INFORMATION TECHNOLOGY

УДК 004.048

ДМИТРИЙ ЛАНДЭ,
ВАДИМ ДОДОНОВ,
ТАРАС КОВАЛЕНКО

КОРПОРАТИВНАЯ СИСТЕМА МОНИТОРИНГА СЕТЕВЫХ ИНФОРМАЦИОННЫХ РЕСУРСОВ НА ОСНОВЕ МУЛЬТИАГЕНТНОГО ПОДХОДА

В работе представлена мультиагентная модель распространения информационных сообщений, содержащих ссылки на информационные ресурсы в сети Интернет. Результаты моделирования подтверждены путем исследования реальной сети микроблогов Twitter. Описаны этапы создания корпоративной системы мониторинга сетевых информационных ресурсов, состав которых определяется ссылками в микроблогах. Представленный подход к формированию баз данных на основе учета ссылок в микроблогах на информационные ресурсы обеспечивает такие преимущества: оперативность – информационное сообщение попадает в базу данных информационно-аналитической системы в режиме реального времени по мере того, как на него сослался первый же пользователь; охват главных информационных материалов по теме. Учет мнения аудитории заинтересованных пользователей, контент сообщений которых удовлетворяет широким корпоративным запросам; компактность баз данных, а, следовательно, удобство доступа конечных пользователей. Предсказуемость объемов баз данных, динамики их наполнения; технологическая совместимость с существующими информационно-аналитическими системами и системами контент-мониторинга; возможность выявления информационных кампаний, операций.

Ключевые слова: социальная сеть, моделирование, мониторинг информационных ресурсов, мультиагентная система, распространение информации.

Постановка проблемы. Объемы информационных ресурсов в веб-пространстве сегодня затрудняют оперативное получение необходимой пользователю информации даже с помощью мощнейших сетевых поисковых систем (Google, Baidu, Яandex и др.) [1]. Одним из путей решения этой проблемы является подключение мнения большого числа людей, экспертов в предметной области. Такая возможность, которую можно назвать краудсорсинговой, открывается при помощи содержательного анализа социальных сетей, где пользователи “голосуют” за те или иные информационные материалы, устанавливая на них гиперссылки. Особенно это актуально в сегментных предметных областях, соответствующих потребностям корпоративных пользователей. Несмотря на то, что анализ социальных сетей сам по себе является сложной научно-технической задачей, существующие поисковые возможности некоторых из них, дают надежду на решение названной проблемы.

В рамках данной работы предлагается подход к созданию корпоративной системы мониторинга сетевых информационных ресурсов, состав которых определяется ссылками в социальных сетях, в частности, сети микроблогов Twitter [2].

Вместе с тем, процессы распространения информации в сетях, содержащей ссылки на информационные ресурсы, требуют детального анализа. Моделирование распространения информации позволяет исследовать соответствующие информационные процессы, выявлять закономерности, которые могут использоваться как при изучении механизмов передачи информации в таких сетях, так и уровня ее воздействия на людей [3].

Мультиагентная модель распространения сообщений. Для создания мультиагентной модели распространения информации, прежде всего, необходимо сформировать близкое к реальности виртуальное информационное пространство, населенное виртуальными агентами, с которыми ассоциируются отдельные сообщения в социальной сети, которые инкапсулируют в себя ссылки на информационные ресурсы сети Интернет [4-5]. Предполагается, что отдельные агенты могут [6]:

1) самозарождаться;
2) порождать новых агентов путем репостинга (repost);
3) ”умирать” – исчезать из пространства агентов;
4) получать лайки (like) от других агентов.

Агент обладает ”потенциалом”, зависящим от его возраста, авторитетности (ссылок на него) и плодовитости (количества порожденных им агентов, репостов). Варьирование четырьмя параметрами управления позволили смоделировать профили поведения информационных сюжетов. В результате проведенных исследований была реализована программа эволюции пространства агентов, исследована эволюция мультиагентной системы, найдены аналогии с реальными тематическими информационными потоками. Были выявлены статистические закономерности, относящиеся к жизненному циклу отдельных сообщений, распределение которых, соответствует распределению Вейбулла [3]. Данные моделирования были проверены путем исследования реальной сети микроблогов Twitter. Совпадение результатов моделирования и параметров распределения реальной сети позволяют говорить о закономерности, присущей реальным сетям, а также об адекватности модели.

Моделирование динамики всего информационного потока начинается с одного агента. Появление нового агента возможно двумя способами. Первый заключается в копировании существующего агента с помощью операции репост. Также возможно самозарождение агента, что отвечает публикации нового сообщения. Таким образом, в каждый момент времени с определенными вероятностями, с каждым из агентов, может произойти любое из событий. Также в любой момент времени с вероятностью p_s может появиться новый агент в результате самозарождения.

Рассмотрим жизненный путь одного агента. Агент появляется с начальным значением энергии E_0 и далее его энергия изменяется в зависимости от событий, которые с ним происходят. Будем считать, что возможны два события: лайк и репост. За единицу времени может произойти одно из этих событий, оба одновременно или не произойти ни одного.

Обозначим ε_t значение энергии агента в момент времени t. Тогда значение энергии в следующий момент времени можно записать таким образом

$$\varepsilon_{t+1} = \varepsilon_t + \delta_t,$$

где δ_t является случайной величиной со значениями в {-1, 0, 1, 2}. Согласно с правилами изменения энергии, введенными выше, увеличение энергии на 2 соответствует тому, что произошли одновременно лайк и репост; увеличение на 1 – произошел только репост; энергия не меняется, если был лайк; и уменьшается на 1, если не произошло ни одно из событий. Следовательно, можно указать условное распределение δ_t при известной энергии ε_t:

$$P\left(\delta_t = 2 \mid \varepsilon_t = E\right) = p_{like}^{(E)} p_{repost}^{(E)};$$

$$P\left(\delta_t = 1 \mid \varepsilon_t = E\right) = \left(1 - p_{like}^{(E)}\right) p_{repost}^{(E)};$$

$$P\left(\delta_t = 0 \mid \varepsilon_t = E\right) = p_{like}^{(E)} \left(1 - p_{repost}^{(E)}\right);$$

$$P\left(\delta_t = -1 \mid \varepsilon_t = E\right) = \left(1 - p_{like}^{(E)}\right)\left(1 - p_{repost}^{(E)}\right).$$

Данные формулы справедливы при $E > 0$. Далее будем использовать обозначение $P_{\Delta}^{(E)} = P\left(\delta = \Delta \mid \varepsilon = E\right)$. Процесс изменения энергии агента можно рассматривать как целочисленное случайное блуждание с переходными вероятностями

$$p_{ij} = \begin{cases} P_{j-i}^{(i)}, & (j-i) \in \{-1,0,0,1,2\}, \quad i > 0; \\ 1, & i = j = 0; \\ 0, & \textit{иначе.} \end{cases}$$

Так как значение энергии в следующий момент времени зависит только от значения энергии в предыдущий момент времени, то стохастическая последовательность $(\varepsilon_0, \varepsilon_1, ..., \varepsilon_t, ...)$ является марковской цепью с переходными вероятностями p_{ij}.

Обозначим τ_{E_0} длину жизни агента с начальным значением энергии E_0 или, что то же самое, время, за которое из E_0 попали в 0. В реалистичной модели хотелось бы иметь оценку $P(\tau_{E_0} > T_{\max}) < \varepsilon$ для малого ε не очень большого значения $T_{\max}$, для того, чтобы можно было вместо бесконечных последовательностей $(\Delta_1, \Delta_2, ..., \Delta_t, ...)$ рассматривать конечные.

Рассмотрим функцию $\rho_T(E) = P(\tau_E > T)$. Справедливо рекуррентное соотношение:

$$\rho_T(E) = P_2^{(E)} \rho_{T-1}(E+2) + P_1^{(E)} \rho_{T-1}(E+1) + P_0^{(E)} \rho_{T-1}(E) + P_{-1}^{(E)} \rho_{T-1}(E-1).$$

Систему таких рекуррентных соотношений можно решить, используя начальные условия:

$$\rho_0(E) = \begin{cases} 0, & E = 0; \\ 1, & E \neq 0. \end{cases}$$

При начальных параметрах $p_{l_0} = 0.4$, $p_{r_0} = 0.1$ из решения рекуррентного уравнения можно получить оценку $P(\tau_{E_0} > 1.5E_0) < 10^{-3}$. То есть время жизни агента ограничено $1.5E_0$ с большой вероятностью и, следовательно, для получения достаточно точных оценок распределения для $(\delta_0, \delta_1, ..., \delta_t, ...)$ можно рассматривать вектора конечной длины $T_{\max} = 1.5E_0$.

В результате моделирования было определено, что распределение среднего время жизни, количества лайков и репостов в данной модели соответствует плотности распределения Вейбулла [3]:

$$f(x) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k}, & x \geq 0; \\ 0, & x < 0. \end{cases}$$

Параметры распределения Вейбулла k и λ были получены методом максимального правдоподобия. При указанных начальных параметрах, полученные значения $k = 1.9$, $\lambda = 3.8$.

Полученные результаты моделирования сравнивались с проведенными результатами исследования жизненного цикла новостных сообщений в сети микроблогов Twitter, где, в частности, анализировались характеристики роста количества специальных репостов (ретвиттов) выбранных сообщений. Распределение лайков и ретвиттов в этом случае, как и в модели, соответствовало стандартному распределению Вейбулла, причем параметр k с высокой точностью совпал с модельным (см. рис. 1).

Формирование базы данных актуальных информационных материалов из сети Интернет. Авторами в начале 2016 года проводился эксперимент по сбору сообщений из сети микроблогов Twitter, для чего в поисковом интерфейсе этой сети на периодической основе обрабатывался пакет из 100 запросов по бизнес-тематике.

В результате были получены следующие количественные данные, связанные с количеством информационных ресурсов сети Интернет, на которые были указаны ссылки из Twitter-сообщений:

1. Отсканировано около 100000 сообщений по 100 элементарным запросам к Twitter за март 2016 г.

2. 58 % сообщений содержат гиперссылки на веб-ресурсы сети Интернет.
3. Количество уникальных гиперссылок 48 %.
4. Функция распределения количества гиперссылок на одни и те-же источники – степенная.
5. Имеются проблемы идентификации внешних ссылок, связанная, прежде всего, с использованием “коротких ссылок” – переадресации с такими базовыми адресами:
   - http://migre.me/
   - http://bit.ly/
   - http://ow.ly/
   - http://tinyurl.com/
   - https://lnkd.in/
   - https://goo.gl/
   - http://wp.me/
   - http://j.mp/
   - http://dlvr.it/
6. Наиболее часто цитируемые ресурсы сети Интернет:
   - https://youtu.be (https://www.youtube.com)
   - http://fb.me/ (https://www.facebook.com/) – публичные страницы
   - https://vk.com/
   - https://twitter.com/ - ссылки на туже социальную сеть
   - https://plus.google.com/
   - http://livejournal.com/

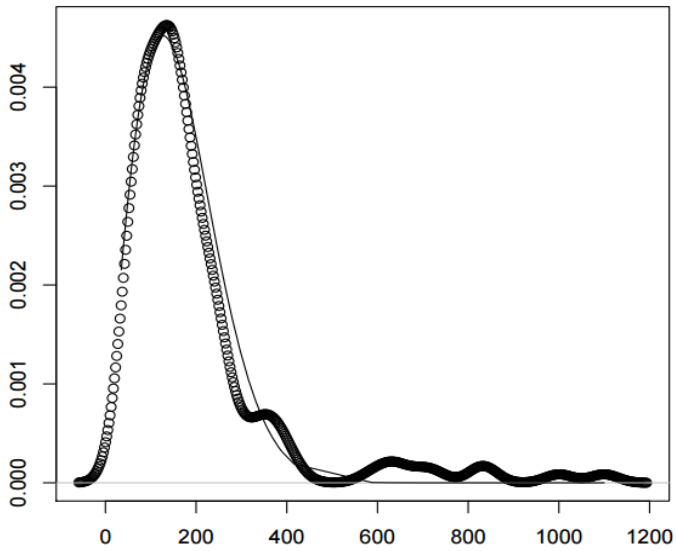

Рисунок 1 – Плотность распределения количества ретвиттов, полученных из реальной сети (аппроксимация распределением Вейбулла при $k = 1.9,\ \ \lambda = 180$)

Таким образом, практика показала, что распределение инкапсулированных в сообщения социальных сетей ссылок на информационные ресурсы сети Интернет соответствует скорее степенному распределению (см. рис. 2). В соответствии с этим фактом, приведено дополнение к указанной выше модели, а именно, изменение вероятности репостинга сообщений, содержащих ссылки на внешние информационные ресурсы.

На основе полученной информации о распределении сообщений, на которые реализованы ссылки из сети микроблогов, предлагается следующая ”краудсорсинговая” схема формирования базы данных корпоративной системы мониторинга сетевых информационных ресурсов (см. рис. 3).

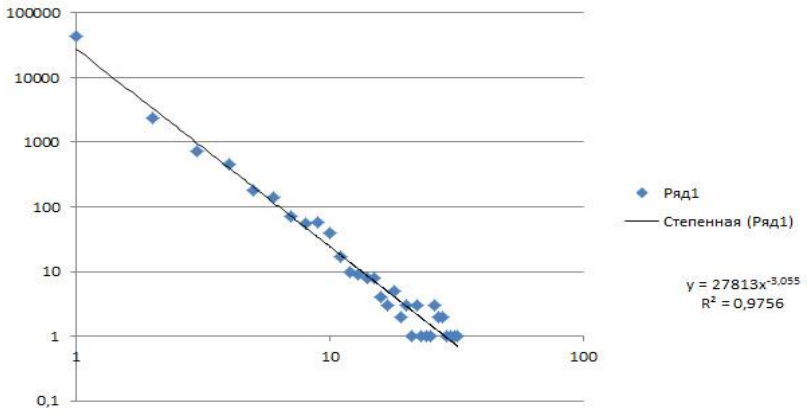

Рисунок 2 – Распределение количества репостов сообщений, содержащих ссылки на внешние информационные ресурсы

Рисунок 3 – Общая схема формирования базы данных корпоративной системы мониторинга сетевых информационных ресурсов

Приведем основные этапы этой процедуры:

1. В интересах корпоративного пользователя создаются запросы к сети микроблогов, например:

- ***a) Банки Украины;***
- ***b) Диамантбанк;***
- ***c) Укрсиббанк;***
- ***d) ПриватБанк;***
- ***e) Банк Хрещатик;***
- ***f) Платинум Банк;***
- ***g) Кредит Днепр.***

2. Сформированный ”широкий” пакет запросов передается программе сканирования сети микроблогов, в результате чего на корпоративный сервер на регулярной основе поступают сообщения, формально релевантные этим запросам.

3. Извлечение из отсканированных сообщений сети микроблогов гиперссылок на внешние сетевые информационные ресурсы.

4. Обработка гиперссылок, раскрытие ”коротких адресов”, сортировка гиперссылок, ранжирование отдельных документов и внешних источников.

5. Сканирование внешних информационных ресурсов, соответствующих выделенным гиперссылкам, первичная обработка полученных документов, приведение их к входному формату используемой корпоративной информационно-аналитической системы.

Загрузка сформированного информационного потока в корпоративную информационно-аналитическую систему, предоставление в доступ корпоративным пользователям.

Выводы. В результате описанных исследований построена мультиагентная модель распространения информационных сообщений, содержащих ссылки на информационные ресурсы в сети Интернет. Результаты моделирования проверены путем исследования реальной сети микроблогов Twitter.

Найденные закономерности могут использоваться при формировании баз данных информационно-аналитических систем, при изучении аномалий в статистике ссылок на отдельные информационные материалы, а соответственно, и в выявлении информационных операций, искусственно поддерживаемых информационных кампаний [7].

Представленный подход к формированию баз данных на основе учета ссылок в микроблогах на информационные ресурсы обеспечивает, наряду с существенным сокращением охвата информационного пространства, такие преимущества:

1. Оперативность – информационное сообщение попадает в базу данных информационно-аналитической системы в режиме реального времени по мере того, как на него сослался первый же пользователь.

2. Охват главных информационных материалов по теме. Учет мнения аудитории заинтересованных пользователей, контент сообщений которых удовлетворяет широким корпоративным запросам. Возможность ранжирования информационных материалов исходя из интересов пользователей социальных сетей.

3. Компактность баз данных, а, следовательно, удобство доступа конечных пользователей. Предсказуемость объемов баз данных, динамики их наполнения.

4. Технологическая совместимость с существующими информационно-аналитическими системами и системами контент-мониторинга.

5. Возможность выявления информационных кампаний, операций.

Публикация содержит результаты исследований, которые проводились при грантовой поддержке Государственного фонда фундаментальных исследований по конкурсному проекту Ф73 № 23558 “Разработка методов и средств поддержки принятия решений при обнаружении информационных операций”.

СПИСОК ИСПОЛЬЗОВАНОЙ ЛИТЕРАТУРЫ

[1] Д.В. Ландэ, А.А. Снарский, и И.В. Безсуднов, *Интернетика: Навигация в сложных сетях: модели и алгоритмы*. Москва, Россия: Либроком (Editorial URSS), 2009.

[2] R. Li, K. Lei, R. Khadiwala, and K. Chang, “TEDAS: A Twitter-based Event Detection and Analysis System”, in *Proc. IEEE 28th Int. Conf. Data Engineering (ICDE)*, Washington, USA, 2012, pp. 1273-1276.
doi: 10.1109/ICDE.2012.125.

[3] А.Г. Додонов, Д.В. Ландэ, В.В. Прищепа, и В.Г. Путятин, *Конкурентная разведка в компьютерных сетях*. Киев, Украина: ИПРИ НАН Украины, 2013.

[4] J. Woo, and H. Chen, “Epidemic model for information diffusion in web forums: experiments in marketing exchange and political dialog”, *Springerplus*, no. 22, pp. 5-66, 2016.
doi: 10.1186/s40064-016-1675-x.

[5] K. Lerman, “Information Is Not a Virus, and Other Consequences of Human Cognitive Limits”, *Future Internet*, vol. 8, iss. 2, pp. 1-11, 2016.
doi: 10.3390/fi8020021.

[6] Д.В. Ландэ, А.Н. Грайворонская, и Б.А. Березин, “Мультиагентная модель распространения информации в социальной сети”, *Реєстрація, зберігання і обробка даних*, т. 18. № 1, с. 70-77, 2016.

[7] А.Г. Додонов, Д.В. Ландэ, и В.А. Додонов, “Распознавание информационных операций: мультиагентный подход”, на *VI междунар. науч.-техн. конф. Открытые семантические технологии проектирования интеллектуальных систем (OSTIS-2016)*, Минск, 2016, с. 253-256.

Статья поступила в редакцию 25.03.2016.

REFERENCE

[1] D.V. Lande, A.A. Snarskii, and I.V. Bezsudnov, *Internetika: Navigation in complex networks: models and algorithms.*Moscow, Russia: Librokom (Editorial URSS), 2009.

[2] R. Li, K. Lei, R. Khadiwala, and K. Chang, “TEDAS: A Twitter-based Event Detection and Analysis System”, in *Proc. IEEE 28th Int. Conf. Data Engineering (ICDE)*, Washington, USA, 2012, pp. 1273-1276.
doi: 10.1109/ICDE.2012.125.

[3] A.G. Dodonov, D.V. Lande, V.V. Prishchepa, and V.G. Putiatin, *Opponent intelligence in computer networks*. Kyiv, Ukraine: IIR NAS of Ukraine, 2013.

[4] J. Woo, and H. Chen, “Epidemic model for information diffusion in web forums: experiments in marketing exchange and political dialog”, *Springerplus*, no. 22, pp. 5-66, 2016.
doi: 10.1186/s40064-016-1675-x.

[5] K. Lerman, “Information Is Not a Virus, and Other Consequences of Human Cognitive Limits”, *Future Internet*, vol. 8, iss. 2, pp. 1-11, 2016.
doi: 10.3390/fi8020021.

[6] D.V. Lande, A.N. Graivoronskaia, and B.A. Berezin, “Multi-agent model of information dissemination in social networks”, *Registration, storage and processing*, т. 18. № 1, pp. 70-77, 2016.

[7] A.G. Dodonov, D.V. Lande, and V.A. Dodonov, “Information Operations Recognition: multiagent approach”, in *Proc. IVth Int. Conf. Open Semantic Technology for Intelligent Systems (OSTIS-2016)*, Minsk, 2016, pp. 253-256.

ДМИТРО ЛАНДЕ,
ВАДИМ ДОДОНОВ,
ТАРАС КОВАЛЕНКО

КОРПОРАТИВНА СИСТЕМА МОНІТОРИНГУ МЕРЕЖЕВИХ ІНФОРМАЦІЙНИХ РЕСУРСІВ НА ОСНОВІ МУЛЬТИАГЕНТНОГО ПІДХОДУ

У роботі представлено мультиагентну модель розповсюдження інформаційних повідомлень, що містять посиланням на інформаційні ресурси в мережі Інтернет. Результати моделювання підтверджено дослідженням реальної мережі мікроблогів Twitter. Описано етапи створення корпоративної системи моніторингу мережевих інформаційних ресурсів, склад яких визначається посиланнями в мікроблогах. Запропонований підхід до формування баз даних на основі обліку посилань в мікроблогах на інформаційні ресурси характеризується такими перевагами: оперативність – інформаційне повідомлення заноситься в базу даних інформаційно-аналітичної системи в режимі реального часу після того, як на нього послався перший користувач; охоплення головних інформаційних матеріалів за темою. Врахування думки аудиторії зацікавлених користувачів, контент повідомлень яких задовоольняє широким корпоративним запитам; компактність баз даних і, як наслідок, зручність доступу кінцевих користувачів; передбачуваність обсягів баз даних, динаміки їх наповнення: технологічна узгодженість з існуючими інформаційно-аналітичними системами та системами контент-моніторингу; можливість виявлення інформаційних компаній, операцій.

Ключові слова: соціальна мережа, моделювання, моніторинг інформаційних ресурсів, мультиагентна система, розповсюдження інформації.

DMYTRO LANDE,
VADYM DODONOV,
TARAS KOVALENKO

CORPORATE SYSTEM OF NETWORK INFORMATION RESOURCES MONITORING BASED ON MULTI-AGENT APPROACH

In article the multi-agent model of distribution of the information messages containing references to information resources on the Internet is provided. Characteristics of growth of quantity of special retweets of the chosen messages were analyzed. Distribution of retweets in this case, as well as in model, corresponded to standard distribution of Weibull. Stages of corporate system of network information monitoring resources creation which structure is determined by references in microblogs are described. As a result of the described researches the multi-agent model of distribution of the information messages containing references to information resources in the Internet is constructed. Results of modeling are checked by research of a real network of microblogs Twitter. The found regularities can be used when forming databases of information and analytical systems, when studying anomalies in statistics of references to separate information materials, and respectively, and in identification of information transactions, artificially supported information campaigns. The provided approach has such advantages: Efficiency – the information message is included in the database of information and analytical system in real time; Spanning of the principal information materials on a subject; A possibility of ranging of information materials proceeding from interests of users of social networks; Compactness of databases, convenience of access for ultimate users. Predictability of volumes of databases, dynamics of their filling; Technological compatibility with the existing information and analytical systems and systems of content monitoring; Possibility of identification of information campaigns, operations.

Keywords: social network, modeling, information resources monitoring, multi-agent system, information dissemination.

Дмитрий Владимирович Ланде, доктор технических наук, старший научный сотрудник, заведующий отделом специализированных средств моделирования, Институт проблем регистрации информации Национальной академии наук Украины, Киев, Украина.
E-mail: dwlande@gmail.com.
Вадим Александрович Додонов, ведущий инженер отдела специализированных средств моделирования, Институт проблем регистрации информации Национальной академии наук Украины, Киев, Украина
E-mail: dodonov.vadim@gmail.com.
Тарас Васильевич Коваленко, ведущий инженер отдела специализированных средств моделирования, Институт проблем регистрации информации Национальной академии наук Украины, Киев, Украина
E-mail: 2005ste@ukr.net.

Дмитро Володимирович Ланде, доктор технічних наук, старший науковий співробітник, завідувач відділом спеціалізованих засобів моделювання, Інститут проблем реєстрації інформації Національної академії наук України, Київ, Україна.
Вадим Олександрович Додонов, провідний інженер відділу спеціалізованих засобів моделювання, Інститут проблем реєстрації інформації Національної академії наук України, Київ, Україна.
Тарас Васильович Коваленко, провідний інженер відділу спеціалізованих засобів моделювання, Інститут проблем реєстрації інформації Національної академії наук України, Київ, Україна.

Dmytro Lande, doctor of technical science, senior researcher, head of the specialized modeling tools department, Institute for information recording of National academy of science of Ukraine, Kyiv, Ukraine.
Vadym Dodonov, lead engineer of the specialized modeling tools department, Institute for information recording of National academy of science of Ukraine, Kyiv, Ukraine.
Taras Kovalenko, lead engineer of the specialized modeling tools department, Institute for information recording of National academy of science of Ukraine, Kyiv, Ukraine.