S < \frac{cc_i}{L_i} < 1 - S\right\}$, $Im = \left\{i : \frac{cc_i}{L_i} > S\right\}$

x – значение равномерно распределенной случайной величины $X\left(0, \sum_j cc_j\right)$;

S – критическое значение;

Co_A^t, Cf_A^t, Cmf_A^t – пороговые значения, которые определяются моделью поведения автора *A* в момент времени *t* и описывают количественное распределение комментариев автора по ветвям дискуссии.

В обоих случаях идентификатор ветви, выбранной для генерации комментария, а также сама ветвь хранятся в переменных ***neededwayid*** и ***cur_way*** соответственно.

Таким образом, с использованием данных правил поведение каждого агента-автора в случае централизованной модели может быть описано следующим образом (*рисунок 2*).

В распределенной многоагентной модели данное поведение усложняется и определяется для каждого агента-автора индивидуально на основе иерархической марковской цепи [23] (*рисунок 3*). Однако в целом этапы (уровни марковской цепи) применения данной модели совпадают с централизованной моделью.

Важно отметить, что результатом работы централизованной модели является набор марковских цепей агентов-авторов, который используется как входной массив данных при работе многоагентной версии модели. Именно за счет такой взаимосвязанной организации работы моделей мы можем провести их последующий сравнительный анализ на идентичность.

Рис. 2. Концептуальная обобщенная схема поведения агента–автора

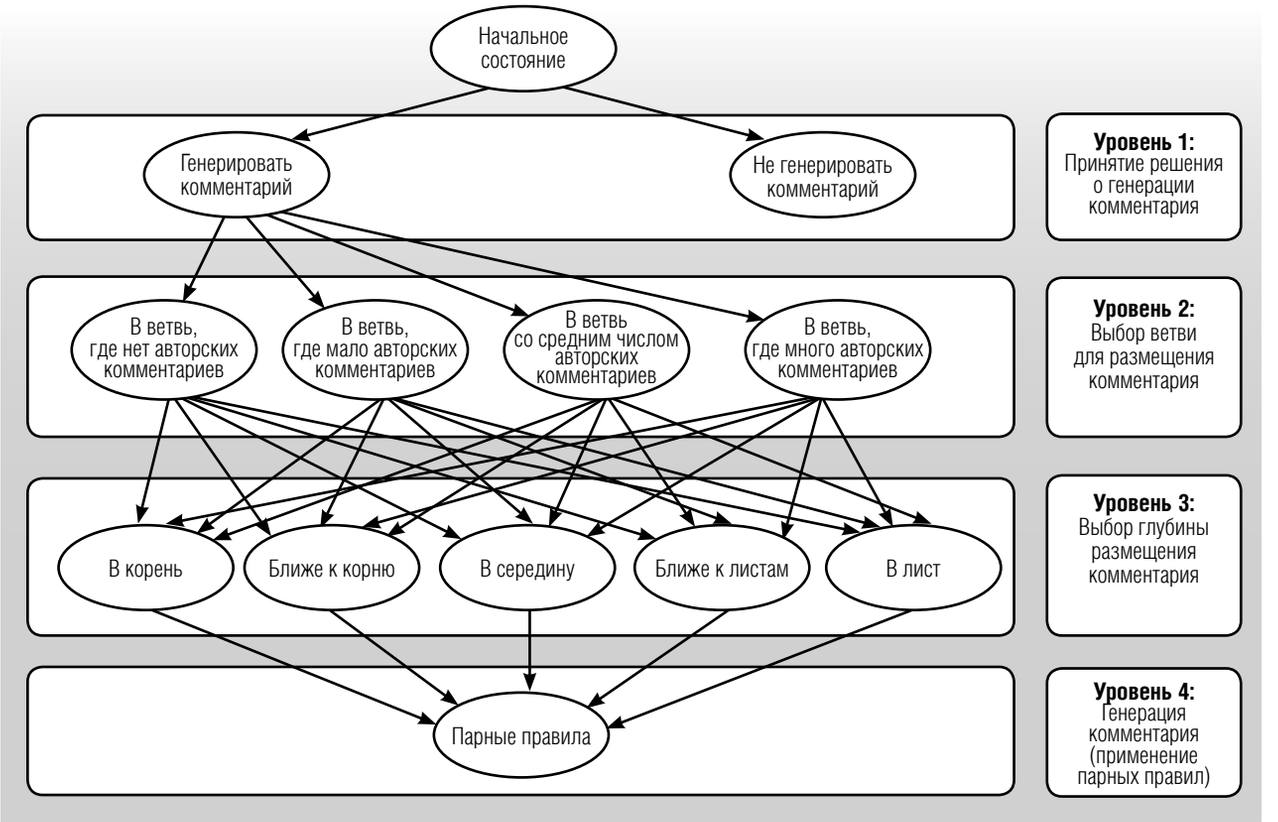

Рис. 3. Концептуальная схема марковской цепи поведения агента–автора

3. Программная реализация системы имитационного моделирования

Для программной реализации предложенных моделей дискуссии использовалась среда многоагентного моделирования NetLogo (версия 6.0.1+) [24]. Для визуализации и анализа графов марковских цепей, описывающих индивидуальное поведение агентов-авторов в распределенной модели, был разработан программный модуль анализа на языке программирования Java с использованием библиотеки с открытым исходным кодом для визуализации графов JUNG (версия 2.0.1+) [25].

	00;з4;1
5	00;е4;1
[1 0 4 0]	ш3;ш3;1
[3 0 0 0 2]	00;н4;1
	н4;и4;1
Файл с мета–данными	Файл с парными правилами

Рис. 4. Примеры выходных файлов централизованной модели (в первой строке зафиксировано общее количество комментариев автора, в двух других – их распределение: во второй – по глубине, в третьей – по ветвям дерева дискуссии)

Созданная программная реализация централизованной многоагентной модели не требует никаких входных файлов, а на выходе с заданной периодичностью генерирует пары файлов для каждого автора. Примеры содержимого таких файлов приведены на *рисунке 4.*

Эти файлы являются входными для программной реализации распределенной модели, которая, в свою очередь, сохраняет результаты своей работы в виде идентичных по структуре пар файлов с заданной периодичностью для последующего их сравнения с результатами работы централизованной модели.

Программный модуль анализа принимает на вход группу файлов с парными правилами и строит графическое отображение заданных парных правил (*рисунок 5*).

Как видно из рисунка, при визуализации в центре отражены все вершины-родители, которые возникали на протяжении всей работы модели, а в разных кластерах (секторах схемы) сгруппированы вершины-комментарии, являющиеся потомками, возникавшими на разных этапах работы модели, демонстрируя изменения в марковской цепи с течением времени.

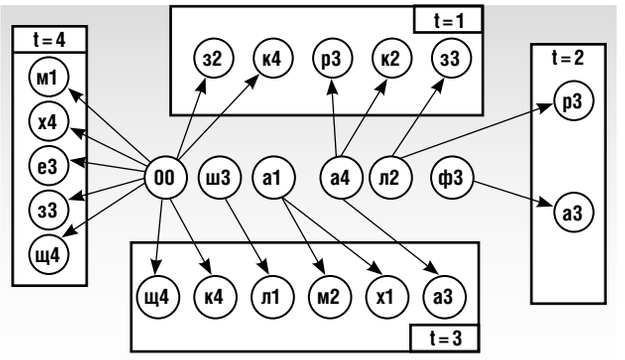

Рис. 5. Примеры визуализации выходного графа марковской цепи агента–автора (окружности – комментарии (их интент и референциальный объект), в прямоугольники объединены комментарии, сгенерированные автором на каждом периоде работы модели)

Сначала проведем эксперимент с малой дискуссией (*рисунок 6*). Определим двух агентов-авторов, которые на каждом шаге генерируют комментарий с вероятностью 0,9 и относят созданный комментарий к нелистовым и листовым вершинам с вероятностями 0,25 и 0,75 соответственно. Период создания записи в файлах результата и общее время работы модели составляют 5 и 25 тактов соответственно.

По итогам работы централизованной модели проведем запуск распределенной версии (граф конечного состояния представлен на *рисунке 7*) и сравним полученные деревья дискуссии и их основные показатели междоу собой.

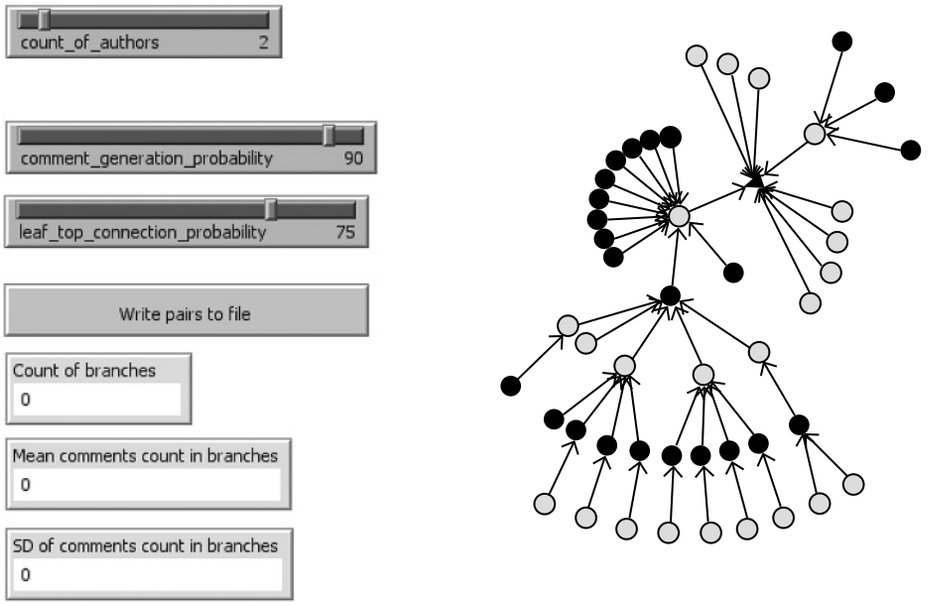

Рис. 6. Начальные параметры панели управления и конечное состояние графа дискуссии централизованной модели (разными цветами выделены комментарии, сгенерированные разными авторами)

4. Анализ результатов экспериментов

Для того, чтобы оценить качество разработанной модели и ее соответствие данным реальных графов онлайн-дискуссий, проведем несколько серий экспериментов для централизованной и многоагентной версий модели соответственно. Будем рассматривать эксперимент на малых (количество комментариев ~25) и на больших (количество комментариев ~200) размерах дискуссии. Такие показатели взяты, исходя из размеров реальных деревьев онлайн-дискуссий (см. раздел 1), что позволит в дальнейшем провести сравнительный анализ результатов работы моделей с фактическими данными.

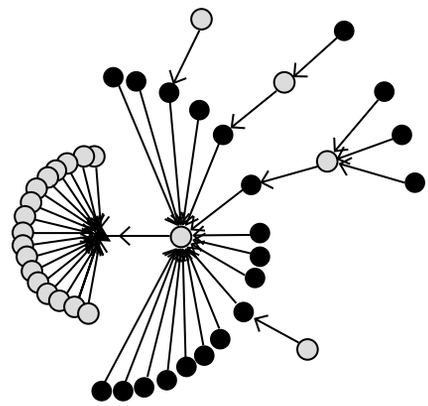

Рис. 7. Конечное состояние графа дискуссии распределенной модели

Как видно из *рисунков 6* и *7*, по своему внешнему виду данные графы имеют ряд схожих свойств. Например, большая часть комментариев сконцентрирована вокруг оригинальной статьи (вершина №0) и одного из комментариев к ней (комментарий №3). Данные вершины являются центральными вершинами графа (по количеству входных/выходных вершин): корневая вершина имеет степень 18 и 9 во втором и первом случаях соответственно, а комментарий №3 – 17 и 11 во втором и первом случаях соответственно.

Чтобы подтвердить согласованность результатов имитационного моделирования с помощью обеих моделей между собой и фактической онлайн-дискуссией, проведем более детальный анализ с помощью сравнения графов марковских цепей поведения агентов-авторов в аспекте правил генерации интент-контентных характеристик комментариев.

Для первого агента-автора данная марковская цепь была представлена ранее (*рисунок 5*). Если сопоставить ее с моделью для того же автора, полученной в ходе работы централизованной модели (*рисунок 8*), то можно обнаружить сходство: в центральной части содержатся часть совпадающих между собой вершин. Это свидетельствует о сохранении общих принципов поведения автора и совпадении индивидуальной модели с результатами работы централизованной модели. Аналогичные результаты можно получить и для второго агента-автора, участвующего в данной модели.

Теперь рассмотрим поведение моделей в случае дискуссий большого размера. Сохраним ранее заданные параметры централизованной модели (*рисунок 6*), но установим длительность 230 тактов. На выходе получим следующую марковскую цепь поведения первого агента-автора (*рисунок 9*).

Используя эту марковскую цепь для получения входных параметров распределенной версии модели, на ее выходе также получим марковскую цепь поведения первого агента-автора (*рисунок 10*).

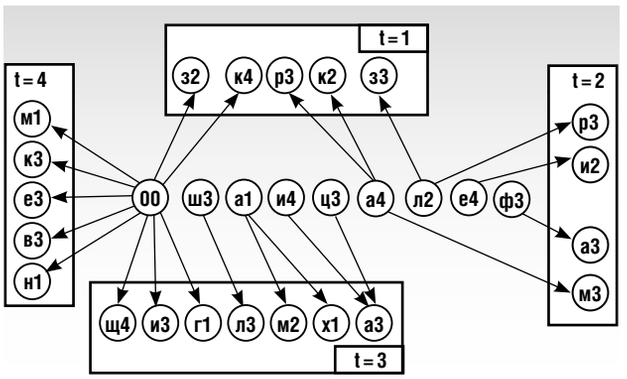

Рис. 8. Марковская цепь первого агента–автора по итогам работы централизованной модели

Сопоставляя полученные по итогам двух моделей схемы поведения первого автора, можно заметить, что они в значительной степени похожи между собой. Так, центральная клика вершин графа в обоих случаях включает в себя вершины 00 (оригинальная статья), а также комментарии с интенциями а3, р3, г4 и др. Кроме того, сходство по интенциям обнаруживается и при сравнении тактовых сегментов марковских цепей, что говорит об идентичности поведения автора в моделях. Подтверждение этому можно получить, если сравнить между собой файлы с мета-данными по иерархической марковской цепи поведения автора (*рисунок 11*).

Так, по итогам работы распределенной модели агент-автор сгенерировал 296 комментариев, из которых 42 попали в ветви без комментариев данного автора, 115 – со средним количеством комментариев автора (от 20% до 80% общего количества в ветви) и 75 – с малым количеством комментариев автора (<20% общего количества в ветви), остальные попали в ветви с большим количеством комментариев автора (>80% общего количества в ветви). При этом 67 из данных комментариев были сгенерированы в корень дискуссии, 70 – в листы, 66 – ближе к листам дискуссии, а оставшиеся – в среднюю часть ветви дискуссии.

Важно также отметить тот факт, что, поскольку при разработке централизованной модели использовались данные о парных правилах из реальных онлайн-дискуссий, можно говорить о соответствии между централизованной моделью и реальными графами дискуссии. Подтверждение этому также можно найти, сравнивая марковскую цепь централизованной модели (*рисунок 9*) с марковской цепью реальной дискуссии (*рисунок 12*).

Таким образом, можно отметить, что разработанные модели не противоречат данным реальных онлайн-дискуссий и, следовательно, могут использоваться для анализа и прогнозирования развития последних. Также можно отметить, что для моделирования поведения автора может быть использована модель конечного автомата с памятью, поскольку в централизованной модели поведение автора зависит только от его предыдущих действий и состояния графа дискуссии к настоящему моменту времени.

Заключение

В настоящей работе было проведено исследование проблемы имитационного моделирования

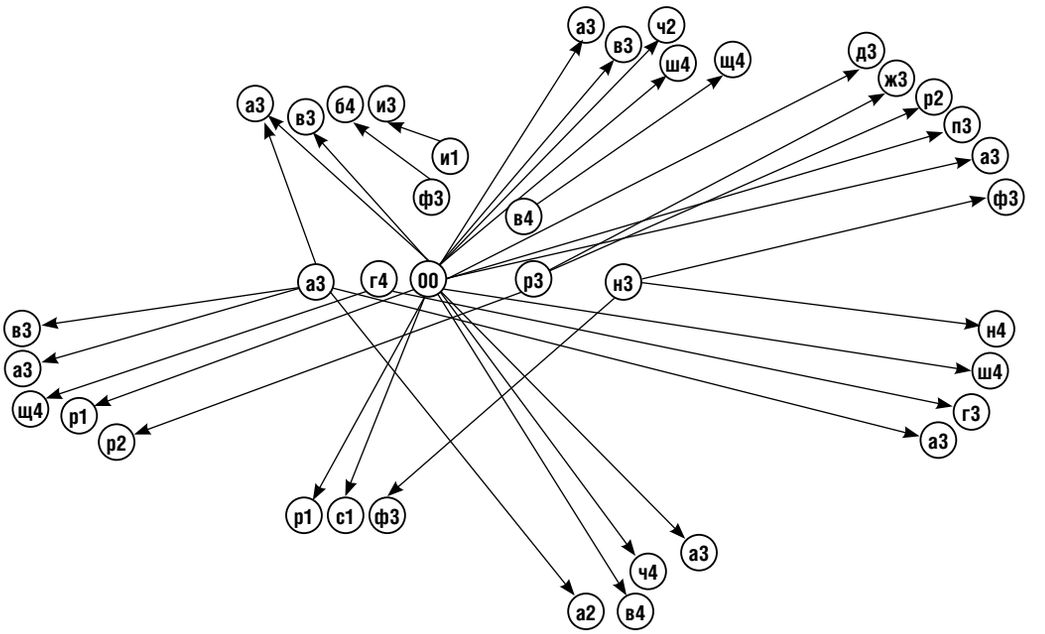

Рис. 9. Фрагмент марковской цепи первого автора централизованной модели (230 тактов)

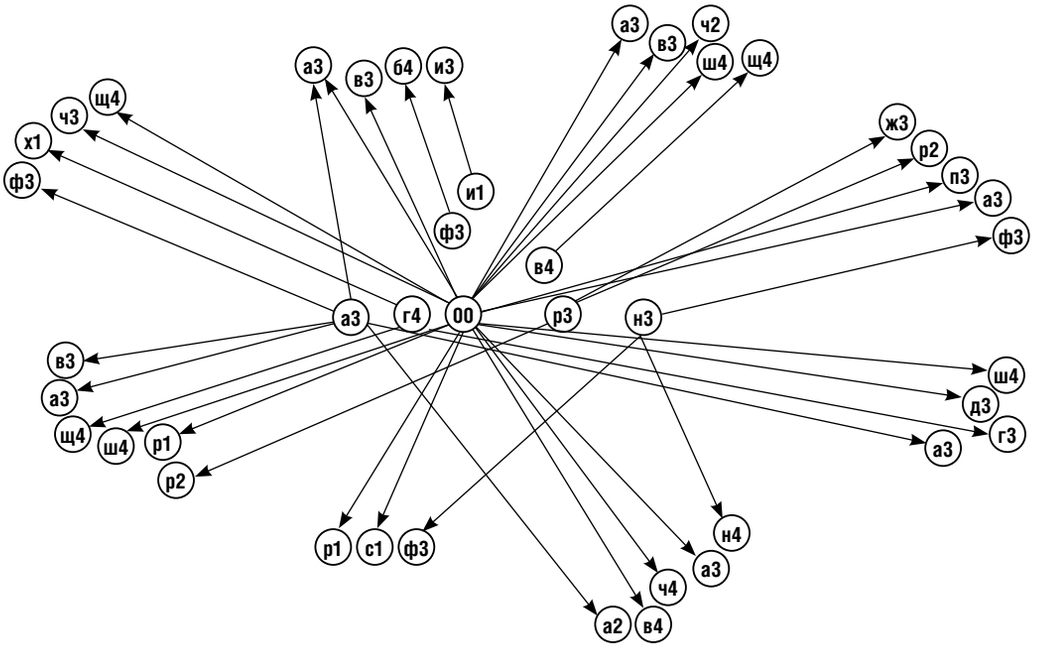

Рис. 10. Фрагмент марковской цепи первого автора распределенной модели (230 тактов)

296 [427511554] [6736376670]	326 [828710354] [7736776670]
Индивидуализированная модель	Обобщенная модель

Рис. 11. Файлы мета–данных первого автора по итогам работы моделей

поведения авторов в ходе онлайн-дискуссий. Особенностью постановки исследовательской задачи стала комбинация количественной оценки большого объема фактических данных из электронных СМИ и последующего многоагентного моделирования структуры и динамики дискуссии в терминах стохастической многоагентной системы с применением иерархической марковской сети особого вида. Для нее была выполнена программная реали-

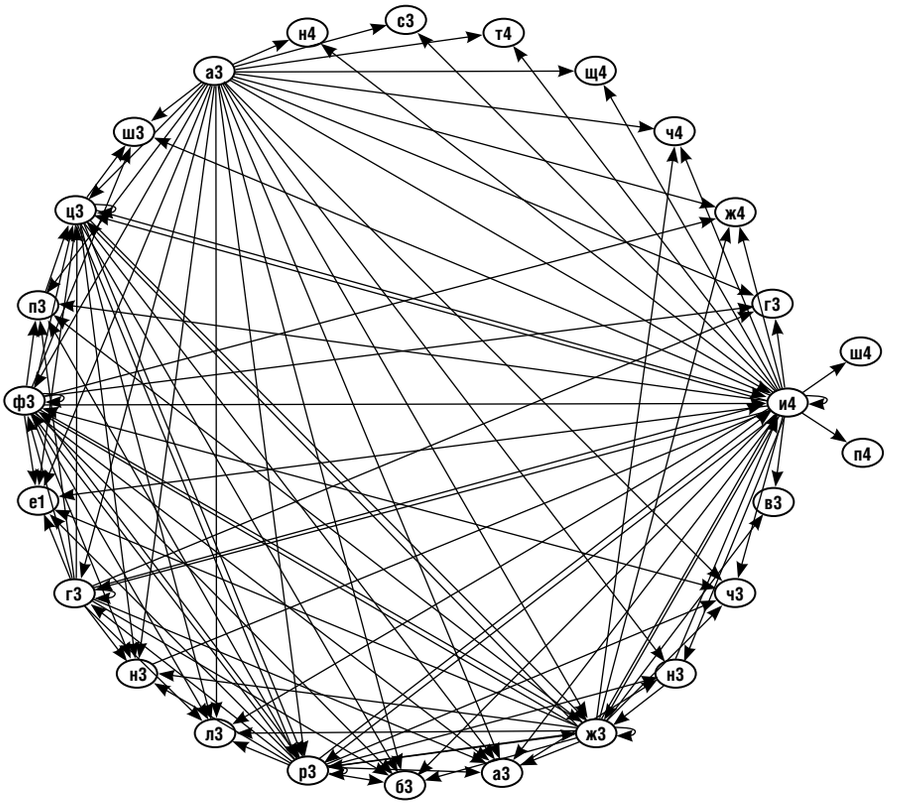

Рис. 12. Фрагмент марковской цепи агента–автора реальной дискуссии

зация двух видов моделей и проведен сравнительный анализ полученных в ходе имитационного моделирования характеристик дискуссии.

Можно сделать вывод о том, что разработанные правила поведения агентов-авторов адекватно отражают основные принципы формирования структуры дискуссии в целом и характеристики отдельных сообщений в терминах референциальных объектов, интент- и контент-анализа. Ключевые характеристики результатов двух моделей совпадают между собой и с результатами обработки реальных онлайн-дискуссий, полученных из сети Интернет. Таким образом разработанные средства имитационного моделирования могут стать эффективным средством прогнозирования развития онлайн-дискуссий для широкого круга практических задач в области Интернет-маркетинга, политологии и электронного бизнеса. По сравнению с другими исследованиями в области моделирования дискуссии (в частности [6, 7]) предлагаемый подход позволяет имитировать последовательное формирование структуры дискуссии группой авторов в виде графа с такими важными характеристиками отдельных сообщений как интенция, референциальный объект, контент-код.

Говоря о будущем развитии построенных многоагентных моделей дискуссий, следует обратить внимание на необходимость ее видоизменения с целью подтверждения гипотезы о достаточности и адекватности предельной модели поведения агента-автора в виде конечного автомата с памятью. Для этого можно существующую модель перенести на многоагентную основу, представив каждого агента в виде классической тройки «предпосылки – стремления – намерения» (belief–desire–intention, BDI) [26], где в качестве первых будет использоваться модель автомата, построенная на состоянии графа дискуссии, а в качестве последних – стохастический модуль, описывающий поведение агента.

Также положительным является внедрение в модель полноценного контентного дополнения для завершения полноты созданной модели, позволяющей анализировать только интентный аспект дискуссии, без соотнесения ее с той предметной областью, в которой данная дискуссия проводится. Это позволит осуществить более тонкую настройку модели, а также изучить ограничения на ее применения в различных предметных областях. ■

Литература

1. Yang D., Huang C., Wang M. A social recommender system by combining social network and sentiment similarity: A case study of healthcare // Journal of Information Science. 2017. Vol. 43. No. 5. P. 635–648.
2. Mining the minds of customers from online chat logs / K. Park [et al.] // Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM 2015), Melbourne, Australia, 19–23 October 2015. P. 1879–1882.
3. Yano T., Cohen W.W., Smith N.A. Predicting response to political blog posts with topic models // Proceedings of the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2009). Boulder, Colorado, 31 May – 5 June 2009. P. 477–485.
4. Кудрявцева Е.Н., Радина Н.К. Антисоциальное поведение в online-дискуссиях // Материалы 3-й Международной научно-практической конференции «Актуальные проблемы исследования массового сознания», г. Пенза, 24–25 марта 2017 г. С. 86–89.
5. Радина Н.К. Интент-анализ онлайн-дискуссий (на примере комментирования материалов Интернет-портала ИноСМИ.ру) // Медиаскоп. 2016. № 4. С. 25.
6. Hoang T.A., Lim E.P. Modeling topics and behavior of microbloggers: An integrated approach // ACM Transactions on Intelligent Systems and Technology. 2017. Vol. 8. No. 3. P. 44.
7. Ritter A., Cherry C., Dolan B. Unsupervised modeling of Twitter conversations // Human Language Technologies // Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2010), Los Angeles, California, 1–6 June 2010. P. 172–180.
8. Diversionary comments under blog posts / J. Wang [et al.] // ACM Transactions on the Web. 2015. Vol. 9. No. 4. P. 18.
9. Karpov N., Demidovskij A., Malafeev A. Development of a model to predict intention using deep learning // Proceedings of the 6th International Conference “Analysis of Images, Social Networks and Texts” (AIST 2017), Moscow, Russia, 27–29 July 2017; Lecture Notes in Computer Science, Revised Selected Papers. Springer, 2017. P. 69–78.
10. Референциальный выбор как многофакторный вероятностный процесс / А.А. Кибрик и [др.] // Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции «Диалог 2010». Т. 9. № 16. М.: РГГУ, 2010. С. 173–180.
11. Сулейманова Е.А. О референциальных аспектах задачи извлечения фактов // Программные системы: теория и приложения. 2012. Т. 3. № 3. С. 41–56.
12. Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet allocation // Journal of Machine Learning Research. 2003. No. 3. P. 993–1022.
13. Jelodar H., Wang Y., Yuan C., Feng X. Latent Dirichlet Allocation (LDA) and topic modeling: models, applications, a survey // arXiv preprint, arXiv:1711.04305. 2017.
14. Paisley J., Wang C., Blei D.M., Jordan M.I. Nested hierarchical Dirichlet processes // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2015. Vol. 37. No. 2. P. 256–270.
15. Balazs J.A., Velásquez J.D. Opinion mining and information fusion: A survey // Information Fusion. 2016. Vol. 27. P. 95–110.
16. Bele N., Panigrahi P.K., Srivastava S.K. Political sentiment mining: A new age intelligence tool for business strategy formulation // International Journal of Business Intelligence Research. 2017. Vol. 8. No. 1. P. 55–70.
17. Олешков М.Ю. Моделирование коммуникативного процесса. Нижний Тагил: НТГСПИ, 2006.
18. Macal C.M. Everything you need to know about agent-based modelling and simulation // Journal of Simulation. 2016. Vol. 10. No. 2. P. 144–156.
19. Garcia D., Garas A., Schweitzer F. An agent-based modeling framework for online collective emotions // Cyberemotions / Eds. J. Holyst. Cham: Springer, 2017. P. 187–206.
20. Sun J., Tang J. A survey of models and algorithms for social influence analysis // Social network data analytics / Eds. C. Aggarwal. Boston, MA: Springer, 2011. P. 177–214.
21. Babkin E., Abdulrab H., Babkina T. AgentTime: A distributed multi-agent software system for university’s timetabling // From system complexity to emergent properties / Eds. M.A. Aziz-Alaoui, C. Bertelle. Berlin, Heidelberg: Springer, 2009. P. 341–354.
22. Satunin S., Babkin E. A multi-agent approach to intelligent transportation systems modeling with combinatorial auctions // Expert Systems with Applications. 2014. Vol. 41. No. 15. P. 6622–6633.
23. Ausin M.C. Markov chain Monte Carlo, Introduction. Wiley StatsRef: Statistics Reference Online, 2015. [Электронный ресурс]: https://www.sciencedirect.com/science/article/pii/B0080430767004691 (дата обращения 02.04.2018).
24. Sklar E. Software review: NetLogo, a multi-agent simulation environment // Artificial Life. 2007. Vol. 13. No. 2. P. 1–9.
25. Yan-Biao B. The JUNG (Java Universal Network/Graph) Framework. Technical Report UCI-ICS 03-17. Irvine, CA: University of California, 2004. [Электронный ресурс]: http://www.datalab.uci.edu/papers/JUNG_tech_report.html (дата обращения 02.04.2018).
26. Fasli M. Interrelations between the BDI primitives: Towards heterogeneous agents // Cognitive Systems Research. 2003. Vol. 4. No. 1. P. 1–22.

Multi-agent simulation modeling of online Internet discussions[2]

Eduard A. Babkin
Professor, Department of Information Systems and Technologies
National Research University Higher School of Economics
Address: 25/12, Bolshaya Pecherskaya Street, Nizhny Novgorod, 603155, Russian Federation
E-mail: eababkin@hse.ru

Tatiana S. Babkina
Senior Lecturer, Department of Information Systems and Technologies
National Research University Higher School of Economics
Address: 25/12, Bolshaya Pecherskaya Street, Nizhny Novgorod, 603155, Russian Federation
E-mail: tbabkina@hse.ru

Boris I. Ulitin
Senior Lecturer, Department of Information Systems and Technologies
National Research University Higher School of Economics
Address: 25/12, Bolshaya Pecherskaya Street, Nizhny Novgorod, 603155, Russian Federation
E-mail: bulitin@hse.ru

Abstract

Joint analysis of the general structure of online Internet discussions and different attributes of particular text comments becomes an important scientific task in theoretical and applied aspects. Although methods of machine learning facilitate stochastic analysis of text messages, appropriate modeling of dynamics of online discussion and psycho-linguistic characteristics of comments in the presence of multiple individual authors remains the unresolved problem. In this article, the authors suggest applying the methods of multi-agent simulations for resolution of that problem. This work offers two models of online discussion which allow us to take into account characteristics of individual comments and the presence of multiple authors with individual models of behavior. The behavior models are designed in the result of analysis of actual online Internet discussions. The first model is centralized and represents the behavior of each author in the same manner, using a set of fixed parameters. In comparison to the centralized model, the multi-agent distributed model states the individualized behavior for every author through the Markov chain of the special form. Such individualized structure allows us not only to approach the real dynamics of the discussion, but also to compare the models with the actual online Internet discussions. Using pre-processed factual data of real discussions from various Internet portals became an important feature of the suggested approach to simulation modeling. Pre-processing includes expert evaluation of psycho-linguistic characteristics (intent and content analysis), as well as methods of mathematical statistics. Therefore, this research is a positive example of inter-disciplinary research of Internet communication phenomena.

Key words: communication models, online discussions, multi-agent systems, simulation modeling.

Citation: Babkin E.A., Babkina T.S., Ulitin B.I. (2018) Multi-agent simulation modeling of online Internet discussions. *Business Informatics*, no. 2 (44), pp. 17–29. DOI: 10.17323/1998-0663.2018.2.17.29.

References

1. Yang D., Huang C., Wang M. (2017) A social recommender system by combining social network and sentiment similarity: A case study of healthcare. *Journal of Information Science*, vol. 43, no. 5, pp. 635–648.
2. Park K., Kim J., Park J., Cha M., Nam J., Yoon S., Rhim S. (2015) Mining the minds of customers from online chat logs. Proceedings of the *24th ACM International on Conference on Information and Knowledge Management (CIKM 2015), Melbourne, Australia, 19–23 October 2015*, pp. 1879–1882.
3. Yano T., Cohen W.W., Smith N.A. (2009) Predicting response to political blog posts with topic models. Proceedings of the *2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2009), Boulder, Colorado, 31 May – 5 June 2009*, pp. 477–485.

[2] This work was supported by the Russian Foundation for Basic Research (project No. 16-06-00184 A)

4. Kudryavtseva E.N., Radina N.K. (2017) Antisotsial'noe povedenie v online-diskussiyah [Antisocial behavior in online discussions] // Proceedings of the *3rd International Scientific and Practical Conference "Actual problems of mass consciousness research", Penza, Russia, 24–25 March 2017*, pp. 86–89 (in Russian).
5. Radina N.K. (2016) Intent-analiz onlayn-diskussiy (na primere kommentirovaniya materialov internet-portala InoSMI.ru) [Intention analysis of online discussions (based on the example of comments on the materials of the Internet portal InoSMI.ru)]. *Mediascope*, no. 4, p. 25 (in Russian).
6. Hoang T.A., Lim E.P. (2017) Modeling topics and behavior of microbloggers: An integrated approach. *ACM Transactions on Intelligent Systems and Technology*, vol. 8, no. 3, p. 44.
7. Ritter A., Cherry C., Dolan B. (2010) Unsupervised modeling of Twitter conversations // Human Language Technologies. Proceedings of the *2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2010), Los Angeles, California, 1–6 June 2010*, pp. 172–180.
8. Wang J., Yu C.T., Yu P.S., Liu B., Meng W. (2015) Diversionary comments under blog posts. *ACM Transactions on the Web*, vol. 9, no. 4, p. 18.
9. Karpov N., Demidovskij A., Malafeev A. (2017) Development of a model to predict intention using deep learning. Proceedings of the *6th International Conference "Analysis of Images, Social Networks and Texts" (AIST 2017), Moscow, Russia, 27–29 July 2017; Lecture Notes in Computer Science, Revised Selected Papers*. Springer, pp. 69–78.
10. Kibrik A.A., Dobrov G.B., Zalmanov D.A., Linnik A.S., Lukashevich N.V. (2010) Referentsial'nyy vybor kak mnogofaktornyy veroyatnostnyy protsess [Referential choice as a multi-factor stochastic process]. *Computer linguistic and intelligent technologies (according to "Dialog 2010" International Conference)*, vol. 9, no. 16. Moscow: RSUH, pp. 173–180 (in Russian).
11. Suleymanova E.A. (2012) O referentsial'nyh aspektah zadachi izvlecheniya faktov [On the referential aspects of the facts extracting task]. *Program Systems: Theory and Applications*, vol. 3, no. 3, pp. 41–56 (in Russian).
12. Blei D.M., Ng A.Y., Jordan M.I. (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research*, no. 3, pp. 993–1022.
13. Jelodar H., Wang Y., Yuan C., Feng X. (2017) Latent Dirichlet Allocation (LDA) and topic modeling: models, applications, a survey. *arXiv preprint, arXiv*:1711.04305. 2017.
14. Paisley J., Wang C., Blei D.M., Jordan M.I. (2015) Nested hierarchical Dirichlet processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 2, pp. 256–270.
15. Balazs J.A., Vel squez J.D. (2016) Opinion mining and information fusion: A survey. *Information Fusion*, vol. 27, pp. 95–110.
16. Bele N., Panigrahi P.K., Srivastava S.K. (2017) Political sentiment mining: A new age intelligence tool for business strategy formulation. *International Journal of Business Intelligence Research*, vol. 8, no. 1, pp. 55–70.
17. Oleshkov M.Y. (2006) *Modelirovanie kommunikativnogo protsessa* [Simulation of the communication process]. Nizhny Tagil. NTSSPI (in Russian).
18. Macal C.M. (2016) Everything you need to know about agent-based modelling and simulation. *Journal of Simulation*, vol. 10, no. 2, pp. 144–156.
19. Garcia D., Garas A., Schweitzer F. (2017) An agent-based modeling framework for online collective emotions. *Cyberemotions* (eds. J. Holyst). Cham: Springer, pp. 187–206.
20. Sun J., Tang J. (2011) A survey of models and algorithms for social influence analysis. *Social network data analytics* (eds. C. Aggarwal). Boston, MA: Springer, pp. 177–214.
21. Babkin E., Abdulrab H., Babkina T. (2009) AgentTime: A distributed multi-agent software system for university's timetabling. *From system complexity*
Satunin S., Babkin E. (2014) A multi-agent approach to intelligent transportation systems modeling with combinatorial auctions. *Expert Systems with Applications*, vol. 41, no. 15, pp. 6622–6633.
23. Ausin M.C. (2015) *Markov chain Monte Carlo, Introduction*. Wiley StatsRef: Statistics Reference Online. Available at: https://www.sciencedirect.com/science/article/pii/B0080430767004691 (accessed 02 April 2018).
24. Sklar E. (2007) Software review: NetLogo, a multi-agent simulation environment. *Artificial Life*, vol. 13, no. 2, pp. 1–9.
25. Yan-Biao B. (2004) *The JUNG (Java Universal Network/Graph) Framework. Technical Report UCI-ICS 03-17*. Irvine, CA: University of California. Available at: http://www.datalab.uci.edu/papers/JUNG_tech_report.html (accessed 02 April 2018).
26. Fasli M. (2003) Interrelations between the BDI primitives: Towards heterogeneous agents. *Cognitive Systems Research*, vol. 4, no. 1, pp. 1–22.